Annales UMCS Informatica AI 4 (2006) 279-290

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

http://www.annales.umcs.lublin.pl/

KTDA: emerging patterns based data analysis system

Roman Podraza*, Krzysztof Tomaszewski

Institute of Computer Science, Warsaw University of Technology,
Nowowiejska 15/19, 00-665 Warsaw, Poland

Abstract

Emerging patterns are kind of relationships discovered in databases containing a decision attribute. They represent contrast characteristics of individual decision classes. This form of knowledge can be useful for experts and has been successfully employed in a field of classification. In this paper we present the KTDA system. It enables discovering emerging patterns and applies them to classification purposes. The system has capabilities of identifying improper data by making use of data credibility analysis, a new approach to assessment data typicality.

1. Introduction

Knowledge discovery or data based inference is one of the most important purpose of accumulating data and maintaining large, often only growing, databases. Emerging patterns (EPs) [1] are examples of special relationships observed on attribute values of items. EPs can be then analyzed by experts (supported by computer systems) to discover new rules or relations in a given domain to understand it better. For instance EPs can be exploited for classification purposes. It seems nowadays almost no one has to be convinced of benefits of data mining and knowledge discovery, especially in the business world.

But all data analysis and knowledge discovery make sense only if processed data are credible. At first, to ensure the most possible data credibility, validity and consistency checks are used at the data gathering stage. Then in most cases, a large number of processed records are analyzed to gain some generalized information, facts, rules. There is an unspoken assumption that most of data are correct, thus a minor, not credible part of considered dataset will not disrupt discovered knowledge too much. Often there is so much of data that we can reject some of them by applying some data cleaning procedures without much information loss. However, there exist still some applications where such

*Corresponding author: *e-mail address*: R.Podraza@ii.pw.edu.pl

approach is inappropriate. As an example one can point medicine [2], where a single record of a database can often represent an individual patient. In such a case no records can be removed, even if there are indications that data may be corrupted. Moreover, in such sensitive domains data credibility gets its special significance. If the data based inference can have any influence on medical decisions, it is obvious that a particular care must be taken to ensure or at least assess data credibility. One of possible approaches is to employ some data credibility estimation mechanism which will pay expert's attention to records, which seem to be most incredible.

In this paper we present the KTDA (shortening for *KT Data Analysis*) system. It is a user-friendly tool for discovering emerging patterns in data. The KTDA system implements two different algorithms of discovering emerging patterns, proposed in [2] and [3], but with some extensions and improvements. EPs enable data classification for which the CAEP algorithm [4] is applied. Moreover, with the KTDA system it is possible to assess data credibility using the credibility coefficient, as proposed in [5] and [6]. In the KTDA system an original credibility coefficient computing algorithm was implemented. It takes into account data characteristics expressed by discovered emerging patterns. Its details are going to be published elsewhere. The paper is organized as follows. In Section 2 a short description of emerging patterns is given. Then, in Section 3, a brief introduction into data credibility analysis is submitted. After presenting in Section 4 an overall view of the KTDA system and its capabilities the paper is completed with some conclusions.

2. Emerging patterns

Emerging patterns are closely related to frequent patterns, widely known as *frequent itemsets* [7]. Both are kinds of relations on attribute values discovered in datasets and both have the same form. In this paper we define a dataset as a set of data records, each described with the same set of attributes which can be continuous (numeric) or nominal (discrete).

A pattern consists of some terms which, in fact, are individual conditions or, in other terminology, true-false tests. Each condition refers to a single dataset attribute and determines a set of values of this attribute satisfying this condition. In most cases conditions for continuous attributes check whether the attribute value is less-equal or greater than the given thresholds. A condition for a nominal attribute checks if its value is equal or not equal to a certain constant. A particular record agrees with the whole pattern if and only if it satisfies all conditions contained in this pattern. Then we say the pattern matches to this record. The ratio of the number of records matched by the pattern to the number of all records in the considered dataset is named the pattern *support*.

If we are interested what attribute values often appear jointly we would like to discover in our dataset some patterns with a support high enough (greater or

equal then a specified support threshold). These are frequent itemsets and they describe some characteristic features, states or relations in the dataset (at the given support threshold). Now let us assume that our dataset contains a decision attribute. This is a typical nominal attribute but its value denotes association of the given record to a group of records with the same value. These disjoint groups of records create decision classes. For example, *diagnosis* can be a good decision attribute dividing some patients' dataset into two decision classes: *healthy* ones and *ill* ones. Now if we are curious to know what is distinguished in one of these decision classes the frequent itemsets are not sufficient. Some of these patterns could be common to both decision classes (high support in both classes) and do not represent knowledge describing only *healthy* class or only *ill* class. Really interesting are these patterns which have a high support in one decision class and at the same time a low support in the other one. To distinguish two decision classes it is desirable to find out such patterns which are frequent itemsets in one class and are infrequent in the other one. These patterns are just called emerging patterns. The decision class in which an EP has a higher support is referred to as a *target class* for this EP. In more general situation there are N decision classes and we are interested in discovering EPs for each decision class as their target class. In this case for each decision class we compose a temporary division of the dataset into two subsets, the first one consisting only of records belonging to the decision class and a second one consisting of all other records (the rest of the dataset). The ratio of the pattern support in its target class to the pattern support in the rest of the dataset is a *growth rate* for this pattern.

How high should be EP's support in its target class and how low in the rest of the dataset? Actual values of support are not important. The EP's growth rate is essential. Larger values of the growth rate denote more characteristic EPs for its target class. In the approach proposed in [2] a growth rate threshold (greater than 1) is arbitrary chosen and only these EPs which have the growth rates greater or equal to that threshold are discovered. As a result we can obtain many EPs with quite low values of both supports and still having satisfactory value of the grow rates.

The other approach [3] is to detect only EPs with sufficient statistical significance. In this methodology the growth rate threshold is of no importance and a significance level value parameterizes the set of results (EPs). The significance level value is used then in a process of statistical hypothesis testing to assess statistical significance of each inferred EP and not significant EPs are rejected. In consequence, we can acquire many EPs with lower growth rates but with higher supports and we have got the guarantee that they are all statistically significant at the specified level.

These two approaches lead to different sets of EPs generated from the same dataset although obviously many patterns are the same or similar. In the KTDA system both methods of discovering Emerging Patterns have been implemented.

The first of them utilizes maximal frequent itemsets approach [7]. Details of the algorithm can be found in [2]. The second method makes use of decision trees [8]. The exact Fisher's test [9] is employed in the procedure of decision tree construction as a statistical test for assessing significance. This algorithm was proposed in [3].

3. Data credibility analysis

Data credibility analysis is a new research area in a domain of knowledge acquisition. The main goal of the research is estimating credibility of individual records of analyzed datasets and applying this expertise for ensuring maximal data credibility. Evaluation of data credibility is done by specialized heuristic algorithms. Some of them were described in [5] and [6]. The most important aspect of these algorithms is unawareness of meaning of the processed data. This makes them general, universal and ready to operate on any data. Based on a given dataset they assign to each data record the relative credibility estimation known as a credibility coefficient. This is just a real number in range [0, 1]. Lower values indicate lower estimated credibility. The intention of the proposed data credibility assessment algorithms is to assign lower credibility coefficients to less typical record. They are commonly invalid, outlying or abnormal data. In any of these cases it is good to identify such records. Invalid data are obviously incredible and outlying data do not match well to typical schemes so they cannot be used to infer general knowledge. For example if in a medical application an outlying patient record denotes a special case, he or she is going probably to be treated with some extra care and most likely will get slightly different remedies. Since calculated credibility coefficients are relative to the analyzed dataset the system itself cannot decide how low coefficient value denotes an incredible record. Nevertheless an expert can revise a chosen number of records (for example: 10% of the dataset) which were given the lowest credibility coefficient. Then he/she can make the decision how significant are the records and what to do with them (e.g. neglect, correct, start thorough investigation of cases).

The KTDA system contains our two new, general algorithms of computing credibility coefficients: *the Voting Classifier Method* and *the Multi Credibility Coefficient Method*. They are general because their parameters are other algorithms. They will be described elsewhere in details.

The Voting Classifier Method computes credibility coefficients by using a voting classifier. In the KTDA system it uses the CAEP (Classification by Aggregating Emerging Patterns) classifier [4], which is a voting one. In this way EPs can be exploited in data credibility analysis. Some other kinds of voting classifiers, such as neural network, SVM, k-NN, Bayesian classifiers, etc., are planned to be exploited as well for the Voting Classifier Method.

The Multi Credibility Coefficient Method allows to obtain credibility coefficients as an aggregation of many credibility coefficients computed by an arbitrary number of algorithms. The main idea of proposing this solution was to gain all advantages of various approaches. Different credibility coefficient computing algorithms produce better results in different cases. Usually it is impossible to choose the best one of them. Instead of choosing one such algorithm it would be better to use them all and benefit from their individual advantages. This is exactly what the Multi Credibility Coefficient Method performs. Our initial experiments have shown that this approach allows to obtain even better results than the best outcome of a single method, which is incorporated into the Multi Credibility Coefficient Method. In the current version of the KTDA system the Multi Credibility Coefficient Method has a fixed configuration consisting of two Voting Classifier Methods based on CAEP and differing in algorithms they use to discover EPs.

4. System overview

The KTDA system has been developed for 1.5 years. It has a comfortable graphic user interface and its source code level portability (C++) enables implementations under many different operating systems. The KTDA system has been successfully used under Linux and MS Windows.

The KTDA system has multi-window interface architecture but its main window plays a key role in controlling the execution of the program and managing other information windows. The KTDA system main window is shown in Figs. 1 and 2. It has a very simple and intuitive interface consisting of the main menu and two views: *Object* and *Windows*. In most cases only the *File* menu from the main menu is used for opening and closing datasets.

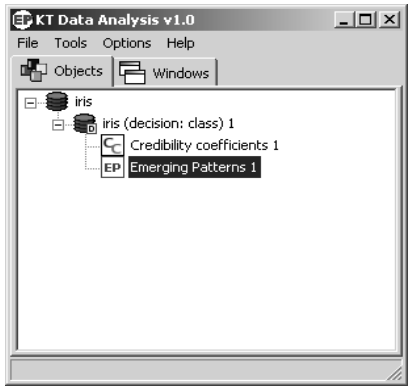


Fig. 1. Main window of the KTDA system under MS Windows. Object view contains: dataset object (*iris*), decision system object (*iris (decision: class) 1*), credibility coefficients object (*Credibility coefficients 1*) and Emerging Patterns object (*Emerging Patterns 1*)

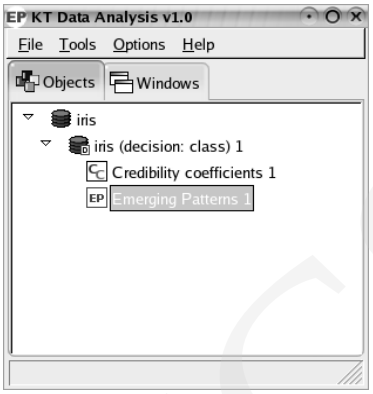


Fig. 2. Main window of the KTDA system under Fedora Core (Linux) with GNOME window manager. Object view contents as described for Fig. 1

The other functions in the main menu of the KTDA system cover experiments associated with the data credibility analysis. There are also some options which do not affect KTDA results anyhow. The *Windows* view plays only a supporting role and allows to bring up and down or closing other KTDA windows. Thus the most important element of the main window is the *Objects* view. It shows a hierarchical view of all objects created and processed during applying the KTDA system: opened dataset, defined decision systems (dataset with set decision attribute), discovered emerging patterns, computed credibility coefficients, CAEP classifiers and classification results.

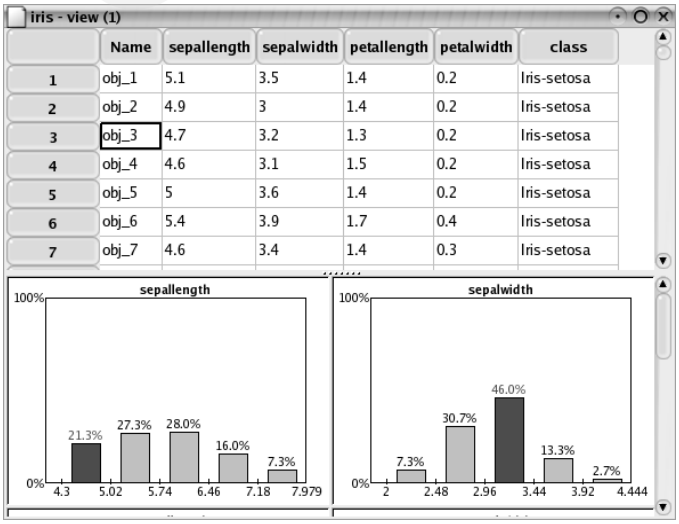


Fig. 3. Dataset view window. The bottom part of the window contains histograms showing the value distribution of individual attributes. The histogram bins related to a record chosen in the table are marked

Each of these object types has its own icons so the view is very comprehensible. All operations the user can accomplish on a given object are accessible from a context menu appearing after clicking the right button of the mouse while pointing to the object. For most object types the context menu contains the *View* and the *Properties* items. Choosing the *View* item the user opens a new, dedicated view window presenting information characteristic of the selected object. Depending on the type of the selected object view window provides an additional context-dependent functions. The examplary view windows are shown in Figs. 3, 4 and 5. Choosing *Properties* item from the context menu the user gets access to some detailed information on the given object.

4.1. Loading a dataset

To start working with the KTDA system one has to decide on a dataset to be analyzed. There are two possibilities: a dataset can be open from a file or generated by the KTDA system itself (the KTDA system supports two types of synthetic datasets). The second case is related mainly to performing comparative experiments, with artificial datasets having the required and known characteristics. The KTDA system can be used, for example, as a generator of datasets with multivariate Gaussian distribution. KTDA can read data files in the following formats: ARFF (*WEKA* program files) [11], CSV (compliant with spreadsheets like MS Excel), DATA (UCI Repository) [12] and TAB (*RSES 2* program files) [13]. This allows comparative studies with other classification results of many other systems as well as processing of already existing datasets.

Finally, the user must choose a decision attribute which will divide the loaded dataset into decision classes. The operation is commenced by choosing the *Create a decision system* option from the dataset context menu. In the KTDA system one can define many decision systems with different decision attributes which allows data analysis from many perspectives.

4.2. Discovering Emerging Patterns

Discovering EPs is available through *Discover Emerging Patterns By...* item from the decision system object context menu. There are two algorithms to choose: *Maximal frequent itemsets based algorithm* and *Decision tree based algorithm*. Selecting one brings up a particular configuration dialog. The algorithm based on maximal frequent itemsets requires the four parameters:

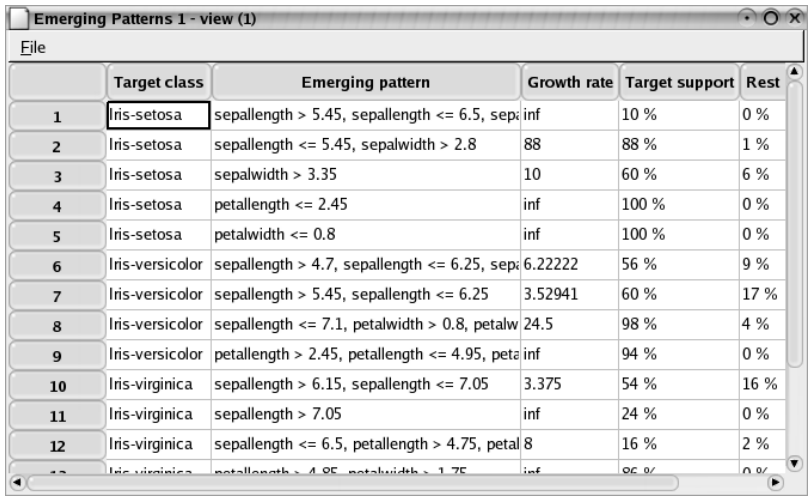
- *Minimal EP Growth Rate* – the growth rate threshold for mined EPs,
- *Minimal EP support in target class* – specifies an initial support threshold in EPs' target classes,
- *Minimal-EP-support increase per iteration* – specifies a support threshold increase per main algorithm iteration. Each iteration runs with the EP support threshold in target class calculated as the support threshold from

previous iteration (starting with the value equal to *Minimal EP support in target class*) increased by this parameter value. Smaller this parameter is, more iterations are performed and more EPs can be discovered,
- *Reduce discovered EPs* – a two-stage switch whether to reduce set of discovered EPs or not.

Default values of these parameters should give the best results with relatively short computing time for most cases. All parameters in the KTDA system can be set through comfortable and easy to use dialog windows.

The EP discovering algorithm based on a decision tree has a much simpler parameterization. Moreover, our experiments have shown that this algorithm is significantly insensitive to values of the parameters, so the default ones should be sufficient almost in every case. These parameters are as follows:
- *Split significance level* – determines a significance level used in checking the significance of splits considered during a decision trees constructing,
- *EP significance level* – a significance level used to test if EPs extracted from decision trees are statistically significant.

The user can examine discovered EPs with their growth rates and supports in target classes and in the rest of the dataset. EPs view window is shown in Fig. 4. The KTDA system allows exporting them to a CSV file (through menu *File* in the view window). By choosing *Create a CAEP* option in the context menu of EPs object one can obtain a CAEP object. It may be used to conduct a classification of dataset objects. It may be also used to compute credibility coefficients through the Voting Classifier Method. But the KTDA system provides much shorter and more practical way to do this. It is described in the next section.


Emerging Patterns 1 - view (1)

File

	Target class	Emerging pattern	Growth rate	Target support	Rest
1	Iris-setosa	sepallength > 5.45, sepallength <= 6.5, sepa	inf	10 %	0 %
2	Iris-setosa	sepallength <= 5.45, sepalwidth > 2.8	88	88 %	1 %
3	Iris-setosa	sepalwidth > 3.35	10	60 %	6 %
4	Iris-setosa	petallength <= 2.45	inf	100 %	0 %
5	Iris-setosa	petalwidth <= 0.8	inf	100 %	0 %
6	Iris-versicolor	sepallength > 4.7, sepallength <= 6.25, sepa	6.22222	56 %	9 %
7	Iris-versicolor	sepallength > 5.45, sepallength <= 6.25	3.52941	60 %	17 %
8	Iris-versicolor	sepallength <= 7.1, petalwidth > 0.8, petalw	24.5	98 %	4 %
9	Iris-versicolor	petallength > 2.45, petallength <= 4.95, peta	inf	94 %	0 %
10	Iris-virginica	sepallength > 6.15, sepallength <= 7.05	3.375	54 %	16 %
11	Iris-virginica	sepallength > 7.05	inf	24 %	0 %
12	Iris-virginica	sepallength <= 6.5, petallength > 4.75, petal	8	16 %	2 %

Fig. 4. View window for the discovered Emerging Patterns

4.3. Computing credibility coefficients

To calculate the credibility coefficients one simply chooses the *Compute credibility coefficients* item from a decision system context menu. Then there are two choices of algorithm to be used: *Voting classifier method based on CAEP classifier* or *Mutli credibility coefficient method*. In the first method there is one more dialog consisting of choosing EP discovering algorithm and configuration parameters of this algorithm. It was described in the previous section. Since in the current implementation of the KTDA system *Mutli credibility coefficient method* has a fixed configuration, in the second case there is nothing more to set up. After computations a new credibility coefficients object appears in the *Objects* view of the main window. The *View* menu item from the credibility coefficients object context menu launches a specialized view window (Fig. 5). Marking of records with the lowest values of credibility coefficients attracts attention of the user to the data requiring a special care and/or handling. There are two modes of record marking. The user can select marking of all records that have credibility coefficient values less or equal to a given threshold. The second option is marking a specified part of the dataset, consisting of records with the lowest credibility coefficients. Especially the latter mode seems to be useful as we would rather like to inspect some minor fraction of all records that are probably the most incredible.

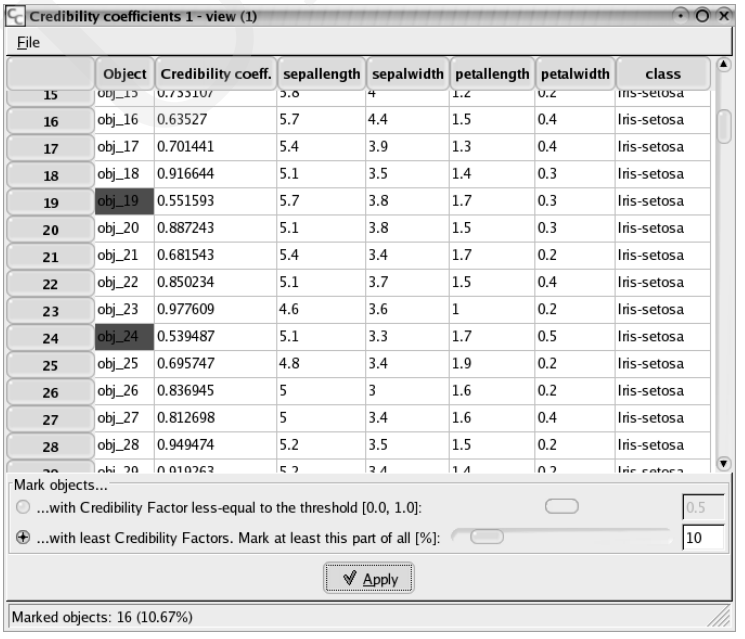


Fig. 5. Credibility coefficients view window. Among the visible ones records *19* and *24* are marked according to marking condition specified in the bottom panel of the window

As it was described above the whole process of data credibility analysis using the KTDA system is simple and easy and does not demand any sophisticated knowledge from the user. Even quite inexperienced user can process data with the KTDA system to recognize the records seeming to be not typical and having the lowest estimated credibility. The decision what to do with such records is up to the user.

4.4. Other Functions

The KTDA system has some more auxiliary functions. They help in carry out many experiments with discovering EPs, classification based on EPs and credibility coefficient calculation algorithms. They have been used in different research undertakings. For example, these auxiliary capabilities maintain adding some false, randomly generated records to the loaded dataset and checking whether they were properly identified by relatively low credibility coefficient values.

The KTDA system can be used as a data analysis system or as a research and educational tool. It supports performing the following automatic experiments:

- *classification experiment* – it can be carried out to observe how modifications of a given parameter of a particular EPs discovering algorithm influence the quality of classification accomplished by the CAEP classifier on a basis of the revealed patterns,
- *false object detection experiment* – it is purpose is to test how many generated false records are successfully identified by a certain credibility analysis method in respect to a number of false records inserted to a genuine dataset and parameters for the false record generator,
- *credibility coefficient and probability experiment* – it is performed to analyze correlations between the credibility coefficient values and the probability values for records of generated synthetic datasets, in which the probabilities are known. Such experiments are carried out to prove and/or assess correctness of algorithms for evaluation of the credibility coefficients – lower credibility coefficients should be assigned to less probable (more unusual) records.

These are *automatic* experiments, since each of them can be automatically repeated a required number of times and the results from all iterations are averaged to circumvent influence of random fluctuations caused by applying a pseudorandom number generator. Other (non-automatic) experiments require some planning and user assistance.

4.5. Technology

The entire KTDA system was implemented in the ISO C++ programming language [14] which benefited in high performance and source code level

portability. The portability is preserved even by the graphic user interface as it utilizes *wxWidgets* [10] library–a portable and open-source GUI toolkit. The system can be compiled on almost any platform that has a contemporary C++ compiler and the standard C/C++ library. Implementations of the KTDA system were run under MS Windows and Linux (Fedora Core 3) operating systems.

All tools and libraries needed to compile KTDA are free and open-source. By choosing Linux operating system and GCC compiler one obtains absolutely free and stable platform for using the KTDA system. Moreover the KTDA system has relatively low hardware requirements. For quite a long time it has been developed on a machine with only 64 MB of RAM and a CPU of 400 MHz.

5. Conclusions

The KTDA system general description and its capabilities were presented in the paper. The KTDA system is technologically advanced but easy to use and user-friendly tool for data analysis. Its fundamentals are based on emerging patterns concept, a relatively novel form of knowledge discovered in databases. Some introductory information on emerging patterns was submitted in Section 2. Two different algorithms for discovering emerging patterns were put into practice in the KTDA system. Comparative studies of these two approaches can be very beneficial for researchers and experts.

The KTDA system is also a tool for the data credibility analysis. The paper presents essentials of the research and briefly explains its target and a methodology of credibility coefficients. The KTDA system supports our two innovative algorithms for computing credibility coefficients: *Voting Classifier Method* and *Multi Credibility Coefficient Method*. The first former employs emerging patterns in generating the measure of credibility. The latter algorithm is much more general and applies cooperation of many credibility coefficient calculating methods to obtain better results of credibility coefficients. The KTDA system only partially utilizes its advantages as in a current version it supports only *Voting Classifier Method* with different parameterizations (EP discovering algorithm). We believe that *Multi Credibility Coefficient Method* used with a broader set of credibility coefficients computing algorithms will increase data credibility analysis quality.

The system can be employed to work with almost all data having a tabular form, for example stored in a CSV file. The presence of predefined decision attribute is not required as the program allows to define one temporarily. The ability to define many different decision attributes enables to discover emerging patterns related to different aspects of processed data. Although the medicine was the primary inspiration for data credibility analysis research the KTDA system is suitable not only for medical applications. It is universal and can be applied in almost every domain.

To evaluate advantages and drawbacks of the KTDA system fairly some more experience has to be gained. The perspectives are promising. The KTDA system is an interesting novelty in the field of data classification. The rules inferred from the dataset can be supplemented by the exceptions identified by credibility assessment tools. Experiment-oriented bias of the KDTA system makes it attractive for research and educational purposes.

References

[1] Dong G., Li J., *Efficient Mining of Emerging Patterns: Discovering Trends and Differences,* Proceedings of the SIGKDD (5th ACM International Conference on Knowledge Discovery and Data Mining), San Diego, USA, (1999) 43.

[2] Podraza R., Ryszkowski P., Podraza W., *Ignoring improper data in decision support system for medical applications,* Annales UMCS Informatica, AI 2 (2004) 163.

[3] Boulesteix A., Tutz G., Strimmer K., *A CART-based approach to discover emerging patterns in microarray data,* Bioinformatics, Oxford University Press, (19)18 (2003) 2465.

[4] Dong G., Zhang X., Wong L., Li J., *CAEP: Classification by Aggregating Emerging Patterns*, Proceedings of 2nd International Conference on Discovery Science, Tokyo, Japan, (1999) 30.

[5] Podraza R., Jurkowski A., *Coefficient of Credibility in Rough Set System,* The 22nd IASTED International Conference on Artificial Intelligence and Applications (AIA), Innsbruck, Austria, (2004) 776.

[6] Podraza R., Walkiewicz M., Dominik A., Credibility Coefficients in ARES Rough Set Exploration System, Rough Sets, Fuzzy Sets, Data Mining and Granular Computing 10th International Conference (RSFDGrC), Regina, Canada, (2005) 29.

[7] Gouda K., Zaki M.J., *Efficiently Mining Maximal Frequent Itemsets,* ICDM, San Jose, (2001) 163.

[8] Shafer J., Agrawal R., Mehta M., *SPRINT: A Scalable Classifier for Data Mining,* Proceedings of the 22nd VLDB Conference, Bombay, India, (1996) 544.

[9] Weisstein E.W., *MathWorld−A Wolfram Web Resource.* CRC Press LLC, 1999. Wolfram Research Inc. 1999-2004. http://mathworld.wolfram.com (2004).

[10] Smart J., Roebling R., Zeitlin V., Dunn R., et al: *wxWidgets – a portable C++ GUI toolkit* (2005) http:://www.wxwidgets.org

[11] *WEKA – Waikato Environment for Knowledge Analysis*, Department of Computer Science, University of Waikato, New Zeland, (2004) http://www.cs.waikato.ac.nz/ml/weka

[12] Blake C.L., Merz, C.J., *UCI Repository of machine learning databases*, Irvine, University of California, Department of Information and Computer Science (1998), http://www.ics.uci.edu/~mlearn/MLRepository.html

[13] *RSES2-Rough Set Exploration System*, Institute of Mathematics, Warsaw University, (2004) http://logic.mimuw.edu.pl/~rses

[14] Stroustrup B., *The C++ Programming Language*, Third Edition, Addison-Wesley, (1997).