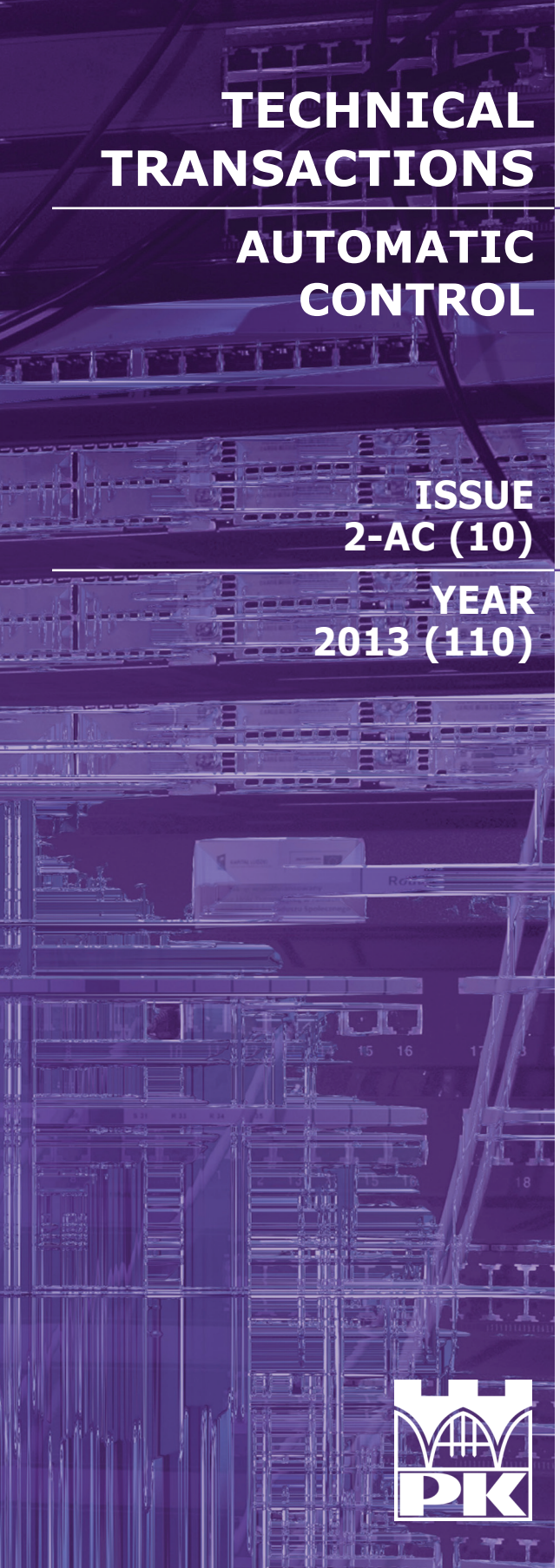

TECHNICAL TRANSACTIONS

AUTOMATIC CONTROL

CZASOPISMO TECHNICZNE

AUTOMATYKA

ISSUE 2-AC (10)

YEAR 2013 (110)

ZESZYT 2-AC (10)

ROK 2013 (110)

WYDAWNICTWO POLITECHNIKI KRAKOWSKIEJ

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE

AUTOMATIC CONTROL | AUTOMATYKA

ISSUE 2-AC (10) | ZESZYT 2-AC (10)
YEAR 2013 (110) | ROK 2013 (110)

Pierwotną wersją każdego Czasopisma Technicznego jest wersja on-line
www.czasopismotechniczne.pl www.technicaltransactions.com

Technical Transactions
series *Automatic Control*
vol. 2-AC/2013

System Analysis and Computer Science in Decision Support

Eds. Andrzej Myśliński, Sławomir Zadrożny, Piotr Kulczycki

Editorial Board

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
2-AC/2013

CARLOS A. DONIS-DIAZ*, RAFAEL BELLO*, JANUSZ KACPRZYK**

LINGUISTIC DATA SUMMARIZATION USING AN ENHANCED GENETIC ALGORITHM

LINGWISTYCZNE PODSUMOWANIA DANYCH Z UŻYCIEM ULEPSZONEGO ALGORYTMU GENETYCZNEGO

Abstract

This paper presents work is presented an enhanced Genetic Algorithm (GA) specifically designed for the production of linguistic data summaries. The model is able to obtain not a set of 'good linguistic summaries' but a 'good set' of summaries. The model incorporates an operator and fitness function specially designed to fulfil this aim. Experiments show how the enhanced model is able to improve results obtained with the classical model of GA and to guarantee a summary with high diversity and good values for the quality measures in individual summaries

Keywords: *Linguistic Data Summarization, Data mining, Fuzzy Logic, Genetic Algorithms*

Streszczenie

Przedmiotem niniejszego artykułu jest ulepszony algorytm genetyczny (GA) zaprojektowany z zamiarem użycia głównie do tworzenia lingwistycznych podsumowań danych. Zaproponowany model pozwala na uzyskanie nie tyle zbioru „dobrych podsumowań lingwistycznych", co „dobrego zbioru" tych podsumowań. Cel ten uzyskano przez zastosowanie w modelu odpowiedniego operatora i funkcji przystosowania. Przedstawione w artykule eksperymenty obliczeniowe potwierdzają, że autorski model wpływa na poprawę wyników otrzymanych dla klasycznego modelu opartego na algorytmie genetycznym oraz gwarantuje podsumowania charakteryzujące się dużą różnorodnością oraz dobrymi wartościami miar jakości poszczególnych podsumowań.

Słowa kluczowe: *lingwistyczne podsumowania danych, eksploracja danych, logika rozmyta, algorytmy genetyczne*

* Carlos A. Donis-Diaz, Rafael Bello, e-mail: cadonis@uclv.edu.cu, Computer Science Department, Universidad Central Marta Abreu de Las Villas, Santa Clara.

** Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw; Department, of Automatic Control and Information Technology, Faculty of Electrical and Computer Engineering, Cracow University of Technology.

1. Introduction

Linguistic data summarization has for a long time been a subject of intensive research, and various tools and techniques from computational linguistics, natural language generation etc. have been proposed. The traditional approaches do not however provide means for the representation and processing of imprecision and vagueness that is inherent in natural language. Fuzzy logic with linguistic quantifiers is one of the most conceptually simple, developed and used approaches for the linguistic summarization of numerical data (LDS), and – what is crucial for our application, provides simple and natural means it dealing with imprecision. The concept of a linguistic data summary, using fuzzy logic with linguistic quantifiers, which will be employed in this paper, was introduced by Yager in [1–3], considerably advanced in [4–7] and presented in an implementable way in [8].

The process of generating linguistic data summaries for a given set of numerical data, usually a relational numerical database, can conveniently be represented as an optimization problem in which the best summaries from a large set of candidates are selected, and the basic objective function is assumed to be the truth degree of a linguistic summary that is equated with a degree of truth of a linguistically quantified proposition that is conceptually equivalent to a linguistic summary in question. Several works to deal with this problem have been developed [9–13]. Much of them use a genetic algorithm, in principle in its basic version.

In the present work, an enhanced genetic algorithm specifically designed for finding not a set of 'good ('best') linguistic summaries', is proposed as it was the case in virtually all works done so far, but a 'good ('best') set' of linguistic summaries. A specific genetic operator and fitness function are proposed to deal with some problems observed in searching the solution space in the specific context of linguistic summarization.

The experiments were carried out on *creep* data. The creep rupture stress (*creep*) is one of the most important mechanical properties considered in the design of new steels used in industries like aeronautical, energy and petrochemical. Basically, (sometimes called cold flow too) is a tendency of a solid material to slowly deform permanently under mechanical stresses, and is increased due to high temperature, notably close to the melting point of the material in question, i.e. the *creep* measures the stress level in which a steel structure fails when exposed to quite aggressive conditions (like high steam temperatures) over periods of time as long as 30 years.

2. Linguistic data summaries for *creep* data: the use of a genetic algorithm

In this section are presented the main theoretical aspects needed to introduce the proposed enhanced model.

2.1. Linguistic data summarization

In this paper, the linguistic data summarization approach that uses the fuzzy logic with linguistic quantifiers proposed in [4] is considered. A basic description is presented here.

Having: $Y = \{y_1, ..., y_n\}$ a set of objects (records) in a database, e.g., the set of workers, and $A = \{A_1, ..., A_m\}$ a set of attributes (fuzzy variables) characterizing objects from Y, e.g., salary, age, etc. in a database D of workers, and $A_j(y_i)$ denotes the value of attribute A_j for object y_i. A linguistic summary from D consists of:

- a summarizer S, i.e. an attribute together with a linguistic value (fuzzy predicate) defined on the domain of attribute A_j (e.g. 'low salary' for attribute 'salary');
- a quantifier Q (a linguistic quantifier), i.e. a fuzzy set with universe of discourse in the interval [0, 1] expressing a quantity in agreement, e.g. *most*;
- a truth degree T (validity) of the summary, i.e. a number from the interval [0, 1] assessing the truth of the summary (e.g. 0.7); usually, only summaries with a high value of T are of interest;
- optionally, a qualifier R, i.e. a fuzzy filter determining a fuzzy subset of Y; can be composed for one or a set of triplets <fuzzy variable, operator, linguistic value> (e.g. 'young' for attribute 'age').

Thus, linguistic summaries may be exemplified by:

$$T \text{ (most employees earn low salary)} = 0.7 \tag{1}$$

$$T \text{ (most young employees earn low salary)} = 0.7 \tag{2}$$

and their foundation is Zadeh's [14] linguistically quantified proposition corresponding to either, for (1) and (2):

$$Qy's \text{ are } S \tag{3}$$

$$QRy's \text{ are } S \tag{4}$$

The T, i.e., the truth value of (3) or (4) may be calculated by using either Zadeh's original calculus of linguistically quantified statements [14], or other interpretations of linguistic quantifiers. In this work is used the first where a (proportional, nondecreasing) linguistic quantifier Q is assumed to be a fuzzy set in [0, 1] and the values of T are calculated as:

$$T(Qy's \text{ are } S) = \mu_Q\left[\frac{1}{n}\sum_{i=1}^{n}\mu_S(y_i)\right]$$

and defining:

$$r = \frac{\sum_{i=1}^{n}\mu_R(y_i) \wedge \mu_S(y_i)}{\sum_{i=1}^{n}\mu_R(y_i)}$$

we have:

$$T(QRy's) \text{ are } S = \mu_Q\left(r\right)$$

Besides using T (truth), other validity measures have been proposed to determine the quality of summaries. In [5] can be found: the truth degree ($T1$) that corresponds with the mentioned T; the degree of imprecision ($T2$) that only depends on the form of the summarizer and expresses the fuzziness of the summary due to its description S; the degree of covering ($T3$) that says how many objects in the database corresponding with the qualifier R, are covered by the summary, i.e. by the particular description S; the degree of appropriateness ($T4$) that describes how characteristic for the particular database the summary found is; finally the length of the summary ($T5$) that is defined using the number of terms of the summarizer S.

As mentioned, several approaches have been used to mine the best summaries from a large set. The definition of *protoforms*, as a set of templates used to specify the form of summaries to be mined reducing the search space, is one of them. In [15–17] the concept of a *protoform* as a more or less abstract prototype of a linguistically quantified proposition is presented. The most abstract protoform corresponds to (3) and (4), while (1) and (2) are examples of fully instantiated protoforms. Thus, protoforms form a hierarchy where higher/lower levels correspond to more/less abstract protoforms. Going down this hierarchy, one has to instantiate particular components of (3) and (4), i.e., Q, S and R.

In the present work, the term *proposition* is used, to be more specific a linguistically quantified proposition, to refer a linguistic summary and the term (linguistic) *summary* refers to a set of propositions. This is basically consistent with [4, 8], and in particular the modern natural language generation (NLG) based approach [18].

2.2. The use of genetic algorithms in linguistic data summarization

The basic principles of GAs were first laid down rigorously by Holland [19], and they are well described in many books, such as [20]. The basic idea is to maintain a population of chromosomes, which represents candidate solutions to the concrete problem being solved, which evolves over time through a process of competition and controlled variation. Each chromosome in the population has an associated fitness to determine which chromosomes are used to form new ones in the competition process. The new ones are created using genetic operators such as crossovers and mutations. GAs have a great measure of success in search and optimization problems.

A GA starts off with a population of randomly generated chromosomes (solutions), and advances toward better chromosomes by applying genetic operators. The population undergoes evolution in the form of natural selection. During successive iterations called generations, a new population of chromosomes is formed using a selection mechanism and specific genetic operators such as crossovers and mutations. An evaluation or fitness function must be devised for each problem to be solved. Given a particular chromosome (a possible solution), the fitness function returns a single numerical fitness, which is supposed to be proportional to the utility or adaptation of the solution represented by that chromosome.

The classic GA model has been used previously on linguistic data summarization to efficiently handle the search space [11, 12]. The classic form of the model employed in those works refers to the basic form of the GA presented in [20], i.e. using the basic operators: selection, crossover and mutation. Other works, mainly developed for time series applications, use GA or related evolutionary heuristics with specific adaptations [9, 10].

3. An enhanced genetic algorithm for discovering linguistic data summaries

The proposed model uses an approach that permits obtaing the 'best set' of propositions (i.e. the 'best summary') what is different to the approach that looks for the set of 'best propositions' (i.e. the summary of 'best propositions'). The former approach has the advantage of getting, not only a summary with high quality propositions, but also with 'desirable characteristics' for the interaction among all propositions. The model was designed following two main aims: (1) to obtain propositions with a high goodness; (2) to get a final set of propositions with a high diversity in the sense of comprising propositions with sufficient differences in their structures and components. This is in line with recent developments in the linguistic data summarization, cf. [16, 17].

3.1. The genetic representation

In the present work, the chromosome represents a whole linguistic summary (i.e. a set of linguistically quantified propositions) and each gene codifies just one such proposition. Making a parallel with approaches for rules discovery used by evolutionary models, the proposed model falls into the Pittsburgh approach [21]. In this approach, each individual corresponds to a complete set of rules and each run of the evolutionary procedure discovers one set of rules, probably the best one between all iterations. An important feature is that in order to guide the discovery process, a complete set of propositions is evaluated instead of a single proposition so the model can consider the importance of the interaction between propositions; i.e. in the present work, the model is able to control the diversity inside the set of propositions to be obtained.

The protoform used to search the propositions has the following form:

Protoform	Given	Sought
QRy's are *S*	structure of *S* (*creep is* <linguistic value>)	*R*, *Q* and linguistic values in *S*

Taking this into consideration, the genes encode the three main components of a linguistically quantified proposition: the quantifier (*Q*), the qualifier (*R*) and the summarizer (*S*).

3.2. The fitness function

To achieve the aims mentioned above, the proposed fitness function contains: (1) four measures to control the quality of the propositions: degree of Truth ($T1$), imprecision degree ($T2$), covering degree ($T3$) and appropriateness degree ($T4$); and (2), a variable called *Diversity* to control the interaction between the propositions inside the summary. The function to be maximized for a chromosome i is defined in the interval [0, 1] as $F_i = m_g G_i + m_d D_i$ where G and D represent the degree of Goodness and Diversity respectively, and m_g, m_d are the importance degrees assigned to the terms. The Goodness (G) of a chromosome j is calculated as the average value of the goodness of genes. The goodness for each gene (proposition) g_j is characterized by a weighted sum of the quality measures ($T1^{St}$, $T2$, $T3$, $T4$) where

$T1^{St}$ represents a term called *Linguistic Strength* that is calculated as: $T1^{St} = T1 \cdot \overrightarrow{St}$ been $\overrightarrow{St}$ a vector of values in the interval [0, 1]) defined as a parameter to express the preference for each linguistic label of the quantifier. In this work it was defined as: St[Most] = 1, St[Much] = 0.75, St[Half] = 0.20, St[Some] = 0.15 and St[Few] = 0.05 so *Most* and *Much* are preferred; this is in line with some indications of a special role played by the linguistic quantifier '*most*' which can be found in many contributions shown in, e.g., [22]. The Goodness of a gene j is calculated as $g_j = 0.4 \cdot T1^{St} + 0.1 \cdot T2 + 0.25 \cdot T3 + 0.25 \cdot T4$.

The Diversity (D) degree is calculated taking into account the number of clusters of genes (C) existing in the chromosome: $D_i = C_i/n$ where C is obtained by a clustering process using a similarity function (L) to determine if two propositions are similar or not. L is defined as:

$$L(p1, p2) = \begin{cases} Yes & if \ \sum_{k=0}^{m} H\left(p1_k, p2_k\right) < 2 \\ No & otherwise \end{cases}$$

where $p1$ and $p2$ are vectors of size m representing the propositions to be compared. Its components refer to the linguistic values (labels) used in the propositions. There is one component for each fuzzy variable; if a specific fuzzy variable is not used in the proposition, the respective component is equal to zero.

The function $H(p1_k, p2_k)$ is defined as:

$$H(p1_k, p2_k) = \begin{cases} 1 & if \ \left|p1_k - p2_k\right| > round(20\% \ of \ V_k) \ \text{ or } \ if \ \ p1_k = 0 \text{ and } p2_k \neq 0 \ \text{ or} \\ & if \ \ p1_k \neq 0 \text{ and } p2_k = 0 \\ 0 & otherwise \end{cases}$$

where V_k is the number of labels of the fuzzy variable represented by the k-th component of the vector. This function determines if there is a difference between the labels used in $p1$ and $p2$ for a specific fuzzy variable.

3.3. The genetic operators

When using the normal operators (selection, crossover and mutation) in the classical GA the following problem arises: the crossover operator improves the degree of diversity of the summary but not the degree of goodness of individual propositions due to this operator has not direct influence over the genes (propositions) in the chromosome. On the other hand, the mutation operator does not guarantee a sufficient perturbation inside the chromosome to solve the situation.

To deal with this problem, the use of an additional operator is proposed to be added to the evolution process in the enhanced model. This operator, called the *Propositions Improver*, implements a local search based on a best first strategy [23] when looking for a better variant in the neighborhood of the proposition. It implements a greedy random procedure based on six possible transformations of the proposition. Four of these transformations occur on the qualifier, one on the summarizer and one on the quantifier. The transformations are:

– Change in R (the qualifier) a randomly selected fuzzy predicate by another randomly generated.
– Change in R the linguistic value of a randomly selected fuzzy predicate by another randomly generated.
– Add in R a new randomly generated fuzzy predicate.
– Delete in R a new randomly selected fuzzy predicate.
– Change in Q (the quantifier) its linguistic value by a 'nearby' one, i.e. by the following (backward or forward) linguistic value in the set of terms.
– Change in S (the summarizer) its linguistic value by a 'nearby' one.

The stopping criterion for the local search occurs when at least one of the following values is reached:

– the total number of new generated propositions is equal to 8,
– the number of continuously generated propositions without improvement is equal to 5 or
– the value of T (truth) is equal to or greater than 0.85

4. Experiments results and analysis

Linguistic data summarization has proven effective to describe *creep* trends regarding specific variables used in the process of designing new ferritic steels as experimented in [24]. However, its use in this work was reduced to obtain a small set of fully instantiated propositions to be compared with results obtained with an artificial neural network model. Unlike this, a complete mining of propositions is performed in the present work. The creep data and the fuzzy modeling used is the same as in [24].

Several experiments were performed to measure and compare the performance of the proposed enhanced model (Enhanced) in relation to the classical model (Classical) for obtaining a good linguistic summary from *creep* data. Ten runs of the models were used for each experiment and each run was limited to a fixed number of generated propositions: a maximum of 250.000 what represent an insignificant amount of possible propositions for this problem (the 6.91E-15 percent). In this way, can be ensured that one model does not take advantage over the other with respect to the amount of propositions considered to find the best solution.

Is important to note that for the *creep* problem, the propositions having *Most* or *Much* as quantifiers are more interesting, that is why the parameter $\overrightarrow{\mathrm{St}}$ was set preferring these values in both models.

For a better interpretation and analysis, the Wilcoxon's test and Monte Carlo's technique were used to compare the results pairs to pairs and to calculate a more precise signification of the differences respectively. A value less than 0.05 in Monte Carlo's technique were considered as significant for the differences.

Table 1 shows the results obtained. In this table, the rows represent the results obtained with each model. Columns refer the parameters used to measure the quality of the obtained summary:

– Columns (1) to (4) represent the mean values obtained by models for the indicators used in the fitness function to measure the quality of the propositions composing the summary:

  - column (1) represents the linguistic strength,
  - column (2) represents the degree of imprecision,
  - column (3) represents the degree of covering,
  - column (4) represents the degree of appropriateness.
- Column (5) represents the mean goodness value of propositions that compose the summary. This value is calculated as a weighted sum of the previous indicators.
- Column (6) represents the mean value of diversity.
- Column (7) represents the mean value of fitness.
- Column (8) represents the mean number of propositions in the summary that: (8A) have the desired linguistic values for the quantifier (*most* and *much*); (8B) do not have the desired values of truth ($T < 0.85$).

Table 1

Behavior of the two variants of the GA model

GA model	Mean values of							Mean number of propositions with (8)	
	$T1^{St}$ (1)	$T2$ (2)	$T3$ (3)	$T4$ (4)	G (5)	D (6)	*Fitness* (7)	Quantifier (*most*, *much*) (A)	$T1 < 0.85$ (B)
Classical	0.1962	0.8918	0.3158	0.3805	0.3417	1.000	0.5392	12.0	15.5
Enhanced	0.5157	0.8960	0.5287	0.5343	0.5616	1.000	0.6931	16.3	1.8

When analyzing the indicators used to measure the quality of the propositions (columns 1 to 4), it can be observed that the enhanced model provides better results in all values except for the imprecision degree. This is because this indicator depends only on the linguistic terms used in the summarizer and in the present application the summarizer has only one fuzzy predicate and its linguistic terms are randomly selected with the same probability. For the rest of values, the enhanced model presents significant differences in relation to values obtained with the classical model. As a direct consequence, the value of goodness (column 5) obtained with the enhanced model is better than the value of the classical model and presents significant differences.

The column (8A) indicates how the enhanced model is able to find a better mean number of propositions with the desired quantifier. Column (8B) shows the number of propositions on the summary (from a total number of 30, the size used for the chromosomes) with a degree of truth less than a value considered good for this application. This column reflects how the classical model is unable to evolve a big number of the propositions towards better ones using the normal operators.

Those results show how the local search implemented in the enhanced model through the *Propositions Improver* operator is able to obtain individual propositions with an improved quality.

When analyzing the values of the diversity degree shown in column (6) it can be noted that both models obtain good values. This result to remark allows us that the effectiveness of the crossover operator for fitting the desired behavior in the diversity degree between the propositions composing the summary.

Finally, column (7) shows the better overall behavior of the enhanced model in comparison to the classical model. The differences in the fitness values are significant.

5. Conclusions

A model to obtain linguistic data summaries using an enhanced GA has been proposed. The use of a local search in the form of an additional operator in the classical GA has shown an improvement in results. The proposed fitness function including the term *Diversity*, the parameter St and the used quality measures guarantees a summary with high quality propositions, a good degree of diversity and many propositions with the desired quantifiers. Our future works will be further explored possibilities to use local search in GA type schemes, exemplified by memetic algorithms, as well as the use of new approaches for the derivation of linguistic data summaries based on natural language technology, notably natural language generation (NLG) [18].

References

[1] Yager R.R., *A new approach to the summarization of data*, Information Sciences, 28, 1982, 69-86.

[2] Yager R.R., *On linguistic summaries of data*, [in:] G. Piatetsky-Shapiro, W.J. Frawley (Eds.) *Knowledge Discovery in Databases*, AAAI Press/The MIT Press, Menlo Park, 1991, 347-363.

[3] Yager R.R., *Database discovery using fuzzy sets*, International Journal of Intelligent Systems, 11, 1996, 691-712.

[4] Kacprzyk J., *Intelligent data analysis via linguistic data summaries: a fuzzy logic approach*, [in:] R. Decker, W. Gaul (Eds.) *Classification and Information Processing at the Turn of the Millennium*, Springer-Verlag, Heidelberg and New York 2000, 153-161.

[5] Kacprzyk J., Yager R.R., *Linguistic summaries of data using fuzzy logic*, International Journal of General Systems, 30, 2001, 133-154.

[6] Kacprzyk J., Yager R.R., Zadrożny S., *A fuzzy logic based approach to linguistic summaries of databases*, International Journal of Applied Mathematics and Computer Science, 10, 2000, 813-834.

[7] Kacprzyk J., Yager R.R., Zadrożny S., *Fuzzy linguistic summaries of databases for an efficient business data analysis and decision support*, [in:] W. Abramowicz, J. Żurada (Eds.) *Knowledge Discovery for Business Information Systems*, Kluwer, Boston 2001, 129-152.

[8] Kacprzyk J., Zadrożny S., *Computing with words: towards a new generation of linguistic querying and summarization of databases*, [in:] P. Sinčak, J. Vaščak (Eds.) *Quo Vadis Computational Intelligence?*, Physica-Verlag, Heidelberg and New York 2000, 144-175.

[9] Castillo-Ortega R. et al., *Linguistic Summarization of Time Series Data using Genetic Algorithms*, 7th Conference of European Society for Fuzzy Logic and Technology – EUSFLAT 2011, Atlantis Press, Aix-les-Bains 2011, 416-423.

[10] Castillo-Ortega R. et al., *A Multi-Objective Memetic Algorithm for the Linguistic Summarization of Time Series*, 13th Annual Genetic and Evolutionary Computation Conference – GECCO' 2011, ACM, Dublin 2011, 171-172.

[11] George R., Srikanth R., *Data summarization using genetic algorithms and fuzzy logic*, [in:] F. Herrera, J.L. Verdegay (Eds.) *Genetic Algorithms and Soft Computing*, Physica-Verlag, Heidelberg 1996, 599-611.

[12] Kacprzyk J., Wilbik A., Zadrożny S., *Using a Genetic Algorithm to Derive a Linguistic Summary of Trends in Numerical Time Series*, International Symposium on Evolving Fuzzy Systems, Ambleside 2006, 137-142.

[13] Kacprzyk J., Wilbik A., Zadrożny S., *Linguistic summarization of time series using a fuzzy quantifier driven aggregation*, Fuzzy Sets and Systems, 159, 2008, 1485-1499.

[14] Zadeh L.A., *A computational approach to fuzzy quantifiers in natural languages*, Computers and Mathematics with Applications, 9, 1983, 149-184.

[15] Kacprzyk J., Zadrożny S., *Protoforms of linguistic data summaries: towards more general natural-language-based data mining tools*, [in:] A. Abraham, J.R.D. Solar, M. Koeppen (Eds.) *Soft Computing Systems*, IOS Press, Amsterdam 2002, 417-425.

[16] Kacprzyk J., Zadrożny S., *Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools*, Information Sciences, 173, 2005, 281-304.

[17] Kacprzyk J., Zadrożny S., *Protoforms of Linguistic Database Summaries as a Human Consistent Tool for Using Natural Language in Data Mining*, International Journal of Software Science and Computational Intelligence, 1, 2009.

[18] Kacprzyk J., Zadrożny S., *Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries and natural language generation*, IEEE Transactions on Fuzzy Systems, 18, 2010, 461-472.

[19] Holland J.H., *Adaptation in natural and artificial systems*, MIT Press, Cambridge 1992.

[20] Goldberg D.E., *Genetic algorithms in search*, Optimization and Machine Learning, Addison-Wesley, Reading 1989.

[21] Smith S., *Flexible learning of problem solving heuristics through adaptive search*, 8th International Conference on Artificial Intelligence, Morgan Kaufmann, 1983, 422-425.

[22] Zadeh L.A., Kacprzyk J., *Computing with Words in Information/Intelligent Systems*, Physica-Verlag (Springer-Verlag), Heidelberg and New York 1999.

[23] Russell S.J., Norvig P., *Artificial Intelligence: A Modern Approach*, Third ed., Prentice Hall, 2009.

[24] Díaz C.A.D., Perez R.B., Morales E.V., *Using Linguistic Data Summarization in the study of creep data for the design of new steels*, 11th International Conference on Intelligent Systems Design and Applications – ISDA 2011, Cordoba 2011, 160-165.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
2-AC/2013

ALEX TORMÁSI*, LÁSZLÓ T. KÓCZY**

CONCEPT AND DEVELOPMENT OF A FUZZY-BASED MULTI-STROKE CHARACTER RECOGNIZER

KONCEPCJA I ROZWINIĘCIE ROZPOZNAWANIA WIELOLINIOWEGO PISMA ODRĘCZNEGO NA PODSTAWIE LOGIKI ROZMYTEJ

A b s t r a c t

In this paper, the latest member of the FUzzy-BAsed character Recognizer (FUBAR) algorithm family with multi-stroke character support is presented. The paper summarizes the basic concept and development of multi-stroke FUBAR and compares the single-stroke, multi-stroke FUBAR algorithms with the most similar methods found in literature.

Keywords: *fuzzy systems, fuzzy grid, fuzzy-based character recognition*

S t r e s z c z e n i e

W niniejszym artykule opisano najnowsze rozwiązanie z rodziny algorytmów rozpoznawania pisma odręcznego na podstawie logiki rozmytej, wspomagające wykrywanie wieloliniowych liter. W artykule przedstawiono podstawowe pojęcia oraz rozwój autorskiego algorytmu opartego na logice rozmytej, a także porównano go – zarówno w wersji dla jednoliniowych oraz wieloliniowych liter – z podobnymi metodami znalezionymi w literaturze.

Słowa kluczowe: *systemy rozmyte, rozmyte siatki, rozpoznawanie pisma na podstawie logiki rozmytej*

* M.Sc. Alex Tormási, e-mail: tormasi@sze.hu, Department of Information Technology, Faculty of Engineering Sciences, Széchenyi István University, Győr.

** Prof. D.Sc. Ph.D. László T. Kóczy, Department of Automation, Faculty of Engineering Sciences, Széchenyi István University Győr; Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics.

1. Introduction

The classification problem is a well researched area covering several methods and applications [1, 2]. Character recognition is a multi-dimensional subfield in classification problems, which has also been investigated and researched for a long time by academics and industrial partners as well.

Most of the recognition methods in literature are using various computational intelligence methods and other special solutions to reach high accuracy recognition with a low computational cost [3, 4]. Despite the high accuracy, these methods are not always usable (on devices with limited hardware resources) for on-line (real-time) handwriting recognition as a result of their high computational complexity and processing time. It is very important to find a recognition engine, which is able to process the input strokes within a short time period with an acceptable level of accuracy even on devices with limited resources such as tablets.

LaLomia defined the user acceptance threshold at 97% [5], however, most multi-stroke character recognition methods known from the literature that are applicable for 26 symbols are well below that, on the other hand, with a strict set of symbols (16 gestures) the $N recognizer reached 96.7% [3].

In this paper, we present a new attempt to recognize multi-stroke letters (26 symbols) with a rather good recognition rate (however definitely below LaLomia's 97% threshold). As a starting point, the FUBAR algorithm that was very successful for single-strokes is used, with extensions and modifications towards multi-stroke symbols (up to 3 strokes).

After the introduction, basic concept of the FUBAR algorithms [6, 7] is overviewed. In Section 3, the design and development of the new method with the capability of recognizing multi-stroke symbols is presented. In Section 4, results are presented and the average accuracy of the single-stroke and multi-stroke FUBAR algorithms are compared, for the case of the same methods with a hierarchical rule-bases [8, 9]. The results are summarized and future directions are discussed in the last section.

2. Concept of the FUBAR Algorithm Family

2.1. Features, Goals and Limitations of FUBAR

During the design of the concept of FUBAR methods, four key important features were identified as the necessary but not sufficient condition for modern and acceptable recognition engines, which are the following:

1. Acceptable accuracy: The algorithm must reach the user acceptance threshold.
2. Efficiency: The designed methods must fit to the user's requirements in response time and even in hardware with limited resources such as tablets. This means that complex geometrical transformations and other complex mathematical functions should be avoided.
3. Flexibility of the alphabet: The model of the alphabet must be easily modifiable to support various alphabets and context-sensitive recognition.
4. Learning: The designed system should be able to learn user-specific writing styles.

These features were considered as main goals to achieve; all the used techniques and the model of the designed system were selected based on their properties and abilities to reach the listed goals.

The focus during the development was on the recognition engine itself in order to reduce the influences caused by other sub-problems of the recognition process such as segmentation. The designed recognition engine is on-line (the algorithm uses digital ink information to describe strokes) and personalized (it recognizes the handwriting of a specific person).

2.2. Input Handling and Processing

The FUBAR method collects the positions of the digital pen used during the writing process (a three-dimensional continuous input signal is shown in Fig. 1), which is represented by a list of two-dimensional coordinates in chronological order.

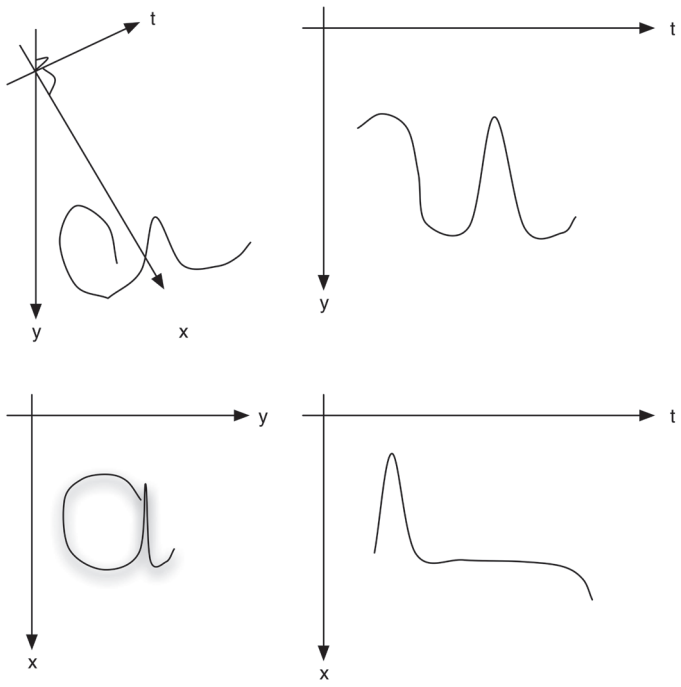

Fig. 1. The multi-dimensional input stroke from various aspects

The received stroke contains empty spaces (it is non continuous) depending on the writing speed and the BUS and CPU usage due to hardware and bandwidth limitations. It is more difficult to process the stored stroke as a result of the stochastic properties of the point distribution.

For further processing, the input stroke should be re-sampled, which provides a low-level anti-aliasing for the stroke and an almost equal distance between the sampled points of the stroke as seen in Fig. 2.

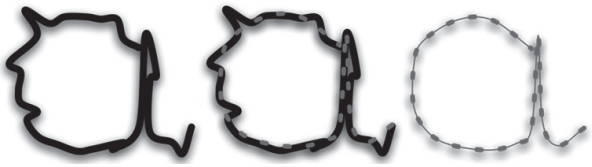

Fig. 2. Comparing the original and re-sampled input stroke

2.3. Character-Feature Extraction

The next step of the FUBAR method is feature extraction. The method use two kinds of features: (1) the width/height ratio of the stroke; and (2) the average number of points in rows and columns of a grid drawn around the stroke.

The first FUBAR algorithm used general grids with sharp borders, but the method reached a low average recognition rate as a result of the italic writing style of the test subjects as seen in Fig. 3.

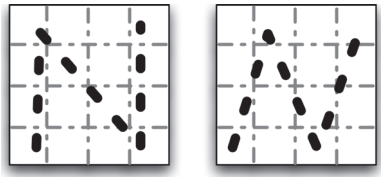

Fig. 3. Strokes with normal and italic writing style in a grid

Other similar recognizers are using mathematical transformations such as rotation to resolve the italic writing style problem, this increases the complexity of the algorithm. In the FUBAR algorithms, a fuzzy grid [6] is used, where the columns and rows of the grid are defined by fuzzy sets [10] to keep low the computational cost of the method. The points of a stroke may belong to more than one row or column with various membership values as seen in Fig. 4.

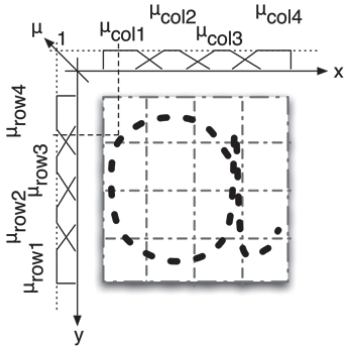

Fig. 4. A fuzzy grid example

2.4. Inference

In FUBAR algorithms, a discrete Takagi-Sugeno method [11] is used for the inference phase. The fuzzy sets in the initial rule-base of the system were determined statistically from previously collected samples.

The rule-base represents the alphabet; each character is defined by a single rule. The antecedent of the rules consists of the previously collected stroke-features, while the consequent part of the rule defines the degree of matching between the features of the input stroke and the character represented by the given rule.

The algorithm returns the character assigned to the rule with the highest matching value.

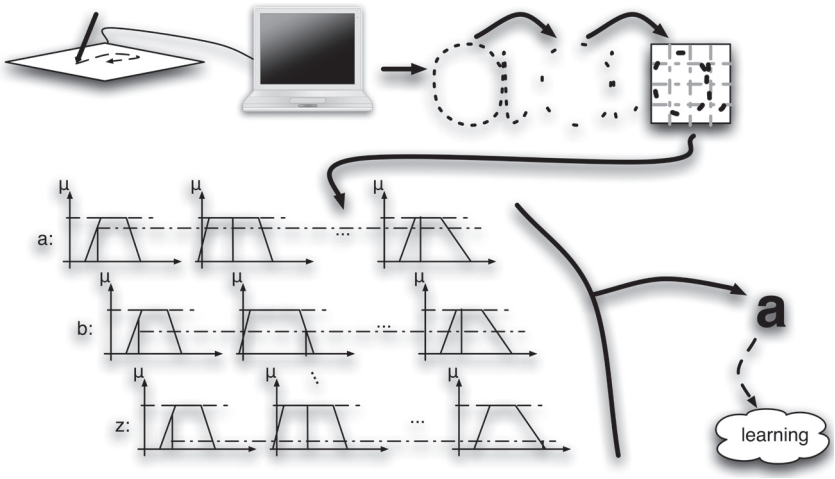

Fig. 5. Concept of FUBAR algorithms

After the inference step, the system is able to change the initial rule-base according to the features of the input stroke, this gives the ability to learn user specific writing styles. The learning phase was disabled during the tests to avoid its influence on the results caused by the heuristic properties of the used algorithm.

3. Multi-Stroke Character Support of FUBAR

3.1. Motivation and Concept

The first members of the FUBAR algorithm family supported only single-stroke symbols, but the handling of multi-stroke characters was the next step in the development of the recognition engine.

The designed method is independent from the order or the number of sub-strokes representing a multi-stroke character, but the basic shape must be identical to the model stored in the rules. This provides freedom to the users and the possibility to use any permutations of the sub-strokes representing the character with the same look.

The algorithm merges together each input sub-stroke to represent a multi-stroke character, which means that it will be represented by a single list of x and y coordinates in chronological order like a single-stroke character; this merged stroke is used during the same steps of recognition as described in the previous section.

3.2. Determining the Initial Rule Base

In the first members of FUBAR, the parameters extracted from a collection of 60 single-stroke character samples were used to determine the rule-base for the fuzzy system. The quartiles of the character-features were used to calculate the break points of the trapezoidal shaped fuzzy sets. The same method was used during the determination of the initial rule-base for the multi-strokes to have a better basis for the comparison of the algorithms.

3.3. Reducing Computational Complexity with Hierarchical Rule Base

There are many papers dealing with the use of hierarchical rule-bases in fuzzy systems in different areas [8, 9]. The use of hierarchical fuzzy rule-bases could reduce the computational cost of the single-stroke FUBAR method by decreasing the number of the evaluated rules. The details of building the hierarchical rule structure by rule input parameters for the single--stroke alphabet were presented in [12].

In the multi-stroke FUBAR algorithm, the number of sub-strokes representing a character is also used during the inference. Only those rules are evaluated, which have exactly the same number of sub-strokes as the input character has. This means that each character must be written with a predetermined number of sub-strokes, which is a limitation compared to the original multi-stroke FUBAR with a flat rule-base, but the number of the evaluated rules were significantly reduced.

4. Results

Both single-stroke and multi-stroke FUBAR algorithms were tested using the same context and conditions. A training set with 60 samples per character from various test subjects was used to determine the initial rule-base by calculating (using the quartiles of the dataset as breakpoints of trapezoid membership functions) the fuzzy sets in antecedents of the rules describing the ‘template symbols’ (the knowledge is stored as a single fuzzy rule per characters, also known as the template symbol). Another 120 samples per character (validation set) were used to determine the average accuracy of the methods.

The best result for the multi-stroke FUBAR was achieved by the algorithm using a 3×4 fuzzy grid; the letter-wise average recognition rates are listed and compared with the results of the similar single-stroke FUBAR method (with both flat and hierarchical rule-bases) in Table 1.

Table 1

Letter-wise average recognition rates of various FUBAR algorithms

Symbol	FUBAR algorithms with various properties			
	Single-Stroke FUBAR with 6 × 4 fuzzy grid	Multi-Stroke FUBAR with 3 × 4 fuzzy grid	6 × 4 Single-Stroke FUBAR with hierarchical rule-base	3 × 4 Multi-Stroke FUBAR with hierarchical rule-base
A	100	96.1111	100	96.1111
B	92.7778	89.4444	92.7374	89.4444
C	97.7778	76.6667	97.7654	76.6667
D	98.8889	96.6667	98.8827	96.6667
E	98.8889	92.2222	97.2067	92.2222
F	100	96.1111	100	96.1111
G	100	97.2222	100	97.2222
H	100	95.5556	100	95.5556
I	100	98.3333	100	98.3333
J	100	97.7778	100	97.7778
K	99.4444	96.6667	99.4413	96.6667
L	100	98.3333	100	98.3333
M	100	96.1111	100	96.1111
N	100	96.6667	96.0894	96.6667
O	93.8889	92.7778	93.8548	92.7778
P	100	92.7778	100	92.7778
Q	100	97.7778	100	97.7778
R	100	87.7778	100	87.7778
S	100	97.7778	100	97.7778
T	100	96.1111	100	96.1111
U	100	93.3333	97.2067	93.3333
V	100	88.3333	98.3240	88.3333
W	100	92.2222	100	92.2222
X	98.3333	89.4444	98.3240	89.4444
Y	100	91.1111	99.4413	91.1111
Z	100	85	100	85
Average	99.23	93. 4	98.82	93.4

All the mistakes made by single-stroke FUBAR were related to false-positive results. It was caused by the overlap of fuzzy sets describing rule input parameters; the membership values of some input variables could not been distinguished between different (true and false-positive) symbols (similarly to over fitting).

The results have been analyzed in depth including the search for the reason of the false results in the multi-stroke FUBAR. The mistakes of the multi-stroke FUBAR were caused

by inconclusive results. All rules had 0 as the degree of matching for the input, which means that it could not find a rule with parameters (describing a letter) similar to the input. Each recognition process returning with an error could be traced back to the fuzzy sets describing the rule antecedents. The sources of all the false results in the multi-stroke system were pointing at uncovered areas in the antecedents of the fuzzy rules. This means that there was at least one antecedent in each rule with a 0 membership degree for at least one input parameter. This could be solved by tuning the fuzzy sets, covering a wider range over the universal set of the given dimensions or by using a sample based model identification technique like a meta-heuristic optimization algorithm.

The single-stroke FUBAR could reach 99.23% accuracy with the initial rule-base on the validation set, which is well over the user acceptance threshold of 97% [5]. The average recognition rate of the single-stroke FUBAR with a hierarchical rule-base was below the accuracy of the same method using flat a rule-base. The 0.41% drop in the average recognition rate was caused by the used meta-level rules in the hierarchy. The meta-rules are determining which (predefined) subset of the rules should be evaluated in the given case. The used parameter for the meta-rules was the average fuzzified number of points in the third row of the fuzzy grid drawn around the symbol. This influence of the meta-rules could be ruled out or minimized by defining more complex meta-rules, which are able to select the subset more accurately.

Both multi-stroke FUBAR methods with flat and hierarchical rule-bases reached the same 93.4% average recognition rate. This result is below the 97% user acceptance threshold and the results of the $N recognizer [3], but still better, than the accuracy of the Graffiti 2 [4]. In the multi-stroke FUBAR the input parameters of meta-level rules were the number of the sub-strokes representing the characters; the number of sub-strokes is strictly specific for the letters compared to the parameter used in the single-stroke FUBAR. This is why the results with the multi-stroke system did not change like it did in the single-stroke system. It is important to highlight that the results for multi-stroke FUBAR (both flat and hierarchical) could be higher with a better initial rule-base (or an algorithm which identifies it).

5. Conclusions and Future Work

It was shown that after the new FUBAR algorithm was able to recognize multi-stroke alphabets also with a 93.4% average recognition rate. The results indicate that the accuracy should be further increased by the redefinition of the initial rule-base.

Finally in this work, a similar method with the multi-stroke alphabet support using a hierarchical rule-base was presented. The topology of the hierarchy was built based on the number of used strokes. The modified system reached the same accuracy as the original one with the flat rule-base, but in this case, the computational cost of the recognition process was considerably reduced by the limited number of rules to evaluate.

The accuracy of the FUBAR method is moderate compared to the average recognition rate of the modified Palm Graffiti with limited multi-stroke support (known as Graffiti 2) was studied by Költringer and Grechenig in [4] and the $N multi-stroke recognizer introduced by Anthony and Wobbrock in [3].

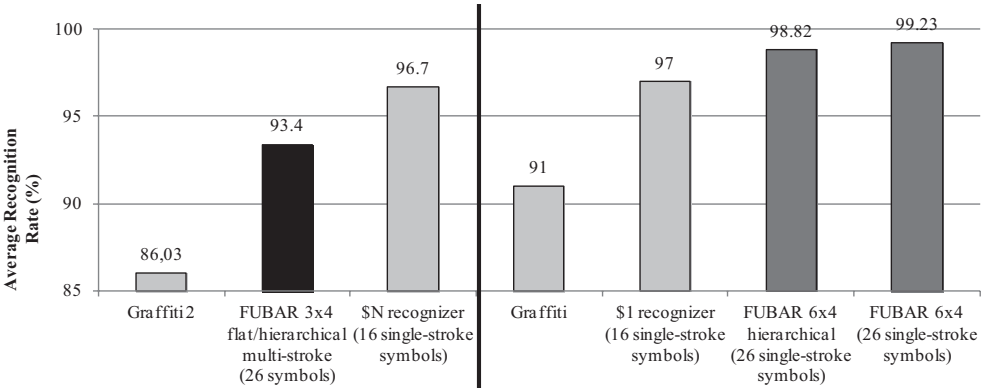

Fig. 6. Average accuracy of various recognition engines

The Graffiti 2 reached only 86.03% accuracy, while 96.7% average recognition rate was achieved for only 16 different single-stroke symbols with the $N recognizer. Both the single--stroke and multi-stroke versions of FUBAR performed well over the results of Graffiti 2. The $N algorithm reached a better average recognition rate compared to the FUBAR method, but the number of symbols was much less and the symbols used during the evaluation of the system were single-stroke. The results are shown in Fig. 6.

This paper was supported by the Hungarian Scientific Research Fund (Hungarian abbreviation: OTKA) K105529, K108405 and TÁMOP-4.2.2.A-11/1/KONV-2012-0012.

References

[1] Kowalski P.A., Kulczycki P., *Data Sample Reduction for Classification of Interval Information using Neural Network Sensitivity Analysis*, Lecture Notes in Artificial Intelligence, Vol. 6304, Springer-Verlag, 2010, 271-272.

[2] Lilik F., Botzheim J., *Fuzzy based Prequalification Methods for EoSHDSL*, Technology, Acta Technica Jaurinensis, Vol. 4, No. 1, Győr 2011, 135-144.

[3] Anthony L., Wobbrock J.O., *A Lightweight Multistroke Recognizer for User Interface Prototypes*, Proc. GI 2010, Ottawa 2010, 245-252.

[4] Költringer T., Grechenig T., *Comparing the Immediate Usability of Graffiti 2 and Virtual Keyboard*, Proc. CHI EA '04, New York, 2004, 1175-1178.

[5] LaLomia M.J., *User acceptance of handwritten recognition accuracy*, Companion Proc. CHI '94, New York 1994, 107.

[6] Tormási A., Botzheim J., *Single-stroke character recognition with fuzzy method*, New Concepts and Applications in Soft Computing SCI, Vol. 417, V.E. Balas et al. (eds.), 2012, 27-46.

[7] Tormási A., Kóczy T.L., *Comparing the efficiency of a fuzzy single-stroke character recognizer with various parameter values*, Proc. IPMU 2012, Part I. CCIS, Vol. 297, S. Greco et al. (eds.), 2012, 260-269.

[8] Sugeno M., Griffin F.M., Bastian A., *Fuzzy hierarchical control of an unmanned helicopter*, Proc. IFSA ’93, Seoul 1993, 1262-1265.

[9] Kóczy T.L., Hirota K., *Approximate inference in hierarchical structured rule-bases*, Proc. IFSA ’93, Seoul 1993, 1262-1265

[10] Zadeh L.A., *Fuzzy sets*, Inf. Control, Vol. 8, 1965, 338-353.

[11] Takagi T., Sugeno M., *Fuzzy identification of systems and its applications to modeling and control*, IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-15, 1985, 116-132.

[12] Tormási A., Kóczy T.L., *Improving the Efficiency of a Fuzzy-Based Single-Stroke Character Recognizer with Hierarchical Rule-Base*, Proc. 13th IEEE International Symposium on Computational Intelligence and Informatics, Óbuda 2012, 421-426.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
2-AC/2013

MARCIN WRÓBLEWSKI*

QUANTUM PHYSICS METHODS IN SHARE OPTION VALUATION

METODY FIZYKI KWANTOWEJ W WYCENIE OPCJI NA AKCJE

Abstract

This paper deals with European share option pricing using quantum physics methods. These contingent claims are usually priced using the Black-Scholes equation. This nonlinear parabolic equation is based on geometric Brownian motion model of the stock price stochastic process. Similar processes also appear among quantum particles and are described by the time-dependent Schrödinger equation. In this paper, the option pricing based on the Schrödinger equation approach is proposed. Using Wick transformation, the Black-Scholes equation is transformed into the equivalent Schrödinger equation. The Fourier separation method is used to find analytical solutions to this equation. The last square method is used to calibrate the Schrödinger model based on real market data. Numerical results are provided and discussed.

Keywords: *option pricing, econophysics, quantum physics methods*

Streszczenie

Artykuł dotyczy wyceny europejskich opcji na akcje z użyciem metod fizyki kwantowej. Tego typu obliczenia zazwyczaj przeprowadza się z wykorzystaniem równania Blacka-Scholesa. To nieliniowe, paraboliczne równanie, oparte jest na geometrycznym modelu ruchu Browna procesu stochastycznego cen akcji. Podobne procesy dotyczą także cząstek kwantowych i są opisane zależnym od czasu równaniem Schrödingera. Zaproponowano wycenę opcji na akcje z wykorzystaniem równania Schrödingera. Używając transformacji Wicka, równanie Blacka-Scholesa przekształcone jest do równoważnej postaci równania Schrödingera. W celu znalezienia analitycznego rozwiązania tego równania, zastosowano metodę separacji zmiennych Fouriera. Metoda najmniejszych kwadratów została użyta w celu kalibracji modelu Schrödingera dla danych giełdowych. Dostarczono i przedyskutowano wyniki numeryczne.

Słowa kluczowe: *wycena opcji, ekonofizyka, metody fizyki kwantowej*

* M.Sc. Marcin Wróblewski, Doctoral Studies, e-mail: wrobel@e-wrobel.pl, Systems Research Institute, Polish Academy of Sciences, Warsaw.

1. Introduction

Collective phenomena appearing in economics, social sciences or ecology pose intriguing theoretical challenges for researchers. In view of the empirical abundance of non-trivial fluctuation patterns and statistical regularities, they are very attractive not only for economists or sociologists but also for physicists, especially statistical physicists. In recent years, many physical theories like theories of turbulence, scaling, random matrix theory or renormalization group were successfully applied to economy giving a boost to modern computational techniques of data analysis, risk management, artificial markets, and macro-economy.

The term econophysics was introduced by H.E. Stanley in the mid-nineties to describe the host of papers written by physicists to explain economic and financial phenomena. Econophysics is regarded as an interdisciplinary research field, applying theories and methods originally developed by physicists in order to solve problems in economics, usually those including uncertainty or stochastic processes (random and disspersion models) and nonlinear dynamics (chaos, criticality, power laws). Both areas involve the study of complex systems formed by a large number of smaller subsystems.

The term econophysics can also be understood to mean the physics of finance, because it attempts to understand the global behavior of financial markets [13] from a scientific standpoint. It has its roots in ancient history. N. Copernicus and I. Newton were two luminaries who applied statistical physical concepts to economic problems. Physics involves trying to understand how macroscopic effects are bought about by a huge number of microscopic interactions, so some of the tools used by statistical physicists can be used to more accurately understand market dynamics. For instance, studies of entropy have been applied to gain a better understanding of salary distribution in a free market. Many similarities have been found between data on stock markets and earthquakes. This could help economists better understand and perhaps even predict stock market crashes.

The following problem is considered in many papers that try to describe economy using quantum mechanics. One of the most pioneering works is presented by [8] and is based on the Black-Scholes [2] transformation into time-dependent Schrödinger [1] equation. The solution is given by applying semiclassical methods, of common use in theoretical physics, to find an approximate analytical solution of the Black-Scholes equation. The semiclassical approximation is performed for different arbitrage bubbles (step, linear, parabolic). This model can be interpreted as a Schrödinger equation in imaginary time for a particle of mass $\frac{1}{\sigma^2}$ with a wave function in an external field force generated by the arbitrage potential.

This paper introduces similar Black-Scholes into the Schrödinger equation transformation, but the final solution of option price is not given using a semiclasscal limit but is given by an analytical function. The asset price function is approximated by *n*-degree polynomial. Arbitrage existence is not considered. The option price is given by the wave function that is solved for the Schrödinger equation in real time for free particle of mass equal to $\frac{2}{\sigma^2}$. A particle interacts with constant potential $\widehat{U} = \frac{r}{2}$.

1.1. Schrödinger equation

The time-dependent Schrödinger equation [1] is a linear partial differential equation, first order in time, second order in the spatial variables. This equation is often written in the form:

$$\widehat{H}\psi(x,t) = i\hbar\frac{\partial\psi(x,t)}{\partial t} \tag{1}$$

where:

$$\widehat{H} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + \widehat{U}(x,t) \tag{2}$$

$\widehat{H}$ is called Hamiltonian and is interpreted as system energy (kinetic and potential energy of the particles constituting the system). $\widehat{U}(x,t)$ is the potential energy operator. In quantum physics, the Schrödinger equation is the same what Newton's second law of motion to classical mechanics. It describes how a physical system will change over time. In classical mechanics, we have positions and momenta of all particles at every time *t*: that give a full description of the system. In quantum mechanics, the information about the system is contained in the solution to Schrödinger's equation, a wave function Ψ. The square of the absolute value of the wave function, $|\psi(x,t)|^2$ gives the probability density for finding the particle at position *x*. But it is also possible to solve Schrödinger's equation for many particle systems and to find wave functions for other observable quantities, for example the momenta of the particles. We want to know if it is possible to describe economic systems in the same way as physical systems. If it is possible to describe economic systems by Schrödinger equation then it is solution – the wave function can explain behavior of economic forms like options.

1.2. Black-Scholes model assumptions

In 1973, Fischer Black and Myron Scholes developed a formula [2] for the valuation of European contingent claims based on a geometric Brownian motion model for the stock price process. Robert Merton [3] developed another method to derive the mentioned formula that turned out to have very wide applicability.

The Black-Scholes model is based on two assets, a (risky) stock with price governed by the stochastic process $S = S_t$, $t \in [0, T]$ and a (riskless) bond with price process $B = B_t$, $t \in [0, T]$.

1.3. Ito rules

Using Ito's [4] formula, we can write the stock price $S(t)$ process and the bond price process equations in the following form:

$$dS_t = \mu(S_t, t)dt + \sigma(S_t, t)dW_t,\ t \in [0, T] \tag{3}$$

where dW_t is a differential of a continuous-time stochastic process called the Wiener [5] process. dW_t is also called Brownian motion. It is one of the best known Lévy processes, so stochastic processes with stationary independent increments, and occurs frequently in physics. It means that process *S* should be treated as a diffusion process. Moreover:

$$\mu(S_t, t) = \mu S_t \quad \text{and} \quad \sigma(S_t, t) = \sigma S_t \tag{4}$$

as well as:

$$dB_t = rB_t dt,\ t \in [0, T] \tag{5}$$

where μ is stock price average value (expected value), σ is standard deviation, and σ^2 is variance.

The key to the correct interpretation of dW_t is to interpret $dW_t = W_{t+dt} - W_t$ as a random variable with a mean of 0 and an infinitesimal small variance dt:

$$dW_t \approx N(0, dt) \tag{6}$$

where $N(\mu, \sigma)$ is a Gaussian distribution [6] of random variable x and is defined as:

$$N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{7}$$

For $\mu = 0$ and $\sigma = dt$ Gaussian distribution becomes a normal distribution of random variable x and is defined:

$$dW_t \approx N(0, dt),\ dt > 0 \tag{8}$$

Now let us write some basic relations called Ito rules, between dW_t and dt:

$$E(dW_t^2) = dt, \quad Var(dW_t^2) = dt^2, \quad E(dW_t) = 0 \tag{9}$$

$$E(dW_t^2) = dt, \quad E(dW_t) = 0 \tag{10}$$

$$E(dW_t^2 dt^2) = 0, \quad dW_t dt = 0, \quad dW_t^2 = dt \tag{11}$$

where E symbol means expected value, and Var is variance.
We assume that dt is infinitesimally small, thus any power of dt can be omitted, in particular:

$$dt^2 = 0 \tag{12}$$

Let us define:

$$dS^2 = dSdS \tag{13}$$

Consider continuous and differentiable function $V(S, t)$. Using Taylor expansion [7], and omitting higher than second order parts, we obtain:

$$dV(S,t) = \frac{\partial V(S,t)}{\partial t}dt + \frac{\partial V(S,t)}{\partial S}dS + \frac{1}{2}\frac{\partial^2 V(S,t)}{\partial^2 t}dt^2 + \tag{14}$$

$$\frac{1}{2}\frac{\partial^2 V(S,t)}{\partial^2 S}dS^2 + \frac{\partial^2 V(\mathrm{s},t)}{\partial t \partial S}dtdS + \ldots$$

$V(S, t)$ is interpreted as the option price which depends on the asset price $S(t)$ and time t. We know by (12) that $dt^2 = 0$, so equation (14) simplifies to:

$$dV(S,t) = \frac{\partial V(S,t)}{\partial t}dt + \frac{\partial V(S,t)}{\partial S}dS + \frac{1}{2}\frac{\partial^2 V(S,t)}{\partial^2 S}dS^2 + \frac{\partial^2 V(s,t)}{\partial t \partial S}dtdS \tag{15}$$

Now, at the basis of (3) and (13) let us calculate dS^2 and $dtdS$:

$$dS^2 = \mu^2(S,t)dt^2 + 2\mu(S,t)\sigma(S,t)dtdW + \sigma^2(S,t)dW^2 = \sigma^2(S,t)dt \tag{16}$$

$$dtdS = \mu(S,t)dt^2 + \sigma(S,t)dtdW = 0 \tag{17}$$

Inserting (16) and (17) into (15) we get:

$$dV(S,t) = \left(\frac{\partial V(S,t)}{\partial t} + \mu(S,t)\frac{\partial V(S,t)}{\partial S} + \frac{1}{2}\sigma^2(S,t)\frac{\partial^2 V(S,t)}{\partial S^2}\right)dt + \sigma(S,t)\frac{\partial V(S,t)}{\partial S}dW \tag{18}$$

For Brownian motion using (4) we obtain from (18):

$$dV(S,t) = \left(\frac{\partial V(S,t)}{\partial t} + \mu S\frac{\partial V(S,t)}{\partial S} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 V(S,t)}{\partial S^2}\right)dt + \sigma S\frac{\partial V(S,t)}{\partial S}dW \tag{19}$$

1.4. Riskless portfolio

Let us define portfolio Π which consists of one option in short the term, and n shares in the long term. The portfolio can be written as:

$$\Pi = -V(S,t) + nS \tag{20}$$

or in differential form:

$$d\Pi = -dV(S,t) + ndS \tag{21}$$

Substituting (4), (3), (18), into (21) we get:

$$d\Pi = -\left(\frac{\partial V(S,t)}{\partial t} + \mu S \frac{\partial V(S,t)}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V(S,t)}{\partial S^2} - n\mu S \right) dt + \left(n - \frac{\partial V(S,t)}{\partial S} \right) \sigma S dW \quad (22)$$

We assume that our portfolio (22) is riskless, so it must be described by a deterministic equation, so cannot contain any stochastic parts. It means that the following statement should be true:

$$n - \frac{\partial V(S,t)}{\partial S} = 0 \quad (23)$$

and it means that the number of shares that are in the portfolio is given by:

$$n = \frac{\partial V(S,t)}{\partial S} \quad (24)$$

It means that the number of shares in the portfolio is equal to the partial derivative of option price at time t. This derivative is also called *delta* and it measures the sensitivity of the value of an option to changes in the price of the underlying stock, assuming all other variables remain unchanged. If a position is delta neutral it means that its instantaneous change in value, for an infinitesimal change in the value of the underlying security, will be zero [14]. Since delta measures the exposure of a derivative to changes in the value of the underlying, a portfolio that is delta neutral is effectively hedged. The overall value will not change for small changes in the price of its underlying instrument.

Equation (22) simplifies to the following form:

$$d\Pi = -\left(\frac{\partial V(S,t)}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V(S,t)}{\partial S^2} \right) dt \quad (25)$$

As we assumed that our portfolio is riskless then the rate of return on this portfolio must be equal to the rate of return on any other riskless instrument; otherwise, there would be opportunities for arbitrage [8]:

$$d\Pi = \Pi r dt \quad (26)$$

where r is the risk-free rate of return. Comparing (25) and (26) we get:

$$-\left(\frac{\partial V(S,t)}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V(S,t)}{\partial S^2} \right) = \Pi r \quad (27)$$

1.5. Black-Scholes equation

Finally substituting (20) and (24) into (27) we get the Black-Scholes equation:

$$\frac{\partial V}{\partial t} + \frac{\sigma^2 S^2}{2}\frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0 \tag{28}$$

with boundary and initial conditions:

$$\text{for } t = 0 : V(S(0), 0) = V_p \tag{29}$$

$$\text{for } t = T : V(S(T), T) = \max\{S(t = T) - K, 0)\} = V_T$$

$$S \in [S_0, S_T]$$

$$t \in [0, T],\ T > 0$$

$$V, S, t \in R$$

where K is the strike price (exercise price) of an option. Parameter $\sigma \in R$ denotes the volatility of the stock's returns. This is the square root of the quadratic variation of the stock's log price process, $r \in R$ is the annualized risk-free interest rate, continuously compounded (the force of interest), $\mu \in R$ is annualized drift rate of S.

2. Black-Scholes to Schrödinger equation transformation

Quantum [12] mechanics is the theory describing the micro world. We want to to apply quantum mechanics in the stock market in which the stock index/option is based on the statistics of the share prices of many representative stocks. Let us consider the index/option as a macro scale object, it is reasonable to view every stock, which constitutes the index, as a micro system. Stock is always traded at certain prices, which presents its corpuscular property. Stock price fluctuates in the market, which presents the wave property. It means that there is particle – wave dualism, we suppose the micro scale stock as a quantum system. Rules are different between quantum and classical mechanics. In order to describe the quantum characters of the stock, we are going to build a price model on the basic hypotheses of quantum mechanics.

We will use several mathematical operations to transform the Black-Scholes equation into Schrödinger's [1] form. If the transformation is possible, then we will be able to use a quantum physics interpretation to explain the economy's options pricing issues. We are going to use quantum methods, because it seems that the real world of economics might not be describable merely in terms of conventional macroeconomic variables (unemployment

rate, GNP, aggregate demand etc.) that is why conceptual innovation is needed. An economic system is in some ways like a mechanism as is recognized in all theories. In the following, we present some very tentative, preliminary conjectures about what an economic theory based on the quantum physics analogy. In this paper, the transformed equation will be called: the Black-Scholes-Schrödinger (BSS) equation.

2.1. Introducing new variables

Let us start by introducing the new variable:

$$x = LnS,\ x \in R \tag{30}$$

We obtain:

$$\frac{\partial V}{\partial S} = \frac{1}{S}\frac{\partial V}{\partial x} \tag{31}$$

and:

$$\frac{\partial^2 V}{\partial S^2} = \frac{\partial}{\partial S}\left(\frac{1}{S}\frac{\partial V}{\partial x}\right) = \frac{-1}{S^2}\frac{\partial V}{\partial x} + \frac{1}{S}\frac{\partial^2 V}{\partial x^2}\frac{\partial x}{\partial S} = \frac{1}{S^2}\left(\frac{\partial^2 V}{\partial x^2} - \frac{\partial V}{\partial x}\right) \tag{32}$$

Using (31), Black-Scholes equation (28) takes the form:

$$\frac{\partial V}{\partial t} + \frac{\partial^2}{2}\frac{\partial^2 V}{\partial x^2} + \frac{\partial V}{\partial x}\left(r - \frac{\sigma^2}{2}\right) - rV = 0 \tag{33}$$

with boundary and initial conditions:

$$\text{for } t = 0 : V(x(0), 0) = V_p \tag{34}$$

$$\text{for } t = T : V(x(T), T) = \max\{x(t = T) - K, 0)\} = V_T$$

$$x \in [Ln(S_0), Ln(S_T)]$$

$$t \in [0, T],\ T > 0$$

$$V, x, t \in R$$

We want to exclude from (33) the expression that contains $\left(r - \frac{\sigma^2}{2}\right)$ element, because we want to liken (33) to Schrödinger form.

Let us introduce new variable y:

$$y = x - \left(r - \frac{\sigma^2}{2} \right) t \tag{35}$$

Previously we had assumed that x does not depend on time. Data related to the market showed that stock price S is time dependent, so the logarithm of that price described in this paper as x, should be also time dependent. The variation of x using time dependent part, generates a new variable y which is time dependent. This procedure was necessary to reproduce the behavior of the real market.
A similar approach has been used in [8].
Let us calculate new partial derivatives:

$$V_y \overset{\text{def}}{=} \frac{\partial V}{\partial y} = \frac{\partial V}{\partial x} \frac{\partial x}{\partial y} \tag{36}$$

Second partial derivative is given by:

$$\frac{\partial^2 V}{\partial y^2} = \frac{\partial}{\partial y} \left(\frac{\partial V}{\partial y} \right) = \frac{\partial V_y}{\partial y} = \frac{\partial V_y}{\partial x} \frac{\partial x}{\partial y} = \frac{\partial}{\partial x} \left(\frac{\partial V}{\partial x} \frac{\partial x}{\partial y} \right) \frac{\partial x}{\partial y} \tag{37}$$

Because $\frac{\partial x}{\partial y} = 1$ and using (35), equation (37) simplifies to $\frac{\partial^2 V}{\partial y^2} = \frac{\partial^2 V}{\partial x^2}$.
Designating x from (35) and use it to calculate time dependent partial derivative of option price:

$$\frac{\partial V}{\partial t} = \frac{\partial V}{\partial x} \frac{\partial x}{\partial t} = \frac{\partial V}{\partial x} \left(r - \frac{\sigma^2}{2} \right) \tag{38}$$

Inserting (37), (38) into (33), we get:

$$2 \frac{\partial V}{\partial t} + \frac{\sigma^2}{2} \frac{\partial^2 V}{\partial y^2} - rV = 0 \tag{39}$$

and:

$$-\frac{\sigma^2}{2} \frac{\partial^2 V}{\partial y^2} - rV = 2 \frac{\partial V}{\partial t} \tag{40}$$

Let us put (40) into Heat-equation form:

$$\left(-\frac{\sigma^2}{4} \frac{\partial^2}{\partial y^2} + \frac{r}{2} \right) V = \frac{\partial V}{\partial t} \tag{41}$$

We want to transform (41) which represents the Heat-equation [15] into the Schrödinger form, to do this, we have to introduce a new variable τ defined as imaginary time. Using the Wick rotation, we get:

$$\tau = -it \tag{42}$$

Finally (41) equation is represented by Schrödinger's equation:

$$\left(-\frac{\sigma^2}{4}\frac{\partial^2}{\partial y^2}+\frac{r}{2}\right)V = i\frac{\partial V}{\partial \tau} \tag{43}$$

2.2. Quantum interpretation

Equation (43) can be interpreted as Schrödinger's equation for free particle of mass equal to:

$$m = \frac{2}{\sigma^2} \tag{44}$$

and our particle interacts with constant potential:

$$\widehat{U} = \frac{r}{2} \tag{45}$$

Remark: equation (43) is written in natural units. The use of $\hbar = c = 1$ units can simplify particle physics notation considerably.
We have proved that it is possible to use quantum physics to describe the option pricing. We showed that it is possible to write Schrödinger's equation for the selected option. Now we have to find its solution.

2.3. Fourier separation method

We want to find the solution of (43), after using the Fourier separation method [9], we get:

$$V(y, \tau) = \psi(y)\, T(\tau) \tag{46}$$

where the boundary and the initial conditions are below:

$$\Psi(x) \in [\Psi_0 = \Psi(x_0), \Psi_T = \Psi(x_T)] \text{ – boundary condition} \tag{47}$$

$$T(\tau) \in [T_0^i, T_t^i] \text{ – initial condition}$$

We obtain two independent equations that describe dependency on variable y and imaginary time variable τ:

$$\left[-\frac{1}{2m}\frac{\partial^2}{\partial y^2}+\widehat{U}\right]\psi(y)=k\psi(y) \tag{48}$$

$$i\frac{\partial T(\tau)}{\partial \tau}=k \tag{49}$$

where m and $\widehat{U}$ are given by (44) and (45). Constant k can be interpreted as particle energy, and we assume that:

$$k>0 \quad \text{and} \quad k<\widehat{U} \tag{50}$$

The solution for (49) has the form:

$$T(\tau)=T_0^{\,i}\exp(-ik\tau) \tag{51}$$

As we remember, all computations in physics and economy are based on time t that is real. That is why we cannot use (51) form as it is based on imaginary time τ. This has to be replaced by an equivalent equation that uses real time t. Going back to real variables means that (51) turns into:

$$T(t)=T_0\exp(-ik) \tag{52}$$

Our general form (46) moves into the following equation:

$$V(y,t)=\psi(y)\,T(t) \tag{53}$$

with boundary conditions:

$$\Psi(y)\in[\Psi_0=\Psi(y_0),\Psi_T=\Psi(y_T)] \tag{54}$$

and initial condition:

$$T(t)\in[T_0,T_t] \tag{55}$$

Constant T_0 is determinated from the following initial condition:

$$T_0=T(t=0)\neq 0 \tag{56}$$

Function $\psi(y)$ is given by:

$$\psi(y)=D\exp\left(\sqrt{2}\sqrt{m\left(\widehat{U}-k\right)}y\right)+E\exp\left(\sqrt{2}\sqrt{m\left(\widehat{U}-k\right)}y\right) \tag{57}$$

where D and E are constants given by the boundary condition.
Remark: In order to represent a physically observable system, the wave function must satisfy certain constraints [10]:
– it must be a solution to the Schrödinger equation,
– it must be a continuous function of y,
– the slope of the function in y must be continuous. Specifically $\frac{\partial \psi}{\partial x}$ must be continuous,
– it must be normalizable. This implies that the wave function approaches zero as x approaches infinity.

To be consistent with the mentioned assumptions, we have to remove the part that grows to infinity. This means that constant D should be equal to 0. From the economy part it means that if t approaches infinity, the option price function should approach 0. Even though the stock price approaches infinity, the option price will not grow to infinity.

2.4. General solution for the Black-Scholes-Schrödinger equation

From (52), (53) and (57) we have general solution for Black-Scholes-Schrödinger equation:

$$V(y,t) = C\exp(-kt)E\exp\left(\sqrt{2}\sqrt{m\left(\hat{U}-k\right)}y\right) \tag{58}$$

which is equal to:

$$V(y,t) = A\exp\left(-kt-\sqrt{2}\sqrt{m\left(\hat{U}-k\right)}y\right) \tag{59}$$

for given mass (44) and potential (45), equation (59) becomes:

$$V(y,t) = A\exp\left(-kt-\frac{2\sqrt{\left(r/2-k\right)}}{\sigma}y\right) \tag{60}$$

where:

$$A = CE \tag{61}$$

The key to compute BSS equation (60) for the given option is to understand $y(t)$ behavior. We know that $y(t)$ depends on $Ln(S)$ and also contains the linear time dependent part.

Unfortunately we do not know the exact form of $Ln(S)$, that is why $y(t)$ will be interpolated by the first degree polynomial function and then will be used to predict option price using equation (60). Additionally, similar computations will be performed, but $y(t)$ will be interpolated by the second degree polynomial function. Fig. 1, shows $Ln(S)$ behavior for the WIG20 index which is listed on Polish market.

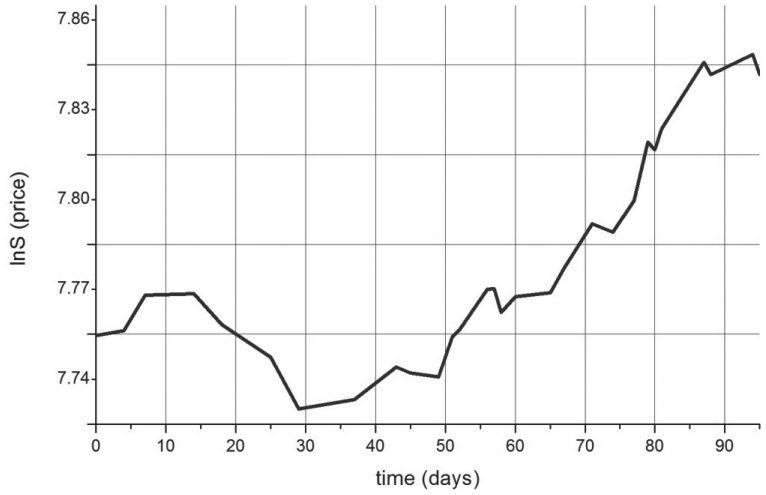

Fig. 1. *Ln*(*S*) time behavior

We will also check how our model fits to market data, we will check this by calculating the residual sum of squares and Pearson's correlation factor.

3. Numerical computations

In this chapter, we want to check how our BSS model (60) fits market data. We will perform computations for cases. The first case assumes that $y(t)$ is modelled by the first order polynomial, the second case assumes that $y(t)$ is modelled by the second order polynomial. Following computations and figures have been generated using scipy library. The mentioned library was implemented into python scripts. The WIG20 index is used as asset price, the OW20F3280 is used as anoption.
Remark: after inserting interpolated $y(t)$ form into BSS equation (60), our $V(y, t)$ becomes $V(t)$ function.

3.1. $y(t)$ is modelled by first order polynomial

Equation (60) can be written in the following form:

$$V(y,t) = \exp\left(-kt - \frac{2\sqrt{(r/2-k)}}{\sigma} y + g \right) \tag{62}$$

where g is constant.
Let us assume that $y(t)$ behavior is interpolated by the first order polynomial:

$$y(t) = f(t) \approx zt + d \tag{63}$$

where z and d are constants that will be designated on the basis of market data using the last squares method.
Let us put (63) into (62):

$$V(t) = \exp\left(-kt - \frac{2\sqrt{(r/2-k)}}{\sigma}(zt+d) + g \right) \tag{64}$$

Additionally we see that (64) can also be interpolated, but using an exponential function with the linear expression:

$$V(t) = \exp\left(-kt - \frac{2\sqrt{(r/2-k)}}{\sigma}(zt+d) + g \right) = \exp(at+b) \tag{65}$$

Constant a and b will be designated on the basis of market data using the last squares method, if z and d are also known, then we can designate constant k by solving following equations:

$$a = -\left(k + \frac{2\sqrt{(r/2-k)}}{\sigma} z \right) \tag{66}$$

and:

$$b = g - \frac{2\sqrt{(r/2-k)}}{\sigma} d \tag{67}$$

We can choose only that values for k that are allowed by (50) equation. Fig. 2 shows the WIG20 index interpolated by the first order polynomial, Fig. 3 shows the OW20F3280 interpolation using $\exp(at + b)$ function. OW20F3280 is the call option for WIG20 asset.

We have calculated residual sum of squares and Pearson's correlation factor for the BSS model for the OW20F3280 option. The residual sum of squares is equal to **210.92**, and the correlation is equal to **0.87**.
Now, we want to check if we should choose the second order polynomial to interpolate the WIG20 index then we will have better results in OW20F3280 option pricing.

3.2. $y(t)$ is modelled by the second order polynomial

Let us assume that $y(t)$ behavior is interpolated by the second order polynomial:

$$y(t) = f(t) \approx zt^2 + td + e \tag{68}$$

where z, d and e are constants that will be designated on the basis of market data using the last squares method.

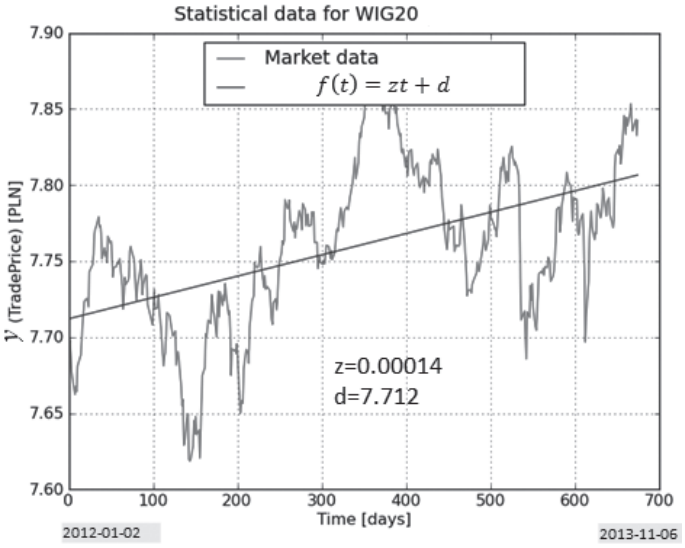

Fig. 2. WIG20 index interpolated by the first order polynomial

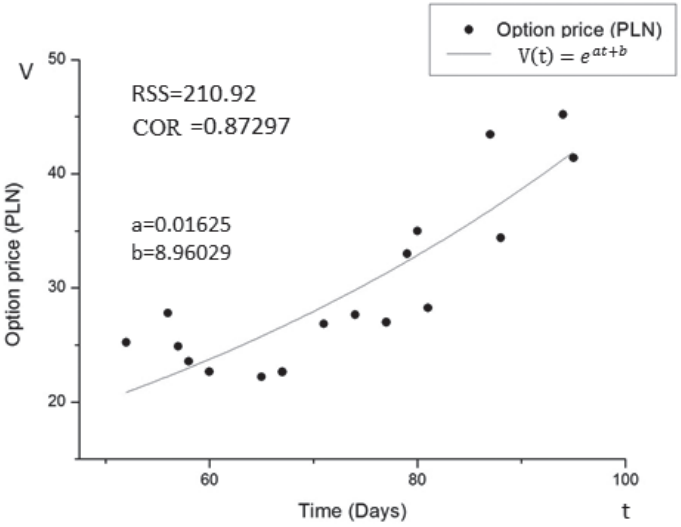

Fig. 3. OW20F3280 interpolation using exp(*at* + *b*) function

Let us insert (68) into (62):

$$V(t) = \exp\left(-kt - \frac{2\sqrt{(r/2-k)}}{\sigma}\left(zt^2 + td + e\right) + g \right) \tag{69}$$

Additionally we see that (69) can be also interpolated, but using an exponential function with quadratic expression:

$$V(t) = \exp\left(-kt - \frac{2\sqrt{(r/2-k)}}{\sigma}\left(zt^2 + td + e\right) + g \right) = \exp\left(at^2 + bt + c\right) \tag{70}$$

Constant a, b and c will be designated on the basis of market data using the last squares method.

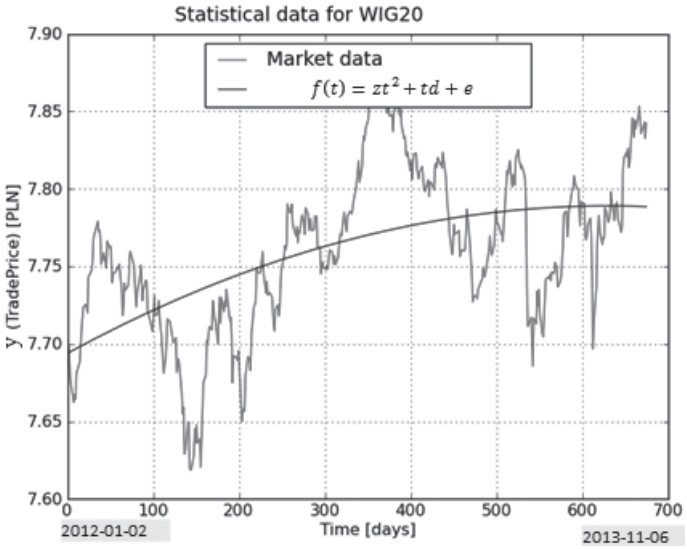

Fig. 4. WIG20 index interpolated by the second order polynomial

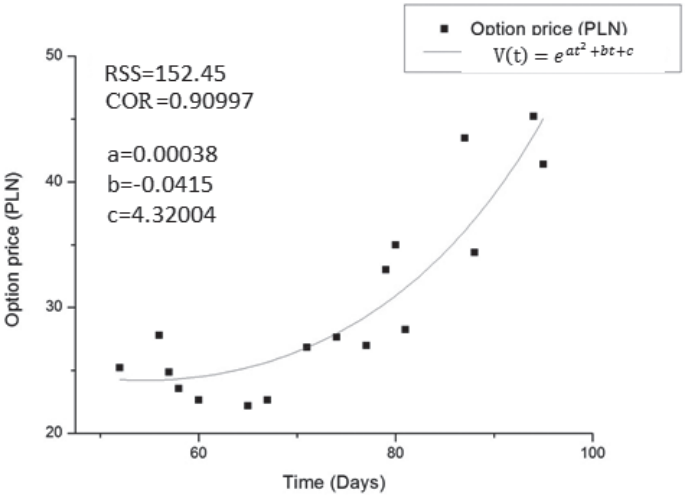

Fig. 5. OW20F3280 interpolation using exp($at^2 + bt + c$) function

Constant k can be designated from:

$$a = -z\frac{2\sqrt{(r/2-k)}}{\sigma} \tag{71}$$

$$b = -\left(k + \frac{2\sqrt{(r/2-k)}}{\sigma}d\right) \tag{72}$$

and:

$$b = g - \frac{2\sqrt{(r/2-k)}}{\sigma}e \tag{73}$$

We can choose only that values for k that are allowed by (50) equation. Fig. 4 shows the WIG20 index interpolated by the second order polynomial, Fig. 5 shows OW20F3280 interpolation using $\exp(at^2 + bt + c)$ function.

We have calculated the residual sum of squares and Pearson's correlation factor for the BSS model for the OW20F3280 option. The residual sum of squares is equal to **152.45**, and the correlation is equal to **0.91**. As we see, choosing the second order polynomial in the WIG20 interpolation, gave better results.

4. Concluding remarks

We have verified that it is possible to transform the Black-Scholes equation into Schrödinger's form. It has been performed by introducing new variables and also by making additional transformations and simplifications. We tried to describe trade price relations in time by using polynomial interpolation. Increasing rhe polynomial order gave us better results. We have developed our own procedures based on Python scripts and this approach will be used in further research.

Our next calculations will be related to searching for the new functions that can describe $y(t)$ behavior. We will test the new list of wave functions with trade price modelled by different functions and we will calculate if that model fits market data. We will use numerical experiments to check if that approach is reasonable. Otherwise we would need to have a fundamental change in our assumptions. One of those assumptions is to take into consideration, that arbitrage [11] exists.

References

[1] Schrödinger E., *Quantisierung als Eigenwertproblem (Erste Mitteilung)*, Ann. Phys. 79, 1926, 361-376.

[2] Black F., Scholes M., *The pricing of options and corporate liabilities*, Journal of Political Economy, 81 (3), 637-654.

[3] Merton R., *Option pricing when underlying stock returns are discontinuous*, Journal of Financial Economics, Vol. 3, Issue 1–2, 1976, 125-144.
[4] Kishimoto M., *On the Black-Scholes Equation: Various Derivations*, MSE 408 Term Paper, 2-3.
[5] Papoulis A., *Wiener-Lévy Process, §15-3 in Probability, Random Variables, and Stochastic Processes*, 2nd ed. McGraw-Hill, New York 1984, 292-293.
[6] Honarkhah M., Caers J., *Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling*, Mathematical Geosciences, 42, 2010, 487-517.
[7] Dzhrbashyan M.M., *On integral representation and expansion in generalized Taylor series of entire functions of several complex variables*, Mat. Sb. (N.S.), 41(83), 3, 1957, 257-276.
[8] Contreras M., Pellicer R., Villena M., Ruiz A., *A quantum model of option pricing: When Black Scholes meets Schrodinger and its semi-classical limit*, Physica A 389, 2010, 5447-5459.
[9] Hazewinkel M., ed., *Fourier method*, Encyclopedia of Mathematics, Springer, 2001.
[10] Peleg Y., Pnini R., Zaarur E., Hecht E., *Quantum mechanics. Schuam's outlines*, (2nd ed.), McGraw Hill., 2010, 68-69.
[11] Shleifer A., Vishny R., *The limits of arbitrage*, Journal of Finance, 52, 1997, 35-55.
[12] Bohr N., *Discussion with Einstein*, [in:] Schilpp P.A., ed., Albert Einstein: Philosopher-Scientist, 235.
[13] Challet D., Marsili M., Zhang Y.C., *Modeling Market Mechanism with Minority Game*, Physica A-276, 2000, 284-315.
[14] McCrary S.A., *Chapter 1: Introduction to Hedge Funds. How to Create and Manage a Hedge Fund: A Professional's Guide*, John Wiley & Sons, 2002, 7-8.
[15] Cherpakov P.V., *Periodic solutions of the heat equation*, Izv. Vyssh. Uchebn. Zaved. Mat., No. 2, 1959, 247-251.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
2-AC/2013

VOLODYMYR FEDAK, ADRIAN NAKONECHNY*

SPATIO-TEMPORAL ALGORITHM FOR CODING ARTIFACTS REDUCTION IN HIGHLY COMPRESSED VIDEO

PRZESTRZENNO-CZASOWY ALGORYTM REDUKCJI ARTEFAKTÓW W WYSOCE SKOMPRESOWANYM FILMIE

A b s t r a c t

Images and video are often coded using block-based discrete cosine transform (DCT) or discrete wavelet transform (DWT) which cause a great deal of visual distortions. Restoration of image sequences can obtain better results compared to restoring each image individually, provided that the temporal redundancy is adequately used. In this article, efficient approach for artifacts reduction has been presented. In order to enhance the overall video quality, the proposed approach uses image sequence redundancy. Spatial and temporal information is used for the video de-noising process.

Keywords: *artifacts, spatio-temporal postprocessing, wavelet transformation*

S t r e s z c z e n i e

Zdjęcia i filmy są często kodowane za pomocą blokowej dyskretnej transformaty kosinusowej (DCT) lub dyskretnej transformaty falkowej (DWT), które powodują duże zakłócenia wizualne. Przywrócenie sekwencji obrazów pozwala uzyskać lepsze wyniki niż przywracanie każdego obrazu osobno, pod warunkiem odpowiedniego użycia redundancji czasowej. W niniejszym artykule zaprezentowano efektywne podejście do redukcji artefaktów. W celu zwiększenia ogólnej jakości obrazu omawiane podejście wykorzystuje redundancję sekwencji obrazu. Do procesu odszumiania filmu wykorzystano informacje czasowe i przestrzenne.

Słowa kluczowe: *artefakty, obróbka przestrzenno-czasowa, transformacja falkowa*

* M.Sc. Volodymyr Fedak, e-mail: vfedak@softserveinc.com, Prof. D.Sc. Ph.D. Andrian Nakonechnyy, Department of Computerized Automatic Systems, Lviv Polytechnic National University.

1. Introduction

Every time we need to obtain, process, and deliver information. This information is not limited to text files or sample messages, nevertheless various visual pieces of information could be transmitted including image and video files. However, transmission channels have limited bandwidth and storage devices have a limited capacity. Digital video is broadcast and stored in an encoded form, so it requires less information (bits) than the original. At low bit-rates, the coarse quantization exploited during compression results in visually annoying coding artifacts [1].

Compression artifact is a particular class of data errors that are usually the consequence of quantization in lossy data compression. These distortions can be classified into the following types:

Blocking artifacts. Such types of image/video distortions are the most visible degradation of all artifacts. This effect is caused by all block-based coding techniques. It is a well-known fact that all compression techniques divide the image into small blocks and then compress them separately. Due to the coarse quantization, the correlation among blocks is lost, and horizontal and vertical borders appear.

Ringing artifacts. The ringing effect is caused by the quantization or truncation of the high frequency coefficients and can also come from improper image restoration operations. Ringing artifacts are visible for all compression techniques especially when image or video is transformed into frequency domain. Moreover, it appears as distortion along sharp edges in the video sequence. This artifact occurs very often when the DWT encoder is used. Furthermore, it may be observed after the image or video has been de-coded using a frequency coder.

Blur effect. Blurring is another artifact resulting from the absence of high frequencies in low bit rate video. It appears around the sharp edges and all image details become blurred. This effect is very similar to the ringing artifact, and sometimes it is hard to distinguish between them.

Flickering is one of the most annoying temporal artifacts that appears in video. As it is widely known, modern algorithms encode video as a sequence of images. The first frame from this sequence is a key frame (I), others are additional (previous [P] and subsequent [B]) frames. All sequences are encoded by motion-compensated algorithms. When an observer watches the de-coded video, the flickering effect is noticeable due to the difference between key frames (I) and other frames (P, B).

Different techniques could be used to reduce most annoying artifacts and all of these techniques could be divided by filtering domain (spatial, frequency, temporal). Different authors provide versatile methods of image/video quality improvements and sometimes the most challenging task is to chose the necessary technique. However, the most promising results are shown by patch-based methods that use image/video self-similarity for the artifacts reduction task.

In general, all post-processing methods (that use an image/video redundancy) could be divided into two types: those that use temporal information; those that only use spatial information. Having several images of the same scene can be greatly beneficial to the restoration results. The first step in exploiting temporal redundancy is inferring the connection between the images. This connection is what sets apart treating an image sequence from treating a random set of images. The connection is usually inferred by estimating the motion

between the frames in the sequence, i.e., detecting the location of each pixel in each image in every other image [3]. In some cases, when the motion is of global nature (e.g., an affine transform), it is relatively simple to accurately estimate the motion trajectories. However, most sequences contain very complex motion patterns of non-rigid shapes and with many occlusions. In such cases, motion estimation is a severely under-determined problem, and is very prone to errors and inaccuracies. Several recent algorithms developed for the de-noising of the image sequences show that example-based methods that are able to bypass the classic explicit motion estimation need [4–6]. Spatial filtering is commonly used for noise and artifacts reduction. However, some artifacts, such as temporal flickering or severe blockiness, cannot be removed efficiently using only spatial techniques. In order to remove highly resistant artifacts, information from adjacent frames should be used. Thus, our algorithm relying on the motion estimation attempt to detect areas where the motion estimation is reliable, and turn to spatio-temporal image sequences processing mechanisms for those areas.

In this article, an efficient algorithm based on spatio-temporal filtering for artifacts reduction has been presented. The proposed algorithm can reduce the most annoying effects such as: 'blockiness', 'flickering' and 'ringing'. In order to diminish artifacts in video sequences, our approach tries to take advantage of the redundancy and self-similarity of the image sequences. A true motion-estimation algorithm is required to effectively use temporal information. Therefore, one existing motion-estimation algorithm is used and functionality is added to determine the quality of each motion vector.

2. Existing approaches to artifacts reduction

In modern digital systems and video broadcast chains, video compression is applied to reduce bandwidth or storage size. Post-processing of the decoded image sequence is an acceptable technique to achieve a better perceived picture quality [10]. Furthermore, modern consumer vision products like televisions and PCs use image enhancement and restoration techniques to improve the objective and subjective picture quality. All postprocessing algorithms and methods can be divided into the following types [1]:

– Spatial filtering;
– Filtering in the frequency/wavelet domain;
– Temporal filtering;
– Hybrid algorithms (mainly combines spatial and frequency filtering).

Many approaches have been proposed in the literature aimed at the alleviation of the blocking artifacts in the images and video. Spatial algorithms modify image pixel values. These approaches are usually used together with the edge detection algorithms to prevent the blurring effect. As nowadays a great number of algorithms have been developed, it would be rational to overview these approaches due to which completely versatile solutions can be reached.

With the purpose of improving image and video quality authors in [12] proposed the algorithm that uses local statistics of transform coefficients. The authors investigated that pixel brightness diversity among blocks is greater than within one block, and border pixels are filtered by the spatial algorithm. This approach reduces the blocking effect from the image and simultaneously introduces the additional blur to the image's edges.

In [10], authors used local statistics as a means of differentiation between monotone and edge blocks and introduced a generic filter for the removal of blocking artifacts and the staircase effect. Monotone blocks contain less spatial details than the edge blocks. They propose to use two-dimensional filtering that is applied for monotone blocks and one-dimensional directional filtering for the edge blocks.

A new pixel classification-based approach for the block artifacts reduction has been proposed in [13]. Instead of classifying each block of fixed size to smooth region or edge region, they distinguish each pixel using the binary edge map from the edge detection process. They reduce grid noise in the smooth region using an adaptive filter.

Most encouraging results could be received using the NLM approach [7]. The efficiency of this algorithm is proven in many different areas and this algorithm tries to take advantage of the redundancy and self similarity of the image. This approach will be discussed in the next sections of this article.

Frequency algorithms transform image or video (sequence of images) to frequency domain and modify DCT (Discrete Cosine Transform) or DWT (Discrete Wavelet Transform) coefficients. These approaches are very efficient but of high complexity because image and video signals have to be transformed from spatial to frequency domain and vice versa. Authors in [10] proposed a adaptive algorithm of blocking artifacts reduction in DCT domain. This algorithm performs filtering using following steps:

– The image is divided into edge and monotone areas. A sobel edge detector [14] is used for this purpose;
– Reduce blocking artifacts in non-edge areas. Horizontal and vertical smoothing filters in the spatial domain is used;
– Apply Filter Tao [15] in the edge areas;
– Transform image to the original format. Quantization Constraints.

The effect of averaging the spatially closest pixels can also be achieved in the Fourier domain. The average of the spatially closest pixels is then equivalent to the cancellation of the high frequencies. As the analogous spatial filter, this cancellation leads to the blurring of the image and a Gibbs effect. The optimal filter in the Fourier domain is the Wiener filter which does not cancel the high frequencies but attenuates them all.

In the wavelet domain, the noise is uniformly spread throughout the coefficients, while most of the image information is concentrated in the few largest ones (sparsity of the wavelet representation) [19–20]. The most straight-forward way of distinguishing information from noise in the wavelet domain consists of thresholding the wavelet coefficients. The soft-thresholding filter is the most popular strategy and has been theoretically justified in [21]. They proposes a three steps denoising algorithm:

– The computation of the forward WT;
– The filtering of the wavelet coefficients;
– The computation of the IWT of the result obtained.

Consequently, regarding the three steps denoising algorithm, there are two tools to be chosen: the WT (Wavelet Transform) and the filter. In [22] the UDWT (Undecimated Discrete Wavelet Transform) was used, in [23] the DTCWT (Dual Tree Complex Wavelet Transforms), and in [24] the DWT.

From the first category can be mentioned the hard-thresholding filter that minimizes the Min-Max estimation error and the Efficient SURE-Based Inter-scales Point-wise Thresholding Filter [24], which minimizes the Mean Square Error (MSE). To the second category belong

filters obtained by minimizing a Bayesian risk under a cost function, typically a delta cost function (MAP estimation [25]) or the minimum mean squared error [22]. The denoising algorithms proposed in [24] exploit the inter-scale dependence of wavelet coefficients. The method proposed in [22] takes into account the intra-scale dependence of wavelet coefficients as well. The statistical distribution of the wavelet coefficients changes from scale to scale. The coefficients of the WT have a heavy tailed distribution.

In [11], the authors introduced the wavelet-based de-blocking and de-ringing the algorithm for artifacts suppression. Based on a theoretical analysis of the blocking artifacts, the proposed algorithm is able to take into account the statistical characteristic of block discontinuities, as well as the behaviour of wavelet coefficients across scales for different image features to suppress both the blocking and ringing artifacts.

Temporal filtering is used to diminish different types of artifacts based on temporal information. Furthermore, these techniques are very often used with spatial and frequency algorithms (hybrid algorithms).

The provided review of different approaches demostrates the level of variance for different postprocessing algorithms and methods that were proposed by the last decade. And the main task is to chose a right filering approach that provides the most promising results. Non-Local means filtering has proven efficiency and provides the most promisng results [26, 27], that's why it's used in this research. Hovewer it is worth conducting additional research to compare this approach with other most promising wavelet based algorithms.

2.1. Image filtering using Non-Local Means

All image and movie filters which are intended to reduce noise by averaging similar pixels are considered to be neighbourhood filters. Noise reduction can thus be achieved by averaging the pixels which have received the same original grey level value. The NLM algorithm removes the noise while keeping all this meaningful image information. For this purpose, the NLM algorithm tries to take advantage of the redundancy and self-similarity of the image. Most image details occur several times; each small window in a natural image has many similar windows in the same image.

As example see Fig. 1 from [7].

Fig. 1. The similar image patches within the same image. Most image elements appear repeatedly. Each different rectangle indicates a squares in the image which are almost indistinguishable from the set of rectangles with the same color

The NLM algorithm is an improvement for bilateral filtering. The bilateral and the NLM filters are two very successful image de-noising filters. Both the bilateral and the NLM filters are based on the assumption that the image contents are likely to repeat themselves within some neighbourhood. Therefore, the de-noising of each pixel is done by averaging all pixels in its neighbourhood.

The NLM algorithm estimates the value of x as an average of the values of all the pixels. The probability that y is similar to x is determined by looking at the difference in the luminance value and the difference in position between x and y in the neighbourhood filters.

Given a discrete noisy image $v = \{v(i) | i \in I\}$, the estimated value $\mathrm{NL}(v)(i)$ is computed as a weighted average of all the pixels in the image:

$$\mathrm{NL}(v)(i) = \sum_{j \in I} w(i, j)\, v(j) \tag{1}$$

The neighbourhood of a pixel x is defined as the set of pixels in a sequence in which each pixel has a surrounding window similar to the window around x. All pixels in this neighbourhood can be used for predicting x. The NLM filter is defined as:

$$\mathrm{NL}_h\{\underline{x}\} = \frac{1}{C(\underline{x})} \cdot \sum_{\underline{y} \in Q(\underline{x})} z(\underline{x}) \cdot e^{-\frac{\left\|\underline{N}(\underline{x}) - \underline{N}(\underline{y})\right\|_2^2}{h^2}} \tag{2}$$

where:

$C(\underline{x}) = \sum_{\underline{y} \in Q(\underline{x})} e^{-\frac{\left\|\underline{N}(\underline{x}) - \underline{N}(\underline{y})\right\|_2^2}{h^2}}$ – a normalizing constant,

$N(x)$ – a vector which contains the pixels in the window surrounding pixel x,

$Q(x)$ – a search window around x, in which the neighbourhood of x is searched,

The window $N(x)$ – contains $Sx \cdot Sy$ pixels,

Search window $Q(x)$ – contains $Ax \cdot Ay$ pixels.

Considering the previous research discussed above (all pros. and cons.), our algorithm will need to meet the following requirements:

- As the most encouraged technique for image/video filtering is the algorithms that use image/video redundancy to restore its content, our algorithms should also use redundancy spatial and temporal to restore the image sequence (video) content;
- As algorithm uses information from the temporal domain, the technique must model the motion compensation;
- Information from the temporal domain shouldn't add additional noise;
- To avoid the modification of existing decoders, the post-processor shouldn't use coding parameters.

Based on the specified assumptions and tasks, a new algorithm for artifacts reduction has been developed and presented in this article.

3. Spatio-temporal algorithm for coding artifacts reduction

The NLM filter is based on the assumption that image content is likely to repeat itself within some neighborhood [7]. All pixels that have similar surrounding windows can be used for predicting the luminance of the original scene. Originally, the NLM was designed as the spatial filter. In this way the NLM takes advantage of redundancy that is presented in the spatial domain.

Extension of the NLM to the temporal domain gives more information for the NLM to retrieve the original frame. This algorithm will take advantage over both temporal and spatial domains.

On the one hand, providing more information gives a grater possibility of the NLM retrieving the original frame with higher quality but, on the other hand, it can cause some other undesirable effects. However, this temporal information should be carefully checked before the filtering process. The main goal is not to provide flawed information from the temporal domain, but only to provide useful data. This step can guarantee that no additional noise has been added to the processed frame.

In order to guarantee that no additional noise is added, a true motion estimation algorithm is used for searching motion vectors [2]. This algorithm uses a custom model to verify the quality of each motion vector.

In [18] we presented an original idea/approach of Spatio-Temporal filtering with the motion vectors quality determination. This algorithm was evolved and some part of the initial approach was simplified due to performance reasons. It was determined that for the majority of video signals, 3DRS Motion estimation is good enough and itial step with initial motion vectors finding (based on Gabor vawelets) very rarely has influence to overal filtering.

The general flow chart of the proposed Spatio-temporal algorithm for artifacts reduction is depicted on Fig. 2. This algorithm can be divided into the following steps (additional information about these steps is presented in the next sub-chapters of this article):

– Motion estimation 3DRS. True motion estimation algorithm is used for searching motion vectors [14];
– Determine type of filtering. If motion vector is consistent (error value less than some Threshold), additional temporal information will be used due to it having at an advantage over spatial information, otherwise only temporal information would be used in the filtering process;
– Filtering process. NLM is used as a core algorithm for filtering.

In case motion vector is consistent, additional temporal information will be used which will have advantage over spatial information. If motion vector quality is turned up within the specific fixed range (it is not a final true motion vector but can be used as a temporal candidate), this area will be filtered in the same way as spatial candidates, otherwise (motion vector is not true) only information from the spatial domain will be used [2]. In this implementation, the previous and the next frames are used.

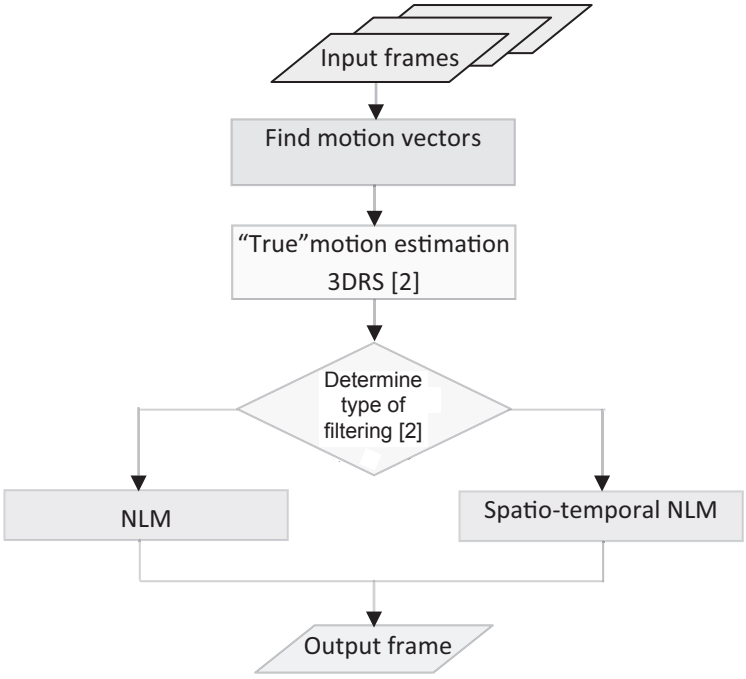

Fig. 2. Flow chart of the spatio-temporal Non-Local Means algorithm

$$\mathrm{NL}_h\{\underline{x}\} = \frac{1}{C(\underline{x})} \cdot \sum_{\underline{y} \in Q(\underline{x})} z(\underline{x}) \cdot e^{-\frac{\|\underline{N}(\underline{x}) - \underline{N}(\underline{y})\|_2^2}{h^2}} \cdot z(\underline{x}) \cdot e^{-\frac{\|\underline{N}(\underline{x}) - \underline{N}_{\pm 1}(\underline{x})\|_2^2}{10 \cdot h^2}} \tag{3}$$

where:

$\underline{N}_{\pm 1}(\underline{x})$ –is the corresponding area of the next frame (+1) or the previous frame (–1).

$$\mathrm{NL}_h\{\underline{x}\} = \frac{1}{C(\underline{x})} \cdot \sum_{\underline{y} \in Q(\underline{x})} z(\underline{x}) \cdot e^{-\frac{\|\underline{N}(\underline{x}) - \underline{N}(\underline{y})\|_2^2}{h^2}} \cdot z(\underline{x}) \cdot e^{-\frac{\|\underline{N}(\underline{x}) - \underline{N}_{\pm 1}(\underline{x})\|_2^2}{h^2}} \tag{4}$$

In case the current block has appropriate patches from the next and the previous frames, two additional patches will be added to corresponding patches from the search area and filtering across all these patches will be performed.

3.1. Find motion vectors between different frames

Motion estimation algorithms calculate the motion between two input images and produce output a field of motion vectors. Block matching is a popular method for estimating motion vectors from the image sequence. It assumes that the motion is uniform over a block of pixels and that the motion can be modelled as the displacement of these blocks. The

maximum possible vertical and horizontal displacement of the block defines the search area, and the best matching block is determined by minimizing the Sum of Absolute Difference (SAD) between the source block and the destination block inside the search area.

Plenty of motion estimation algorithms have been proposed [3], among which the three-dimensional-recursive-search (3DRS) has proved to be efficient in many applications [8]. The 3DRS principle is based on the following assumptions:

1. Objects in the frame are assumed to be larger than blocks (block size that is used in motion estimation);
2. Vectors estimated for neighbouring blocks are good candidates for the current block.

To summarize, the candidate vectors are constructed as follows (5) [17]:

$$c_i = \begin{cases} d(x+\rho_i,\ n); \\ d(x+\rho_i,\ n-1); \\ c_j + u, \quad j \neq i, \quad u \in US \end{cases} \tag{5}$$

The candidate set contains two spatial candidates $d(x + \rho, n)$, one temporal candidate $d(x + \rho_i, n - 1)$ and two update candidates $c_j + u$ (Fig. 3).

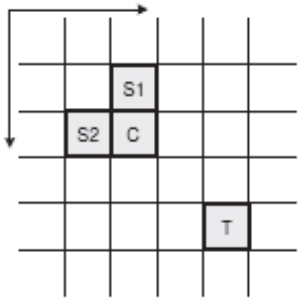

Fig. 3. Candidate set configuration. *C* is the current block; *S*1, *S*2 and T indicate two spatial and temporal candidates. Update candidates are random candidates generated using *S*1 and *S*2. The arrows indicate the scanning direction

There is one additional requirement which is quite different from other applications like quality of the motions vectors. This parameter is described in more details in the next section.

3.2. Determining motion vectors quality during the process of motion estimation

Motion estimation algorithms calculate the motion between two input images and produce output a field of motion vectors. Block matching is a popular method for estimating motion vectors from image sequences. It assumes that the motion is uniform over a block of pixels and that the motion can be modelled as displacement of these blocks. The maximum possible vertical and horizontal displacement of a block defines the search area, and the best matching block is determined by minimizing the Sum of Absolute Difference (SAD) between the source block and the destination block inside the search area.

There has been lot of motion estimation algorithms developed over the last 2 decades. However it is very difficult to find out the motion vectors quality evaluation for detected motion vectors. The main aim of this section is to highlight the calculating of the motion vectors quality and change the original 3DRS algorithm to be steady for the rapidly changing video.

During the block motion estimation process, each frame is divided into small blocks (in 3DRS implementation, each block has a fixed size of 4 × 4 pixels).

The candidate motion vectors c_i are constructed as follows [17]:

$$c_i = \begin{cases} d(x+\rho_i,\ n) & \text{Spatial candidate} \\ d(x+\rho_i,\ n-1) & \text{Temporary candidate} \\ c_j + u,\quad j \neq i,\quad u \in US & \text{Update motion vector candidate} \end{cases} \tag{6}$$

Sum of Absolute Differences (SAD) with additional penalty (7) is used to determine the best candidate motion vector for the current block [17].

$$p_i = \begin{cases} 1/2 \cdot \beta_0 \cdot \beta_1 & \text{Spatial candidate} \\ \beta_0 \cdot \beta_1 & \text{Temporary candidate} \\ 2 \cdot \beta_0 \cdot \beta_1 & \text{Update motion vector candidate} \end{cases} \tag{7}$$

The candidates set used in this study, contains two spatial candidates, two temporal candidates and four update vectors, and in total, eight candidates per block.

In comparison to the original 3DRS implementation, we added two more update vectors in order to adopt this algorithm for rapidly changing video.

A penalty mechanism ensures preferences for those candidates that have the same displacement. The final motion vector is determined as in the original algorithm 3DRS:

$$d(x, n) = \arg\min(\varepsilon(c_i, x, n) + p_i),\ \ c_i \in \mathrm{CS} \tag{8}$$

The quality of the motion vector can be calculated from the following equation:

$$\mathrm{Quality}(c) = \varepsilon(c_i, x, n) + \alpha \cdot (S)^{1/3} + \beta \cdot \mathrm{dev}(x, y) \in \mathrm{CS} \tag{9}$$

where:

- $\mathrm{Quality}(c)$ – is motion vector quality,
- $\varepsilon(c_i, x, n)$ – is a Sum of absolute differences,
- $\mathrm{dev}(x, y)$ – deviation of the neighboring blocks,
- α, β – are balancing coefficients.

So, the final NLM filtering type is determined from the next equtation:

$$\mathrm{Quality}(C) = \begin{cases} < \mathrm{Th1} & \text{Motion vector is clear} \\ > \mathrm{Th1}\ \&\&\ < \mathrm{Th2} & \text{Motion vector is within specified range} \\ > \mathrm{Th2} & \text{Motion is inconsistent} \end{cases} \tag{10}$$

where:

Th1, Th2 – Threashold values for quality coefficients of motion vectors.

These values are defines differently based on image/video clour scheme.

4. Results

In this section, an objective analysis of the original NLM algorithm and spatio-temporal algorithm is performed over sequences with several levels of compression (0.5 Mbps, 3 Mbps are used in this research) to evaluate how efficiently the proposed spatio-temporal algorithm reduces the compression artifacts. The proposed spatio-temporal and original NLM algorithms are applied to sequences which are encoded and decoded using the MPEG-2 codec. The blocking and flickering artifacts in these compressed sequences are strongly visible. The MSE and BIM metrics are used for evaluation of the processed sequences. The calculation of MSE and GBIM and the encoding of test sequences are done using the PTS tool.

The Mean Square Error (MSE) is the error metric used to compare image processing (de-noising, compression) quality. The MSE represents the cumulative squared error between the de-noised and the original image. The lower the value of MSE, the higher the quality of a restored signal.

BIM and PSBIM are used to measure the amount of blocking artifacts in the image/video. These metrics show a strong consistency with the human perception of coding impairments and subjective evaluations is the General Block Impairment Metric (GBIM), introduced in [28, 29]. The lower the value of GBIM the lower the quantity of the blocking artifacts. PSBIM is an improved GBIM metric.

Table 1

Objective metrics results for NLM filtering

Sequence	Metric	0.5 Mbps	3 Mbps	Original
Akio	MSE	6.331692	5.646	0
	BIM	1.277975	1.274	7.521864
	PSBIM	0.738614	0.734	1.749776
Bowling	MSE	4.374	4.143	0
	BIM	1.144023	1.1423	8.792035
	PSBIM	0.545938	0.54585	1.465607
Foreman	MSE	39.809	10.215	0
	BIM	1.024102	1.025591	6.129313
	PSBIM	0.758773	0.695938	1.757120
Bus	MSE	264.278	40.115	0
	BIM	1.263775	1.13198	2.051955
	PSBIM	1.124431	1.01635	1.280819
Claire	MSE	3.737	3.167	0
	BIM	1.716750	1.71634	6.548664
	PSBIM	0.999271	0.99699	1.468032

Five different sequences were used in the tests. The sequences were chosen for having varying content and intensity of motion.

Objective metrics for compressed sequences after spatio-temporal NLM filtering are presented in Table 2.

Table 2

Objective metrics results for spatio-temporal NLM filtering

Sequence	Metric	0.5 Mbps	3 Mbps	Original
Akio	MSE	4.414	4.295	0
	BIM	1.261789	1.261	7.521864
	PSBIM	0.738614	0.734	1.749776
Bowling	MSE	2.814	2.785	0
	BIM	1.136261	1.13612	8.792035
	PSBIM	0.530264	0.52991	1.465607
Foreman	MSE	31.295	10.248	0
	BIM	1.009726	1.02476	6.129313
	PSBIM	0.746233	0.69826	1.757120
Bus	MSE	234.389	30.333	0
	BIM	1.229437	1.11699	2.051955
	PSBIM	1.089562	1.00574	1.280819
Claire	MSE	3.001	2.835	0
	BIM	1.702442	1.70290	6.548664
	PSBIM	0.970346	0.96077	1.468032

For all processed sequences, the MSE values are lower than the MSE of the unprocessed sequences. Sequences processed by spatio-temporal algorithm have slightly less blockiness, meanwhile the BIMs and PSBIMs metrics have slightly better value.

a) b)

Fig. 4. Foreman videosequence: a) processed by NLM, b) processed by proposed spatio-temporal algorithm

Test sequences processed by the spatio-temporal algorithm have lower MSE value, consequently, we received significant improvement and proved that the spatio-temporal algorithm is more effective for highly compressed video. The user can observe essential quality improvements for video processed by means of proposed the spatio-temporal algorithm.

In most cases, proposed the spatio-temporal NLM algorithm can preserve image details in a better way. It is especially visible on highly compressed image/video (Fig. 4a Foreman processed by NLM, 4b by spatio-temporal algorithm).

5. Conclusions

The presented method for artifacts reduction demonstrated that a spatio-temporal approach provides a significant improvement of picture quality at low bitrates compared to spatial filtering only. In case sequences suffering from severe artifacts (e.g. flickering), spatio-temporal filtering proved to be a preferred option. This research also demonstrates benefits that can be achieved by using additional temporal information, especially consistent temporal information (approach for 'consistency' measurement of temporal information also provided in this research).

Temporal filtering is effective mostly for low bitrate videos. High quality sequences simply do not suffer from the severe artifacts propagated to the temporal domain, and therefore do not need much blurring. For those sequences, the most important fact is to differentiate between object details and artifacts, which can be achieved by means of spatial analysis. Therefore, methods based on spatio-temporal analysis and adaptive edge-preserving filtering are the most efficient for high bit-rate videos. Applying spatio-temporal filtering provides better results than original NLM implementation because temporal and spatial information is included in this filtering. Temporal information doesn't propagate additional blurring effects because this information is used only when true motion vector exists. These additional steps to the original NLM algorithm introduce additional complexity, so performance of the proposed algorithm should be enhanced.

True motion estimation is also very important, as it is an initial attempt to introduce some metric for calculating a quality of motion vectors. This technique can be used in other post-processing algorithms or in different directions of image and video processing.

The work described in this article demonstrates advantages of the spatio-temporal filtering approach over the spatial approach and additionally proves the temporal information usage for the coding artifacts reduction in video.

References

[1] Fedak V.I, Nakonechny A.Y., *Artifacts suppression in images and video. Non-Local Means as algorithm for reducing image and video distortions*, Hluboka nad Vltavou, Czech Republic 2009.

[2] Fedak V., Nakonechny A., *Determining motion vectors quality during the process of motion estimation*, 6th International Scientific and Technical Conference 16–19 November 2011, Lviv.

[3] Callico, G.M. et al., *Analysis of fast block matching motion estimation algorithms for video super-resolution systems*, IEEE Transactions on Consumer Electronics, 54(3), 2008, 1430-1438.

[4] Protter M., *Processing Image Sequences Without Motion Estimation*, Technion-Computer Science Department, Ph.D. Thesis.

[5] Boulanger J., Kervrann C., Bouthemy P., *Space-time adaptation for patch based image sequence restoration*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, No. 6, 2007, 1096-1102.

[6] Dabov K., Foi A., Egiazarian K., *Video denoising by sparse 3D transform-domain collaborative filtering*, Proc. of the European Signal Processing Conference, EUSIPCO, 2007.

[7] Buades A., *Image and film denoising bynon-local means*, Journal of Visual Commincation and Image Representation, Vol. 1, No. 1, Ph. D. Thesis, 24 March 2004, Article No. VC970378.

[8] Ramamurthi B., Gersho A., *Nonlinear space-variant postprocessing of block coded images*, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 34, No. 5, Oct. 1986, 1258,1268.

[9] Truong Quang Vinh, Young-Chul Kim, *Block Artifact Reduction Based on Pixel Classification Using Binary Edge Map*, 2008.

[10] Wang C. et al., *Adaptive Reduction of Blocking Artifacts in DCT Domain for Highly Compressed Images*, 2003.

[11] Alan W.-C. Liew, Hong Yan, IEEE *Blocking Artifacts Suppression in Block-Coded Images Using Overcomplete Wavelet Representation*, IEEE Trans Circuits Syst. Video Technol., Vol. 11, No. 14, No. 4, April 2004.

[12] Choy S.O., Chan Y.-H., Siu W.-C., *Reduction of block-transform image coding artifacts by using local statistics of transform coefficients*, IEEE Signal Processing Letters, Vol. 4, No. 1, Jan. 1997, 5-7.

[13] Truong Quang Vinh, Young-Chul Kim Block., *Artifact Reduction Based on Pixel Classification Using Binary Edge Map* 2008.

[14] Vincent O.R., Folorunso O., A *Descriptive Algorithm for Sobel Image Edge Detection*, Proceedings of Informing Science & IT Education Conference, InSITE, 2009.

[15] Tao C., Wu H.R., Bin Q., *Adaptive postfiltering of transform coefficients for the reduction of blocking artifacts*, IEEE Trans. Circuits Syst. Video Technol., Vol. 11, No. 5, May 2001, 594-602.

[16] Buades A., Coll B., Morel J.-M, *A non-local algorithm for image denoising*, 2006.

[17] De Haan, G. et al., *True-motion estimation with 3-D recursive search block matching*, Circuits and Systems for Video Technology, IEEE Transactions on, 3(5), 1993, 368-379, 388.

[18] Fedak V., Nakonechny A., *Spatio-temporal Non-Local Means Algorithm for Coding Artifacts Reduction*, Lviv Polytechnic National University, 2011.

[19] Zhou Z.-F., Shui P.-L., *Contourlet-based image denoising algorithm using directional windows*, Electronic Letters, 43, No. 2, 2007, 92-93.

[20] Olhede S.C., *Hyperanalytic denoising*, IEEE Transactions on Image Processing, 16 (6), 2007, 1522-1537.

[21] Donoho D.L., Johnstone I.M., *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, Vol. 81, No. 3, 1994, 425-455.

[22] Pizurica A., Philips W., *Estimating the probability of the presence of a signal of interest in multiresolution single and multiband image denoising*, IEEE Transactions on Image Processing, 15, No. 3, 2006, 654-665.

[23] Shui P.-L., *Image Denoising Algorithm via Doubly Local Wiener Filtering With Directional Windows in Wavelet Domain*, IEEE Signal Processing Letters, 12, No. 6, 2005, 681-684.

[24] Luisier F., Blu T., Unser M., *A New SURE Approach to Image Denoising: Inter Scale Orthonormal Wavelet Thresholding*, IEEE Transactions on Image Processing, 16, No. 3, 2007, 593-606.

[25] Achim A., Kuruoglu E.E., *Image Denoising Using Bivariate-Stable Distributions in the Complex Wavelet Domain*, IEEE Signal Processing Letters, 12 (1), 2005, 17-20.

[26] Avindra F., *Image Denoising with the Non-local Means Algorithm*, University of Kansas, December 2010.

[27] Buades A., Coll B., Morel J.-M., *A non-local algorithm for image denoising*, Computer Vision and Pattern Recognition, 2005.

[28] Du Y., Kirenko I., *Benchmarking of Coding Artifacts Reduction Techniques*, Philips Restricted Technical Note PR-TN 2007/00681, October 2007.

[29] Wu H.R., Yuen M., *A generalized Block-edge impairment metric for video coding*, IEEE Signal Processing Letters, 4, November 1997, 317-320.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
2-AC/2013

JULIA SIDERSKA*

APPLICATION OF NEURAL NETWORKS FOR SOCIAL CAPITAL ANALYSIS

ZASTOSOWANIE SIECI NEURONOWYCH DO ANALIZY KAPITAŁU SPOŁECZNEGO

A b s t r a c t

The paper investigates the possibility of using soft computing for estimating the value of social capital. Our approach is applied to the case of Red Hat Inc. – the world's leading provider of open source solutions. The objective of the research was to develop an artificial neural network for forecasting the value of social capital. These studies also allow us to identify variables significantly affecting the value of social capital. Computer simulations and assessments were done using software package STATISTICA Automated Neural Networks. The paper concludes with discussion and proposals for further research.

Keywords: *artificial neural network, soft computing, multilayer perceptron, social capital, independent variables, fundamental equation, global sensitivity analysis*

S t r e s z c z e n i e

Głównym celem artykułu jest analiza możliwości zastosowania obliczeń inteligentnych do modelowania kapitału społecznego firmy Red Hat Inc. – światowego lidera rozwiązań *open source* dla biznesu. Zasadniczym celem badań jest zaproponowanie struktury sztucznej sieci neuronowej do analizy wartości kapitału społecznego. Zidentyfikowano zmienne istotnie wpływające na wartość tego kapitału. Wszystkie symulacje komputerowe oraz oszacowania przeprowadzono w pakiecie statystycznym STATISTICA Automatyczne Sieci Neuronowe. W artykule przedyskutowano wyniki testów otrzymanych z zastosowaniem zaprojektowanego modelu oraz zaproponowano tematykę dalszych badań.

Słowa kluczowe: *sztuczna sieć neuronowa, obliczenia inteligentne, perceptron wielowarstwowy, kapitał społeczny, zmienne niezależne, równanie fundamentalne, globalna analiza wrażliwości*

* M.Sc. Julia Siderska, e-mail: j.siderska@pb.edu.pl, Faculty of Management, Bialystok University of Technology.

1. Introduction

Since intangible assets, particularly social capital, are commonly understood to be the most precious assets of any IT company, the appraisal of their value should be essential for chief executive officers and managers. Estimating the value of social capital using tools developed so far, especially those based on financial balance sheets, turns out to be a laborious and time-consuming task. Therefore, we suggest the application of soft computing for the analysis and modeling of the value of the intangible assets of IT companies.

There exists a vast amount of studies in the area of social capital analysis, however, the widely accepted method for extrapolating and predicting the value of social capital has not been proposed so far. The attempts to develop an effective method to estimate the company's intangible assets have been undertaken by a number of scientists, as well as by the companies themselves. The most significant achievements in this area are by Swedes and Americans, whose methods have gained popularity in Poland: Intangible Assets Monitor IAM (Karl E. Sveiby); Balanced Scorecard BSC (Robert S. Kaplan and David P. Norton), Skandia Framework – SF (Edvinsson L., Malone). These are non-financial models, allowing for the identification and evaluation of the most important intangible factors defining the difference between the market value and the book value of a given company.

The literature review revealed that any soft computing tool for modeling the social capital of an IT company has not been developed thus far. This identified research gap was the most essential premise for considerations in this sphere. This paper demonstrates the proposition of an artificial neural network model as an innovative method for such assessments and forecasts.

For many years, artificial intelligence tools, including mainly neural networks, have been increasingly implemented not only in areas of engineering (pattern recognition, signal processing, control, optimization), but also in economics and management. Neural network models are often applied for anticipating economic phenomena, such as predicting trends in the stock market, sales forecasting etc. Artificial neural networks are efficient methods of data analysis and therefore, they are often used as an alternative to traditional analytical methods, providing satisfactory results [11].

The paper is organized as follows. First, a short profile of Red Hat Inc. is presented. Then basic terminology used throughout this paper and some theoretical considerations are introduced. The next section concerns background information about artificial neural networks and their training process. Section 4 contains the description of the training set, consisting of data for 8 years of the history of Red Hat. Then we discuss the structure and parameters of the obtained three-layer perceptron model. Finally, we provide first insights into the developed methodology for anticipating the value of the social capital of Red Hat Inc.

2. Red Hat Inc. in a brief

Red Hat Inc. is a global software company, established in 1998, providing open-source software products. The company offers award-winning support, training, and consulting services as well as operating system, storage servers, virtualization, middleware and cloud technologies. It's corporate headquarter is in Raleigh (North Carolina, USA) with satellite

offices worldwide [7]. Red Hat is the world leader in providing open source solutions to businesses and the creator of the popular Linux distributions – Red Hat Linux, which is disseminated under the GNU General Public License free of charge.

The company benefits mostly from support and services achieving significant successes in the operating system market. Red Hat became the first one-billion dollar open source company in the fiscal year 2012, reaching $1.13 billion in annual revenue [1]. Its total revenue has been constantly growing for more than ten years. Red Hat became a part of the NASDAQ in 2005 – over-the counter, regulated stock market in the United States. The company listed its shares on the New York Stock Exchange (NYSE) under the ticker symbol RHT in 2006. Red Hat boasts a number of powerful customers including Amazon, DreamWorks, and Morgan Stanley [4].

For many years, the company has maintained a commanding lead in the area of Linux kernel development [2]. The Linux kernel is one of the most successful open source projects. The huge rate of change and the number of individual contributors prove that it has an active community, constantly causing the evolution of the kernel. The number of different developers who are doing the Linux kernel development have been increasing over different kernel versions in the last few years. There are about 1300 developers contributing to the newest kernel release [2]. The idea behind the development of open and free applications is common, parallel and creative work of a team of experts. The passion, commitment and enthusiasm of those developers determine the success of such applications. It is worth noting that the employees of Red Hat are responsible for about 11% of the total improvements to each Linux kernel release.

3. Basic definitions

3.1. Social capital

For the purpose of the paper, we define **social capital** as capital composed of formal and/or informal relationships among workers of IT company. These relationships are both positive (trust, cooperation, solidarity etc.) as well as negative (distrust, hypocrisy, inter-personal conflicts etc.) [13].

In the case of Red Hat, these are the relationships between software engineers, programmers, and testers, analysts etc. Red Hat employs those software developers for their talent, skills, experience, knowledge of programming languages and so on. All these features together form the social capital of the entire company and the whole community. These assets are, in the author's opinion, the most precious resources of Red Hat as well as of any IT company. Software is developed as a result of the joint, creative work of many developers. Unquestionably, company prosperity depends mostly on its employees, their codified and tacit knowledge, and their positive relationships mutual collaboration, trust, empathy etc.

Social capital is treated like any other capital and consequently, the question arises of how to estimate the social capital value. These new factors significantly affecting the market value resulted in the necessity to develop methods of analysis and tool for measuring this value. There are many approaches for the mentioned assessments, nevertheless, the widely acceptable method for such estimations have not been yet proposed.

Therefore, we suggest using soft computing for measuring the value of intangible assets.

3.2. Background information about neural networks

Analysis of the applicability of soft computing for modeling the value of social capital should start with some theoretical explications. The artificial neural model is considered as a system processing input signals into a single output [3].

Artificial neurons have *n* inputs and one output. Input information is always numeric and forms the vector of input values. Each input neuron is associated with a numerical factor called weight. Input values of weights are typically different and they are mostly determined automatically in the learning process. The operation of the neural consists of two phases, the first one can be defined as an aggregation of the input value, calculated by multiplying the weights and the corresponding input values. The second step lies in the fact that the aggregated sum of values becomes an argument in the activation function. At the beginning of the training process, the input variable gets initially randomly assigned weight – the strength of its effect on the output variable value. The proper values of the weighting factors are determined in the learning process [3].

The output neuron is given by the relation [3]:

$$y = f(\mathbf{w}^T \mathbf{x}) = f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{1}$$

where **w** is the weight vector defined as [11]:

$$\mathbf{w} \overset{\text{def}}{=} \left[w_1, w_2 \ldots w_n\right]^T \tag{2}$$

x is the input vector:

$$\mathbf{x} \overset{\text{def}}{=} \left[x_1, x_2 \ldots x_n\right]^T \tag{3}$$

The activation function is denoted as $f(w^T x)$. Among the number of activation functions, the most commonly applied are: linear activation function; sigmoid activation function; Gaussian activation function; hyperbolic tangent activation function [10].

4. Data in the example

The main idea of our method is based on the assumption that there are seven input variables significantly affecting the output dependent variable – the value of social capital. Modeling the value of social capital is a problem of regression, that is why only one neuron, characterized by the dependent variable is presented in the output.

The analysis covered the following input, independent variables: market value; book value; stock price; number of shares; employment; total assets; liabilities.

The output (dependent) variable **Y** represents the value of social capital, calculated by the Fundamental Equation. The formula says that in a market economy, under the equilibrium conditions, when demand equals supply, the value of Red Hat equals the aggregate sum of four component values of its capital: financial; physical; human and social at any moment *t* of the firm's past, present and future [14].

The model was built and all the simulations were carried out in the software package STATISTICA Automated Neural Networks.

The market value of Red Hat Inc. ($\mathbf{X_1}$), at a given moment t, is calculated by multiplying the number of issued stocks by their current stock price at the end of each quarter from the first quarter of 2006 to the third quarter of 2013. The book value of Red Hat Inc. ($\mathbf{X_2}$) for a given moment *t*, is calculated as the difference between the sum of total assets and the total liabilities in the mentioned periods. All necessary independent and dependent variables are calculated on the basis of balance sheets published by Red Hat Inc. at the end of each quarter in the examined quarters [5]. The training set consisted of data concerning 31 quarters (from the first quarter of 2006 to the third quarter of 2013).

For instance, the values of input and output variables for the third quarter of 2013 are shown in Table 1. All variables are demonstrated in USD. For simplicity the values of social capital, assets, liabilities, the number of shares, the market value and the book value are converted to millions of USD. The volume of employment $\mathbf{X_5}$ is presented in units. The stock price is presented in USD.

Table 1

Values of input and output variables in third quarter of 2013 (in millions $)

Market value $\mathbf{X_1}$	Book value $\mathbf{X_2}$	Stock price $\mathbf{X_3}$	Number of shares $\mathbf{X_4}$	Employment (in units) $\mathbf{X_5}$	Total asssets $\mathbf{X_6}$	Liabilities $\mathbf{X_7}$	**Social capital** $\mathbf{Y}$
8743	1686	46.14	189.5	5500	2662	0.976	4546

Out of the total samples, 80% samples were chosen for training set, 10% samples constituted the cross-validation and 10% samples were used for the testing set. The training set was used to train the neural network. Measures were determined based on the training set and allowed to assess the capacity for approximation. They point the precision in determining the output variable value for input vectors presented during learning. Much more important is the correctness for such input vectors that were not presented in the training process. The ability to proper operation of the network for data from outside the training set is called the ability to generalize. The measures determined on the basis of the test set enabled the evaluation of the network properties. The set of validation is used to calculate quality measures used to monitor the course of the learning process.

It was assumed that the relationships between variables are non-linear, therefore, the applicability of classical linear models to the analysis was groundless. The traditional linear regression, used for estimating the expected value of the dependent variable, can only be used to analyze linearly related data. It was hypothesized that artificial neural networks solve the problem of non-linearity of the data, significantly affecting the value of the social capital.

5. Neural network model

Undoubtedly, the operation of the artificial neural network depends primarily on the training process. Network learning is an iterative action, repeated many times, step by step, with the fundamental objective to optimize the network parameters – the weighting factors. Initially, each of the input variables gets randomly assigned weight, the strength of its effect on the value of the output variable. The values of the weighting factors are determined in the learning process – the higher the weight, the more important the variable [8].

The model should reflect the existing reality, the link between a set of input variables (independent), and a set of dependent variables (output). Most methods assume the existence of a single dependent variable.

Multilayer perceptrons (MLP) are layered feedforward networks, typically trained with the backpropagation algorithm, as well as one of the most widely implemented neural network topologies. The backpropagation algorithm allows for the learning of input and output mappings from training samples. The network learns the relationship between the set of example patterns, and could be able to apply the same relationship to new input patterns [10].

In the present case, a supervised, learning-with-a-trainer approach was adopted (Fig. 1). It should be noted, therefore, in addition to the input signals, the desired (expected) answers – the output signal should be determined. To make a diagram easier to read not all weights are presented [12]. The network was trained on the basis of the knowledge of the values of social capital, calculated by the Fundamental Equation.

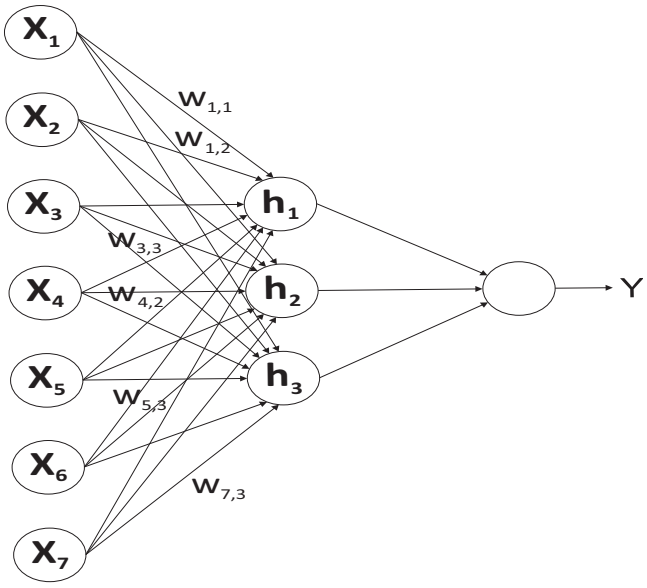

Fig. 1. Supervised learning of three-layered perceptron

The model was trained with the backpropagation algorithm, proposed by Rumerhart in 1986. Backpropagation is one of the most frequently implemented and the most effective learning algorithms of multilayer neural networks. This algorithm is based on the collected

data and modifies the weights and the threshold value so that the network minimizes the error (for example, the prediction error) for all data included in the training set. Errors that occur at the output of the network are propagated in the opposite direction than the signals passing through the network, from the output layer to the input [8].

Before the final construction of the network was chosen, many structures and parameters were checked. The neural networks differed in terms of parameters such as the number of hidden neurons, the activation function, the learning algorithm etc. The software package Automated Neural Networks allows us to implement only one hidden layer. However, the considering problem was not so complicated to use two hidden layers. Three neurons were used in the hidden layer (h_1, h_2, h_3). The choice of the number of neurons in the hidden layer is an essential issue – the excess can cause that the network learns relations on memory; their scarcity may remove the network's capacity for learning. Finally, the following ones were adopted as the activation functions, the hyperbolic tangent in the hidden layer and the linear function in the output layer. This choice did not allow for the loss of the prediction ability and it also improved the ability to extrapolate the results.

6. Results

Table 2 presents a summary of the parameters of the obtained three-layer perceptron:

Table 2

Parameters of obtained neural network model

Neural network	Activation function (hidden neurons)	Activation function (output neurons)	Error	Learning algorithm
MLP 7-3-1	Tanh $y = \mathrm{tg}h\left(\frac{\alpha\varphi}{2}\right) = \frac{1-\exp(-\alpha\varphi)}{1+\exp(-\alpha\varphi)}$	Linear $y(x) = ax + b$	Sum of squares (SOS)	BFGS (Broyden – Fletcher – Goldfarb – Shanno)

All elements belonging to the training set were presented for 131 times (131 epochs) to reach the minimum of network error. The learning error decreased rapidly at first and then tended gently to zero.

Some available regression statistics enabled us to assess the accuracy of estimates. Table 3 presents the values of Data Mean and Data Standard Deviation for data in the training, test and validation sets.

Measures determined based on the training set allowed us to assess the capacity for approximation. They point the precision in determining the output variable value for input vectors presented during learning. Much more important is the accuracy of such input vectors that were not presented in the training process. The ability of the network to properly operate for data outside the training set is called the ability to generalize. The measures determined on the basis of the test set enabled the evaluation of the network properties. The set of validation is used to calculate quality measures used to monitor the course of the learning process.

Table 3

Data statistics in the training, test and validation sets

	Data Mean			Data Standard Deviation		
	Training set	Test set	Validation set	Training set	Test set	Validation set
Market value $\mathbf{X_1}$	6475001	7137511	6473114	2695142	2388960	2438974
Book value $\mathbf{X_2}$	1331295	1302916	1426957	244994	352167	566777
Share price $\mathbf{X_3}$	34.06	37.27	31.80	13.69	12.43	10.05
Share number $\mathbf{X_4}$	190898.2	191413.3	191043.3	2904.7	1024.6	5059.2
Employment $\mathbf{X_5}$	3428	3666.66	3666.66	1154.88	1850.22	2286.19
Assets $\mathbf{X_6}$	1943948	1880415	2161915	398011	713250	1191906
Liabilities $\mathbf{X_7}$	612653.2	577499.3	734958	259678.7	361097.9	625209.3
Social capital **Y**	3332266	3807065	3278134	1715346	1363508	1271606

Table 4 presents the results of the predicted value of social capital **(Y*)** of Red Hat for the fourth quarter of 2013. The data concerning the fourth quarter of 2013 were not in the training set, therefore, were not trained in the network learning process. The value of the predicted social capital was provided by trained model MLP 7-3-1. This estimation was possible owing to the ability to generalize the data.

Table 4

The value of social capital in the fourth quarter of 2013 (in millions USD)

Market value $\mathbf{X_1}$	Book value $\mathbf{X_2}$	Stock price (in USD) $\mathbf{X_3}$	Number of shares $\mathbf{X_4}$	Employment (in units) $\mathbf{X_5}$	Total asssets $\mathbf{X_6}$	Liabilities $\mathbf{X_7}$	**Social capital (Fundam. Equation) Y**	**Social capital (MLP 7-3-1-) Y***
10 622	1326	56.04	189.5	6100	2851	1025	5904	5729

The predicted value of social capital **(Y*)** in the fourth quarter of 2013 was anticipated using obtained three-layer perceptron MLP 7-3-1. The value of social capital **(Y)** was calculated using the Fundamental Equation. The difference between those values is only 3%, which authenticates the proposed method as a reliable tool for credible estimations of the value of intangible assets.

7. Conclusions

This contribution examined the possibility of using an artificial neural network for modeling the value of the social capital of IT companies. The results of the conducted survey confirmed the research hypothesis that artificial neural networks solve the problem of non-linearity of the data that significantly affect the value of social capital. On the basis of those investigations and obtained results, it can be concluded that neural networks are sophisticated modeling techniques, capable of mapping extremely complex functions. In particular, neural networks are non-linear models, which significantly extends the capabilities of their applications.

The investigated neural network model is based on data, not on an analyst's knowledge, therefore one does not need to examine the relationships between data in the training set before setting up the neural network model. Artificial neural networks can be implemented in virtually any situation where the objective is to determine an unknown relationship or a set of relationships between dependent and independent variables.

The possibility of using proposed method for predicting the value of social capital constitutes likewise a promising area of upcoming research. Such assessments are likely to be accurate using the function predictions for new data, as well as using the developed model as a regression time series. Further work in the research area of this contribution will involve employing a developed model to identify the relationships among dataset characterizing the world's biggest companies in the information technology industry: Microsoft; IBM; Oracle; Novell; SAP. Next, the goal for future research should be to estimate the social capital value of companies operating on the Polish software market. The subjects of the author's interests are Asseco Poland SA, Comarch, ABC Data etc.

Furthermore, the developed model of the artificial neural network, can also be adopted to predict the value of social capital for completely new data, which did not belong to the training set during the learning process. This could be possible owing to the ability to generalize data and on the basis of the previously learned dependency. Moreover, it can be assumed that the constructed model will be useful for analysis and modeling of the value of social capital of companies operating in other economic sectors.

Author is a beneficiary of the project "Raising the potential of universities as a factor in the development of knowledge-based economy". The project is co-financed by European Social Fund and Polish Government.

References

[1] Babock Ch., *Red Hat: First $1 Billion Open Source Company*, Information Week, 2012.

[2] Corbet J., Kroah-Hartman G., McPherson A., *Linux Kernel Development: How Fast it is Going, Who is Doing It, What They Are Doing and Who is Sponsoring It*, The Linux Foundation, 2012, 7-12.

[3] Lula P., Paliwoda-Pękosz G., Tadeusiewicz R., *Metody sztucznej inteligencji i ich zastosowania w ekonomii i zarządzaniu*, Akademicka Ekonomiczna w Krakowie, Kraków 2007, 79-86.
[4] Munga N., Fogwill T., Williams Q., *The adoption of open source software in business models: a Red Hat and IBM case study*, ACM, New York 2009, 112-121.
[5] Red Hat Corporation, Annual Reports 2005–2013, available at: http:/investors.redhat.com/annuals.cfm (accessed: October 2013).
[6] Red Hat Corporation, available at: http:/investors.redhat.com (accessed: October 2013).
[7] Red Hat Corporation, available at: http:/www.redhat.com (accessed: October 2013).
[8] Rutkowska D., Piliński M., Rutkowski L., *Sztuczne sieci neuronowe, algorytmy genetyczne i systemy rozmyte*, Wydawnictwo Naukowe PWN, Warszawa 1997, 34.
[9] Reuter U., Liebscher M., *Global sensitivity in a view of nonlinear structural behavior*, LS-DYNA Anwenderforum, Bamberg 2008, 1-2.
[10] Sibi P., Allwyn-Jones A., Siddarth P., *Analysis of different activation function using back propagation neural network*, Journal of Theoretical and Applied Information Technology, Vol. 47, No. 3, 2013, 1264-1268.
[11] Tadeusiewicz R., *Sieci neuronowe*, Akademicka Oficyna Wydawnicza, Warszawa 1993.
[12] Walczak S., *Methodological Triangulation Using Neural Networks for Business Research*, Advances in Artificial Neural Systems Volume 2012, Article ID 517234, University of Colorado, Denver 2012.
[13] Walukiewicz S., *Kapitał społeczny*, Systems Research Institute of the Polish Academy of Science, Warsaw 2012.
[14] Walukiewicz S., *Kapitał ludzki*, System Research Institute of the Polish Academy of Science, Warsaw 2010.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
2-AC/2013

ANDRIY LAGUN*

CRYPTOGRAPHIC STRENGTH OF A NEW SYMMETRIC BLOCK CIPHER BASED ON FEISTEL NETWORK

KRYPTOGRAFICZNA ODPORNOŚĆ NOWEGO ALGORYTMU BLOKOWEGO SZYFROWANIA INFORMACJI OPARTEGO NA SIECI FEISTELA

A b s t r a c t

The paper summarizes research on the cryptographic strength of a new symmetric block cipher based on the Feistel network. The classification of cryptographic attacks, depending on the cryptanalyst's input data, is considered. For the purpose of testing, the linear and differential cryptanalysis as well as the Slide attack were used.

Keywords: *block cipher, the Feistel network, cryptographic strength*

S t r e s z c z e n i e

Niniejszy artykul pokazuje wyniki badania kryptograficznych szyfrów blokowych opartych na sieci Feistela. W artykule przedstawiono klasyfikacje ataków kryptograficznych na podstawie danych wejściowych, które ma analityk kryptograficzny. Do badań użyto ataki wykorzystujące metody kryptoanalizy liniowej i różnicowej, a także ataki typu Slide.

Słowa kluczowe: *szyfr blokowy, sieć Feistela, odporność kryptograficzna*

* Ph.D. Andriy Lagun, e-mail: a.e.lagun@gmail.com, Department of Information Security Management, Lviv State University of Life Safety.

1. Introduction

The rapid development of computer technology and open networks, modern methods of storage, processing and the transmission of information has contributed to the emergence of threats relating to the possibility of loss, disclosure and the modification of data belonging to different users. Cryptographic techniques and data protection form the basis of information security in information and telecommunication systems.

Determining the efficiency of cryptographic algorithms is usually a more difficult task than their design, because they requires a higher level of knowledge in this area and are inherently more scientific than engineering problems. This means there exists a large number of cryptographic protection methods, the reliability of which are not defined or guaranteed, because the algorithms on which they are based are insufficient or completely unexplored.

2. Review of cryptographic systems

Cryptographic systems are split into symmetric and public key systems. In symmetric cryptosystems, the same key is used for encryption and decryption. In public-key systems, public and private keys are used, which are mathematically related to each other. Information encrypted with a public key, which is available to everyone, is decrypted using the private key, known only to the recipient of the message.

Symmetric cryptosystems are split into block and stream systems.

Block ciphers are easy to use and allow for the handling of more parts of information compared to stream ciphers, in addition, they can be easy transformed into stream ciphers. Block ciphers are easily deployed. Their advantage over asymmetric ciphers is a greater performance and reliability using smaller keys [1].

Since the article deals mostly with block ciphers, we conducted a thorough investigation of them.

Block ciphers are one form of symmetric ciphers which allows for the handling of plaintext with blocks of multiple bytes per iteration. For a modern block cipher block, the size is 128 or 256 bits. The basic principles used in block ciphers are diffusion and confusion. Diffusion hides the statistical properties of the plaintext and ensures that any change of character in the plaintext or encryption key affects a large number of ciphertext characters. Confusion complicates the tracking of statistical dependencies between ciphertext and plaintext [8].

The main advantage of this class of ciphers is that in most cases, data encryption and decryption procedures differ only in the order of operations. This feature greatly simplifies the creation of software and hardware tools for encryption and enables the use of the same tools to both encrypt and decrypt data.

Block ciphers are two mutually related algorithms – an algorithm for encryption and a converse decryption algorithm which are demonstrated by the formulas (1) and (2) [1]. Input data are the blocks of plaintext (ciphertext) and the encryption key, encrypted (decrypted) data block of a similar size appear on the output. For a cipher of this type, an equation must be performed (3) that provides an unambiguous data encryption and decryption. Some of the encryption algorithms are based on transformations which are an involution. In this case, the encryption algorithm can be used for decryption without additional changes and modifications.

$$C = E(M, K) \tag{1}$$

$$M = E^{-1}(C, K) \tag{2}$$

$$M = E^{-1}(E(M, K)) \tag{3}$$

where:

- M – block of plaintext,
- C – block of ciphertext,
- K – encryption key,
- E – direct cryptographic transformation,
- E^{-1} – inverse cryptographic transformation.

The block cipher consists of the simple transformation of plaintext that is performed in a certain sequence a certain number of times. These transformations with plaintext, or its component parts, or with the encryption key, provide the opportunity to achieve the main goal of encryption – to eliminate or significantly reduce the statistical information and dependence of the plaintext. In other words, it is necessary to increase the entropy of the plaintext to such a value where a relationship between the input and output of the cryptographic algorithm is not observed. In most cases, operations and transformations which are used in the cryptographic algorithm should have an inverse transformation – this is demonstrated in expression (4). In this case, the realization of the operations (which have an inverse) with plaintext will also have an inverse operation. This will be a set of inverse transformations applied in an inverse order, as shown in formula (5).

$$E^{-1}(E(M)) = M \tag{4}$$

$$E_3^{-1}(E_2^{-1}(E_1^{-1}(E_3(E_2(E_1(M)))))) = M \tag{5}$$

Certain operations can be inverse of themselves, namely, involutions. An example of this type of operation is the exclusive *OR* operation (*XOR*), which is most common in cryptographic algorithms. The two main methods used for this purpose are diffusion and confusion. The application of confusion provides a certain property, when the change in one byte of plaintext leads to changes in many bytes of ciphertext – this is the so-called 'avalanche' effect. The easiest way to achieve this effect is to use transposition. Mixing allows you to hide the statistical properties of the plaintext and its redundancy. The simplest variant of mixing is the alphabetical substitutions of different types. As a rule, in modern block ciphers these methods have never been used alone, but only in combination. This fact allows for achieving the best effect.

In addition to these methods, different algebraic operations are used. These operations often belong to different algebraic groups. An example of using this type of operation is the IDEA algorithm, which uses the operations of *the* addition modulo, multiplication and *XOR*. The feature of these operations is their incompatibility in the sense that no two of them satisfy the associative and distributive laws, which greatly complicates cryptanalysis.

In addition to the above, cryptographic algorithms frequently use operations such as logical shifts, multiplication in the Galois field, Hadamard pseudo transformation, and different arithmetic and logical operations.

Since the usage of one of the above operations does not provide the necessary result, these operations should be used in a specific combination and according to certain rules in order to achieve the desired effect. The most common of these combined methods are the Feistel network and the Substitution-Permutation network, as well as their modification or combination.

The Feistel network is a certain structure, which is repeated a certain number of times, for each of which, a different round key is used [2]. Encryption and decryption operations on every stage are quite simple and usually identical, but require the reordering of round keys.

The essence of the Feistel transform is determined using the following algorithm [3]:

– the block of plaintext is split into two equal parts – left and right;
– to the left part and the round key, certain functions are applied and to the execution result of this function and the right part operation, XOR is performed; the result from the previous step is assigned to the new left sub-block, whereas to the right sub-block is assigned the unmodified previous right sub-block;
– the two previous steps are repeated a certain number of times with different round keys.

3. Description of the algorithm

The algorithm is a designed block encryption algorithm based on the Feistel network (Fig. 1) which has the following characteristics and properties:

– data block size: 256 bits;
– variable key size encryption (to 256-bit) and a variable number of rounds;
– orientation to 64-bit architecture;
– uses methods and mechanisms to prevent known types of attacks.

This algorithm is based on the use of direct changes in the encrypted information (Fig. 1a) and the inverse – in the decrypted information (Fig. 1b). Data encryption and decryption in this algorithm is described by formulas (6) and (7), respectively:

$$X_i = E2(E1(X_{i-1}, rK_i), rK_i) \tag{6}$$

$$X_i = D1(D2(X_{i-1}, rK_{n+1-i}), rK_{n+1-i}) \tag{7}$$

where

X_i – block input data (plaintext or ciphertext),
$E1$, $E2$, $D1$, $D2$ – Feistel transformations of different types,
rK_i – round key,
i – round number.

There have been different types of Feistel networks used with different properties in this algorithm. This allows for achieving different levels of dispersion and implicit use of different types of permutations in one round.

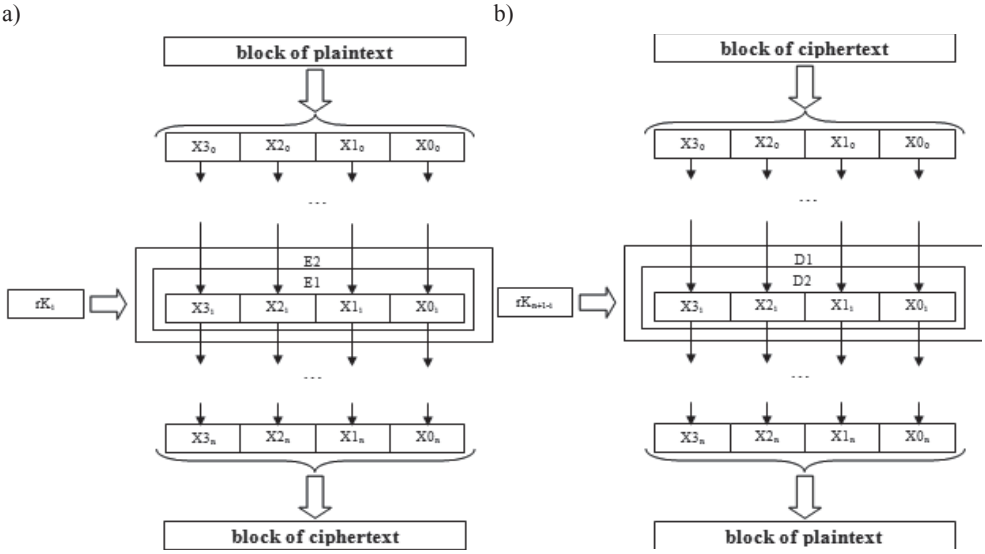

Fig. 1. The general structure of the algorithm: a – encryption scheme; b – decryption scheme

3.1. The Feistel transformation

During the encryption of the information for each pair of blocks, transformations $E1$ and $E2$ are used. These transformations are alternated between themselves. During the decryption, a pair of transformations $D1$ and $D2$, which are inverse of the previous ones, must be applied in reverse order.

Transformation $E1$ is shown in Fig. 2a. The block of information on its input is split into four equal parts $X0_{i-1} - X3_{i-1}$. Part $X3_{i-1}$ and the round key rK_i are input arguments of function $F1$, which has three outputs. These outputs form new values $X1_i - X3_i$ after summation modulo 2 with values $X0_{i-1} - X2_{i-1}$. Part $X0_i$ is formed by cyclic i bit shift of $X3_{i-1}$. This transformation provides a modification of all input parts, moreover, the dependence of the transformation $E1$ on the round number provides another modification of the block every time.

This transformation is described by formulas (8) and (9). To perform the inverse operations, transformation $D1$ is applied, as shown in Fig. 2b. It is performed by a similar principle and has the same properties as transformation $E1$ (formulas (10), (11)).

Transformation $E2$ is shown in Fig. 2c. The block of information on its input is split into four equal parts $X0_{i-1} - X3_{i-1}$. Parts $X0_{i-1} - X2_{i-1}$ and round key rK_i are input arguments of function $F2$ which forms the output part of the transformation $X0_i$ by performing the *XOR* operation with the output function $F2$ and part $X3_{i-1}$. Outputs $X1_i - X3_i$ are not modified parts of $X0_{i-1} - X2_{i-1}$ respectively. This transformation provides a modification of only one part of the input, but this modification depends on all parts of the input block $X0_{i-1} - X3_{i-1}$ and round key rK_i. To perform the inverse operations, transformation $D2$ is applied, as shown in Fig. 2d. It is performed by a similar principle and has the same properties as transformation $E2$. Transformations $E2$ and $D2$ are described by the formulas (12)–(15).

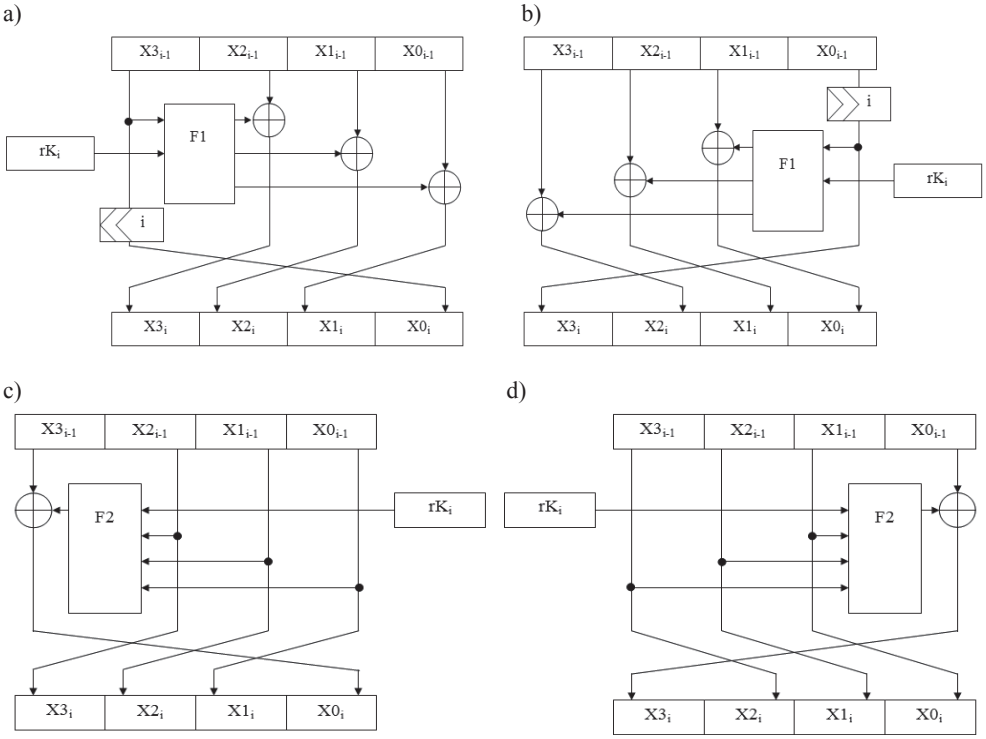

Fig. 2. Structures of used Feistel networks: a – $E1$, b – $D1$, c – $E2$, d – $D2$

$$X0_i = X3_{i-1} << i \tag{8}$$

for $t = 1..3$ $$X(t)_i = F1(i, X3_{i-1}, \mathrm{rK}_i)_1 \oplus X(t-1)_{i-1} \tag{9}$$

for $t = 0..2$ $$X(t)_i = F1(X0_{i-1} >> (n+1-i), rK_{n+1-i})_1 \oplus X(t+1)_{i-1} \tag{10}$$

$$X3_i = X0_{i-1} >> (n+1-i) \tag{11}$$

$$X0_i = F2(i, X2_{i-1}, X1_{i-1}, X0_{i-1}, rK_i) \oplus X3_{i-1} \tag{12}$$

for $t = 1..3$ $$X(t)_i = X(t-1)_{i-1} \tag{13}$$

for $t = 0..2$ $$X(t)_i = X(t+1)_{i-1} \tag{14}$$

$$X3_i = \mathrm{F2}(i, X3_{i-1}, X2_{i-1}, X1_{i-1}, rK_{n+1-i}) \oplus X0_{i-1} \tag{15}$$

where $F1(i, X, rK)_j$, $F2(i, X3, X2, X1, rK)$ are complicated modification functions.

The described transformations contain different degrees of dispersion and implicit permutations of different types through the use of cyclic shift operations and structure changes. Transformations $E1$ and $D1$ are focused on the diffusion of bits of the whole input block and permutations in one of its parts. Transformations $E2$ and $D2$ are focused on the diffusion of bits in one part of the input block and permutations in the whole block. The availability of diffusion and permutation methods, and also striping of the transformations of different types, provide rapid achievement of the 'avalanche' effect.

3.2. Complicated modification function

Function $F1$, the structure of which is shown in Fig. 3, contains three input arguments: round i number, data block X and round key rK sized 64 bits; and returns three output values $F1_1 - F1_3$.

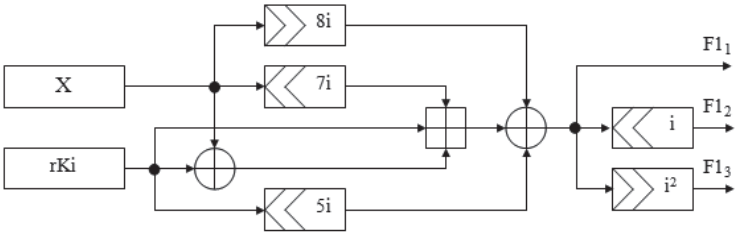

Fig. 3. Structure of complicated modification function $F1$

Function $F1$ uses three types of operations: exclusive OR; addition modulo 2^{64}; and the cyclic shift. Using operations of these types provides protection against linear and differential cryptanalysis provided their alternation. The main result of the transformation appears at the output $F1_1$, and the other two outputs are permutations of this result, which are achieved through the use of cyclic shifts (formulas (16)–(18)).

$$F1_1(X1, rK_i) = ((rK_i \oplus X1) + (X1 << 7i) + rK_i) \oplus (X1 >> 8i) \oplus (rK_i << 5i) \quad (16)$$

$$F1_2(X1, rK_i) = F1_1(X1, rK_i) << i \quad (17)$$

$$F1_3(X1, rK_i) = F1_1(X1, rK_i) << i^2 \quad (18)$$

The use of different types of shifts in a different number of positions after a sufficient number of rounds provides permutation across the set of the input block bits. The scattering of bits of the input block is achieved by imposing the initial and modified input arguments.

Function $F2$, the structure of which is shown in Fig. 4, contains five input arguments – number of round i, data blocks $X1 - X3$ and round key rK sized 64 bits; it returns one output value. This function uses three additional constants const1 – const3, which provides additional scattering of bits. This function is represented by formulas (19)–(22).

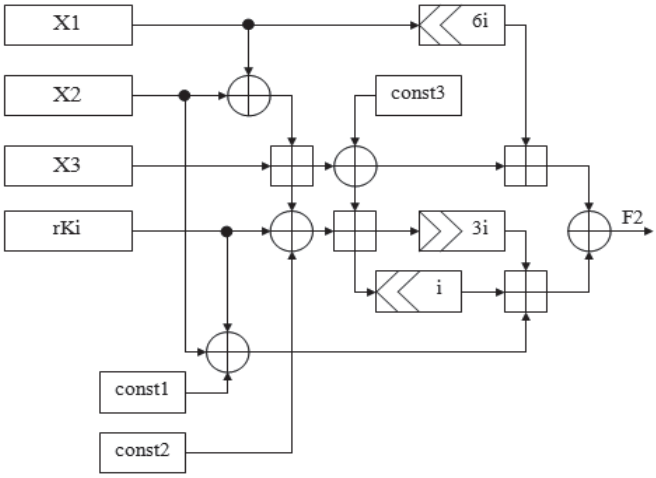

Fig. 4. Structure of complicated modification function *F*2

$$\lambda_1 = (X1 \oplus X2) + X3 \tag{19}$$

$$\lambda_2 = \lambda_1 \oplus \text{const3} \tag{20}$$

$$\lambda_3 = \lambda_2 + (\lambda_1 \oplus rK_i \oplus \text{const2}) \tag{21}$$

$$F2(X1, X2, X3, rK_i) = ((X1 << 6i) + \lambda_2) \oplus ((\lambda_3 >> 3i) + \\ + (\lambda_3 << i) + (X2 \oplus rK_i \oplus \text{const1})) \tag{22}$$

Function *F*2 is applied to modify only one part, but it allows for providing a greater effect of diffusion by the complicated internal structure.

Implementing the algorithm. Constants const1 – const3 can be initialized by the arbitrary values, or used as an additional key information It is desirable that the constants are of a random bit sequence. This would bring more sustainability than a periodic sequence filling.

The following function is applied in transformations *E*2 and *D*2, as well as to form the round keys, which allows to form strong subkeys.

3.3. Mechanism of round keys formation

In order to form the round keys, the input key, if necessary, is complemented by zero bits to the size of 256 bits. The structural diagram of the key formation for the *i*-th round is shown in Fig. 5.

The parts of the current encryption key are supplied as input arguments to the input of function *F*2 in the order determined depending on the number of the round. The round key is the output value. The resulting function is:

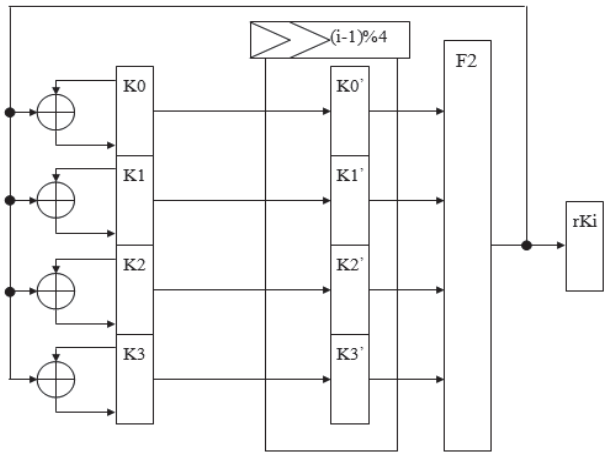

Fig. 5. Mechanism of round keys formation

If:

$$\begin{cases} (i-1)\%4 = 0 & \Rightarrow & rK_i = F2(K0, K1, K2, K3) \\ (i-1)\%4 = 1 & \Rightarrow & rK_i = F2(K1, K2, K3, K0) \\ (i-1)\%4 = 2 & \Rightarrow & rK_i = F2(K2, K3, K0, K1) \\ (i-1)\%4 = 3 & \Rightarrow & rK_i = F2(K3, K0, K1, K2) \end{cases} \quad (23)$$

To form strong round keys, the input key value is modified by operation XOR with its parts and the round key itself.

4. Overview of cryptanalysis methods

The appearance of newly encrypted algorithms leads to the development of their hacking methods. If the purpose is the disclosure of as many ciphers as possible, then the best strategy is to develop universal methods of cryptanalysis.

At this time, there are many methods of block ciphers analysis such as the brute force method, the statistical method, the method of meeting in the middle, linear and differential cryptanalysis, the boomerang method, the slide – attack and others [4, 5].

The brute force method provides the iteration of all possible variants of encryption keys. To search for a key that has a size of n bits, there are 2^n variants. After the iteration of all possible keys, the encryption key will be found. On average, this search requires 2^{n-1} test operations of encryption.

Protection against attacks of this type is an increase of the key size, because an increase of the key size at one bit leads to an increase twice in the number of the encryption key variants.

To increase the efficiency of this method, paralleling is used if the required resources are available, special devices for exhaustive key search and others are applied.

The task of the statistical method is the development of algorithms that determine an unknown key or part of this key. Implementations of this method for particular block ciphers are more efficient than the brute force method.

To the input of the algorithm, a certain amount of pairs (X_i, Y_i) $i = 1 .. n$ of the plaintext and ciphertext is supplied. These pairs are derived from the application of mapping F with key k. It is assumed that the plaintexts are chosen randomly, equiprobably and independently from the whole aggregation. The idea of statistical analysis is that if the results of observations are different, then after a sufficiently large number of observations, it is possible to determine, with a certain probability, the law of observations, that is the searched value.

Linear cryptanalysis combines the search of linear statistical analogues for encryption equations, statistical analysis of the plaintexts and ciphertexts, and methods of coordination and busting. This method examines statistical linear links between the various bits of vectors of the plaintext, ciphertext and the key, and uses these links to determine some bits of the key by the statistical methods. This method uses linear approximations for describing of the work of the cryptographic algorithm.

This method is carried out in two stages:

– forming links between plaintext, ciphertext and key that occur with high probability;
– these links with known pairs of the plaintext – ciphertext are used for getting key bits.

These links are called linear approximations. It is necessary to define the links, the probability of which is not equal 1/2, and use them to find appropriate bits of the key.

To protect against attacks with using linear cryptanalysis it is necessary to achieve that with any change of the plaintext or key, each of the bits of the ciphertext will change with probability Differential cryptanalysis uses pairs of the ciphertext with some differences. The essence of this method is to analyze the evolution of this difference in the process of passing the plaintext through the stages of encryption with one and the same key

The two plaintexts with a fixed distinction are selected. After passing of all stages of encryption and analyzing the distinction in the obtained ciphertexts, different probabilities are assigned for different keys. During further analysis of the following pairs, one of the keys will be more probable – it will be the encryption key.

This method of cryptanalysis has different versions. One of them is the use of impossible differentials (with zero probability). The hacking procedure is as follows: the required number of pairs of the plaintexts with the required distinction is selected; the appropriate ciphertexts are found; the analysis of the received data is performed and all versions of the encryption keys that lead to impossible differentials are considered incorrect and discarded.

Thus, some set of possible keys is received that do not lead to impossible differentials. Differentials with zero probability can be replaced by differentials with minimal probability. At the same time, the procedure of attack on the algorithm is similar to the procedure that is used in the impossible differentials.

There is another kind of attack – slide-attack. The feature of this attack is that its successful application does not depend on the number of rounds of the algorithm in which it is applied. The only requirement for its application is that rounds of the algorithm must be identical [6].

Let it is applied the multiround cryptographic algorithm with round function $E(P, k)$ and the condition that the round keys of the algorithm are identical. Then for attack the pair of the

plaintexts is applied, one of which is chosen randomly by the text P, and the other P' is the result of the oneround transformation of the text P:

$$P' = E(P, k) \tag{24}$$

The ciphertexts that refer to the plaintexts, are similarly related to each other:

$$C' = E^{-1}(C, k) \tag{25}$$

Having the pairs of the plaintexts and ciphertexts, which are related by only one round of the encryption, the round key can be obtained, this allows the disclosue of the full algorithm. The slide-pair *capital 'S'?* is defined as follows [7]:

- for the text P and for each plaintext P' the appropriate ciphertexts C and C' are received;
- the value of round key k_1 is calculated by the formula (24);
- the value of round key k_2 is calculated by the formula (25);
- the coincidence of these keys means that the required plaintext is found and $k = k_1 = k_2$ is the round key.

5. Evaluation of the algorithm stability

For the attack on the cipher by the method of exhaustive search, it is necessary to check up to 2^{256} key options. Verification of imitative sustainability was carried out by determining the impact of a change of one of the bits in the encrypted data block to the decrypted block. The average value of the difference between blocks is approximately 50%, regardless of position of the changed bit. To check availability of the avalanche effect it the impact of change of one of the bits of the key or the plaintext block to the ciphertext was investigated. Regardless of the position of the changed bit and its location (in the key or plaintext), the average value of the changes in the ciphertext is approximately 50% [9].

To check the described characteristics, tests with combinations of different types of input data were performed (block of plaintext and key): completely zero vector; completely unit vector; pseudo randomly generated vector.

Sustainability against attacks, based on the methods of differential and linear cryptanalysis, results in difficulty in determining the key knowing the input and output values of the round functions. The functions of complicated transformation are designed in such a way that selection of the round key among them is a difficult task. Since for transformations in the functions, bitwise shift operations are used, including with the key, the part of the input data, which enters the input of such an operation is rejected and therefore, performance of the inverse operation is impossible. In the best case, certain dependencies (26) for each of the round keys, but not the key in an explicit form, can be received. Knowing these dependencies you can distinguish the part of bits of the key, but not the complete key. For example, the dependence (27) shows that you can be sure in only two higher bits of the key and other bits are unknown.

$$K_r \oplus (K_r << n) \oplus \ldots = X \tag{26}$$

$$K_r \oplus (K_r << 2) = X \tag{27}$$

The application of this approach for masking the value of the round key provides the inability to obtain the key value even with known input and output function values. The selection of bits on which the shift is performed, is providing an opportunity to mask the round key along its length.

Creating a table of approximations for the linear cryptanalysis and creating a table of differentials for the differential cryptanalysis is complicated by the fact, that there is no possibility of abstraction of the key value for each concrete case, because there are no operations performed only with the input value. All operations are dependent on the input value as well as on the key – this is why the creation of tables is needed for each new key value.

Implementation of the slide attack for this algorithm is meaningless, because in addition to the above complexities in the allocation of the key given algorithm has the method of forming round keys, that provides different round keys for each round. Therefore, it is impossible to share cryptographic algorithm to the equal parts that are repeated.

Testing of the *cryptographic algorithm using statistical tests.*

Since any encryption algorithm can be considered as a pseudorandom function which depends on the input text and the key, then to analyze other statistical characteristics of the cipher testing with various sets of statistical tests (DIEHARD, NIST) was carried out [10]. Some test results using the NIST test are shown in Table 1.

Table 1

NIST test results for cryptographic algorithm [11]

No.	Name of test	Result (*p-value*)
1	Frequency (Monobits) Test	0.739918
2	Cumulative Sum (Cusum) Test	0.122325
3	Runs Test	0.122325
4	Test For The Longest Run Of Ones In A Block	0.911413
5	Random Binary Matrix Rank Test	0.350485
6	Discrete Fourier Transform (Spectral) Test	0.066882
7	Non-Overlapping (Aperiodic) Template Matching Test	0.350485
8	Approximate Entropy Test	0.213309
9	Serial Test	0.534146
10	Linear Complexity Test	0.739918

These tests are various statistical tests, results of which are interpreted according to the methodology of determining the probability *p-value*. At small values of probability, the sequence is not considered random and for successful testing, the probability should be higher than 0.01 [11].

According to the testing results, the value of the variables of *p-value* greater than 0.05 was obtained, that is quite a good result

6. Conclusions

In the article, the methods and approaches of the construction of block ciphers have been analysed – several issues related to the security of the cryptographic algorithm as well as the concept of cryptographic strength have been examined. These facts allowed us to determine the requirements for modern cryptographic systems and for block ciphers.

The block encryption algorithm based on the Feistel transformation that is resistant to cryptographic attacks using modern methods of cryptanalysis has been presented. Constructive solutions that allow masking the value of round keys even with the known input and output values of the function have been offered for this algorithm. It can be used in block ciphers to provide resistance to cryptographic attacks.

The most prevalent methods of cryptanalysis of block ciphers have also been reviewed, such as the brute force method, meeting in the middle, statistical method, linear cryptanalysis, differential cryptanalysis and slide attack.

The results of the investigation of the designed cryptographic algorithm have been presented by means of NIST tests, and its resistance to cryptographic attacks has been demonstrated.

References

[1] Sneier B., *Applied Cryptography*: Protocols, Algorithms, and Source Code in C, 1996.

[2] Feistel H., *Cryptography and Computer Privacy*, Scientific American, Vol. 228, No. 5, 1973.

[3] Hoang V.T., Rogaway P., *On Generalized Feistel Networks*, Dept. of Computer Science, University of California, Davis, USA. 2010, 26.

[4] Biham E., Shamir A., *Differential cryptanalysis of DES-like cryptosystems*, Journal of Cryptology, Vol. 4, No. 1, 1991, 3-72.

[5] Wagner D., *The boomerang attack*, U.S. Berkeley.

[6] Chalermpong Worawannotai, Isabelle Stanton. A Tutorial on Slide Attacks.

[7] Ciet M., Piret G., Quisquater J., *Related-Key and Slide Attacks: Analysis*, Connections, and Improvements, 002.

[8] Blaze M., Diffie W., Rivest R.L., Schneier B., Shimomura T., Thompso E., Wiener M., *Minimal Key Lengths for Symmetric Ciphers to Provide Adequate Commercial Security*, 1996.

[9] Stamp M., Low R.M., *Applied Cryptanalysis: Breaking Ciphers in the Real World*, Wiley-IEEE Press, 2007.
[10] Instructions for using DIEHARD: a battery of tests of randomness, 1997.
[11] www.csrc.nist.gov/groups/ST/toolkit/rng/stats_tests.html.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
2-AC/2013

ŁUKASZ SOSNOWSKI*

APPLICATIONS OF COMPARATORS IN DATA PROCESSING SYSTEMS

ZASTOSOWANIA KOMPARATORÓW W SYSTEMACH PRZETWARZANIA DANYCH

Abstract

This paper shows practical examples of compound object comparators and the application of the theory in various fields related to data processing systems. One can also find the necessary theoretical background needed to understand the examples.

Keywords: *compound objects, compound object comparators, networks of comparators, fuzzy sets, rough mereology*

Streszczenie

Niniejszy artykuł przedstawia praktyczne zastosowania teorii komparatorów obiektów złożonych w różnych aspektach dotyczących systemów przetwarzania danych. Dodatkowo został umieszczony skrót materiału teoretycznego pozwalającego na zrozumienie przykładów oraz ogólnej tematyki komparatorów.

Słowa kluczowe: *obiekty złożone, komparatory obiektów złożonych, sieci komparatorów, zbiory rozmyte, mereologia przybliżona*

* M.Sc. Łukasz Sosnowski, Ph.D. Doctoral Studies, e-mail: l.sosnowski@dituel.pl, Systems Research Institute, Polish Academy of Sciences.

Symbols

$C(a, B)$ – compound objects comparator
A – set of input objects
B – set of reference objects
a – input object, an element of A
b – reference object, an element of B

1. Introduction

Data processing systems employ many techniques that allow the synthesis of analysis, and interpretation of data. One of the subsets is a class of decision support systems. They implement solutions based on neural networks [2], evolutionary algorithms [13], immune systems, energy based [7] solutions, etc. All of these techniques are suitable for solving a selected class of problems. They require a specific approach to the construction of the solution for particular problems. The theory of compound objects as well as a network of comparators is an alternative and complementary approach to the already used methods for solving decision problems. It is a kind of methodology which helps to construct one's own solution which solves these kinds of problems. It is characterized by a common approach to every kind of problem, which greatly facilitates the management of unknown cases. Optimal solutions are not guaranteed in many cases, but in a reproducible and easy implemented way, they give the opportunity to obtain satisfactory results. The theory of comparators was described in earlier papers [15, 16, 17, 21, 22, 27] and it has evolved into the form of networks [23, 26]. This network has the ability to learn from using the known techniques [9].

The article is organized as follows: the introduction is located in this section; chapter 2 consists of an abstract of the theoretical basis used in the case of the problems described; chapter 3 presents various examples of applications developed by means of the methods presented. These examples exhibit different degrees of complexity, from very simple to advanced practical problems. The final part contains a section providing a summary of issues presented and the bibliography upon which this article is based.

2. Preliminaries

2.1. Ontology

The terms are borrowed from philosophy, but it is now also frequently found in the field of artificial intelligence. The formal definition (one of many) was introduced in 2001, see [4, 8]. Its meaning is as follows: Ontology is a system marked as O which specifies the structure of concepts, relationships between them, as well as the theory defined on a model. It is defined in the following form:

$$O = \{C, R, H_c, \text{rel}, A, L\} \tag{1}$$

where:

- C – is the set of all concepts of the model and the concept is called the idea of representing a group of objects with common characteristics,
- R – set of non-taxonomic relations defined as signed connections between concepts [1],
- H_c – collection of taxonomic relationships between the concepts,
- rel – defined non-taxonomic relationships between the concepts,
- A – set of axioms,
- L – lexicon specifies how to understand concepts (including relations).

L is the following set:

$$\{L_c, L_r, F, G\} \tag{2}$$

where:

- L_c – lexicon definitions for concepts,
- L_r – lexicon definitions for a set of relationships,
- F – references to concepts,
- G – references to relationship.

In this case, ontology will be used as a set of concepts describing objects and its structure using relations. It will help to describe the features of objects as well as designate a features reduct [25]. It will be a necessary tool for the process of recognition and identification.

2.2. Simple and compound objects

It can be said that the information-based world consists of objects that can be both the source and recipient of information. Objects that exist in the surrounding area can be divided into compound objects (X_c) and simple objects (X_s). Simple objects will be called atomic, indivisible objects which are ordinary entities that have certain characteristics, but are not composed of other objects. While compound objects have their structure and may consist of other objects, either simple or compound. The simple object is any element of the real world having its representation capable of being expressed by the adopted ontology (O). In addition, the following properties arising from their ontological representation may be assumed:

1. An object always belongs to a certain class or a fixed number of classes in ontology. A single object may belong to several classes.
2. An object has a property within a class. Features may vary from class to class.
3. An object may be related to other objects in the same ontology.

The compound object is composed of other objects defined by means of ontologies (connects them) and forms a new entity. The compound object has its specification describing structure, relations and connections between sub-objects. Compound objects satisfy the following additional properties [26]:

1. A minimum of two objects can be extracted from them and can be independent entities.
2. Component objects are interrelated with ontology O.

2.3. Compound object comparator

Compound object comparators are multi-layered structures consisting of several closely interlinked components for determining the similarity between objects. A comparator can be formally described with the following function:

$$C_B : A \to 2^{B\times[0,1]} \tag{3}$$

where:

A: is a set of input objects,

B: is a set of reference objects.

Comparator outcomes take the form of weighted subsets of reference objects:

$$C_B(a) = F\left(\{b, g(\mu(a,b)) : b \in B\}\right) \tag{4}$$

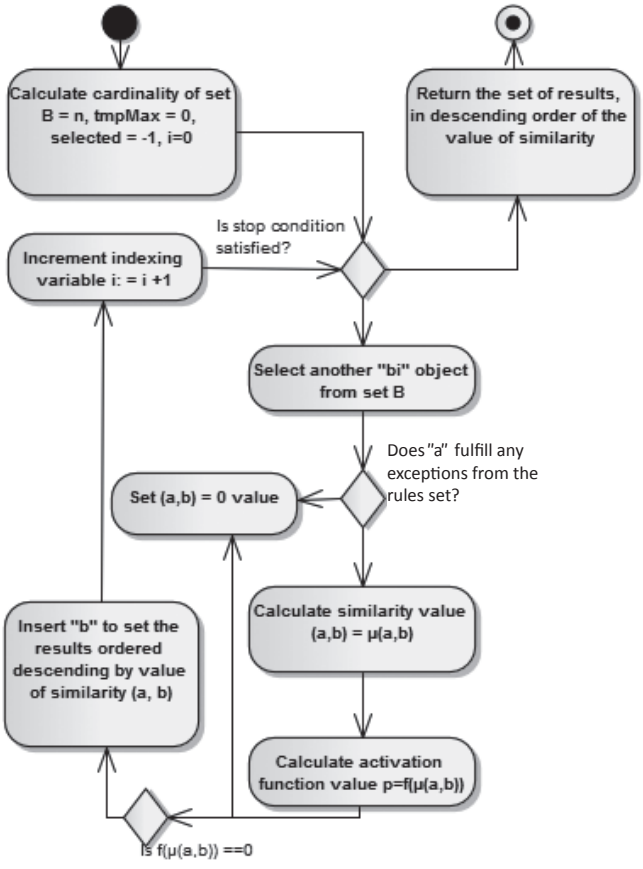

Fig. 1. UML activity diagram of compound object comparator

where:

F – is a function responsible for filtering partial results, e.g. min, max, top,
$\mu(a, b)$: – is a membership function of the fuzzy relation [5], which returns a similarity degree between $a \in A$ and $b \in B$,
$g(x)$: – is an activation function which filters out results that are too weak.

The following is put:

$$g(x) = \begin{cases} 0 : x < p \\ x : x \geq p \end{cases} \tag{5}$$

where:

p – denotes the lowest acceptable similarity [15].

One may also introduce some constraints which make $\mu(a, b_i) = 0$ based on the so-called exception rules [23].

2.4. Information granule

The first definition of information granules was proposed by Lotfi Zadeh [34, 35] in the 70's and read as follows: *'An information granule is a clump of objects of some sort, drawn together on the basis of indistinguishability, similarity, or functionality'* [15]. In this case, the granule is used to represent input object and its closest environment built from reference objects. We will build them through different phases of comparisons.

2.5. Rough mereology

The theory of mereology was developed by S. Leśniewski in the second and third decades of the twentieth century. The basic idea is the relation of 'being part of the whole'. Rough mereology [11] is an extension of classical mereology. It answers the question 'to what degree X is a part of Y' by using the rough inclusion defined as:

$$\mu(X, Y) : U \rightarrow [0, 1] \tag{7}$$

where:

U – is a finite set of objects called 'Universe'.

This function has been developed as a tool to approximate reasoning and has the following properties:

1. $\mu(x, y) \in [0, 1]$
2. $\mu(x, y) = 1$, $\forall x \in U$
3. if $\mu(x, y) = 1$ and $\mu(x, y) = 1$ then $\mu(x, z) \geq \mu(y, z)$ for any triple $x, y, z \in U$
4. there is a null object n, in the case of which $\mu(n, x) = 1$, $\forall x \in U$

The most common form of the rough inclusion function is based on the quantity of elements. It is defined as:

$$\mu(X, Y) = \frac{\text{card}(X \cap Y)}{\text{card}(X)} \tag{8}$$

Here, mereology is a very important part of the solution. In many cases, the entire compound object cannot be compared. The structure of objects is often very complicated. The difficulty lies in choosing appropriate measures of distance. It turns out that during decomposition of the object into sub-objects (sometimes repeatedly), a level of complexity of the structure is obtained. Therefore measuring the distance between sub-objects becomes easy. Use of mereology based on the similarity of sub-objects facilitates the determination of a degree of resemblance of the objects at higher layers (more complex).

2.6. Network of comparators

Networks of comparators are one-way networks designed for examining the similarity of compound objects in an adaptive way. This means that further layers can take into account results obtained in the previous layers. At the same time, they allow for the building of complex (composite) structures based on the captured specification of the compound object. There are two basic types of networks categorized according to their approach to the object:

Type 1 – network of monolithic objects called the homogeneous network.

Type 2 – network of structural objects called the heterogeneous network.

The network is composed of layers. They communicate through connections of their elements. Connections between neighbouring layers may be formed in several ways, depending on the type of network in question.

At the beginning of the process, the element *a* in question is put at the input of the network. It is a point to start creating an information granule around it (in accordance with the definition 2.4). The granule consists of reference objects connected through the similarity relation with input object *a*.

Elements located in a layer are not connected. Each comparator examines the selected features and returns the independent results (individual comparators perform their calculations independently). The network structure is based on groups of independent features and characteristics or analysis of the structure of object. It depends on the type of network used. The network layer can be derived either from the context associated with a group of features (treating the object as a single entity) or from the analysis of the structure of object and its relations (processing object as part of another object or considering the the object to be a set of smaller sub-objects). In the latter case, it is possible to make generalizations of object *a* as well as the decomposition of sub-objects. Depending on the position of the object in the structure (context), various object relationships are considered. Both relationships make the input granule (around *a*) grow with sub-objects or over-objects. For details about subtypes of networks see [23, 24].

2.6.1. Elements of the network

The general scheme of the network is shown in Figure 2. The network in question consists of many independent elements, i.e.:

1. *Layer* – part of a network linked to a common context of comparison. It can include common features examined by different comparators. There are three types of layers: input; intermediate; output. Input and output layers are necessary in the network, while the intermediate layer is optional. Each layer of the network has a different context in respect of the object. This context depends on the domain knowledge of the object, namely its relationship with other objects or sub-objects, e.g. both the relationship for results of decomposition of the object, or the fact that the object (in question) is a sub-object of a different object.

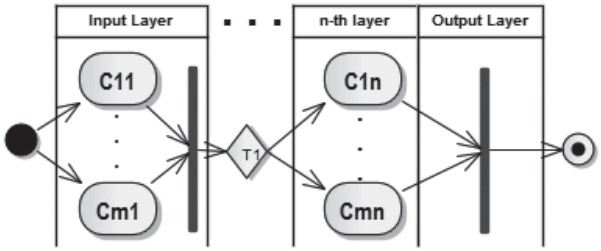

Fig. 2. General scheme of the network. C_{ji} – comparators, T_j – translators, element in output

2. *Input* – place where the input object *a* is converted into its set of representations. This is created for each comparator from the whole network. It is also a starting point of processing.
3. *Compound object comparator* – the unit enables the execution of compound object comparison, as defined in Section 2.3. There are many types of compound object comparators. Classification is listed in [18]. There may be many different types of comparators in one layer. Their selection depends on reasons, efficiency or the chosen solution to a particular problem.
4. *Translator* – component of a network associated with the adaptation of results to the context of the layer to be fed with. In other words, this element expresses the results of the previous layer (reference objects) using reference objects used in the next layer, taking into account relationships between the objects of these two layers. This element is optional due to the fact that not every network will be capable of using it or its use would be redundant in certain cases (for the same set of reference in various layers).
5. *Aggregator* – mandatory part of the network responsible for the synthesis of results obtained by comparators. This element is always present in the output layer. However, it may also be included in other layers, depending on the network structure. Aggregators are functions that operate on partial results of comparators.
6. *Output* – point where the processing results of the object in question over the network are obtained. Results are obtained in the form of a subset of reference objects which were processed by the aggregator.

2.5. Signal granule

The signal granule is based on the Information granule. In this case, the granule will be used to represent an input object and its closest surrounding area built from reference objects. It will be a representation of a signal moving through the network from layer to layer. The content will differ at every step, depending on comparisons already made and layers of the network visited. The outline of the signal granule is presented in Figure 3. The shape of the granule depends on the kind of network used and the reference set. The simplest one is built on the same reference set, so that subsequent parts of similar object are just subsets of the same reference set. In other situation they can come from totally different sets representing completely different features or groups of features.

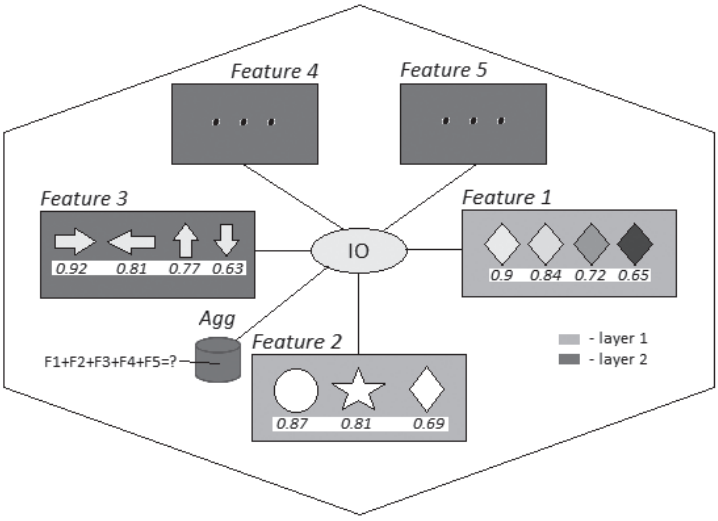

Fig. 3. General schema of signal granule used for transfer of information between comparators, aggregators, layers, in the network of comparators. Feature *n* – processed feature, IO – input object, Agg – aggregated data

3. Examples of applications

3.1. Standardization of pharmaceutical data – the introductory example

The task is to match the text data from different data sources (pharmaceutical warehouses). Each system might have a different ontology as well as local dictionaries of pharmaceutical products [15]. Input objects and reference objects have a common structure (according to the international standard – brand, form, dose etc.). Let us assume that the data sets $A = \{a_1\}$, $B = \{b_1, b_2\}$ and $Z_1 = \{z_{11}\}$, $Z_2 = \{z_{21}\}$ are the following:

a_1 – 'Vitamin C, 100 mg, Acme, tablets',
b_1 – 'Vitamin E 100 mg, Acme tablets',
b_2 – 'Vitaminum C 100 mg, Acme tablets',
z_{11} – 'C',
z_{21} – 'E'.

and there is a factor $p = 0.6$. This simple example illustrates the usage of a single compound object comparator. Let us consider the similarity of elements a_1 and b_1, b_2 using the elements z_{11}, z_{21} of the sets of exceptions. The fuzzy relation, which is the base of the comparator in question, is the following:

$$\mu(a, b) = 1 - \frac{D_L(a, b)}{\max\{n(a), n(b)\}} \tag{9}$$

where:

$D_L(a, b)$ – Levenshtein distance between a and b,

$n(a)$ – length of string a,
$n(b)$ – length of string b,
and the characteristic function of comparator is max().

First of all, Levenshtein distance for the corresponding pairs is calculated and hence the following is obtained: $D_L(a_1, b_1) = 5$, $D_L(a_1, b_2) = 5$, then the value of the maximum function of the length of the string is calculated and gives respectively:

$$\max(n(a_1), n(b_1)) = 31 \text{ and } \max(n(a_1), n(b_2)) = 31$$

The following formula with the membership function is obtained: $\mu(a_1, b_1) = 0.8387$ and $\mu(a_1, b_2) = 0.8387$. One may note that as far as the terms of membership (similarity) are concerned, the two pairs are equivalent. The exceptions criterion is the only thing to be checked. Thus, for a pair (a_1, b_1) it has to be checked whether a_1 contains a sub-string defined by z_{11} (rule condition). In this case, it can be noted that the rule is satisfied, so item b_1 must be rejected for item a_1. The only thing that remains to be checked is whether the rule is satisfied for the pair (a_1, b_2). In this case it is clear that the string represented by a_1 does not contain a string represented by z_{21} ('E'). Only the pair (a_1, b_2) is to be considered further in accordance with the algorithm. Finally, the value of the activation function is checked. For the pair in question, the function value is 0.8387, which is within the range [0.6, 1], which sets the parameter p. Extract the maximum value of membership function. In this case, it is 0.8387. Thus, in this example, element $((a_1, b_2), 0.8387)$ is the result of standardization.

The above mentioned example shows how easy it is to use comparators for resolving data processing problems. This particular usage is very simple, based only on a single comparator. It can be compared with single neuron solutions known from neural networks [2]. Despite the fact that this is only a single comparator, it is a very powerful solution. We can choose many different membership functions and in this way, solve many other processing problems only with this change.

3.2. Contour map identifications

This example shows the practical implementation of the comparators theory in query by example solution of searching contour area maps in Poland (voivodships, counties, communes). The algorithm devised to do that, introduced in [21], consists of three phases: segmentation, granulation [10] and identification. Firstly, objects are acquired, e.g., a given image of contour map or a drawn map. Secondly, granules of objects are extracted, their characteristics are computed, and synthesized into granular descriptions of objects. Only two types of characteristics are considered, although the framework is open for adding more types. In the case of the first type, referred to as coverage, the histogram technique is used (see [3]) to compute a vector of overlaps of image's granules with area that the image represents. In the case of the second type, referred to as contour, linguistic description [28] of directions of lines connecting the extrema points of the area's contour within each granule is produced. In the third phase, similarities between the input and reference objects are computed with respect to each type of granular description, and similarity scores are synthesized. For linguistic descriptions of contours, the following membership function is defined:

$$\mu_{\text{contour}}(a,b) = 1 - \frac{D_L(a,b)}{\max\{n(a), n(b)\}} \tag{10}$$

where:

$D_L(a, b)$ – Levenshtein distance between linguistic descriptions of objects a and b,
$n(a)$ – length of string a,
$n(b)$ – length of string b.

With regard to the granules' coverage, the following may be considered:

$$\mu_{\text{cov}erage}(a,b) = 1 - \frac{\sum_{i}^{n\times m} |\,\text{cov}_i^a - \text{cov}_i^b\,|}{n \times m} \tag{11}$$

where:

cov_i^a – factor of number of pixels belonging to interior area of granule i of object a to all pixels of granule,
n, m – granulation parameters.

For the purposes of this example, the aggregated similarity will be used:

$$\mu_{agg}(a,b) = \frac{1}{2}\left(\mu_{\text{contour}}(a,b) + \mu_{\text{cov}erage}(a,b)\right) \tag{12}$$

Finally, an additional procedure is implemented for cases that meet the following conditions:

1. If a given reference object was not chosen for any investigated object, then it should be used for the most similar unidentified object, even if its degree of similarity is not greater than the activation threshold.
2. If a given reference object was chosen for many investigated objects, then it should be used for the most similar of them and the remaining ones should be re-identified, excluding the already used reference objects.

The method presented above introduces the use of comparators in networks. There are 2 comparators in one layer and an aggregator. There are a lot of possible variations of implementing aggregators, e.g. using an elections algorithms [20]. The advantages of this method is the ease of construction solutions and that it is a human readable solution. We can easily modify the solution by adding more comparators that are specifically defined tasks. The weak point is the manual selection of a feature to examine. It requires a domain knowledge in form of an expert or well-built ontology and instances described by concepts. In practice, mixed solutions are preferred, that there is an ontology already built and also access to expert providing guidance.

3.3. Identification of authors of scientific publications

This example is based on the data processed in the SYNAT project (abbreviation of Polish 'SYstem NAuki i Techniki'), which was exported and prepared in the form of a flat file for the identification process. There is a set of authors, but saved in various publication standards

with possible minor errors (mistypings). The task is to identify the same objects inside the collection and produce a unique set of references. The input data are in the form of a table with three columns (id_instance, name, surname). The target set of input objects is very large (millions of records). To solve this problem, the previously described homogeneous network [23] can be used. This network has 3 layers: input; output; intermediate. The input layer consists of three comparators running concurrently: C_{sfl} – surname first letter comparator; C_{nfl} – name first letter comparator; C_{slen} – surname length comparator. The first two examine the first letter of the name and surname. The third one examines the length of surname with certain tolerances in both directions (e.g. one character). All of these comparators examine computationally simple features (so they can be efficiently computed), but they will be able to limit the search space for the next layer.

Table 1

Distribution of data in the input layer of the network of authors' identification

Initial of surname	Quantity	Initial of name	Quantity	Length of surname	Quantity
s	987	m	1027	6	1820
m	843	j	1014	7	1567
b	812	a	875	5	1515
c	645	s	814	8	1337
k	607	c	625	4	1016
l	588	r	560	9	886
h	577	d	518	10	494
g	507	p	463	3	439
p	489	k	435	11	285
w	479	g	432	12	170
d	449	t	428	2	153
r	438	l	407	13	119
t	388	e	369	14	83
a	355	h	355	16	56
f	315	b	304	15	43
v	293	f	252	17	36
n	262	n	243	19	21
j	213	y	219	18	20
e	178	i	167	20	13
z	170	w	167	21	9
o	154	v	141	23	7
y	145	o	92	22	5
i	94	z	57	24	2
u	55	x	57	1	1
x	34	u	57	26	1
q	21	q	20	–	–

The intermediate layer uses the results of the input layer by feeding the reference set with the results. In this layer, there are five comparators investigating the similarity of attributes: name, surname, sorted acronym, *nlm* category publications and co-authors. Results of these calculations are synthesised by the global aggregator. At the end, the result is obtained and is

interpreted as follows: if the network indicates similarity to a reference object, and it meets the minimum requirements for quality (activation parameter p), then the input object is matched with an already existing element of the reference set. If the answer of the network is empty, then the input object is treated as a unique object (new), and it is added to the set of reference objects. Such proceedings shall be set for each element of the input.

The scheme used in the network is shown in Figure 4. The sample of experiments consisted of 10098 records. Distribution of data for the comparator input layer is presented in Table 1. It can be noted that the largest class of objects amounts respectively to: 's' – 987, 'm' – 1027, '6' – 1820. The aggregator of the input layer selects data to be passed to the second layer. It calculates the arithmetic mean of the values of partial similarities (coming from a single comparator) for each pair (a, b_i), where b_i is a reference object. Subsequently, only those with the highest value are selected. The last two comparators from the intermediate layer will be mentioned. In this case, some additional domain knowledge of an input object is needed. Information about the source publication and its assigned category is needed, too. The comparator examines whether the defined publication falls into the categories of publications assigned to authors who were objects in the reference set. In the future, one may consider relevant categories and explore partial membership in a particular category (as a membership in the related class). The last comparator examines the co-authors, e.g., whether a given author has publications with co-authors from the set of reference.

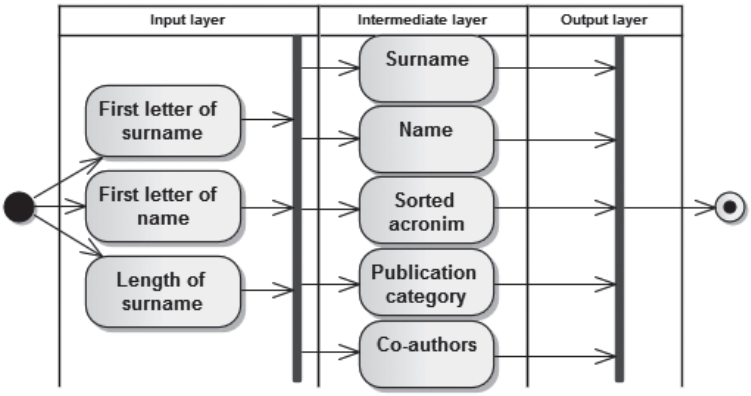

Fig. 4. Network of compound object comparators for identification of authors of scientific publications

This example is quite similar to hash algorithms known in literature to handle problems with huge data cluster processing. It consists of the pre-selection of objects, among which the final result is searched. The preselection is performed by means of simple operations that can be done very efficiently, while the exact match requires more processing power. The strength of this solution is the ability to build complex processing solutions composed of simple parts, fully comprehensible and manageable by humans.

3.4. SYNAT – identification of 'semantic blobs'

This section refers to an academic project in the area of semantic search over a repository of scientific publications. One of the tasks was parsing and interpreting publication references.

Interpreting means understanding which part of reference signifies an author, title, year of publication etc. Such items should be interpreted, parsed and loaded as publications themselves, but the online text analysis algorithms may not handle them credibly enough. In the SYNAT project, there is the strategy of loading such descriptions into the repository as unparsed strings with no properties and related instances assigned. Then, there is a place for using a specific module based on the network of compound objects comparators for extracting important information from such strings. The idea is to discover internal structures of input bibliography descriptions by comparing their dynamically identified components with various types of objects already stored in the repository. Such components can then be treated as new entries in the system, ready for further analysis. The heterogeneous network is considered in Figure 5.

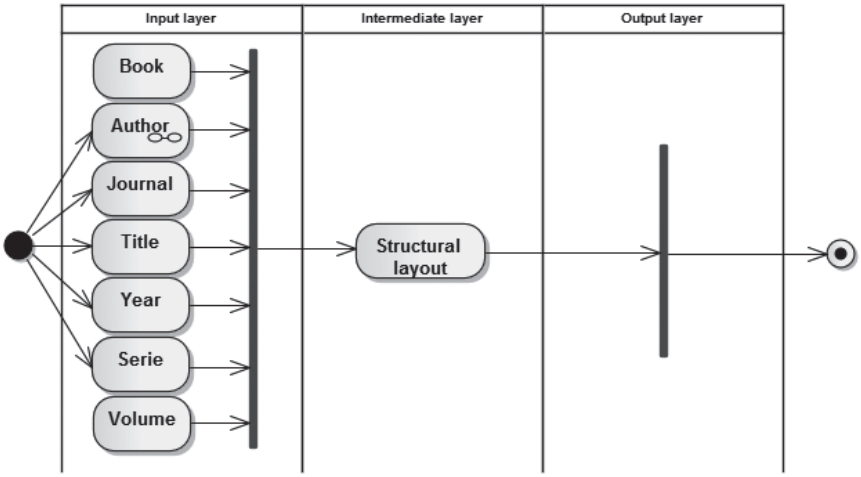

Fig. 5. Network for the identification of publication authors

The process begins with cutting the input text into pieces, e.g. by means of characters such as dots, commas, etc. For it is not known which pieces correspond to particular types of bibliography item components (more formally, properties and relations of the publication in question with some other objects), the first network's layer focuses on the computation of resemblance degrees of all pieces to reference sets maintained in the repository for scientists, authors, publication titles, years, journals and others [24].

Table 2 shows a simplified example of the outcome of stages of text decomposition and resemblance calculation. As in certain cases, the assignments of sub-strings to particular component types are problematic, the identification process is strengthened by the analysis of a bibliography item's structure at the next network layer. Table 3 presents several examples of reference structural objects, i.e. structures of bibliography items already stored in the repository. The structural layout comparator in Figure 5 determines the most reasonable hypothesis about the input's structure basing on such reference objects, in combination with information obtained in the previous layer. The final assignment is retrieved by following the most probable structure. This last step can be implemented in various ways. One of the possible solution is using a rough mereology approach on the final aggregation stage.

Table 2

Input and its decomposition with degrees of resemblance of its components to the corresponding reference sets

Example of the input object	Comp.	Substring	*A*	*T*	*Y*	*P*	*J*
M. Szczuka, Ł. Sosnowski, A. Krasuski,	1	M. Szczuka	0.83	0	0	0	0
K. Kreński, 'Using Domain Knowledge	2	Ł. Sosnowski	0.83	0	0	0	0
in Initial Stages of KDD: Optimization	3	A. Krasuski	0.85	0	0	0	0
of Compound Object Processing',	4	K. Kreński	0.84	0	0	0	0
Fundam. Inform., 2014, to appear.	5	Using (…)	0	0.75	0	0	0
	6	Fundam. (…)	0	0.42	0	0	0.55
	7	2014	0	0	1	0	0
	8	to appear	0	0	0	0	0

Table 3

Sample of reference structures

Abbreviation dictionary		Reference structural objects
A – authors	*T* – title	*A T B S V Y P*
J – journal	*B* – book	*E T B S Pb Y V*
Y – year	*Pb* – publisher	*A T B Y*
P – pages	*S* – series	*A T J Y*
V – volume	*E* – editors	*A T S Pb Y*
N – note ...		...

The solution described above is a popular scientific publication processing problem. There have been many approaches to solve this problem. One of them is the network of comparators. This example shows that comparators provide a universal approach to many processing problems using the same techniques and tools. In this solution, it should be noted that there is an incremental algorithm used which adds unknown cases to the reference set. It is one of the methods of learning classified within the class of lazy learning.

3.5. Optical character recognition of digits

In order to achieve automatic character recognition, a network of compound object comparators shown in Figure 6 has been developed. Before the proper processing is initiated [6], using the designed network, the pre-processing stage for every object has to be ensured. Consequently, a segmented string as a set of individual images is obtained. Each image represents one digit (as a single object). Next, each of these objects must be converted to binary scale (only black & white). After that each image is cut in such a way that each edge of the newly created image is one pixel further from the black edge of the font.

The network constructed is a type I [23] (for monolithic objects), made up of three layers representing particular contexts. The first layer covers a very general nature (rough) features. They can significantly reduce the cardinality of the reference set. The second layer relates to an in-depth analysis of the image [12], thanks to which the final answer is obtained (decision).

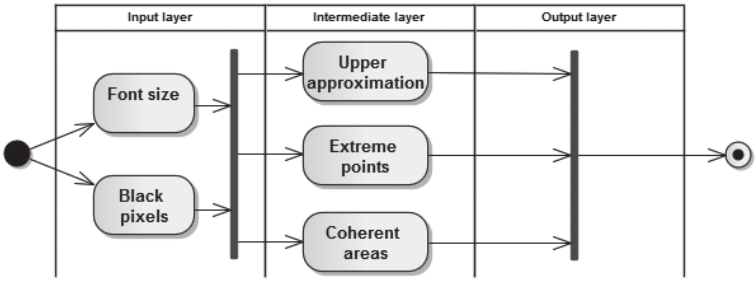

Fig. 6. Network of comparators for the OCR task

The last layer in the classical way only contains an aggregator that performs the synthesis of previous results. The reference set consists of objects which are images of searched characters (digits) grouped in terms of different sizes and font types. In this example, the following structure of the set is used: the size of the font (10, 14, 18, 24, 36, 48, 60); types of fonts (Times New Roman, Arial, Verdana, Courier). The cardinality of the entire reference set is 280 elements. The purpose of the first layer is to limit the cardinality of the reference set. The method applied is to analyse the size of the font and distribution of pixels in the image.

A set of fonts with the size closest to the input object is selected. Then synthesis of the results returned by both comparators is performed. After that, reference objects which meet the imposed specific qualifying criteria are selected. The result of the aggregator feeds the next layer, as a set of reference. In the intermediate layer, there are three comparators: Upper approximation; Extreme points; Coherent area. The first comparator compares images arising from the granules as a result of the granulation process, but only those in which there are black pixels (activation of the granule). This means that the comparison of images is converted to a very low resolution (m×n – granulation parameters).

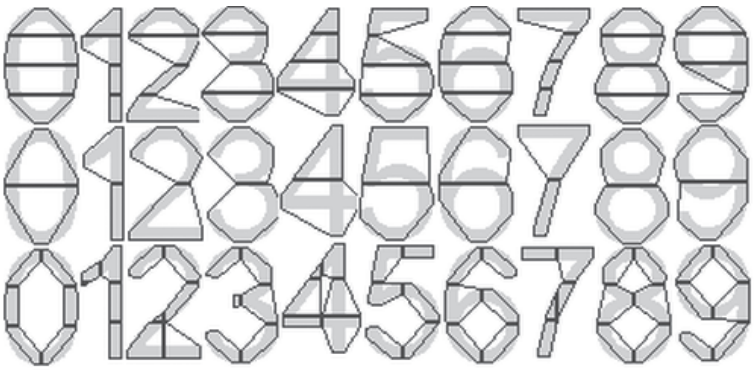

Fig. 7. Reference digits prepared for comparisons with respect to their shapes (with granulation resolution parameters 1 × 3, 2 × 2, 2 × 3 displayed in consecutive rows)

Another comparator deals with the comparison of geometric obtained by means of the connection of extreme points of subsequent granules (see Figure 7). Further action is to take contours of subsequent shapes and make comparisons with methods developed and published

earlier [21]. The third comparator is prepared to compare the coherent areas within objects. In this way, easier recognition of following digits: 0,4,6,8,9 is supported. These figures represent 50% of the possible characters defined for the task. This comparator not only detects their presence, but also indicates the most resembled areas in objects of reference. Consequently, the calculated results of the subsequent comparator feed output layer as input data of the aggregator. The aggregator calculates an arithmetic mean of resemblance for specific pairs (the input object and the reference object) and the subsequent comparators in question. In consequence, results are obtained in the form of a reference object or objects. In this way, identification of individual objects is performed. It is worth noting that individual comparators operate at different representations of the input object and reference objects, although each time they refer to the same signal granule (information granule [14]).

The method presented above is one of the many known and available methods in the range of digits recognition problems. This conclusion confirms the possibility of using it in many different fields of applications. The method in this regard does not discover new things, but allows using well-known methods from other applications. This gives the possibility to use the experience gained previously to instantly develop new solutions.

3. Summary

This article presents compiled examples of different applications of comparators in the field of data processing systems, putting special emphasis on decision support systems. The examples collected show how easy the implementation is for various problems (e.g. identification, matching, searching, etc.). In each of the reported cases, the procedure is analogous, which is undoubtedly a very big strength of this solution. The examples presented are achievements of earlier researches. This made it possible to collect them in this work in ahigh concentration. This allows for agood understanding of these applications and the theory in question.

Future work focuses on other applications of the framework. One of them is the identification of risks in the fire rescue actions used in the ICRA [19] project (www.icra-project.org). This is a wide field of possible applications of the said method.

The next step will also be to develop the learning methods with regard to the solution in question. One of them will be an evolutionary algorithm for searching optimal weights of comparator importance in the layer. Another will be based on the back propagation algorithm for improving the basic parameters of single comparator connections between layers.

Another possible step is the publication of the framework's implementation written in Java as a freeware licence for greater and easier access for researchers. This is important in the case of gaining popularity for this method of AI.

Acknowledgements:
The research was supported by the Polish National Centre for Research and Development (NCBiR) – Grant No. O ROB/0010/03/001 in the frame of Defence and Security Programmes and Projects: 'Modern engineering tools for decision support for commanders of the State Fire Service of Poland during Fire&Rescue operations in the buildings'.

References

[1] Dean Allemang and James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[2] Bishop C.M., *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[3] Deb S. (ed.), *Multimedia Systems and Content-Based Image Retrieval*, IGI Global, 2004.

[4] Gliński W., *Ontologies – an attempt to put terminological chaos in order*, [in:] *From scientific information to the technology of information society: collective work*, Miscellanea Informatologica Varsoviensia, SBP, 2005.

[5] Kacprzyk J., *Multistage Fuzzy Control: A Model-Based Approach to Fuzzy Control and Decision Making*, Wiley 1997.

[6] Kuznetsov S.O., *Pattern Recognition and Machine Intelligence*, 4th International Conference, PReMI, 2011.

[7] LeCun Y. et al., *A Tutorial on Energy-Based Learning*, [in:] Bakir G. et al., Predicting Structured Data, MIT Press, 2006.

[8] Maedche A., Staab S., *Comparing Ontologies – Similarity Measures and a Comparision Study*, Internal report No. 408, University of Karlsruhe, 2001.

[9] Mitchell T.M., *Machine Learning*, McGraw-Hill, 1997.

[10] Pedrycz W., Kreinovich V., Skowron A. (Eds.), *Handbook of Granular Computing*, Wiley 2008.

[11] Polkowski L., *Approximate Reasoning by Parts an Introduction to Rough Mereology*, Springer 2011.

[12] Reed T., *Digital Image Sequence Processing, Compression and Analysis*, CRC Press, 2005.

[13] Rutkowski L., *Computational Intelligence*, Methods and Techniques, Springer 2008.

[14] Skowron A., Stepaniuk J., Peters J.F., *Towards Discovery of Relevant Patterns from Parametrized Schemas to Information Granule Construction*, [in:) (Eds.) I. Masahiro, H. Shoji, T. Shusaku, *Rough Set Theory and Granular Computing*, Springer Berlin Heidelberg 2003, 97-108.

[15] Sosnowski Ł., *Inteligentne dopasowanie danych przy użyciu teorii zbiorów rozmytych w systemach przetwarzania danych*, Analiza systemowa w finansach i zarządzaniu, T. 11 pod redakcją prof. J. Hołubca, 2009.

[16] Sosnowski Ł., *Budowa systemu porównywania obiektów złożonych*, Analiza systemowa w finansach i zarządzaniu, T. 12 pod redakcją prof. J. Hołubca, 2010.

[17] Sosnowski Ł., *Identification with compound object comparators – technical aspects*, Techniki informacyjne teoria i zastosowania, T. 1 pod redakcją prof. J. Hołubca, 2011.

[18] Sosnowski Ł., *Identyfikacja obiektów złożonych przy użyciu komparatorów*, Techniki informacyjne teoria i zastosowania, T. 2 pod redakcją prof. Myślińskiego, 2012.

[19] Sosnowski Ł., Pietruszka A., Krasuski A., Janusz A., *A resemblance based approach for recognition of risks at the fire ground*, [in:] Active Media Technology Proceedings, 2014, to appear.

[20] Sosnowski Ł., Pietruszka A., Łazowy S., *Election Algorithms Applied to the Global Aggregation in Networks of Comparators*, [in:] FedCSIS 2014 Aaia Workshop, to appear.

[21] Sosnowski Ł., Ślęzak D., *Comparators for Compound Object Identification*, [in:] Proc. of RSFDGrC, LNAI 6743, Springer 2011, 342-349.

[22] Sosnowski Ł., Ślęzak D., *RDBMS Framework for Contour Identification*, [in:] Proc. of the international workshop CS&P 2011, Białystok University of Technology, 2011, 487-498.

[23] Sosnowski Ł., Ślęzak D., *Networks of Compound Object Comparators*, [in:] FUZZ-IEEE 2013, IEEE International Conference on Fuzzy Systems, Hyderabad, India, July 7–10 2013, Proceedings. IEEE, 2013.

[24] Sosnowski Ł., Ślęzak D., *How to design a network of comparators*, [in:] Brain and Health Informatics, 2013, 389-398.

[25] Stawicki S., Ślęzak D., Recent advances in decision bireducts: *Complexity, heuristics and streams*, [in:] Lingras P., Wolski M., Cornelis C., Mitra S., Wasilewski P., eds.: RSKT, Volume 8171 of Lecture Notes in Computer Science., Springer 2013, 200-212.

[26] Szczuka M., Sosnowski Ł., Krasuski A., Kreński K., *Using Domain Knowledge in Initial Stages of KDD: Optimization of Compound Object Processing*, Fundamenta Informaticae, 2014.

[27] Ślęzak D., Sosnowski Ł., *SQL-Based Compound Object Comparators: A Case Study of Images Stored in ICE*, [in:] Proc. of ASEA, CCIS 117, Springer 2010, 304-317.

[28] Szczepaniak P., *Computational Intelligence and Applications*, Springer-Verlag, Heidelberg, New York 1999.

[29] Zadeh L., *Towards a theory of fuzzy information granulation and its certainty in human reasoning and fuzzy logic*, Fuzzy Sets Systems, 90, 1997, 111-127.

[30] Zadeh L., *Key roles of information granulation and fuzzy logic in human reasoning, Concept formulation and computing with words*, Proceedings of IEEE 5th International Fuzzy Systems.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
2-AC/2013

FERENC LILIK*, PÉTER SIMONYI**, LÁSZLÓ T. KÓCZY***

COMPUTATIONAL INTELLIGENCE IN PERFORMANCE EVALUATION AND FAULT PROGNOSIS IN TELECOMMUNICATION ACCESS NETWORKS

METODY INTELIGENCJI OBLICZENIOWEJ W OCENIE WYDAJNOŚCI I PROGNOZOWANIA USZKODZEŃ DOSTĘPOWYCH SIECI TELEKOMUNIKACYJNYCH

Abstract

Telecommunication connections are highly reliable and manageable, however, the handling of several parts of the networks is problematic. One of these parts is the access network. The variegation of the applied technologies and the individual connections to the customers in access networks makes the preliminary estimation of the performance of the telecommunications services and troubleshooting difficult. There are existing methods which can handle such problems, but the telecommunications companies (TELCO) are continuously looking for newer and more efficient methods. In this paper some existing methods for performance evaluation and the prediction of the probable failures of the wire pairs of telecommunications access networks are reviewed and novel methods that are based on the measurements of the wire pairs and use computational intelligence, fuzzy inference methods and evolutionary models are introduced.

Keywords: *telecommunications access networks, performance evaluation, fault prediction, fuzzy rule bases*

Streszczenie

Połączenia telekomunikacyjne są z reguły wysoce niezawodne i łatwe w zarządzaniu, jednak obsługa pewnych typów sieci, w tym tzw. sieci dostępowych, może przysparzać problemów. Zarówno różnorodność stosowanych technologii, jak też specyfika indywidualnych podłączeń klientów sprawiają, że wstępna ocena wydajności usług telekomunikacyjnych oraz wykrywania uszkodzeń napotyka trudności. Choć dostępne są metody rozwiązywania tego typu problemów, to firmy telekomunikacyjne stale poszukują nowych, bardziej skutecznych rozwiązań. W niniejszym artykule zawarto przegląd istniejących metod oceny wydajności i prognozowania uszkodzeń par przewodów w dostępowych sieciach telekomunikacyjnych, a także zaprezentowano nowe metody oparte na pomiarach tych przewodów, z użyciem technik inteligencji obliczeniowej, rozmytych metod wnioskowania oraz algorytmów ewolucyjnych.

Słowa kluczowe: *telekomunikacyjne sieci dostępowe, ocena wydajności, predykcja uszkodzeń, baza reguł rozmytych*

* M.Sc. Ferenc Lilik, e-mail: lilikf@sze.hu, Telecommunications Department, Faculty of Engineering Sciences, Széchenyi István University, Györ.

** M.Sc. Péter Simonyi, Invitel – Hungary.

*** Prof. D.Sc. Ph.D. László T. Kóczy, Department of Automation, Faculty of Engineering Sciences, Széchenyi István University Györ; Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics.

1. Introduction

Telecommunication connections are highly reliable and manageable, however, handling of several parts of the networks is problematic. One of these parts is the access network. The variegation of the applied technologies and the individual connections to the customers in access networks makes the preliminary estimation of the performance of the telecommunications services and the troubleshooting difficult. There are existing methods which can handle such problems, but the telecommunications companies (TELCO) are continuously looking for newer and more efficient methods. In this paper some existing methods for performance evaluation and the prediction of the probable failures of the wire pairs of telecommunications access networks are reviewed and novel methods that are based on the measurements of the wire pairs and use computational intelligence, fuzzy inference methods and evolutionary models are introduced.

1.1. Access networks and data transmission technologies

In the communication between two endpoints, plurality of telecommunications technologies and networks are used. The applied transmission technologies of optical and microwave (microwave chains and satellite) connections of the core/backbone network between the exchanges or nodes are very reliable. As these systems serve for the transmission of the communication of numerous customers at the same time, technical safety has a big role during the planning, installation and upkeep. Besides the substantial operation safety, these systems are continuously managed, thus the technical staff observe faults immediately, and dispose of the substitution and the reparation of the errored session or system. The parameters of the technologies used in core or backbone networks are well described, so the performance of its connections are known at the moment of installation, moreover, they are already known at the moment of designing.

The customers are connected to the local exchanges or data communication nodes by the access networks. 'The access network is an implementation comprising those entities (such as cable plant, transmission facilities etc.) which provide the required transport bearer capabilities for the provision of telecommunication services between a Service Node Interface and each of the associated User-Network Interfaces' [21]. In contrast with the connections of the core networks which perform mass communication, the individual connections of the access networks are dedicated to single customers. This results in such a variety of physical connections and lines of the access networks that significantly encumbers the unificability of the handling of singular lines. The operating and handling methods are the same for each connection, however each line is individual with specific physical performance and operational parameters. At this rate, the performance evaluation of the singular connections that can be changed also physically during the time by the replacement of their several sessions, without the installation of the technological equipment of the services is not easy. Moreover, the exact connection between the measurable parameters and the performance does not exists, performance can only be estimated. The situation of the prediction of probable faults is very similar. Lots of physical parameters of huge and complex access networks have to be continuously monitored in order to determine the declination of the connection's quality and forecast the possible faults. The exact mathematical formulas are missing in this case as well, forecasts can be given by previous experiences.

Nowadays, metallic cables can hardly be found in core networks, or are totally missing. They are replaced by modern fibre optical cables. However, due to economical reasons, symmetrical copper wire pairs based cables dominate the access networks and by present trends, these cables will be in operation for decades. Plain old telephone service (POTS) is more and more displaced by new types of communication services, as for example IPTV, VoIP or Internet access. The previous technologies of symmetrical copper wire based networks were made for the demands of POTS and are not able to provide new types of services, therefore the family of DSL (Digital Subscriber Line) technologies were evolved. The members of the DSL technologies enable the provision of different modes of data transmissions.

Asymmetric communication can be performed by ADSL (Asymmetric Digital Subscriber Line) technology [6]. Its data transfer rate (bit rate in other words) is different in upload and download. In accordance with the average use of the Internet, the upload bit rate is lower than the download. Internet connection, VoIP and also IPTV services are provided by this technology mainly to residential customers and SOHO (Small Office/ Home Office) users.

The symmetrical data transmission method of the access networks is SHDSL (Single-pair High Speed Digital Subscriber Line) [7]. This technology is used by those customers whose amount of data to upload is equal to or more than the amount of downloaded data.

These new transmission and communication systems are effective, however, they have high expectations of the networks they use. xDSL technologies are quite efficient in data transmission, but as they use a wider bandwidth and higher frequencies, the possibility of failure is also higher. However, not only this fact is the reason for the failures. Symmetrical cables which are used in this type of transmission are very sensitive to the effects from the environment and as most of them are rather old, failure of the wire pairs is common.

Towards the optimization of the operation, TELCOs are continuously looking for new methods that help them to perform effective performance evaluation and fault prediction in access networks. There are no exact mathematical models and solutions for these problems, handling is possible by expert systems for example. Such incorrectly determined problems can be managed by fuzzy reasoning methods.

1.2. Fuzzy systems

Some of the technological and logistical problems can hardly be handled by traditional logic systems. Fuzzy set theory and fuzzy logic, which is close to the human thinking and describes the phenomena of the world in an easy way, was proposed by L.A. Zadeh in 1965 [20]. Zadeh's logic system is such a many-valued logic in that the number of the values is infinite between the maximal 1 and the minimal 0, as opposite to Aristotle's two valued logic, which uses only the 1 and 0 (YES and NO) as permitted values.

Using fuzzy sets and fuzzy logic, rule systems defined by human language, that contain linguistic variables instead of precise numbers and values can be mathematically handled easily. Such linguistic variable is for example the height of the body if its values are not numerical, but linguistic, e.g. short, normal and tall. [11] The rules are IF...THEN typed as it can be seen in Section 3.3. and the connections between them are OR (ELSE). As the expressions in the rules contain fuzzy sets, substituting the logical operations with fuzzy set operations, fuzzy reasoning system is resulted, that allows the mathematical handling of the rules and rule bases defined by human language.

Fuzzy inference methods are successfully applied in many fields of science from automatic vehicle control [12] to handwriting character recognition [13]. The subject of fuzzy set theory and fuzzy logic is expansively described in [10].

2. Performance evaluation of the wire pairs in access networks using fuzzy rule bases

Performance evaluation in telecommunications is the process of the determination of the transmission capacity of a system or a connection. Its aim is the determination of the maximal available data transfer rate. TELCOs need this data in order to give preliminary offers for their respective customers. Service Providers use several methods of performance evaluation, however, these methods are expensive or imprecise. In this section, some current techniques will be briefly presented and a fuzzy based novel method will be introduced.

2.1. Widely used methods of performance evaluation [15]

In overviewing the related literature, we found several essentially different approaches to the performance evaluation of wire pairs. The most important methods will be briefly overviewed whilst pointing out problems involved with all these methods which essentially motivate our new approach. The key problem of performance evaluation may be found in the contradictory requirements of accuracy and low cost. The main types of such evaluation are described. In each case, the order of magnitude of the cost and an estimation of the achievable accuracy will be indicated.

Physical parameters of the first layer of the OSI (Open System Interconnection) reference model [14] seriously affect the possibilities of the higher layers. It is obvious that all of the performance evaluation techniques have to be based on the physical parameters of the wire pairs. Each technique uses these parameters, however, some of them measure them precisely while others use experimental information in the evaluation. Certainly, there are different performance evaluation methods for different types of telecommunications services. In this paper, we present techniques of DSL services. These techniques are grouped by the applied data and methods.

Evaluation by expert's estimation. In the periods when singular DSL technologies were launched, evaluation by expert knowledge was prevalent in the practice of telecommunications companies. In this method, technicians who had already had experiences with DSL installations in a single area were asked about their opinion on the available bit rates at given geographical addresses. Technicians tried to remember the results of former installations within the same geographical area and gave a rough estimation. In some cases their answers were totally wrong because real parameters of the connections were not taken into consideration. The cost of this technique is only the cost of a phone call in each case but the precision of its result is doubtful and accidental.

Evaluation by distance. Evaluation by distance is more precise than the previous method. This technique is based on the distance between the customer's premises and the telecommunications node. In this way, experiences of previous installations are taken into

consideration, however this knowledge is recorded to geographical maps. Concentric circles around the nodes approximate the borders of areas with different available bit rates. This method is cheap but the drawn borders only approximate the real performance.

Evaluation with data from technical inventories. More accurate evaluation can be performed by data from technical inventories. In this case, data regarding the cable length and wire diameter are used. The drawback of the system is that the data are not controlled, so if they are wrong then the estimation will completely fail. Circumstances relating to other parameters, which influence the performance, e.g. noise, are not taken into consideration. Moreover, in cases when data are missing from the inventories the method cannot give any result. This method has reasonably low price, but its accuracy remains under the current expectations.

Measurement based methods. Accurate evaluation methods are based on instrumental measurements of the wire pairs of the access networks. As the transmission methods and also the used frequency bands of various DSL techniques differ from each other, their instrumental methods are also variable, however, their theoretical bases are similar. During the evaluation generally attenuation, noise, insertion loss, capacitance and symmetry are measured. In some cases, reflections of impulses injected into the wire pairs are used. Measurements are usually performed in batches. According to the practice of telecommunications service providers measurements are not performed wire by wire but batches of pairs are evaluated in each case when the instruments are connected to the cable. In this way, all of the pairs of an MDF are evaluated in a period of the time. However, after a particular time, TELCO have all performance data of its network, changes are not followed. These methods are suitable to deliver a rather accurate evaluation of the expected performance but the necessary measurements involve lots of physical wire parameters, and the evaluation is also performed for unnecessary wire pairs and because of this, the cost is high [16, 17].

Technological pre-survey. The technological pre-survey is the most precise technique for the determination of performance, because in this method, the equipment of the technology is temporarily installed and the maximal available bit rate is measured. Actually, this method is not a performance evaluation but measurement of real performance and it is performed only in case of VIP or industrial customers.

Because of the problems indicated in the above critical remarks, it is necessary to look for a performance evaluation method that delivers acceptably accurate results while keeping the total cost of the evaluation within a reasonable range. We propose a totally novel approach satisfying the double criteria as it will be presented in the next sections of the paper.

2.2. A novel, fuzzy based SHDSL performance evaluation

The lack of the precise connection between the wire pairs' physical parameters and the available data transfer rate, and the numbers of parameters make the usage of fuzzy inference methods desirable in this problem. For the construction of such a system, the smallest set of influencing physical parameters were determined, then fuzzy rule bases were constructed. In the evaluation, a Mamdani type fuzzy inference method was used [18].

2.2.1. Determination of the parameters that affect the data transmission power

SHDSL technology is described in the ITU-T recommendation No. G.991.2 [6]. The recommendation lists the performance primitives, however two of them are in connection

with the physical parameters of the wire pair. These are the loop attenuation defect and the SNR (Signal to Noise Ratio) margin defect. The formula of the loop attenuation is the following:

$$LA_{\mathrm{SHDSL}}(H) = \frac{2}{f_{\mathrm{sym}}}\left\{\int_0^{\frac{f_{\mathrm{sym}}}{2}} 10\log_{10}\right\}\left[\sum_0^1 S(f - nf_{\mathrm{sym}})\right]df - \left(\int_0^{\frac{f_{\mathrm{sym}}}{2}} 10\log_{10}\left[\sum_0^1 S(f - nf_{\mathrm{sym}})\left|H\left(f - nf_{\mathrm{sym}}\right)^2\right|\right]df\right) \quad (1)$$

In Formula (1), f_{sym} is the symbol rate, $1/H_f$ is the insertion loss of the loop and S_f is nominal transmit PSD [6]. The only physical line parameter in the formula is insertion loss. The other performance primitive is the SNR margin defect which has relationship with the noise of the wire pair. In accordance with the recommendation, noise is the other influencing physical parameter.

The set of influencing parameters were checked by instrumental measurements. During the work insulation resistance, loop resistance, near- and far-end crosstalk, noise, signal to noise ratio, line impedance, attenuation to crosstalk ratio, insertion loss and longitudinal balance were measured. Frequency dependent values were measured form 10 kHz to 2 MHz in 10 KHz wide steps. The maximal bit rates were also measured by the installation of SHDSL equipment to singular wire pairs. Bit rates were measured up to 5.7 Mbit/sec.

At first sight, it seemed that there is a connection between the measured line parameters and the measured bit rate values only in the case of the insertion loss. Some examples of it can be seen in Fig. 1. Fig. 1 shows the cumulated results of measurements carried out for eight wire pairs in different cables but all in the same region, namely in Györ-Moson-Sopron County in Hungary. Similar measurements have been done in five more regions. Wire pair parameters were measured at 200 discrete points of the frequency band from 10 kHz to 2 MHz in 10 kHz steps. As the actually observed transmission method (SHDSL) uses only a frequency up to 1.5 MHz, the measured values in higher frequencies are not relevant, so those are not pictured in the figures. Similar to the practice of TELCOs, as they offer the bit rates in ranges, measured data transmission rates were divided into five groups. These groups are marked by S1, S2, ..., S5, where S1 means the range of the lowest and S5 the range of the highest bit rates. In Fig. 1 it can be seen that in the same region, the groups of the maximal available bit rate can be differentiated by the level of insertion loss. The higher the insertion loss values, are the lower the available bit rate is.

The situation regarding the noise is more difficult. Fig. 2 presents the results of similar measurements on the sample set of wire pairs representing the noise data. It is somewhat surprising to observe that insertion loss clearly marks several groups of wires but noise data is rather homogeneous within the same region while the results of similar measurements carried out in a different region of the same county resulted in a set of characteristics essentially different from the first ones, being however similarly homogeneous within the same region. This means that noise does not make differences between the bit rate of the measured wire pairs of the same region. As the noise is almost the same, it modifies the available bit rate at the same rate for all connections in the same area.

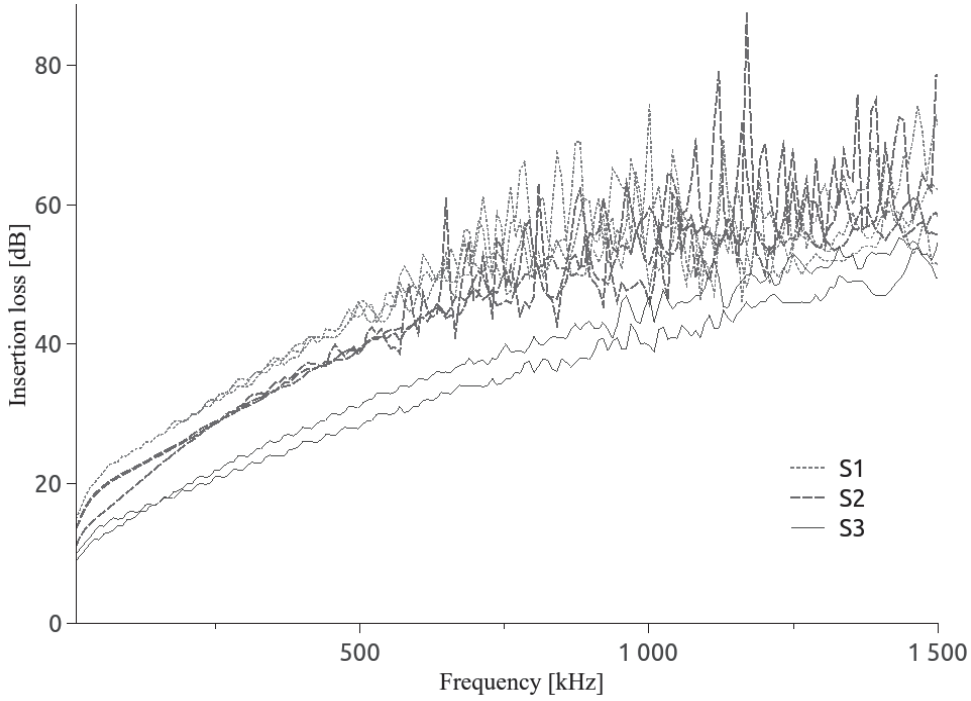

Fig. 1. Examples of measured insertion loss values [8]

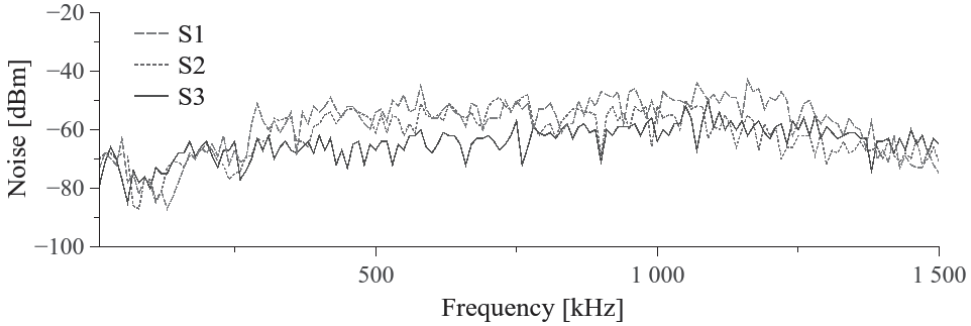

Fig. 2. Noise values of different wire pairs (region *A*) [8]

Fig. 3 shows examples of measured noise values of another region. These graphs are similar to each other, but are different from the ones in Fig. 2. These experiences have two issues. One of them is that the noise is an area dependent parameter has the same effect on the bit rate of each line in the same region. The other one is that performance evaluation can be performed area dependently based only on the values of the insertion loss, while the shape of the noise is handled as a steady parameter of the area and is previously known.

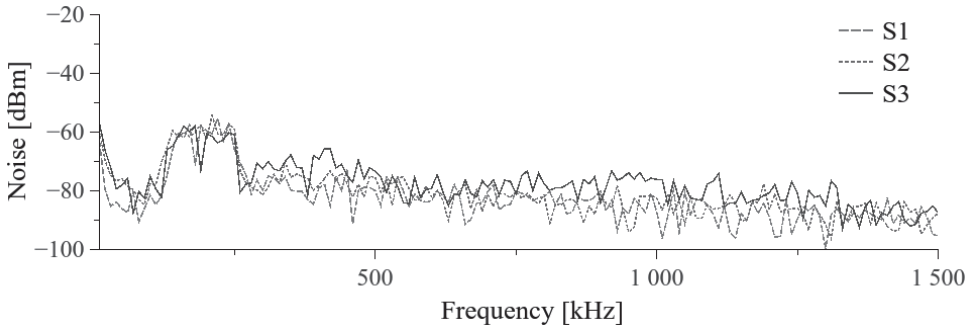

Fig. 3. Noise values of different wire pairs (region B) [8]

2.2.2. Fuzzy rule bases of the evaluation

As was seen in the previous section, the data transfer rate of the SHDSL connections are influenced by two physical parameters, these are the insertion loss and the noise. It was also shown that the values of noise are area dependent, thus in the same or similar noise environments the performance can be evaluated by the exclusive usage of the measured insertion loss values. If the domain of the noise is taken into consideration during the construction of the rule bases (as it takes place automatically during the data collection from single areas), area dependent rule bases can be constructed. These rule bases are based on the measured data transfer rate and insertion loss of the relating areas. Antecedent (input) parameters are the values of insertion loss measured at six discrete frequencies and the consequents (outputs) are the groups of the available bit rates. Different rule bases were created for the same areas.

For swift tests, statistical type rule bases were constructed. These rule bases contain five rules according to the five states of the output, and have six antecedent dimensions with triangular fuzzy sets. Measured data were separated by the measured bit rates, thus groups were created that belonged to singular bit rate groups. The average of the collected insertion loss values of the same frequency and belonging to the same group were selected as the core points of the antecedent fuzzy sets, and the closures of the supports were the minimum and maximum measured values. A graphical example of the creation of an antecedent set can be seen in Fig. 4 and an example of the statistical type rule can be seen in Fig. 5.

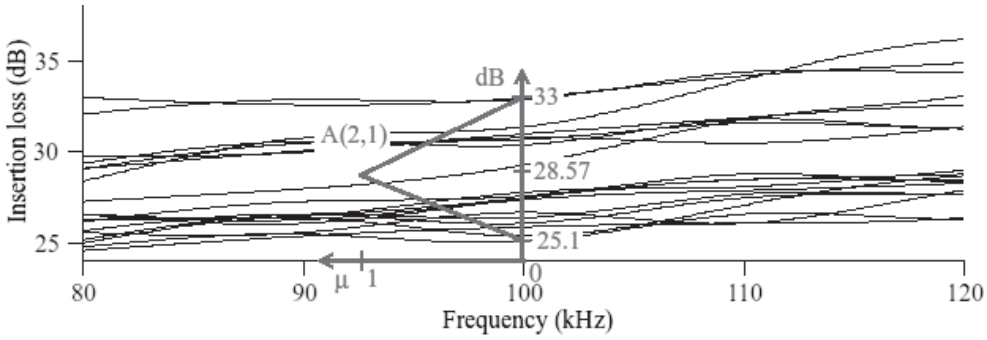

Fig. 4. Creation of an antecedent set [8]

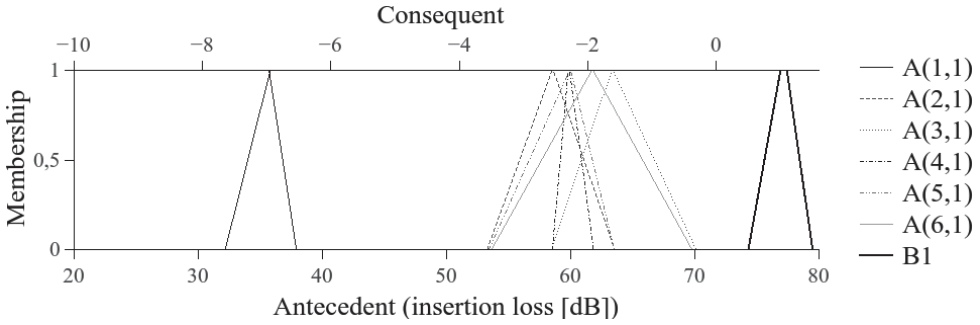

Fig. 5. The first rule of the statistical typed rule base [8]

Evolutionary algorithm [19] was used during the creation of the final rule base. Measured and collected data were used as teaching samples for the algorithm. Final rule bases contain ten rules and the same six antecedent dimensions as the statistical typed ones. All of the antecedent and consequent membership functions are trapezoid. A graphical example of the consequent side of one of the rules of the evolutionary rule base can be seen in Fig. 6.

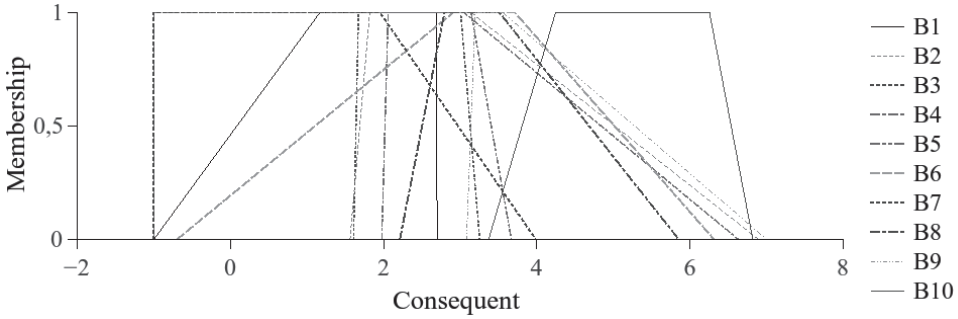

Fig. 6. Example of the consequent sets of the evolutionary rule bases [8]

Both types of rule bases were designed for the Mamdani inference method [18] that uses Zadeh's norms during the evaluation and for centroid defuzzification.

2.2.3. Comparison with other performance evaluation methods

Tests of both types of rule bases were performed. Our results were compared to the results of other performance evaluation methods. For the determination of the accurate maximal bit rate, SHDSL equipment (node and transceiver unit) was installed to the tested lines and the available bit rates were measured. All evaluation methods use ranges of bit rates in the evaluation. These ranges can be different in singular evaluation techniques. Experts use more but smaller ranges and e.g. the distance based methods use less but wider ones. These circumstances make the comparison difficult, so in Table 1 the discrepancies are given in terms of the ranges of the methods used. As an example, in the case of wire the pair signed by F the length based method (4th column) failed the estimation. It rated the pair F into the

lower neighboring bit-rate range instead of the correct one. The examined connections are signed by capital characters from A to R.

Negative aberrations are dominant, positive failure was only seen in one case. This behavior fits to the logic of performance evaluation. Underestimation is safer than overestimation. However, results differing from the measured values are not right.

As Table 1 shows, fuzzy based evaluation methods are not worse than other ones, moreover, the fuzzy evaluation by the evolutionary rule base is definitely better than the others.

There are some 'inputs out of range errors' among the results of the evaluation by the statistical rule base. This is because this statistical rule type base is sparse. If some of the values of the observations are in uncovered areas of the rule base, the inference system is not able to evaluate the performance of the observed line correctly. Another edification of the tests is that sparse rule bases cannot be directly used in the performance evaluation of telecommunication connections.

Table 1

Comparison of different performance evaluation methods [15]

Wire pair	Measured bit rate [kbit/sec]	Discrepancy by expert	Discrepancy by length from inventory	Discrepancy by fuzzy method (statistical rule base)	Discrepancy by fuzzy method (evolutionary rule base)
A	2376	–2	0	1	0
B	2056	–1	–1	0	0
C	4040	1	–1	–1	–2
D	4040	–1	–1	–1	–2
E	3144	0	–2	0	0
F	3144	0	–1	0	0
G	5704	–2	–2	input out of range	0
H	2056	–1	–1	0	0
I	3144	–1	0	input out of range	1
J	4936	–1	–2	0	0
K	4040	0	–1	–1	–2
L	4360	0	–1	–1	–1
M	1736	0	0	0	0
N	2056	0	–1	input out of range	0
O	4804	0	–1	0	0
P	4360	0	–1	0	0
R	4616	–3	–1	0	0

3. Fuzzy based fault prognosis [22]

With few exceptions, some decades ago telecommunication meant phoning or in case of business or government customers, using telex services. Telecommunication networks were designed and installed in order to fulfil these demands. The appearance of the digital technologies in telecommunications, the spread of the data communication services and mainly the big boom of the internet made the use of new, high speed technologies necessary.

One of the ways of raising the data transmission rate is the use of optical cables. The optical cables displaced the old metallic cables in the core networks or in the interexchange networks and these cables are used in cases of installing new access networks. However, telecommunication providers are not able to renew their whole network at the same time, and it would also be economically wrong to ignore the existing and useful copper wire pair based access networks that are able to produce a large amount of income without significant investment.

Another way of installing high-speed connections is the use of new technologies in the old, metallic networks. Such technologies are the members of the xDSL (Digital Subscriber Line) family that are definitely evolved for copper based access networks [4]. These are able to create relatively high-speed connections by use of several transmission methods. SHDSL systems (Symmetrical High-speed Digital Subscriber Line) [7] are able to make symmetrical digital data transmission by using TC-PAM (Trellis Coded Pulse Amplitude Modulation) modulation, where the bit rate of upload and download are equal. This type of transmission is mainly used to fulfil the demands of business customers. In the field of data transmission of private customers in copper wire based access networks, ADSL technology (Asymmetric Digital Subscriber Line) [6] is used. It can provide a sufficient bit rate with the use of a single wire pair, however, in accordance with the mode of the communication of the customers, the bit rates of the different directions (upload and download) are different.

In the communication between the endpoints, not only DSL technology is used, in different sessions different transmission methods are applied. In the core networks and in the interexchange networks for example, SDH and ATM systems carry the packages of data. However, faults can happen in the whole connection, according to empirical observations most of the faults occur in the access network used by the DSL systems, furthermore, these faults are individual so each of them needs separate handling and troubleshooting.

Upkeep and troubleshooting of access networks are time consuming and in many cases, problematic. Errored connections are not able to produce income and if the failures are not fixed within a certain time, telecommunication companies have to pay penalties to the customers. The fact the troubleshooting process generally starts after the fault report leads to logistics problems. Other expenses derive from the unplannable routes and the difficulties of the organization of daily work. Workforce management is not optimal as urgent and important troubleshooting tasks turn up during an organized shift. Problems can be decreased if the faults are handled by planned upkeep of the networks or if they are predictable. Telecommunication companies (TELCO) use several techniques for failure prevention.

The connections of the business customers are continuously monitored by the telecommunication service providers to keep the level of service at a given rate. The failures are often observed by the TELCOs before the customer would notice them. These so called proactive systems are generally able to detect the possible reasons of the failure [5], however the failure is not predicted, so that these systems do not avert the failures. There is no such kind of proactivity in the case of private customers. The procedure of the repair is started by

the service provider after the complaint of the customer is received. However, the forehand reparation can have some economic advantages. On the one hand, there are no deadlines for the reparation and so the work can be performed in a planned way bounded with other tasks, on the other hand, the cancellation of the income in the errored period can be avoided.

Hereinafter, some existing automatic fault detecting methods, and results of such work until now are presented that aims to create a method that is really proactively able to forecast the line failures of ADSL connections.

3.1. DSM (Dynamic Spectrum Management) based systems and copper pair troubleshooting by Principal Component Analysis

The performance of the different DSL systems is mainly influenced by two physical factors. One of them is the insertion loss (in connection with the attenuation) and the other one is the noise [8]. The insertion loss is principally depending on the length and the diameter of the cable, so that in particular cases, it must be considered as a steady parameter. The effect of the insertion loss can be reduced by the reduction of the length of the cable – in this case, the DSLAM is installed closer to the subscriber – or the rise of the diameter of the wire. The other influencing parameter is the noise. It comes from different sources. Its parts that originate from the wire pair itself are the thermal noise and the effects of reflections and echoes. Environmental noise effects are the impulse noise or the noise coupled into the wire pair by the broadcast radio channels and the crosstalk [1]. At the higher frequencies, the effect of the crosstalk is significant [2].

A powerful tool of the cancellation of crosstalk effects is Dynamic Spectrum Management (DSM). These methods use different power spectral masks in the case of ADSL connections that use DMT modulation [3] in order to raise the throughput of the whole system.

Telecommunication providers use DSM based management systems in order to improve the quality of their ADSL services. During iterative interferences, these systems are able to keep the bit rate continuously at the maximal level that is available for the given connection, however, they are not qualified for the forecast and the prevention of the probable failures separately.

Another method proposes the use of DSLAM data in correlation with humidity data [9]. Humidity changes the properties of symmetrical wire pairs, e.g. symmetry or insulation resistance. These changes have effect on noise. The worse the wire pair parameters are the higher the noise is. As noise is increasing, transmission parameters of DSL systems are declining. This declination is observed and recorded by DSLAMs. The method in [9] suggests the use of this recorded information to discover the problematic cables and launch the troubleshooting process proactively before the customers report faults.

3.2. A novel method for improving the ADSL service level

Telecommunication providers perform the reparation of the ADSL connections in a reactive way. The search of the reasons for the failure and the reparation of the failure starts after the customer's complaint is received. In this case, the provider is expected to keep the deadlines for the reparation, so that in lots of cases, he is not able to perform his operation in a planned way. It is disadvantageous mainly in rural areas where the completion of the tasks that come up in the same area would be practical to be made at the same time.

Proactive work needs a system that is able to predict the probable failures. The creation of such systems is greatly helped by the management and monitoring systems, which control the telecommunication networks and connections [5] (Fig. 7).

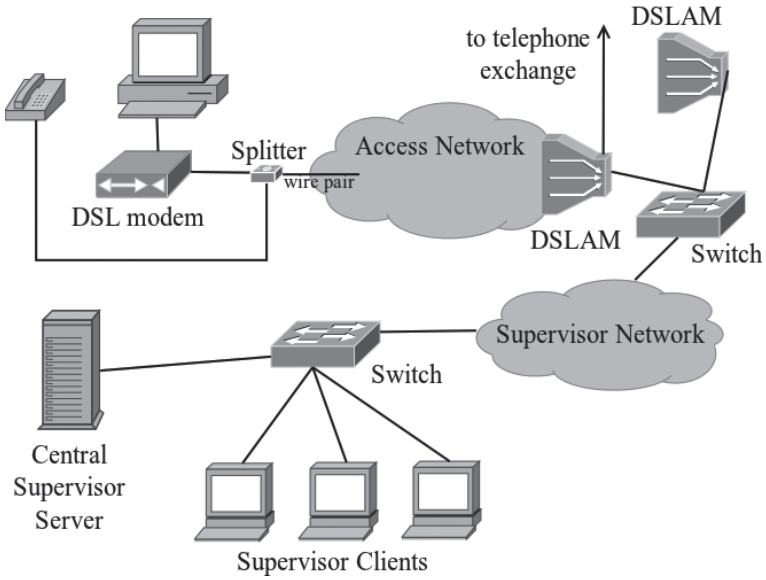

Fig. 7. ADSL model and its supervision system

The two main functions of a supervision system are the setup of the transmission parameters of the connections and the monitoring of the operational parameters of the terminals. Because of the wide range of monitorable parameters and the variance of the monitoring systems of the different manufacturers, only the most important parameters from the point of view of the problem are mentioned here.

Output Power. In downstream, it is the output power of the DSLAM and in upstream, it is the output power of the modem. The output power of the different DMT channels can be different. These values are modified in order to find the best transmission spectral mask during the DSM process that was mentioned previously.

Attenuation. The attenuation is the difference in dB between the received and the transmitted power for a single DMT channel [6]. The average of the measured values of all channels is indicated by the supervision system separately for upstream and downstream.

SNR Margin. The Signal to Noise Ratio Margin in dB refers to the power of the received noise compared to the noise power that the system is designed to tolerate [6]. Also, this value is an average and is separately given for upload and download.

Current rate. The Current rate is the actual bit rate of the upload and download transmissions. In the case of connections in good conditions, the actual rate reaches the bit rate fixed by the network operator.

Attainable Rate. This is the maximal available bit rate. This value is calculated from the line parameters. Also, it is calculated for upload and download. The small gap between this value and the current rate refers to the bad conditions of the wire pair of the access network.

FECS. The number of seconds when at least one received frame was corrected.

ES. The number of seconds when at least one received frame was uncorrectable.

LOSS. Loss is the number of seconds when no frame was received by the DSLAM.

UAS. UAS is the number of seconds without payload.

FIC. (Full Init Count) FIC is the number of resynchronizations in a given period of time. Practically it is equal to the number, the connection lost errors.

FFIC. (Failed Full Init Count) FFIC is the number of unsuccessful attempts for resynchronization.

Table 2

Examples of some measured parameters

Line ID	Attenuation (downstream) [dB]	SNR margin (downstream) [dB]	Attainable rate [kbit/s]	Current rate [kbit/s]	Connection lost errors (previous day)
XT006361	13	23	3508	2560	0
XT009560	56	8	1648	1105	526
XT009564	41	8	1408	842	3
XT006375	33	20	10432	2925	2
XT006458	8	16	12544	5888	0
XT006383	28	14	13540	6746	0

We have examined real ADSL connections in operation. During the tests, 400 connections were examined over a two months period. The values of the parameters mentioned above were measured and recorded hourly. Some of them are shown in Table 2. In case of line numbered XT009560, the high values of connection lost errors show that this line is out of order. The ADSL connection is lost more than 500 times a day. However, this error ratio is very high, the customer did not reported the problem. It can be because of his Internet using habits. The other lines in Table 2 are well operable, however, there are 2 connection lost errors in the case of line XT006375 and 3 in the case of XT009564.

Another cautionary symptom is the unstable current rate. If this parameter is continuously changing from one measurement to another, the physical parameters of the line are presumably wrong.

The measurements suggest two conclusions. The first is that there is a possibility for creating a quick proactive method based on the measured FIC values and the variation of the current rate. The second can be based on the variation of the attenuation and the SNR margin, however, it needs further examinations.

3.2.1. FIC and current rate based proactive method

TELCOs do not use proactive methods in their ADSL maintenance processes, however in several cases, e.g. in cases of connections leased from other telecommunications companies, it would be reasonable. We propose a method that can terminate several types of errors and ensure good quality of the service for a long time before the customers would notice a errors.

In cases of the deterioration of the wire pairs the connection becomes unstable. 'Connection lost' errors arise that last for short periods of time. These errors can occur also in case of lines in good condition because of various reasons, e.g. if the user restarts the modem or if there is a momentary blackout. The variation of the current rate indicates well that the ADSL connection works at top of the technical possibilities of the wire pair. A small rise in the noise or a small change in other parameters worsens the conditions so that the bit rate cannot be held.

It was shown by our examinations that the consideration of these two parameters is enough to predict upcoming failures of the connection. In this case, we recommend two types of action.

The first is a quick action in order to avoid further errors, however, it does not eliminate the reasons for the error. In this case, the target of the SNR margin is raised, so the bit rate of the connection is lowered by the network operator. Because of this action, small failures of the physical line cannot destroy the connection.

The other type of action is the reparation of the physical line. As the reparation of the line can be started before the customer's complaint, it can be executed in a planned way and there are no deadlines and risks of penalty. Fig. 9 shows the result of one of the tests were performed before the reparation of the telecommunications line.

Table 3 shows some numerical data of several wrong connections.

Table 3

Data of wrong connections

Line ID	Attenuation [dB]	Variation of SNR Margin [dB]	Variation of Current rate [kbit/s]	Max. FIC
XT008419	22.6	4	568	84
XT007320	41.0	7	1288	45
XT007476	59.0	6	912	37
XT007990	41.0	5	294	31
XH006699	54.5	5	812	13
XT007998	51.0	7	3136	25

Recorded data of line XT007998 can be seen in Table 3. The measured values were bad and varying until the action. After this time, the connection became steady. Although the maximal available bit rate was declined from 5000 to around 1500 kbit/sec, the number of the FIC fell to zero. The declination of the bit rate is acceptable during the days of the physical reparation of the line.

3.3. Fuzzy based failure prediction

Although, at the moment there is not yet enough recorded data for rule based construction, we plan to create a fuzzy based failure prediction system.

By the experiences of operation of telecommunication networks and DSL systems, it can be said that the erosion of metallic cables based connections goes hand in hand with

the erosion of the attenuation and/or noise. These effects are in connection not only with the aging of the cables themselves, their main reasons are the failures of the junctions or physical failures of the cables (e.g. leaks of the cable jacket and soaking). As DSL systems are highly sensitive to attenuation and noise, these failures of the transmission medium lead to the failure of the service. In most of the cases, the failures do not happen suddenly, this phenomenon is a slower flow. It means that the erosion of these parameters can be observed before the full failure of the DSL connection. Taking also economic and logistic points in consideration, in the case of the first emergence of the failures mentioned, there is no need for sudden repair in all cases, proposed repairs can be performed in view of available failures. For giving forecast of the available failures the next fuzzy based failure prediction system is planned.

The worst but yet acceptable values of the attenuation and the noise will be diagnosed for all levels of service. These values will serve as limits of action (LI) (Reaching LI an immediate repair is needed). Regularly monitoring the attenuation and the noise parameters of the connections the observed values will be recorded and compared to the LIs. Analysing the speed of the declination of these values the rates of failure (RF) will be calculated. Having the rates of failures and the momentary distance from the LI (DLI) the next rules can be used for the conclusion.

*R1. IF DLI(attenuation) is **small** AND RF(attenuation) is **high** THEN action is needed.*
*R2. IF DLI(noise) is **small** AND RF(noise) is **high** THEN action is needed.*
*R3. IF DLI(att.) is **high** AND RF(att.) is **low** AND DLI(noise) is **high** AND RF(noise) is **low** THEN action is NO needed.*

The curiosity of this rule base is that noise as antecedent dimension is missing from rule R1 and attenuation as an antecedent dimension is missing from rule R2, although both of them are present in rule R3. The reason for this is that action is needed in cases of the deterioration of attenuation OR in case of the deterioration of noise. There is OR connection between

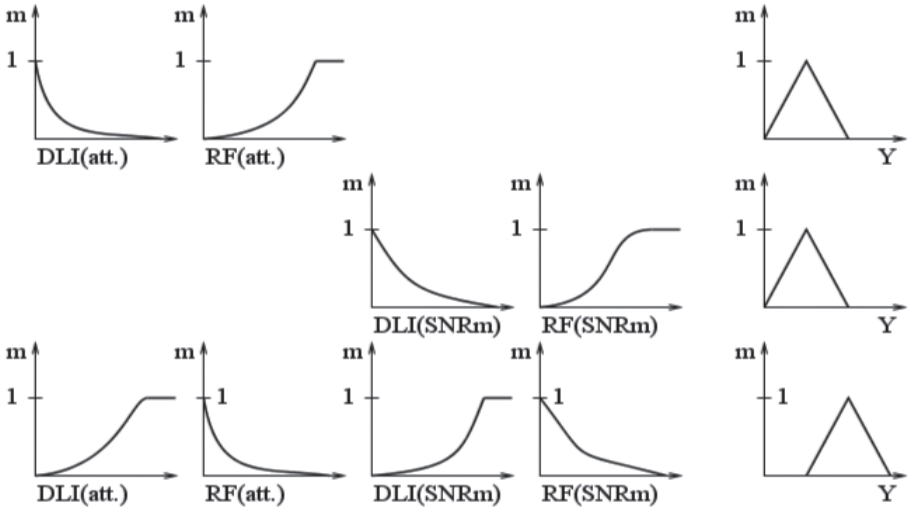

Fig. 8. A graphical example of the rule base described in Section 3.3

them, so they cannot be handled in the same rule where there are only AND connections in the antecedent side. Fig. 8 shows this odd structure of the rule base. The fuzzy sets of the linguistic variables of high and low are planned to be constructed from the recorded data, so the graphical example of the rule base (Fig. 9) is only a fictive illustration, however these shapes are expected.

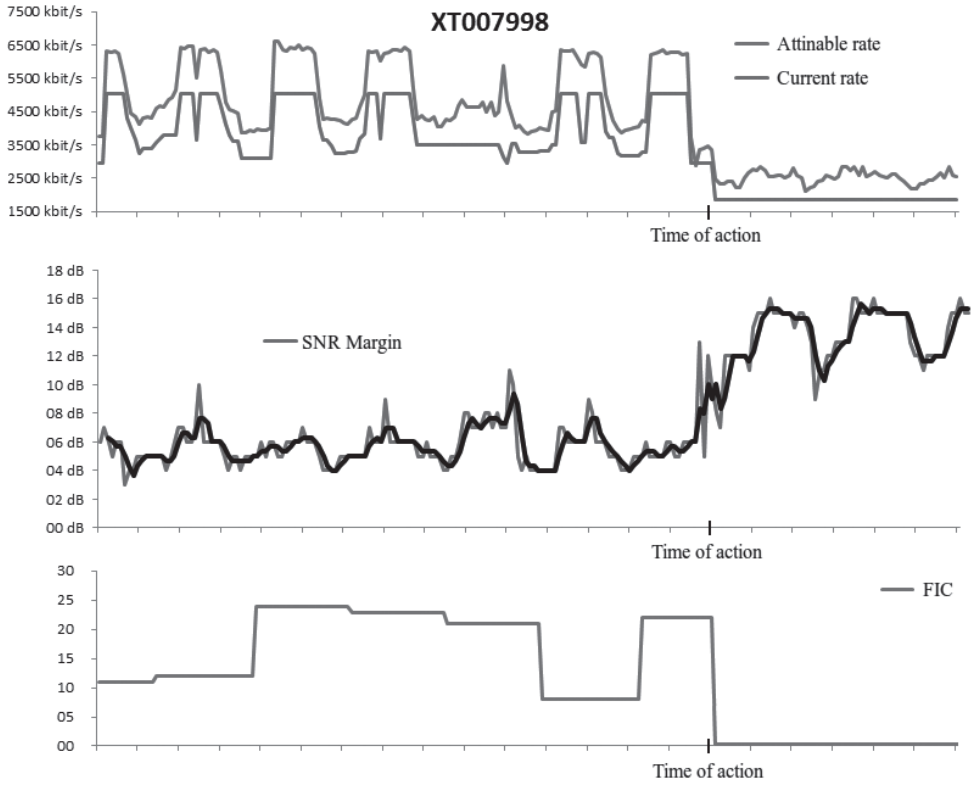

Fig. 9. The behavior of the line XT007998 before and after the interposal

For the final conclusion, the Mamdani inference method is to be used. The crisp conclusion means not only the YES or NO for the necessity of the action but it refers also to the urgency of the action in cases of bad telecommunications connection or the possibility of long term failure in cases of a good connection.

4. Conclusions

Novel fuzzy based approaches were introduced in telecommunications fields.

First, one is a performance evaluation technique that applies evolutionary algorithm created rule bases, based on measured data. The results are better than other methods that were used during the comparison.

The second presented method is a fault prediction technique. Supervision systems monitor lots of operational parameters of the networks. These data can be used in preliminary error recognition, and thus early repair of the elements of the networks can be performed. It has more advantages. In addition to detecting that several parts of the network are in bad condition, the harmful effects of the predicted error can be avoided. The results of the tests of this method are successful, however this method is able only for the prediction of errors which are expected in the very near future.

As for further plans, in the near future the limits of action of each level of DSL service will be diagnosed by results of long-term measurements and fuzzy rule bases will be constructed for giving failure predictions of ADSL connections.

This paper was supported by the Hungarian Scientific Research Fund OTKA K105529, K108405 and TÁMOP-4.2.2A-11/1/KONV-2012-0012.

Appendix

ADSL. Assymetric Digital Subscriber Line. A broadband digital telecommunications transmission technique for the symmetrical wire pairs of access networks used for connecting to the Internet.
Access network. Telecommunications network which connects the subscribers to their respective service provider.
ATM. Asynchronous transfer mode. A broadband digital telecommunications concept for transmitting video, voice and data.
Bit rate. Data transfer rate. The number of the transmitted bits in a second (or in a unit of time).
DMT. Discrete Multi-tone Modulation. A digital multi-carrier method to transmit the data in small amounts in paralell subchannels of the whole frequency band.
DSL. Digital Subscriber Line. A group of transmission methods for connecting the subscribers to the Internet via the symetrical wire pairs of the access network.
DSLAM. Digital subscriber line access multiplexer. A piece of equipment which connects multiple DSL lines to higher level communications channels.
DSM. Dynamic Spectrum management. A method to increase the maximal available bit rate of DSL connections.
IPTV. Internet Protocol Television. Television services using IP networks.
MDF. Main Distribution Frame. A distribution frame which crossconnects the wire pairs of the symmetrical cables of access networks.
OSI. Open Systems Interconnection. A joint standard of ITU-T and ISO.
POTS. Plain Old Telephone Service. The traditional telephone service for voice communications with 3.4 kHz wide bandwidth.
PSD. Power Spectral Density. The power of a signal per unit of frequency.
SDH. Synchronous Digital Hierarchy. Telecommunications protocol for transmitting multiple bit streams over optical connections.
SHDSL. Single-pair High Speed Digital Subscriber Line. A symmetrical transmission member of the DSL family.

SNR. Signal to Noise Ratio.
TC-PAM. Trellis Coded Pulse-Amplitude Modulation. A modulation method used in SHDSL described in ITU-T G-Series.
TELCO. Telecommunications Company.
VoIP. Voice over IP. Protocol for voice transmission over IP networks.
xDSL. An alternative name for DSL.

References

[1] Cook J.W., Kirkby R.H., Booth M.G., Foster K.T., Clarke D.E.A., Young G., *The noise and crosstalk environment for ADSL and VDSL systems*, Communications Magazine, IEEE, Vol. 37, Issue 5, 1999, 73-78.

[2] Huberman S., Leung C., Le-Ngoc T., *Dynamic Spectrum Management (DSM) Algorithms for Multi-User xDSL*, Communications Surveys & Tutorials, IEEE, Vol. 14, Issue 1, 2012, 109-130.

[3] Kalet I., *Multitone Modulation, Subband and Wavelet Transforms*, The Kluwer International Series in Engineering and Computer Science, Vol. 340, 1995, 391-412.

[4] Sorbara S.M., Cioffi J.M., Silverman P.J., *DSL Advances*, Prentice Hall, 2003.

[5] Csányi K.P., Kóczy L.T., Tikk D., *Intelligent Solutions for Umbrella Systems in Telecommunication Supervision Systems*, International Journal of Information and Communication Engineering, 1, 7, 2005, 346-351.

[6] ITU-T, *Recommendation G.992.1, Asymmetric digital subscriber line (ADSL) transceivers*, 06/1999.

[7] ITU-T, *Recommendation G. 991.2, Single-pair high-speed digital subscriber line (SHDSL) transceivers*, 02/2001.

[8] Lilik F., Kóczy L.T., *Performance Evaluation of Wire Pairs in Telecommunication Networks by fuzzy and Evolutionary Models*, IEEE Africon 2013 Conference Mauritius, 9th–12th September 2013, 712-716.

[9] Spahija B., Deljac Z., *Proactive copper pair troubleshooting utilizing Principal Component Analysis*, 18th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2010), Brač, Croatia, 23–25.09.2010.

[10] Kóczy L.T., Tikk D., Botzheim J., *Intelligens rendszerek*, Széchenyi István Egyetem, Györ 2008.

[11] Zadeh L.A., *The Concept of a Linguistic Variable and its Application to Approximate Reasoning-I*, Journal of Information Science, Vol. 8, Issue 3, 1975, 199-249.

[12] Sugeno M., Hirano I., Nakamura S., Kotsu S., *Developement of an intelligent unmanned helicopter*, Proc. IEEE Int. Conf. Fuzzy Systems, Vol. 5, 1995, 33-34.

[13] Tormási A., Botzheim J., *Single-stroke character recognition with fuzzy method*, New Concepts and Applications in Soft Computing SCI, Vol. 417, V.E. Balas et al. (eds.), 2012, 27-46.

[14] ITU-T, *Recommendation X.200, Data Networks and Open System Communications, Open System Interconnection – Model and Notation*, 07/94.

[15] Lilik F., Kóczy L.T., *The Determination of the Bitrate on Twisted Pairs by Mamdani Inference Method*, Studies in Computational Intelligence, L.T. Kóczy et al. (eds.), Springer International Publishing Switzerland, 2014, 59-74.
[16] Goralski W., *xDSL loop qualification and testing*, IEEE Communications Magazine, Vol. 37, Issue 5, 1999, 79-83.
[17] Faulkner R., Schmidt K.E., Zhang Y., *Method and apparatus for qualifying loops for data services*, United States Patent, Pat. No. US 6,385,297 B2, May 7, 2002.
[18] Mamdani E.H., Assilian S., *An Expriment in Linguistic Synthesis with a Fuzzy Logic Controller*, International Journal of Man-Machine Studies, Vol. 7, 1975, 1-13.
[19] Balázs K., Kóczy L.T., *Constructing Dense, Sparse and Hierarchical Fuzzy Systems by Applying Evolutionary Optimization Techniques*, Applied and Computational Mathematics, Vol. 11, No. 1, 2012, 81-101.
[20] Zadeh L.A., *CFuzzy sets*, Information and Control, Vol. 8, Issue 3, June 1965, 338-353.
[21] ITU-T, *Recommendation G. 902, Framework recommendation on functional access networks (AN)*, 11/1995.
[22] Lilik F., Simonyi P., Kóczy L.T., *A Novel Method of Improving the Service Level of ADSL Connections*, International Congress on Control and Information Processing 2013, ICCIP 2013, Cracow 12.07–12.08.2013, 25.

CONTENTS

TREŚĆ