ASSESSING THE PERFORMANCE OF NON-BANKING FINANCIAL INSTITUTIONS – A KNOWLEDGE DISCOVERY APPROACH

Assoc. Prof. Adrian Costea Ph. D
Bucharest University of Economics
Faculty of Cybernetics, Statistics and Informatics in Economy
Bucharest, Romania

Abstract: This paper proposes a framework for assessing the performance of non-banking financial institutions (NFIs). Firstly, we present an overview of the non-banking financial institutions' sector in Romania and, then, the CAAMPL system which is used to evaluate the performance of banks. We argue that this system is suboptimal when applied to assessing NFIs' performance and that the Knowledge Discovery in Databases (KDD) process could offer specific methods that may be used to developing better systems. Next, we discuss different concepts that are closely related with the KDD process: data, information and knowledge. Finally, we present the KDD process and we show how our research problem can be formalized as a KDD process.

JEL classification: D80, D83, G21, G23, G24

Key words: knowledge discovery; data mining; prudential supervision; non-banking financial institutions; financial performance

1. INTRODUCTION

Non-banking financial institutions (NFIs) are financial entities which carry on different lending activities such as: granting of credits, including, without limitation: consumer credits, mortgage credits, real estate credits, microcredits, financing commercial transactions, factoring, discounting, and forfeiting operations (Romanian Parliament's Law No. 93/2009: Section 5). At the same time, NFIs carry on other lending activities: financial leasing, issuance guarantees, undertaking financing commitments, etc. From a prudential supervision perspective (National Bank of Romania's Regulation No.13/2010: Chapter IV), it is necessary to develop an evaluation system for the performance of non-banking financial institutions (NFIs) in order to increase the efficiency of the prudential supervision activity. By differentiating the NFIs that are good performers from the others, the rating system would allow the supervision authority to better allocate its scarce resources so that the propagation of the individual disequilibria to the whole system is prevented.

In this paper we plan to formalize the process of constructing the models for assessing comparatively the performance of NFIs (we call this process *the NFIs financial benchmarking process*), by considering this business problem as a knowledge discovery problem and by following the formal steps of a well-known discovery process called Knowledge Discovery in Databases (KDD) process (Fayyad *et al*., 1996a; Fayyad *et al*., 1996b; Fayyad *et al*., 1996c).

2. CURRENT SITUATION REGARDING THE SYSTEMS OF ASSESSING THE PERFORMANCE OF NFIS

In Romania all authorised non-banking financial institutions are included in the General Register of non-banking financial institutions. Another register opened and kept by the central bank is the Special Register, which includes only those non-banking financial institutions from the General Register that meet certain criteria of performance in terms of loans and borrowings. Non-banking financial institutions which are included in the Special Register remain entered in the General Register as well. Year 2006 was the year in which the first regulations specific to non-banking financial institutions were issued.

In October 2007 the process of licensing of all non-banking financial institutions that have submitted documentation to the central bank in 2006, 2007 has been completed. Thus, it ended with a number of 38 institutions listed in the Special Register, 218 in the General Register and 4600 entered the Evidence Register. The last register includes pawn shops and credit unions which are also considered non-banking financial institutions.

By these rules it has been established that the central bank will monitor non-banking financial institutions registered in the General Register, will prudentially supervise those in the Special Register and will keep track of those registered in the Evidence Register. Below we will refer only to the non-banking financial institutions registered in the General and Special Registers, given their importance in the total of non-banking financial institutions (NFIs).

The distribution of NFIs from both General and Special Registers based on different lending activities is shown in Figure no. 1.

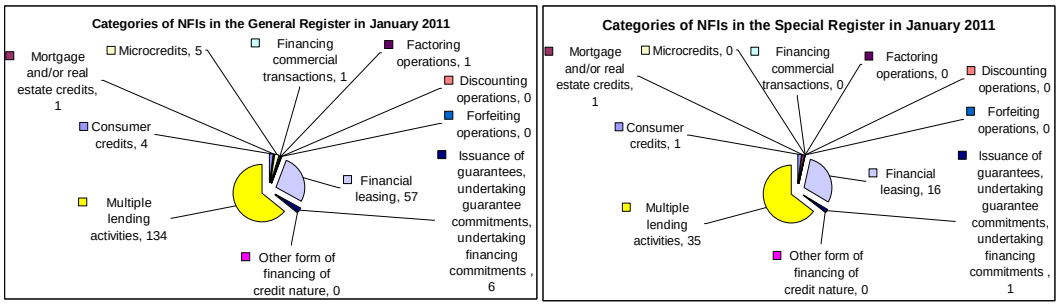


Figure no. 1 The distribution of NFIs from the General and Special Registers based on different lending activities

The problem of assessing comparatively the performance of financial institutions is not new. In Romania, credit institutions (banks) are evaluated based on the Uniform Evaluation System or the CAAMPL system (Cerna *et al.*, 2008), which assesses the performance based on six dimensions: capital adequacy (C), shareholders' quality (A), assets' quality (A), management (M), profitability (P) and liquidity (L). The six dimensions are rated using a 1 to 5 scale, where 1 represents best performance and 5 the worst. Four dimensions (capital adequacy, assets' quality, profitability, and liquidity) are quantitative dimensions and are evaluated based on a number of indicators. The other two dimensions are qualitative dimensions, evaluated based on the textual information provided by the banks as legal reporting requirements at the time of their authorization or as an effect of changes in their situation. At the same time, these two dimensions can be evaluated based on the information obtained during on-site inspections. Finally, a composite rating is calculated as a weighting average of the dimensions' ratings.

Except from being inapplicable for assessing the performance of NFIs, the CAAMPL rating system presents some disadvantages, such as:

- it uses simple linear techniques for discriminating the multidimensional space represented by the independent variables (financial performance ratios);
- the selection of independent variables is not based on scientific rigour, but on the practical experience of the members of the supervision authority;
- it is difficult to substantiate the limits for the independent variables;
- it is based mainly *„on rules"* (IMF, 2010) and does not involve quantitative methods for assessing the performance.

While still in place and useful, the CAAMPL system need to be challenged. This challenge is provided by Computational-Intelligence (CI) methods which come from different fields: *machine learning*, *artificial intelligence, evolutionary computation* and *fuzzy logic*.

The KDD process and its engine, Data Mining (DM), represent the umbrella under which the CI methods operate. There are numerous CI methods available in the scientific literature. However, we restrict the number of CI methods as it would be unfeasible to test all possible solutions (methods). This is in line with Hevner *et al.*'s (2004) sixth guideline for design science research. As a research methodology we employ the constructive (design science) research.

3. THE KDD PROCESS

In this section we discuss the different concepts that are closely related with the KDD process such as: data, information, and knowledge. Even though they are not interchangeable these three terms are related. For organizations (e.g.: non-banking financial institutions) is crucial to clarify what data, information and knowledge mean, which of them is needed, which of them organizations already own, how they differ and how to get from one to the other.

In a general context, ***data*** is a set of discrete, objective facts about events (Davenport & Prusak, 1998). In an organizational context data is seen as a collection of transaction records that has no significance beyond its existence. Data can be considered as a driver for information and knowledge, a means through which information and knowledge can be stored and transferred. Nowadays there is a shift in data management responsibility: from centralized information systems department to individuals' desktop PCs. In other words, the availability of data within organization has increased along with the technology that supports distributed systems. Even though organizations need and sometimes are heavily dependent on data, it does not mean that more data is necessarily better data. As Davenport & Prusak (1998) suggest, the argument that one should gather more data so that the solutions for the organization problems will rise automatically is false from two perspectives: first too much data can *hide* the data that matters and, second, data provides no judgment or interpretation about what has happened.

Information is data that has relevance and purpose (Davenport & Prusak, 1998) or a flow of meaningful messages (Nonaka & Takeuchi, 1995). The information is commonly seen as a message that *"gives shape to"* data. It has a sender and a receiver, but the judgment of the information value – if it really informs the receiver or not – rests with the receiver. According to Davenport & Prusak (1998) there are several ways of transforming the data into information: contextualization – the purpose for what data was gathered is known; categorization – the units of the analysis or key components of the data are known; calculation – transformation of the data using mathematics or statistics; correction – the

data is cleared of errors; condensation – the data is summarized in a concise form. The information can be transmitted using soft or hard networks. Among hard networks we mention: electronic mail-boxes, wires, online instant messengers, satellite, post offices, etc. Soft networks are informal meetings, coffee-breaks, etc. Both information and explicit knowledge can be transmitted via soft networks. In literature there is still confusion about the difference between information and knowledge. Kogut & Zander (1992) present information as a form of knowledge, stating that information is "knowledge which can be transmitted without loss of integrity". Stenmark (2002) presents different definitions of data, information, and knowledge (Stenmark, 2002, Table 1, pp. 2).

Knowledge derives from information as information derives from data. The transformation of information through knowledge is done according to Davenport & Prusak (1998) through human-like activities such as comparison – how does information about this situation compare to other situations that are known; consequence – what are the implications of the information for decisions and actions; connections – how does this piece of knowledge relate to others; and conversation – what do other knowledgeable people think about this information. Quigley & Debons (1999) relate information with who?, when?, what?, and where? question types, and knowledge with why?, and how?. In our thinking all human-like activities through which information can be translated into knowledge can be performed partially using computational intelligence techniques. Comparisons and connections are highlighted using financial benchmarking; consequences of some actions can be traced using the classification models. A society of intelligent software agents can resemble human beings' conversations by sharing and exchanging information about common goals. We agree with the fact that the total human substitution by intelligent systems is neither possible, nor efficient. At the same time, we think that intelligent systems can provide interested parties with accurate, useful and timely knowledge in the decision making process, something that even very experienced people can not provide. We look at our models as a complementary source to support the decision making process.

Knowledge can move also down to the value chain and become information and data. Too much knowledge is hard to disseminate. As the ancient greek playwright Aeschylus said: "Who knows useful things, not many things, is wise".

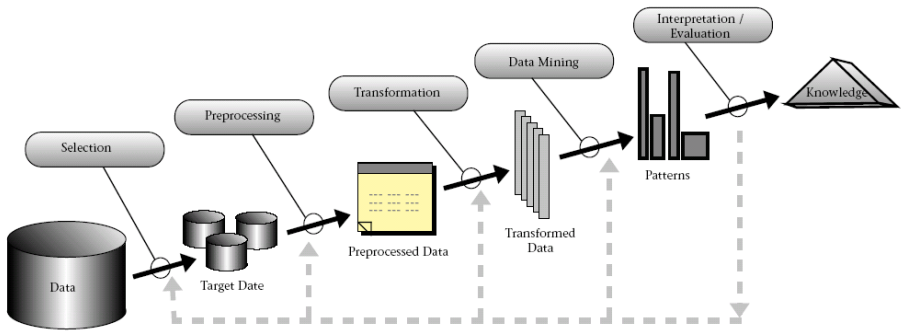
Knowledge is very important asset for organizations especially because the other resources (technology, capital, land, labour) are not anymore sources of sustainable competitive advantage (Davenport & Prusak, 1998, p. 16).

Nowadays, companies are bombarded with tons of data about their market environment. This publicly available data is crucial for their competitiveness. The managers face two problems with regard to this data: information (data) overload and data usefulness. According to Gantz & Reinsel (2010) the estimated volume of digital information created in 2010 has amounted to 1.2 zettabytes (1 zettabyte = 1 trillion gigabytes) and by 2020, our Digital Universe will be 44 times as big as in 2009 (0.8 zettabytes). Data usefulness is closely related with the process of data transformation into knowledge. As useful the knowledge obtained from this transformation is as more useful is the data from which the knowledge was obtained.

KDD addresses both these problems (information overload and data usefulness) by looking at the "new generation of computational theories and tools that can assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data" (Fayyad *et al*., 1996c). KDD is at the confluence of many different disciplines and research fields such us statistics, information theory, databases, artificial intelligence,

machine learning, pattern recognition, fuzzy sets, visualization, and high performance computing. Except these fields, there are some other long-term contributors less mentioned to the KDD growing research field: sciences, logic, and philosophy of science (Klösgen & Zytkow, 2002, p. 22). The link between sciences (quantitative theories) and KDD yields in the usefulness of the empirical demonstrations and generalizations that can be extracted from the data. In science (e.g.: chemistry, physics) basic laws and theories can emerge from a concrete experiment that can be applied to a broad range of situations (Klösgen & Zytkow, 2002, p. 23). In KDD the search for patterns in data can be followed by transformation of the discovered regularities into theories that cover many data sets. The framework of logic is the base for many research disciplines such as mathematics, the theory of databases, artificial intelligence, and, therefore, is linked indirectly with the KDD process. For example, in KDD we may generate some classification rules and treat a minimal number of them as axioms and the other as derived from the axioms, thus, resembling a deductive system. However, KDD is undermining the application of deductive systems by accepting a limited accuracy for the axioms. Inductive logic programming is also present in the emerging KDD field (e.g.: data can be expressed as Prolog literals, while knowledge in the form of Prolog rules). The influence of philosophy of science on KDD is mainly indirect through the introduction of key field concepts and research frameworks that any well-established research field should have.

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad *et al*., 1996a). In other words, KDD is the process of data transformation into knowledge (Figure no. 2).



Source: adapted from Fayyad et al. (1996c)

Figure no. 2 KDD process

In KDD definition, *data* is represented by a set of facts (records in a database table), while *pattern* refers to a subset of the data that share similar characteristics or to some rule that covers a number of observations. Term *process* implies that KDD consists of many steps, which involve data preparation, pattern discovery and knowledge evaluation and refinement. The term *nontrivial* is related with the data mining step of the KDD process in the sense that the methods used to analyze the data are not trivial (e.g.: computing averages), but advanced (CI methods). Fayyad *et al*. (1996c) consider the patterns to be knowledge if they *"exceed some interestingness threshold"* and are determined *"by whatever functions and thresholds the user chooses"*. In other words, knowledge is user oriented and domain specific.

The discovered patterns should be *valid* which means that they should be valid on new data with some degree of certainty (accuracy). Patterns should be *novel* which means that the user could not otherwise find the same patterns, and *understandable* for the users after (if necessary) some post-processing.

The KDD process consists of the following steps (Klösgen & Zytkow, 2002, p. 10):

Definition and analysis of the business problem that is targeted to be solved through KDD process. Among the business problems that can be addressed via knowledge mining in large databases we mention: predicting and analyzing customer behavior, processing loan applications, predict a portfolio's return of investment, optimal shelf space allocation, analysis of exceptions, etc. Our business problem is to *assess comparatively the financial performance of NFIs* (or *NFIs' financial performance benchmarking*). This step matches Fayad *et al*.'s (1996c) first step of the KDD process: "developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer's viewpoint".

Understanding and preparation of data implies selection of the target data set on which the discovery process is to be performed, data cleaning and preprocessing, and data reduction and projection. This step of KDD process is the most time consuming one: according to Romeu (2001) up to sixty per cent of total project time is dedicated to data preparation. When selecting the variables (attributes) we should focus on their relevance to the problem at hand. The data from different tables should be pulled together, because *"the preponderance of discovery tools apply to single tables"* (Klösgen & Zytkow, 2002). Data cleaning task is concerned with the finding odd and missing values and replace them with legitimate values. There are several data preprocessing methods that have to be tested to find the proper one for a particular data set. If the data set is too large for performing a reasonable mining task, it can be reduced (feature selection, elimination of incomplete observations) or transformed (principal component analysis). Our dataset would consist of several financial ratios gathered quarterly for approximately 50 NFIs, from 2006 to 2010. The financial ratios characterize each NFI in terms of capital adequacy, assets' quality and profitability. This step comprises the second (creating the dataset), third (data cleaning and pre-processing), and fourth (data reduction and projection) steps of Fayyad *et al*.'s (1996c) KDD process.

Setup of the search for knowledge. Depending on the data at hand and on the business problem that we attend to solve (goal of the KDD process) we can use a combination of data mining tasks (e.g.: applying clustering task to obtain the performance class – rating – and, then, classification task to model the relationship between the class variable and the independent variables. We call such a mixture a *hybrid* data mining task). The second part of this step is to decide which data mining method(s) and algorithm(s) are better for performing the search for patterns. Romeu (2001) groups data mining approaches in three categories: mathematically based, statistically based and "mixed" algorithms. The last category includes: clustering methods, induction techniques (e.g.: decision trees), neural networks, and genetic algorithms. We introduced these techniques as CI methods. We plan to explore and combine statistically-based and CI methods in order to assess comparatively the performance of NFIs. This step unites steps number five (matching the goal of the KDD process to a particular data-mining task) and six (choosing the data-mining methods for searching for patterns) of Fayyad *et al*.'s (1996c) KDD process.

Data mining (DM) step is the most important step in KDD process. DM is defined as *"a step in the KDD process consisting of applying data analysis and discovery*

algorithms that, under acceptable computational efficiency limitations produce a particular enumeration of patterns over the data" (Fayyad *et al*., 1996b). The term DM has its roots in statistically oriented data analysis research communities. Actually, the correct term for this KDD step should be Knowledge Mining since we mine for knowledge and not for data (as we mine for gold and other precious metals and not for dirt or rock). The user plays an important role at this stage and can help the data mining method by correctly performing the previous steps. It corresponds to step seven of Fayyad's KDD process.

Interpretation and evaluation of the mined patterns or knowledge refinement involves the visualization and interpretation of the extracted patterns/models or of the data covered by the rules extracted. For instance in the case of NFIs' financial benchmarking through clustering this step will consist of looking at the financial performance clusters individually and at the characteristics (variables) of each cluster. This step matches the eighth step of Fayyad's "interpreting mined patterns".

Application of knowledge to the business problems and the consolidation of the discovered knowledge involve incorporating the knowledge in the organization general information system. At this stage predictions can be performed based on the discovered knowledge. For example, in assessing NFIs' financial performance, the obtained financial classification model can be applied for the newly observed data and the information can be documented and reported for interested parties. At this stage we can reveal weaknesses and suggest the best course of actions that an NFI should take so that its financial performance would improve significantly. This step resembles the ninth step in Fayyad *et al*. (1996c), "acting on the discovered knowledge".

Knowledge is not necessarily derived only from numerical structured data (quantitative data). Unstructured, textual (qualitative) data might contain nuggets of knowledge as well. Approximately 90% of the world's data is held in unstructured formats. Tan (1999) claims that 80% of information handled within organizations is of a textual nature. In Table no. 1 we present the main differences of two discovery activities (KDD with quantitative data – DM or data mining and KDD with qualitative data – TM or text mining).

Table no. 1 Differences between text and numeric data processing

	KDD of quantitative data	KDD of qualitative data
Data type	Numeric (structured)	Text, Document (unstructured)
Goal	Find patterns in data	Find patterns in data
Basic unit	record	document
Basic element	At the intersection of records and attributes	At the intersection of documents and terms
Mining steps and technical problems	Numerical data set preparation Data cleaning and preprocessing, Data reduction and projection Task(s) and method(s) selection Algorithm(s) selection Evaluation and interpretation	Documents data set preparation Linguistic preprocessing Term generation Term filtering Complex and subtle syntactic constructions (e.g.: Company's X profit is *not* bad) Synonymy and polysemy Term taxonomy construction Mining the association rules from document collections

Text and numeric data mining can be complementary: text analysis can be used as an input for correctly choosing the parameters of our quantitative models. In other words,

by analyzing financial experts' statements we can focus our quantitative analysis on variables that experts use. For example statements like that of Thomas Kahn, the President and Co-Director of Investment at Kahn Brothers & Company, Inc. who described on CNBC Europe how he values a company: *"We look at companies with strong balance sheets, no debts or low debts and with a lot of cash"* can help us deciding our parameter choice. Moreover, companies from one particular sector can be comparatively analyzed by mining their income statements and balance sheets, and at the same time the textual parts of the annual reports can be scrutinized to find proofs for the rules provided by the quantitative data.

4. DATA MINING AND NFIS' FINANCIAL PERFORMANCE EVALUATION

Data mining step is the core of KDD process, because is the outcome of this step that after evaluation and refinement gives the nuggets of knowledge. Here is the point were KDD process differs from other analytical tools (query and reporting tools, statistical analysis packages, OLAP and visualization tools): the goal of KDD process and DM is to *discover* new patterns in data, while most analytical tools are based on verification where *"the system is limited to verifying user's hypotheses"* (Fayyad *et al*., 1996b). The problem with the verification-based approach is that it *"relies on the intuition of the analyst to pose the original question and refine the analysis based on the results of potentially complex queries against a database"* (Moxon, 1996). DM supports the discovery-based approach since *"one defining data-mining characteristic is that research hypotheses and relationships between data variables are obtained as a result of (instead of as a condition for) the analyses activities"* (Romeu, 2001).

In order to fulfil its role DM could perform a number of tasks such as clustering, classification, regression, dependency modelling, summarisation, and change and deviation detection. The link between these tasks and the real-world applications or business problems (the final goal of KDD is to address these problems) is not straightforward, because real-world applications rarely have a simple single solution. Many different tasks may match a particular application, depending on how one approaches the problem (Smyth, 2002). For example, our real-world application would be to assess NFIs' financial performance. Treating our problem as a supervised learning task implies that we already have financial performance classes for all the observations used to train the classifier. Actually there are no labelled data available, thus, the performance class variable (the rating) has to be created at the beginning, by treating our problem as an unsupervised task.

Only after the class variable has been constructed, can a classifier be trained. Smyth (2002) pinpoints various advices worth consideration when linking real-world applications with the data-mining task. The author states that it is advisable to start with only one task to address a real-world application and, only if necessary, add more complex ones. He also suggests removing the irrelevant details of the original formulation of the problem so that it resembles more closely a standard textbook task description. In order to select the proper task for a given problem, the data miner should have a complete understanding of both the business problem addressed and the task linked to it. Finally, Smyth (2002) states that it is better to approximate the solution to the right problem than it is to solve the wrong problem exactly.

Different authors (Fayyad *et al*., 1996b; Klösgen & Zytkow, 2002; Romeu, 2001) have defined the tasks performed by the means of Data Mining as follows:

Clustering. Traditional clustering methods intend to identify patterns in data and create partitions with different structures. These partitions are called clusters, and elements

within each cluster should share similar characteristics. The partitions can be mutually exclusive (disjoint) or may contain observations that belong in some degree to several clusters (overlapping). The standard application of clustering in business has been consumer behaviour analysis where clusters are constructed with consumers that have similar purchasing characteristics. Clustering is also known as unsupervised classification.

Classification. Bock (2002) presents three approaches related to classification:

- Classification as an ordering system for objects (e.g. classification of books in a library, the ordering of chemical elements in the periodic system, classification of products and merchandise for international standardisation).
- Classification as a class assignment or supervised learning (learning with a teacher). This approach corresponds to the common view of the classification task: a learning function that maps a data item (observation) into one of several predefined classes (Hand *et al*., 2001). In this case, classification models (classifiers) are built with which new observations can be assigned different classes. For example in medicine a disease can be recognised based on patient symptoms, in performance benchmarking NFIs can be classified according to their financial performance, etc.
- Classification as class constructing or clustering or unsupervised learning (learning without a teacher).

Clustering and supervised learning can be combined when class variables are not available to obtain hybrid classifiers. Throughout the research we plan to address business problems by both simplifying them to a single data-mining task and also by matching them with different data-mining tasks when necessary.

Regression is the process of learning a function that maps a data item to a real-value prediction variable and the discovery of functional relationships between variables (Fayyad *et al*., 1996b). Classification can be considered as a particular case of regression analysis where the outcome is a discrete value (class). In regression we try to find a function that links an output (or many) to a number of inputs. These functions range from very simple ones (linear, one input) to very complex (non-linear, many inputs) leading to three different regression models: standard linear model, generalised linear model, and generalised additive model. The standard linear model links the outputs to the inputs with a function that is a linear combination of the inputs. The generalised linear model is applied predominantly to perform classification tasks since the outcome values are constrained to a sensible range. For example the logit function derives expected values between zero and one. The generalised additive models can accommodate the non-linear effects of the original inputs. The standard classic approach to model fitting in regression is called maximum likelihood estimation (MLE). MLEs are estimates that maximise the likelihood function, which is the joint probability density of the data (Rao & Potts, 2002).

Dependency modelling – concerns constructing models that describe significant dependencies between variables. At the structural level the dependency model specifies which variables are dependent on each other, while at the quantitative level the model specifies the strengths of the dependencies (Fayyad *et al*., 1996c). Probabilistic and causal networks (Spirtes, 2002) are two techniques that are increasingly applied to performing this data-mining task.

Summarisation consists of methods for finding a compact description of a subset of data. Among these methods there are: calculation of standard deviation and means for the observations, derivation of summary rules, multivariate visualisation techniques, and discovery of functional dependencies between variables (Fayyad *et al*., 1996c).

Change and deviation detection involves finding the differences between current data and previously measured or normative values. Change detection deals with analysing change (one entity observed at two points of time) or trend (a sequence of equidistant points of time) over the dataset. Deviation analysis starts with identifying the deviating sub-groups (sub-groups where the target variable differs significantly from its expected value in relation to the input values from that particular sub-group) and rely on hypothesis testing to test whether the sub-group is interesting or not. Generally, the rejected null hypothesis assumes an uninteresting, non-deviating sub-group. Klösgen & Anand (2002) call this data-mining task *sub-group discovery*.

The algorithms used to perform data-mining tasks described above are numerous and they come from different research fields (statistics, machine learning, artificial intelligence, fuzzy logic, etc.). Romeu (2001) groups data-mining algorithms in three categories: mathematically based, statistically based and "mixed" algorithms.

Mathematically based (deterministic) algorithms include mathematical programming (linear, non-linear, integer), network methods (link and affinity analysis), and memory-based reasoning approaches (nearest-neighbour classifiers).

Statistically based (stochastic) algorithms include traditional statistics regression, discrimination techniques (linear discriminants, quadratic discriminants, logistic discriminants or logistic regression), statistical time series analysis, factor analysis, etc.

The difference between mathematical and statistical algorithms lies in the approach that they are based upon: mathematical models are deterministic (random phenomena are not involved and these models produce the same output for a given starting condition), while statistical ones are stochastic (based on random trials).

"Mixed" algorithms borrow heavily from both, the algorithmic and the stochastic components. Romeu (2002) includes here: clustering methods, induction techniques such as decision trees, neural networks, fuzzy logic and genetic algorithms. We introduced these techniques as CI methods. In our research we plan to explore and combine statistically-based and CI methods to address the problem of assessing comparatively the performance of NFIs. We match our research problem with both data-mining clustering and classification tasks. For the clustering phase we plan to explore algorithms such as: Self-Organising Maps, C-Means, Fuzzy C-Means and our previously developed algorithm: Weighting FCM algorithm. For the classification phase we plan to explore classification methods such as multinomial logistic regression, Quinlan's algorithm for decision-tree induction, artificial neural networks for supervised learning, and genetic algorithms for learning the weights of an ANN.

Whatever the algorithm we use to perform the data-mining tasks, we need criteria to evaluate its performance to be able to rigorously compare it with other approaches. For our models we plan to use quantitative criteria such as quantisation error, accuracy rate or mean square error, or qualitative ones such as fidelity with real-world phenomena, form and content, and richness of knowledge in the form of class predictions.

5. CONCLUSIONS

In this paper we formalize the problem of assessing the performance of non-banking financial institutions (NFIs) by employing a knowledge-derivation schema, known in scientific literature as Knowledge Discovery in Databases (KDD) process. Firstly, we present an overview of the non-banking financial institutions' sector in Romania. Then, we describe the CAAMPL system used by the supervision authority to evaluate the performance of its supervised entities. The CAAMPL system presents a number of

disadvantages and, in its current form, it is not applicable to NFIs. We argue that there is a need for new systems that rely on other methods than traditional techniques in order to properly assess the performance of NFIs. These methods (which we refer to as Computational-Intelligence – CI – methods) come from different research fields (machine learning, artificial intelligence, fuzzy logic, etc).

Next, we present the Knowledge Discovery in Databases (KDD) process which represents the umbrella under which the CI methods operate. We discuss different concepts that are closely related with this process, namely, data, information and knowledge. Finally, we show how KDD process can be used as a standardized platform in the research activity concerning the development of NFIs' financial performance benchmarking models.

ACKNOWLEDGEMENTS

This work was supported from the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/89/1.5/S/59184 „Performance and excellence in postdoctoral research in Romanian economics science domain".

REFERENCES

1. Bock, H.H. "The Goal of Classification", *Handbook of Data Mining and Knowledge Discovery – Data Mining Tasks and Methods* (Section 16.1.1 – pp. 254-258). Willi Klösgen and Jan M. Zytkow (eds.), Oxford University Press, New York, NY, 2002.
2. Cerna, S., Donath, L., Seulean, V., Herbei, M., Bărglăzan, D., Albulescu, C. and Boldea, B. *Financial Stability*, West University Publishing House, Timişoara, Romania, 2008.
3. Davenport, T.H. and Prusak, L. *Working knowledge: how organizations manage what they know*, Harvard Business School Press, Boston, Massachusetts, 1998.
4. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. "From Data Mining to Knowledge Discovery: An Overview", *Advances in Knowledge Discovery and Data Mining*. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (eds.), AAAI Press, Menlo Park, California, pp. 1-30, 1996a.
5. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. "Knowledge Discovery and Data Mining: Towards a Unifying Framework", *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, E. Simoudis, J. Han and U. Fayyad (eds.), AAAI Press, Portland, Oregon, August 2-4, pp. 82-88, 1996b.
6. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. "From Data Mining to Knowledge Discovery in Databases", *AI Magazine* vol. 17, no. 3, pp. 37-54, 1996c.
7. Gantz, J. and Reinsel, D. "The Digital Universe Decade – Are You Ready?", *IDC – IVIEW*, May 2010, sponsored by EMC Corporation. [http://idcdocserv.com/925]
8. Hand, D.J., Mannila, H. and Smyth, P. *Principles of Data Mining*. The MIT Press, Cambridge, 2001.
9. Hevner, A.R., "Design Science in Information Systems Research", *MIS*

March, S.T., Park, J. and Ram, S. *Quarterly* vol. 28, no. 1, pp. 75-105, 2004.

10. International Monetary Fund (IMF) International Monetary Fund, *Romania: Financial Sector Stability Assessment*, Country IMF Report No. 10/47, February 2010.
11. Klösgen, W. and Anand, T.S. "Subgroup Discovery", *Handbook of Data Mining and Knowledge Discovery – Data Mining Tasks and Methods* (Section 16.3 – pp. 354-367). Willi Klösgen and Jan M. Zytkow (eds.), Oxford University Press, New York, NY, 2002.
12. Klösgen, W. and Zytkow, J.M. *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press, New York, NY, 2002.
13. Kogut, B. and Zander, U. "Knowledge of the Firm. Combinative Capabilities, and the Replication of Technology", *Organization Science* vol. 3, no. 3, pp. 383-397, 1992.
14. Moxon, B. *Defining Data Mining - The Hows and Whys of Data Mining, and How It Differs From Other Analytical Techniques*, 1996.
15. National Bank of Romania (NBR) *Regulation No. 13/2010 concerning the organization and operation of National Bank of Romania*, issued by National Bank of Romania, 2010.
16. Nonaka, I. and Takeuchi, H. *The knowledge-creating company: how Japanese companies create the dynamics of innovation*. Oxford University Press, New York, NY, 1995.
17. Quigley, E.J. and Debons, A. "Interrogative Theory of Information and Knowledge", *Proceedings of SIGCPR '99*, ACM Press, New Orleans, LA., pp. 4-10, 1999.
18. Rao, J.S. and Potts, W.J.E. "Multidimensional Regression Analysis", *Handbook of Data Mining and Knowledge Discovery – Data Mining Tasks and Methods* (Section 16.4.3 – pp. 380-386). Willi Klösgen and Jan M. Zytkow (eds.), Oxford University Press, New York, NY, 2002.
19. Romanian Parliament *Law No. 93/2009 on non-banking financial institutions*, issued by Romanian Parliament, published in Monitorul National al României, Part One, no. 259 of 21 April 2009.
20. Romeu, J.L. "Operations Research/Statistics Techniques: A Key to Quantitative Data Mining", *Proceedings of FCSM (Federal Committee on Statistical Methodology) Conference*, Key Bridge Marriott, Arlington,Virginia, November 14-16, 2001.
21. Smyth, P. "Selection of Tasks", *Handbook of Data Mining and Knowledge Discovery – Task and Method Selection* (Section 17.1 – pp. 443-444). W. Klösgen and J.M. Zytkow (eds.), Oxford University Press, New York, NY, 2002.
22. Spirtes, P.L. "Probabilistic and Causal Networks", *Handbook of Data Mining and Knowledge Discovery – Data Mining Tasks and Methods* (Section 16.6 – pp. 396-409). Willi Klösgen and Jan M. Zytkow (eds.), Oxford University Press, New York, NY, 2002.
23. Stenmark, D. "Information vs. Knowledge: The Role of intranets in Knowledge Management", *Proceedings of HICSS-35*, IEEE Press, Hawaii, January 7-10, 2002.
24. Tan, A. "Text Mining: The state of the art and the challenges", *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, Beijing, China, 1999, pp. 65-70.
[http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf]