Association ACDT as a Tool for Discovering the Financial Data Rules

Jan Kozak
University of Economics
Faculty of Informatics and Communication
Chair of Knowledge Engineering
Katowice, Poland
Email: jan.kozak@ue.katowice.pl

Przemysław Juszczuk
University of Economics
Faculty of Informatics and Communication
Chair of Knowledge Engineering
Katowice, Poland
Email: przemyslaw.juszczuk@ue.katowice.pl

***Abstract*—We present a novel approach based on the original idea of the Ant Colony Decision Tree (ACDT) algorithm used in the problem of building the decision trees. One of the crucial limitations of the canonical ACDT algorithm was its link to strict decision rules. In this paper we transform the algorithm in such way, that it is capable to manage complex association rules. Research is conducted on the various sets of financial data closely related with the swiss frank currency. Evaluation of results was possible on the basis of accuracy measure as well as the proposed fuzzy accuracy. These preliminary studies show, that the proposed algorithm is capable to maintain its effectiveness even in the problems with large number of attribute values.**

***Keywords*—decision trees, rule discovery, financial data**

I. Introduction

Swarm intelligence in last few years has emerged as a one of the major approximate methods effectively used in the most complex problems. Approximate solutions given by the swarm intelligence methods seems to be a good compromise between exact solutions calculated by the slow and complex algorithms and fast methods capable to solve only the simplest problems. Below we give a few examples of the newest articles concerning the swarm intelligence. Omitting earlier publications, which various review can be found in [6], a more recent models involves using the firefly algorithm inspired by the flashing behavior of fireflies. The primary purpose for a fireflies flash is to act as a signal system to attract other fireflies [16]. In the [17] Xin-she Yang and Suash Deb have proposed the concept based on the brood parasitism of some cuckoo species. One of the newest nature inspired algorithms has been proposed in [7] and it is based on the simulation of the herding behavior of krill individuals.

Theoretical background proposed in the above algorithms was confirmed not only on the test functions but also in various real-world applications such as [3], [15], [8]. A very interesting article on the application of the swarm intelligence (more specifically the Ant Colony Optimization algorithm) was a novel approach focused on the problem of building the decision trees. In this case an ACDT (Ant Colony Decision Tree) originally proposed in [1] was used as a tool of building the efficient decision trees capable to accurately solve the various prediction problems. Similar concept was used in the ACDF (Ant Colony Decision Forrest), in which high efficiency of the proposed solution was based on the set of smaller decision trees and the group decisions [10].

978-1-5090-5795-5/17/$31.00 © 2017 IEEE

It should be mentioned, that decision trees were earlier used with connection to the financial data. In the work [9] authors used the decision trees to predict changes on the forex market. In this particular case, an algorithm based on the concept of similarity in the historical data was presented. However such approach should not be possible in the case of association rules proposed in this article.

Purpose of the presented article is the analysis of the efficiency of the new association-ACDT (ASC-ACDT) algorithm. This new concept is tested on the difficult forex data set. We purport that there exists a non-obvious dependency between the forex indicators which may be discovered on the basis of a new version of the ACDT algorithm, which is based on the mining the association rules. Such approach should lead to possibility of effective usage of the the ASC-ACDT approach in various data exploration problems which, until now, were out of reach of classical ACDT approach. At the same time our approach may bring different insight on the concept of building simple rule-based systems of the forex market.

Nowadays a big potential of the automatic transaction systems is still getting visible. It applies to the forex transaction systems which include methods like the High Frequency Trading (HFT) as well as the medium and long term trading approaches focused mostly on building effective portfolios. In both cases a vast majority of proposed methods are based on the concept that technical analysis indicators are capable (with very high efficiency) to point out instruments on the market which should be in the center of interest of the potential decision maker. Rule definitions are given for example in [14]. However in the review article [4] we can find information about approaches related to the technical analysis which were unsuccessful.

To the best of authors knowledge, there is no publications which pinpoint the dependencies between various technical analysis indicators. Thus proposed ASC-ACDT algorithm may be considered as a first step of reducing the complexity of decision rules in the financial data. Described approach

seems to be effectively scalable to the problem of multi-label analysis. At the same time proposed approach may lead to the discovering of the new transaction systems rules.

Outline of the article is as follows: in the section ii we shortly describe basic definitions related with the data exploration, we briefly introduce the concept of accuracy as well as we give short information about the classical decision trees. Section iii brings insights into the original ACDT algorithm which was the canonical version of the approach presented in the article. The section iv describes basic idea behind the ASC-ACDT algorithm. Finally, section v consists all experiments and the data set description. We end with short conclusions.

II. Background

Below we define basic concepts related with the classical data exploration. Let there be a set of objects X, where every object x can be described as a pair of attributes (x_{atr}, x_{class}), where x_{atr} is the set of attributes of the single object represented as an m-dimensional vector in the attributes space:

$$V = [v^1, v^2, ..., v^m], \tag{1}$$

where v^i is the i-th attribute of the object. For every attribute we can define the pair (attribute, value). We assume, that v^i could have numeric or symbolic value. The x_{class} will be the decision class of the given object. In the classification problem the x_{class} has one of the values belonging to the decision class set of values.

One of the most popular classification methods are decision trees. Simple, and intuitive form of the decision tree makes it useful in various real-world decision problems. Example of classical decision tree algorithms are CART, C4.5 and C5.0. Other example can be found in the [13] and extended overview of these algorithms was described in details in [12]. It is especially important in the case of problems in which additional expert knowledge is necessary. Key factors in the process of building the decision tree is the "divide and conquer" rule which allows for multiple, recursive data divisions - which directly leads to dividing the overall problem into smaller sub-problems. Basic division rule is based on the greedy approach, however other division concepts can be found as well. In general, the goal of the classification is to build a classifier capable to assign the decision class value for every new object x_{prime} not belonging to the original data set X. The process of the assigning the decision class to the new object x_{prime} is directly related with the classification accuracy, which is capable to estimate, how good was the classification process:

$$d(DT, D) = \frac{TP}{|D|} \tag{2}$$

where TP is the number of properly classified objects, DT is the decision tree and $|D|$ is the number of all objects in the test set. The above mechanism of accuracy estimation will be used as a core of the experiments in the next sections.

III. ACDT Algorithm

Starting point for the ACDT algorithm was the metaheuristic proposed by Marco Dorigo [5] - Ant Colony Optimization (ACO). Original work, unlike the classical evolutionary algorithms was focused on the combinatorial optimization. In its simplest definition the ACDT is the Ant Colony Optimization algorithm used in the process of constructing the decision trees. A non-deterministic concept allows to generate different classifiers (decision trees) for every algorithm run. Formal definition of the ACDT algorithm is given below:

$$ACDT = \langle (X, A \cup \{c\}), T(S), ants, p_{m,m_{L(i,j)}}(t), S \rangle \tag{3}$$

where X is the set of objects, A is the set of attributes (in which the decision attribute c is included). Above elements are used to define the decision table; $T(S)$ is the decision tree generated by the algorithm, $ants$ is the number of ants in the algorithm, $p_{m,m_{L(i,j)}}(t)$ selection rule used in the decision process, finally S is the set of acceptable objects.

In 1984 in the [2] the splitting rule concept was proposed as the basis of the CART approach. The splitting rule was used to calculate the value of the heuristic function. The split in every single node is calculated on the basis of probability used in the ACO algorithm:

$$p_{i,j} = \frac{\tau_{m,m_{L(i,j)}}(t) \cdot \eta_{i,j}^{\beta}}{\sum_i^a \sum_j^{b_i} \tau_{m,m_{L(i,j)}}(t) \cdot \eta_{i,j}^{\beta}} \tag{4}$$

where $\eta_{i,j}$ is the value for the single split using the attribute i and the value j; t is the step of the algorithm; $\tau_{m,m_{L(i,j)}}$ is an pheromone value currently available at the step t on the connection between nodes m and $m_{L(i,j)}$ (attribute i and the value j are used in the calculation process), whereas relative importance of the heuristic value is denoted by the β.

The pheromone trail values are calculated on the basis of pheromone levels on the edges connecting both: tree node and its parent node (excepting the root):

$$\tau_{m,m_L}(t+1) = (1-\gamma) \cdot \tau_{m,m_L}(t) + Q(T) \tag{5}$$

where $Q(T)$ determines the evaluation function of decision tree (see equation (6)), the a parameter representing the evaporation rate is referred as the γ.

The following equation was used in the process of function evaluation for the decision tree:

$$Q(T) = \phi \cdot w(T) + \psi \cdot a(T, P), \tag{6}$$

where $w(T)$ is the number of nodes (denoted as the size of the decision tree T). Accuracy of the classification of the object from a training set P by the tree T is denoted as the $a(T, P)$. Finally, ϕ and ψ are constants used in the process of determining the relative importance of $w(T)$ and $a(T, P)$.

IV. Problem Formulation

For the purposes of the article, a new concept based on the original work involving the ACDT algorithm was proposed.

In this approach we make use of the association, where the data set X consisting of n-pieces

$$X = \{x_1, x_2, \ldots, x_n\}, \tag{7}$$

described by m attributes

$$A = \{a_1, a_2, \ldots, a_m\}, \tag{8}$$

and without decision classes, each object with a set X, can be described as:

$$x_i = ([v_i^1, \ldots, v_i^m]), v_i^j \in A_j, \tag{9}$$

where v_i^m is the value of the attribute a_j for objects x_i.

Such approach is based on the concept of unsupervised learning, in which there is a need to describe internal properties of the analyzed objects. To simplify: for a given set of attributes we do not define single decision attribute. Instead of this we try to discover connections between attributes (or groups of attributes) describing objects. A new ASC-ACDT algorithm may be formally defined as follows:

$$ACDT = \langle((X, A), T(S), ants, p_{m,m_{L(i,j)}}(t), S\rangle. \tag{10}$$

Table (X, A) represents problem and it is expressed as a two-tuple (X - set of objects and A - set of attributes, without decision attributes). Such form of the algorithm allows to analyze every single attribute in the same manner as the decision attribute in the original ACDT algorithm. Due to specifics of the financial data related with the forex market we assume, that there is possibility of reducing the set of crucial indicators necessary to successfully make a decision on the forex market. Such limitation of the set of attributes should lead to the decreasing the complexity of the problem at the same time maintaining the accuracy of the prediction.

Due to specifics of the analyzed data, we are not only interested in the exact classification, but we assume, that situation, in which object is classified to the neighboring class is as well satisfactory. Assumption is, that the considered decision attribute is the indicator, which value should be predicted. Using the above statement we introduce a new fuzzy accuracy measure, which will allow us to measure efficiency of the proposed ASC-ACDT algorithm in case, where there is a slightly deviation from the target class:

$$d(DT, D)_{fuzzy} = \frac{TP^{-n\%} + TP + TP^{+n\%}}{|D|} \tag{11}$$

where $TP^{-n\%}$ is the number of objects classified to the neighboring class (lower than the target class), and $TP^{+n\%}$ applies for the neighboring class greater than the target class. All decision classes are sorted and labeled from 1 up to N, where N is overall number of classes for the attribute i. Neighboring classes size are defined as:

$$neighbor_{size} = n\% \cdot N \tag{12}$$

On the fig. 1 we can see example classification, in which 1(a) consists the classification matrix - with the assumption, that all decision classes can be calculated and sorted. In this case 10 classes with an fuzziness equal to 0.1 which is equal to 10% can be seen. Thus situation, in which classification to neighboring class is considered as the satisfactory result. All objects classified to the exact expected class are marked with the darker gray, while those classified with 10% error included are marked with the lighter gray color.

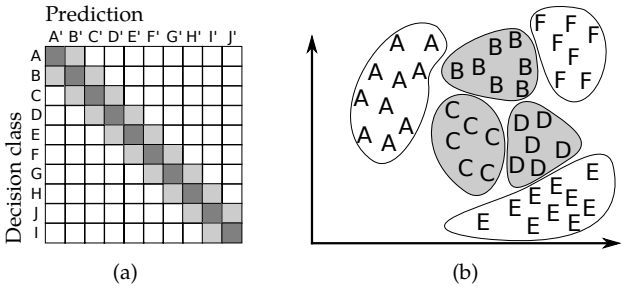

Fig. 1. Example of the fuzzy accuracy used as a one of the measures in the conducted experimental part; (a) an example of the classification matrix, (b) an example for the decision class C in which area considered as a satisfactory classification was marked

The second part of the fig. 1 (b) shows example solution for selected class (in this example C). 5 different sets of objects classified to differed decision classes can be seen. Additionally, with grey color we marked area, which would be considered as satisfactory, if object from the class C would be assigned to one of these classes.

V. Experiments

Below section includes detailed description of experiments conducted on the real financial data related with selected currency pairs involving swiss frank. An analysis included accuracy of the classification, as well as the fuzzy accuracy introduced in the previous section. All conducted experiments were repeated 30 times, and the results were averaged. The method used to calculate the quality of the classifier was the train and test (proportions equal to 50%). This section includes boxplots with detailed statistics related with every analyzed accuracy measure. ACDT algorithm parameters are given bellow:

- number of iterations - 5000;
- population size - 50;
- goal function - accuracy of the classification;
- all remaining parameter values are adequate with these proposed in the [11].

A. Data Sets

To deliver detailed results based on the ASC-ACDT we selected data which included 6 various currency pairs: AUDCHF, CADCHF, EURCHF, GBPCHF, NZDCHF and USDCHF. For every currency pair we selected time interval including 1500 reading. In the analysis we used discrete values of 10 technical analysis indicators: Bears, Bulls, Average True Range (ATR), Demarket, Williams indicator, Commodity Channel Index (CCI), Relative Strength Index (RSI), standard deviation, momentum and Force Index (FI). Number of different values for every indicator is given in the table I.

TABLE I
INDICATORS AND NUMBER OF DIFFERENT VALUES (NUMBER OF ATTRIBUTES VALUES). INDICATORS CRUCIAL IN THE PROCESS OF ANALYSIS WERE EMPHASIZED WITH THE BOLD FONT

Attribute	The number of attribute values
CCI	**27**
CADCHF	12
RSI	**6**
ATR	5
DM	10
FI	12
Williams	11
Bears	**10**
Bulls	**11**
Momentum	12

According to the domain knowledge we selected 4 indicators often selected as a base indicators in the rule-based transaction systems: Bears, Bulls, CCI and RSI - for these indicators we tried to propose an alternative set of indicators capable to replace them. Time interval, for which we calculated all indicators values was equal to 15 minutes, thus every considered data set included overall time interval of approximate 3 weeks.

B. Results of Experiments

Purpose of the conducted experiments was to experimentally verify and measure the efficiency of the proposed ASC-ACDT algorithm in the problem of predicting values of indicators on the forex market. Thus, evaluation, if it is possible to omit selected indicators from further analysis.

Results of all conducted experiments were given in the tables II (accuracy and the fuzzy accuracy of the classification) and III (accuracy and fuzzy accuracy - minimum, maximum and quartiles from all algorithm runs), as well as the 4 (confusion matrix).

As a result of algorithm run we were able to acquire classifiers, which predict values of all available indicators on the basis of remaining indicators. Accordingly to the previous description, we selected four representative indicators (however it is possible to expand the set of analyzed indicators in future). For all selected indicators we were able to estimate the prediction accuracy of their values.

Results given in the table II shows, that the prediction of some indicators gives good results. For example very good classification accuracy near the level $53.4\% - 57.4\%$ was acquired for the problems with 10 and 11 decision classes. More specifically it was possible in the case of Bears and Bulls indicators, which should allow to reduce number of indicators necessary in the rule-based transaction system. Worse results were acquired for the RSI indicator, and finally relatively low accuracy was acquired for the CCI indicator. However in the last example such results are directly related with the large number of decision classed (in this case 27).

However in every case using the fuzzy accuracy with the fuzziness range equal to 10% allows to greatly improve prediction efficiency. It should be noticed, that in case of analyzed data, all classes neighboring to the target class are considered as an acceptable solution. In such case, using the ASC-ACDT algorithm to reducing the number of indicators brings very good results. Likewise, boxplots presented on the fig 2 and 3 confirm stability of such solution. Especially in the case of the fuzzy accuracy, acquired results allowed to achieve similar prediction of the indicator value.

Application of the fuzzy accuracy is well visible on the fig. 4, where classification matrix for every analyzed case are visible. Rows store information about the decision classes, while the columns store information about the predicted decision class (indicator value). Visible focus of values on the diagonal should mean error-less classification. It can be noticed, that prediction in most cases is very close to the diagonal - darker color is means the higher number of proper classifications.

For proposed fuzzy accuracy related to the $n\%$ (in our experiments 10%) proper classification is denoted as assigning the indicator value with possible shift equal to additional given number of decision classes: Bears - 1; Bulls - 1; CCI - 3; RSI - 1. It should be emphasized, that in case of such large number of decision classes, given results are treated as very good.

Despite relatively large number of cases (each with large number of decision classes), acquired classifiers have an acceptable size. Exact values related with the size of the decision tree are given in the table III. As it can be observed, for each analyzed currency pair, each decision trees have similar number of nodes (approximately in the range 270 up to 570 - depending on the currency pair). Their height is similar as well (approximately 19). It means, that the longest path (the longest rules) consists of overall 19 conditions, which allow to estimate value of given indicator on the basis of other elements.

VI. CONCLUSIONS AND FUTURE WORK

Application of the ASC-ACDT in the case of predicting the indicator values on the forex market brings good results. Difficulty of prediction is not equal for different indicators (it is relatively easy to predict values of Bears and Bulls, while the CCI indicator causes visible difficulties). Thus value of given indicator can be almost always predicted with the similar probability of correct prediction. In the analyzed problem all decision classes have been considered as enumerated types, and finally, prediction of slightly different value (than target value) does not visibly affects the analysis, thus an fuzzy accuracy was proposed as well. Fuzzy accuracy allows to predict indicator values with very results. It should be noted, that this feature will be mostly observed only in the specific data and it should not be treated as an universal conclusion.

According to the knowledge domain expert, it is crucial that it was possible to estimate hidden dependencies between technical analysis indicators - which can be main elements of the rule-based trading systems. These dependencies are not clearly visible in every analyzed case, however it brings us to preliminary conclusion, that existing tendency supporting fixed set of indicators in some situations can be insufficient.

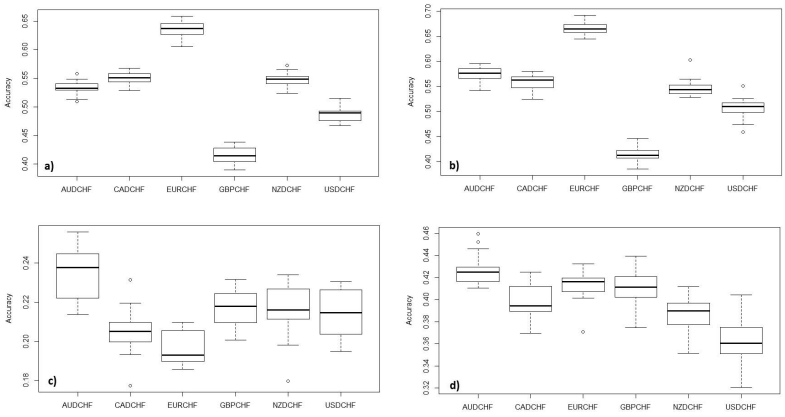

Fig. 2. Boxplot taking into account accuracy of the prediction for indicators values. (Bears indicator prediction; (b) Bulls indicator prediction; (c) CCI indicator prediction; (d) RSI indicator prediction

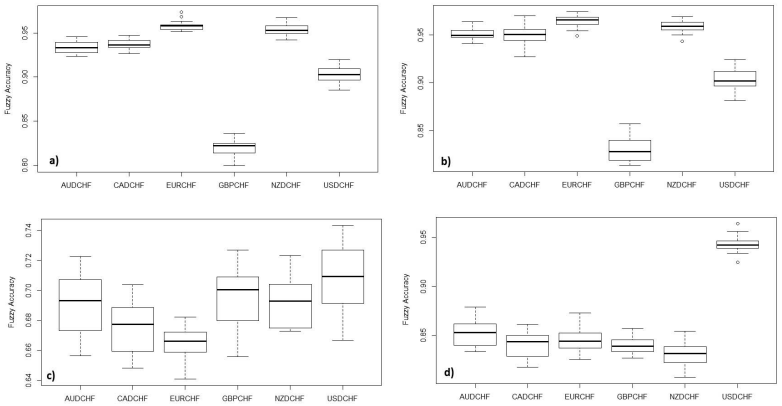

Fig. 3. Boxplot taking into account fuzzy accuracy of the prediction for indicators values. (Bears indicator prediction; (b) Bulls indicator prediction; (c) CCI indicator prediction; (d) RSI indicator prediction

TABLE II
ACCURACY AND FUZZY ACCURACY OF THE ATTRIBUTES PREDICTION.

	CCI		RSI		Bears		Bulls	
Dataset	Accuracy	Fuzzy acc.	Accuracy	Fuzzy acc.	Accuracy	Fuzzy acc.	Accuracy	Fuzzy acc.
AUDCHF	0.234371	0.691513	0.426326	0.851957	0.533773	0.933245	0.574498	0.951205
CADCHF	0.205026	0.674735	0.399079	0.840581	0.550460	0.936859	0.558355	0.949676
NZDCHF	0.196929	0.664352	0.412937	0.845944	0.635367	0.958136	0.665579	0.964248
GBPCHF	0.217206	0.695860	0.410317	0.839418	0.415682	0.820013	0.414597	0.831110
EURCHF	0.216556	0.692088	0.386295	0.830854	0.547572	0.953543	0.547114	0.959329
USDCHF	0.214609	0.707750	0.361968	0.942686	0.486387	0.903097	0.506792	0.903522

Additionally, conducted experiments allowed to show, that in the case of rule-based systems it is possible to introduce some simplifications concerned on the rule sets. Such action should lead to remove redundant indicators, which values can be predicted with high efficiency on the basis of other elements of the transaction systems.

Preliminary results given in the article allow to expect, that proposed ASC-ACDT algorithm is capable to achieve very good results in the multi-label problem. Such analysis related with the forex financial data should result in the visible reduction of the indicators used in the rule-based transaction systems, without decreasing their efficiency.

REFERENCES

[1] U. Boryczka, J. Kozak, *New insights of cooperation among ants in Ant Colony Decision Trees*, 2011 Third World Congress on Nature and Biologically Inspired Computing (NaBIC), pp. 255-260, 2011.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *Classification and Regression Trees*, Chapman and Hall, New York, 1984.

[3] R. R. Bulatovi, S. R. Bordevi, V. S. Dordevi, *Cuckoo search algorithm: a metaheuristic approach to solving the problem of optimum synthesis of a*

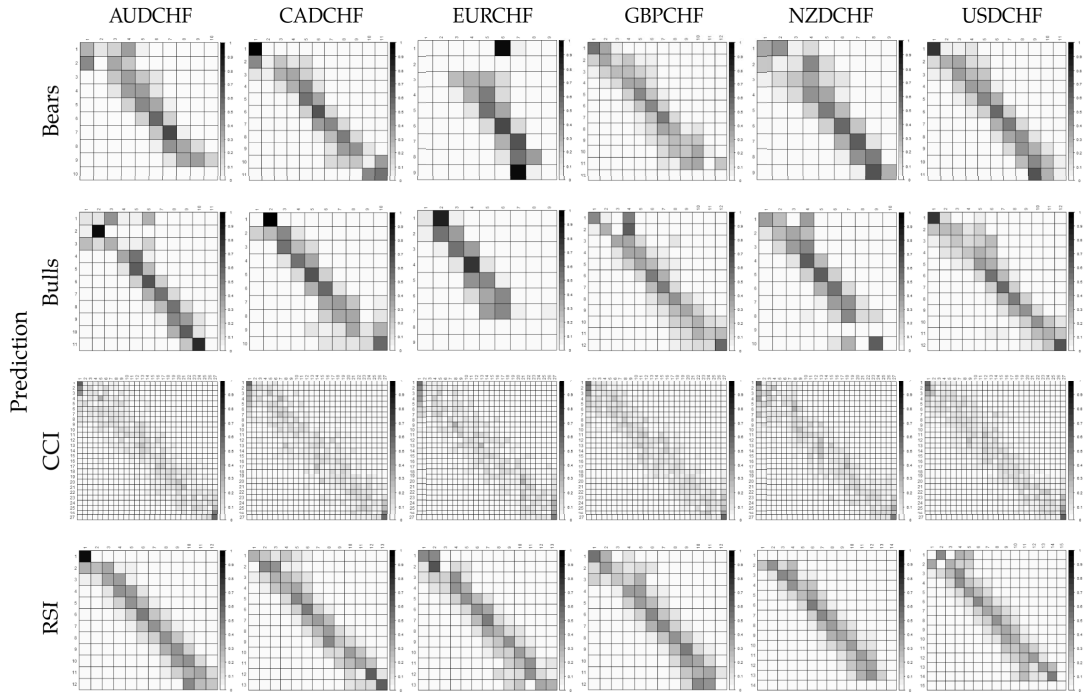

Fig. 4. Classification matrix for analyzed indicators. Darker grey color means larger number of objects assigned to given decision class - rows store the actual value, while columns - predicted indicator value

TABLE III
NUMBER OF NODES AND HEIGHT OF THE DECISION TREE.

	CCI		RSI		Bears		Bulls	
Dataset	Num. of nodes	height	Num. of nodes	height	Num. of nodes	height	Num. of nodes	height
AUDCHF	577.5	18.0	406.0	18.2	338.2	20.7	323.2	19.6
CADCHF	536.0	17.3	419.4	17.8	308.6	18.1	296.3	17.6
EURCHF	529.8	16.6	446.9	17.5	273.4	18.8	251.3	18.2
GBPCHF	514.4	17.4	387.4	18.7	383.4	17.8	367.0	17.4
NZDCHF	555.0	17.8	424.3	18.9	319.4	20.9	313.0	19.0
USDCHF	544.5	18.0	463.6	18.8	343.0	19.2	316.7	17.4

six-bar double dwell linkage, Mechanism and Machine Theory, Vol. 61, pp. 1–13, 2013.

[4] P. Cheol-Ho, S. H. Irwin, *What do we know about the profitability of technical analysis?*, Journal of Economic Surveys, Vol. 21, Is. 4 pp. 786-826, 2007.

[5] M. Dorigo, *Optimization, learning and natural algorithms (in italian)*, Ph.D. thesis, Dipartimento di Elettronica, Politecnico di Milano, IT, 1992.

[6] A. Engelbrecht,*Computational Intelligence: An Introduction*, 2nd Edition, Wiley, pp 237 261, 2007.

[7] A. H. Gandomia , A. H. Alavi, *Krill herd: A new bio-inspired optimization algorithm*, Communications in Nonlinear Science and Numerical Simulation, Vol. 17, Issue 12, pp. 4831 - 4845, 2012.

[8] A. H. Gandomi, S. Talatahari, F. Tadbiri, A. H. Alavi, *Krill herd algorithm for optimum design of truss structures*, International Journal of Bio-Inspired Computation, Vol. 5, Is. 5, pp. 281 - 288, 2013.

[9] Juszczuk P., Kozak J., Trynda K., *Decision Trees on the Foreign Exchange Market*, Intelligent Decision Technologies 2016, Springer, pp. 127 - 138, 2016.

[10] J. Kozak, U. Boryczka, *Multiple boosting in the ant colony decision forest meta-classifier* Knowledge-Based Systems 75, pp. 141 - 151, 2015.

[11] Kozak J., Boryczka U., *Collective data mining in the ant colony decision tree approach*, Information Sciences, Vol. 372, pp. 126 - 147, 2016.

[12] T.-S. Lim, W.-Y. Loh, Y.-S. Shih, *A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms*, Machine Learning 40 (3), pp. 203 – 228, 2000.

[13] J. R. Quinlan, *Introduction of decision trees*, Machine Learning 1, pp. 81-106, 1986.

[14] J. W. Wilder, *New concepts in technical trading systems*, Trend Research, 1978.

[15] X.-S. Yang, S. Deb, *Engineering optimisation by cuckoo search*, Journal of Mathematical Modelling and Numerical Optimisation, Vol. 1, No. 4, pp. 330 - 343, 2010.

[16] X.-S. Yang, *Nature-Inspired Metaheuristic Algorithms*, Luniver Press, 2008.

[17] X.-S. Yang, S. Deb, *Cuckoo search via Levy flights*, World Congress on Nature and Biologically Inspired Computing (NaBIC 2009). IEEE Publication, USA. pp. 210-214, 2009.