JOURNAL OF APPLIED
COMPUTER SCIENCE
Vol. 21 No. 1 (2013), pp. 39-52

Application of FussionClassify for Data Classification

Agnieszka Duraj[1]

[1]*Institute of Information Technology,*
Lodz University of Technology
ul. Wólczanska 215, 90-924 Lodz, Poland
agnieszka.duraj@p.lodz.pl

Abstract. *In the published articles and works there are solutions regarding data fusion. However, there is not any verification as for the efficiency of classifiers in the case of many sources of data given simultaneously. It is, seemingly, a very significant problem to be considered in the case of e.g. data fusion in intelligent traffic control. The intention of the author is to prepare the tools for classification of data which come from various sources. They can be sets (data files) prepared by the user of application, but they can also be one (or many) sets from the UCI machine learning repository (http://archive.ics.uci.edu/ml/).*

Keywords: *data classification, k-NN algorithm, naive bayes algorithm, decision tree, data fussion.*

1. Introduction

Every cognitive process related to information processing is an (inborn) ability of live organisms acquired in a very natural manner. Using all the senses (sight, hearing, taste, touch) enables not only processing of the data which is analyzed by the brain, but also gives the opportunity for a more effective concluding and making particular decisions.

In the traditional decision support system, or expert systems, the decision is made on the basis of one type of information, most often gathered by an expert in one data base. Strong dispersion of the data causes the desire to use the information from multiple sources and only then to designate a particular (global) decision.

The process of simultaneous use of data – information coming from many sources is known in literature as data fusion or information fusion. Other known terms are aggregation, integration, consolidation or amalgamation. Data fusion is a continuous cognitive process occurring in a natural way in the human brain. Initially it was used for military purposes (e.g. steering unmanned vehicles, automatic identification, automatic target recognition). At present it is more and more often used in technological, medical and economic solutions. Simultaneous processing of information from multiple sources is not a new skill, idea or concept; however, in the light of constantly appearing new technologies, accessibility to the Internet, GPS systems, cellular telephone technology, and real time systems, the concept of data fusion must be considered in many new aspects.

The data fusion methods are used in many domains (e.g. radar systems, air traffic control, robot steering systems, systems of observation and forecasting weather, natural science, medicine etc.), and are realized both in the hardware and software scopes. They are based on such knowledge domains as digital signal and image processing, control theories, pattern recognition, neural networks, fuzzy logic, artificial intelligence. The role of artificial intelligence are presented in e.g. [1].

In the published articles and works there are solutions regarding data fusion. However, there is not any verification as for the efficiency of classifiers in the case of many sources of data given simultaneously. It is, seemingly, a very significant problem to be considered in the case of e.g. data fusion in intelligent traffic control. The intention of the author is to prepare the tools for classification of data which come from various sources. They can be sets (data files) prepared by the user of application, but they can also be one (or many) sets from the UCI machine learning repository (http://archive.ics.uci.edu/ml/). The recipient of the presented idea/solution may be broadly understood group of university employees and students, since this tool may be used both in the laboratory exercises in the domain of data mining and artificial intelligence and the conducted research activity. There are many descriptions of algorithms of data fusion and classification algorithms. However, the available tools include usually one of the methods. No methods including both of the aforementioned methods (fusion and classification) are known to the author.

The article has been divided into several sections. In point 2 there are basic terms and definitions related to the topic of the research and the developed application, namely the process of data fusion and the concepts of classification are defined. In point 3 the model and the structure of the developed application are presented, and the classification algorithms (k-NN, and Naive Bayes algorithms, decision tree) implemented in FussionClasify are shown. Point 4 summarizes and presents conclusions.

2. Basic terms and definitions

2.1. Definitions – the concept of classification

The problem of classification can be defined as concluding on unknown qualities on the basis of the known qualities of a given object. The classifying algorithm, which may be described as a certain form of "recipe" for resolving the membership in a given class, is constructed in an inductive way, because a certain set of examples – a training set – is used. The term 'classification' may be also defined as a method of data analysis, the goal of which is prediction of the values of a given attribute on the basis of a certain set of training data. It encompasses not only the method of discovering models (so-called classifiers) or functions describing the relations between the properties of objects and their assigned classification, but also the models of classification which are used for classifying new objects of unknown classification.

Classification is usually understood as multistage division into classes, subclasses or groups according to a predetermined criterion. In the further considerations of the article the following definitions of the space of classification (Definition 1) and the issue of classification (Definition 2).

Definition 1 *The space of classification (the space of objects of classification or simply the space) is the Cartesian product of a finite number of sets* $X = X_1 x X_2 x X_3 ... x X_n$, *each of which is a set of real numbers or a certain finite set of objects.* X_i *are called the qualities or attributes of the space.*

Definition 2 *The problem (or issue) of classification is any finite subset of the set* XxC, *where* X *is the space of classification and* C *is a certain finite set of objects* *called the set of class labels (or set of classes) for this classification problem. For each object $t = (x, c) \epsilon (XxC)$, the value c is called the class of the object t and is designated as $C(t)$.*

The methods of classification of objects are researched and developed in many fields of science, e.g. in signal recognition (recognition of signals of speech, EKG, EEG etc.).

2.2. Definitions of data fusion

As shown in the introduction, data fusion is commonly used in many domains. However, there is no unambiguous and precise definition of the term in the literature. It is probably caused by the variety of domains and fields of knowledge in which fusion domain is present. And so for example in [2], [3] the term fusion was specified for military purposes (Definition 3), and was made to formulate the fusion process (outside military uses) as a concept combining various mathematical methods and techniques (Definition 4).

Definition 3 *Data fusion is a multistage, multilateral process dealing with automatic detection, association, correlation and combination of data and information from one or many sources in order to establish a precise position, route and identity of an object (and its importance and significance).*

Definition 4 *Data fusion is the formal framework in which there are expressed techniques and tools developed in order to combine original data from many sources (e.g sensors, databases, human knowledge) in such a way that the outcoming decision or action taken are better e.g. considering precision, accuracy in quality and quantity.*

In [4, 5] there is an expression of the fusion process based on constructed structures (not based on tools or methods, as it was defined in (Definition 3) and (Definition 4). The authors in (Definition 5) emphasize that the goal of data fusion is obtaining the so-called "better quality". This quality is not precisely defined by the authors, probably due to variety of domains, because depending on the specific use it may mean accurate measurement, efficiency in classification, or more accurate noise detection, etc.

Definition 5 *Data fusion is the formal framework in which there are expressed means and tools allowing for combining data coming from many sources, and its goal is "better quality".*

The variety of definitions of data fusion – the fusion of information – which can be found in literature on the subject, as well as innovative technologies appearing on the market (GPS, cellular telephones etc.), show, according to the author, the constant need for considering the notion in many new aspects. Data fusion, as the process of data processing, automatically becomes information fusion used in the decision support systems, expert systems as knowledge base etc. At the same time it can be said that it is an interdisciplinary term, because it is based on such fields of knowledge as digital signal and image processing, control theories, pattern recognition, neural networks, fuzzy logic, artificial intelligence.

3. FussionClassify - the description of the developed tool

The tool for classifying prepared by the author uses many data files (*.txt, *.csv) generated by uses of applications, it is also possible to use the files from the UCI machine learning repository.

3.1. The model of data fusion process

The variety of applications for the process of data fusion led to the development of six main models. These are Cycle, Boyd Control, JDL, Omnibus, Waterfall and Dasarathy. Intelligence Cycle and JDL are the most popular and the most often modified models.

- Intelligence cycle [6] – the cycle of this model begins with gathering data (hardware and human) and generates “raw data”, which, in the next stage, are initially processed (providing e.g. reports, graphs). Information fusion is done in the next stage – the so called analysis block. Here the detailed process of analysis of data, gathering specific information and detecting deficiencies in data takes place. The last stage is relying information. Depending on use it can be recommendation for further data gathering, but it can also be a decision taken, transfer of the gathered data to an expert system, etc. The characteristic feature of this model is the possibility of returning to the sources of data.

- JDL – developed by [3], it is a very popular model of data fusion process used often with multiple modifications (see e.g.[7]). It is based on 5 levels (0-4). Level 0, which deals with initial processing of input data (organizing, normalizing, compressing, etc.) does not belong to the fusion domain.

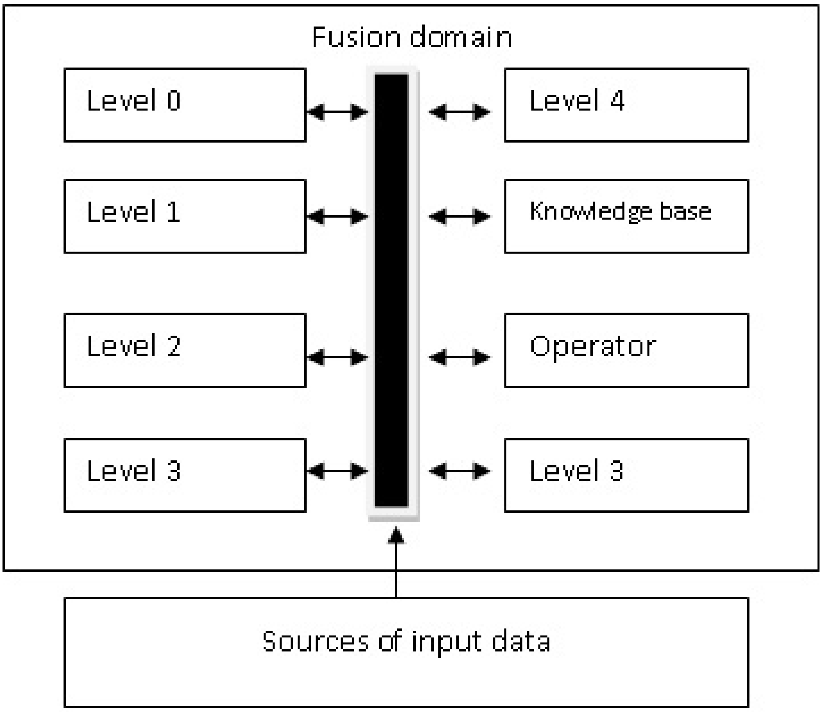

Figure 1. Modification of the JDL model of data fusion process in FussionClassify

> Level 1 iteratively puts the data together, bringing them to one space, dividing according to a given object, checking additional parameters of a given object and defines its identity. Therefore, in the first level the organizing, associating, following and identifying process takes place. Level 2 refers to a complete evaluation of the analyzed phenomenon or object. Therefore, a detailed analysis and interpretation of data is required. Heuristic-based and formal techniques are used. Level 3 uses for analysis all the data gathered in level 2 and from knowledge bases. Level 4 deals with controlling tracks and managing all the stages of the data fusion model.

The application FussionClassify is used a modification of the JDL model of data fusion process. The diagram of modification model JDL is presented in Fig. 1. (See also [8]).

3.2. The architecture of the application

In the case of data fusion process we deal with three types of architecture essential for this process, which define the physical structure of the system. These are the autonomous architecture, centralized architecture and hybrid architecture. Each of the types has both advantages and disadvantages. The advantages of autonomous architecture are:

- flexibility in the selection of the number and type of sensors from which the data come (various files formats) without interference in the fusion algorithm,
- optimization of signal,
- processing of easy addition of another data source (a new sensor).

In the case of centralized architecture we deal with unprocessed or minimally processed data. In this architecture, the central processor (central fusion processor) performs the tasks of organizing, association of data, initial processing. Undoubtedly, the advantage of centralized architecture is the possibility of performing accurate detection and classification already on level 0 of the data fusion model (e.g. on the sensor level). The disadvantage of this architecture is the necessity of transferring big amounts of data to the system directly to the central processor and simultaneous processing of the data in real time. It requires also big computing power. Hybrid architecture is a combination of the two aforementioned types of architecture. The centralized architecture is supplemented with certain characteristic elements and signal processing algorithms, taking into consideration each file (sensor) separately. Hybrid architecture is very flexible, depending on the application of this architecture elements of centralized fusion (centralized architecture) may dominate, in other cases the elements of distributed architecture become more important. Additionally, while applying this type of architecture in the data fusion process, a module responsible for monitoring the whole fusion process should appear. The disadvantage is the hardware and software complexity and big requirements referring to data transmission. Because the application FussionClassify is designed mainly for education, it will not receive data directly from sensors. We do not take into consideration elements related to transmission of big amounts of data. For academic considerations, also online processing is not taken into consideration. Considering the designation of the developed application for educational and academic purposes, out of the discussed briefly above types of architecture

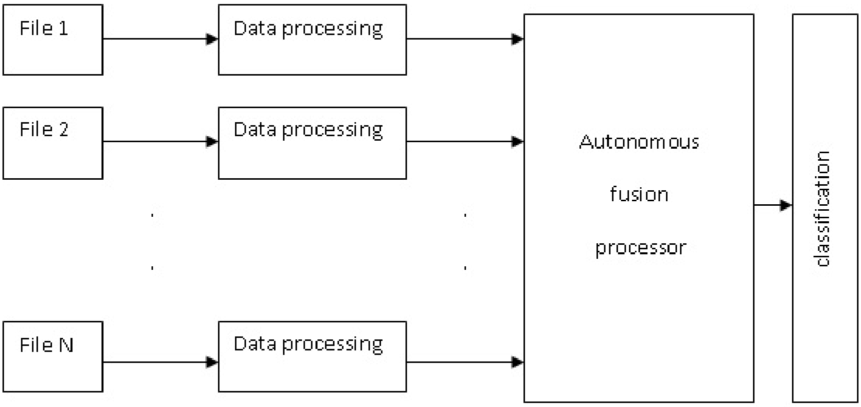

Figure 2. The architecture of data fusion process in FussionClassify

used in data fusion processes, the distributed (autonomous) architecture was chosen. Additionally, in the developed application the autonomous architecture has been modified. It was assumed, that the purpose of FussionClassify is classification of objects, additionally it is not possible to ensure that data is complete, therefore the modification of the distributed architecture was most optimal. The diagram of the architecture is presented in Fig. 2.

4. Utilized classification algorithms

In the literature one can find many different types of models of classification [1, 9, 10, 11], e.g. k-nearest neighbor algorithm (k-NN), Bayes classifiers, neural networks, Metaheuristics (genetic algorithms), approximate sets, decision trees, Support Vector Machine (SVM), etc. The developed application FussionClassify assumes implementation of three most popular classifiers, namely the k-NN, Naive Bayes and decision tree. These algorithms are discussed briefly below. At the same time, the project assumes adding other classifiers.

4.1. K - nearest neighbor method

One of the simplest classification methods is the k-nearest neighbor algorithm [12]. It is based on a simple principle, namely: “data model belongs to this class, to which belong most of its k neighboring models". Therefore, the classification of a model means designating k samples from the training data set, located the closest in the context of the used metric (e.g. Euclidean distance, Manhattan or Hamming distance). Next, the class to which the analyzed sample of data belongs is determined. Here one of the two methods is used:

- Majority Voting – method of voting on equal rights – the sample belongs to the class, to which belong most of k neighbors
- Inverse Distance Voting – method of voting taking into consideration distance – for each of the classes among the found k neighbors the sum of inverse distances from the analyzed sample is calculated. This sample is classified to this given class, for which the calculated sum is the biggest.

4.2. The Naive Bayes algorithm

In many data classification systems linear and non-linear methods of mathematical programming are used. It is assumed that searching for optimal solutions can be limited to the so-called acceptable base solutions, e.g. simplex algorithm. The Support Vector Machines models are also becoming more and more popular. In practical applications simple methods, to which the naive Bayes algorithm belongs, are used most often [13, 14]. It is naive because it assumes independence of changeable variables in a given classification space. It definitely simplifies computation of probabilities and assigning to classes. This algorithm is most often used for discrete data; in the case of continuous space data must be digitized.

4.3. Decision tree

In addition to the algorithms mentioned above, also decision trees play a big role in classification of data. They are very popular because of their numerous advantages, such as:

- representation of complex notions,
- computational efficiency,
- clear representation of the tree,
- easiness in creating rulestent.

In the literature on the subject we can find substantial number of decision tree algorithms and their modification. In the course of conducted research, their drawbacks and faults were recognized. And so, for example, the decision trees created according to ID3 algorithm do not work in case of incomplete data and noise, which means that in ID3 there is lack of tolerance to noise and interference. Troubles also appear in case of complicated problems.

In this case ID3 creates a very big and thus unclear tree. In the ID3 algorithm in case of multivalued values the used information gain measure leads to forming a flat and very wide tree. Quinlan minimized the above problem of the ID3 algorithm by implementing gain ratio in the C4.5 algorithm. Breiman, on the other hand, in the CART (Classifiction And Regression Tree) algorithm, based on statistical analysis, used the Gini index as the measure of the best distribution.

5. Expreriments – own research

In the conducted research, own data sets and the sets coming from the UCI Machine Learning Repository were used. The experiments were to verify the chosen model and architecture in the processes of data fusion and also the discussed above algorithms of k-nearest neighbors, Bayes classifier and decision trees. In the works [15, 16] the influence of the choice of the distance measure on the value of precision and entropy was shown. The obtained results demonstrated that the cosine measure has a significant advantage over the Euclidean measure. The functioning of the k-nearest neighbors algorithm was examined for two methods of voting, namely the methods of voting on equal rights and methods of voting with taking distance into consideration. The results are shown in [17].

It has been noted, that the highest value of entropy was obtained for the k-nearest neighbors algorithm in the case of first own data set analyzed (Z1). Here the algorithm made flawless classification of new values. The k-NN algorithm correctly classified 101 objects to 'No' class, and 2 to 'Yes' class. Naive Bayes'a algorithm showed a significantly greater number of errors. Assigned of probabilities for classes of decision an $'No' = 0.019$ and $'Yes' = 0.013$. (See in Fig. 3, Fig. 4) The collections coming from UCI Repository Machine accurately classify the k-NN algorithm. For example for 'car' file k-NN well qualified 1310=vgood, 384=good, 65=unacc, 69=acc class. See Fig. 5 and Fig. 6.

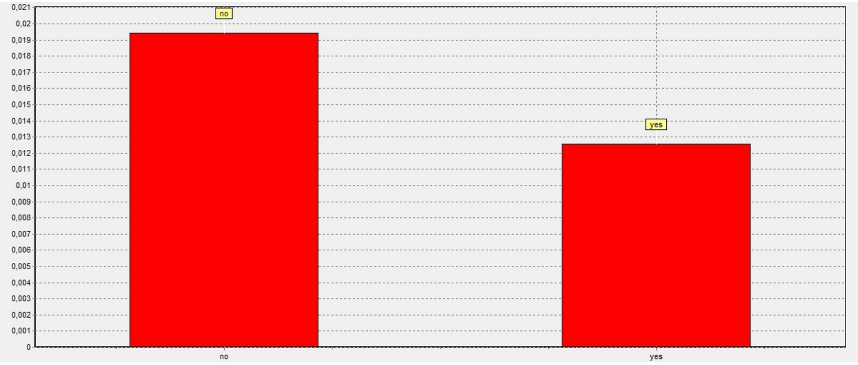

Figure 3. Result of k-NN algorythm for the Z1 file

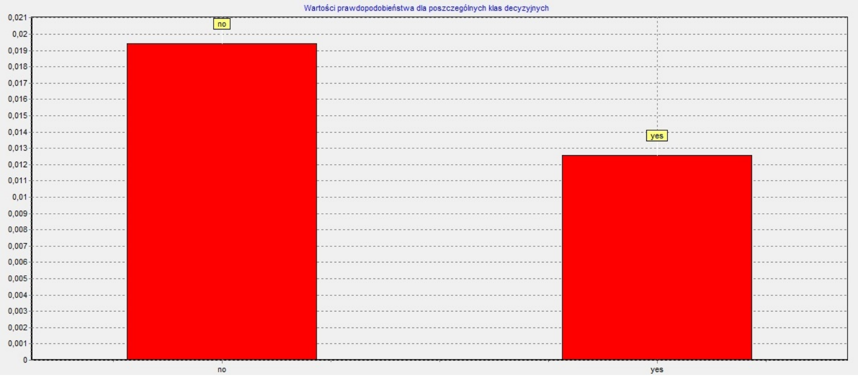

Figure 4. Result of Naive Bayes'a algorythm for the Z1 file

Only for the second set of self (Z2) the naive Bayes'a algorithm proved to be the best as far as classification was concerned. Decision trees did not demonstrate any mistakes during classification for the analyzed data sets and the implemented three criteria of the choice of tests in the form of Gini index, coefficient of information gain and entropy.

The result of classification in the application FussionClassify is presented in Fig. 3, Fig. 4, Fig. 5, Fig. 6.

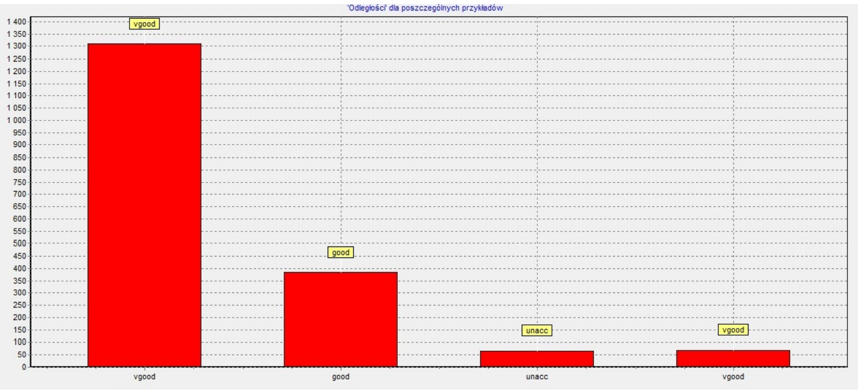

Figure 5. Result of k-NN algorythm for the CAR file in UCI Repository

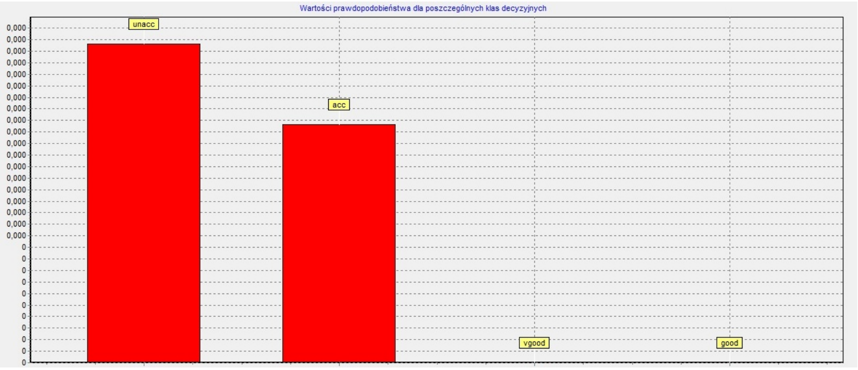

Figure 6. Result of Naive Bayes'a algorythm for the CAR file in UCI Repository

6. Conclusions and summary

In this paper the basic terms of data fusion and classification have been given. Moreover, the models of data fusion process and used architectures (autonomic, centralized, hybrid) have been presented. The advantages and disadvantages of each of them have been indicated. And although undoubtedly the flexibility of hybrid architecture is its significant and key advantage, the autonomous (distributed) architecture, which was proposed in the developed application FussionClassify, proved to be simpler and less demanding, and thus more useful for educational purposes. The implemented classification algorithms (k-NN, naive Bayes, decision tree) supplement the users' knowledge on the issue of data classification. The experiments showed that the popular algorithms employed in the research are good classifiers in the prepared data sets. Research must be continued increasing the number of attributes and the number of classes.

References

[1] Szczepaniak, P. and Tadeusiewicz, R., *The Role of Artificial Intelligence, Knowledge and Wisdom in Automatic Image Understanding*, Journal of Applied Computer Science, Vol. 18, No. 1, 2010, pp. 75–85.

[2] Hall, D., *Mathematical Techniques In Multisensor Data Fusion*, Artech House, London, 1992.

[3] Waltz, E. and Llinas, J., *Multisensor Data Fusion*, Artech House London, 1990.

[4] Wald, L., *A European Proposal for Terms of References In Data Fusion*, International Archives of Photogrammetry and Remote Sensing, Vol. 32, No. 7, 1998, pp. 651–654.

[5] Wald, L., *Definitions and Terms of Reference In Data Fusion*, International Archives of Photogrammetry and Remote Sensing, Vol. 32, No. 7, 1998, pp. 651–654.

[6] Bedworth, M. and O'Brien, J., *The Omnibus Model, A new Model of Data Fusion*, IEEE Aerospace and Electronic systems Magazine, Vol. 15, No. 1, 2000, pp. 30–36.

[7] Klein, L. A., *Sensor Data Fusion Concept and Applications*, SPIE Washington, 1999.

[8] Hall, D., *The implementation of Data Fusion System*, In: Multisensor Fusion, edited by W. E. Hyder, A.K., Vol. 70, Kluwer Academic Pulisher, 2002, pp. 419–433.

[9] Niewiadomy, D., P. A., *Implementation of MFCC vector generation in classification context*, Journal of Applied Computer Science, Vol. 16, No. 2, 2008, pp. 55–65.

[10] Khorissi, N.-E., Mellit, A., Guessoum, A., and Mesaouer, A., *GA-Based Feed-Forward Neural Network For Image Classification: Application For the Grains of Pollen*, Journal of Applied Computer Science, Vol. 17, No. 2, 2009, pp. 83–96.

[11] Sulkowski, G., T. M. W. K., *Implementation of the Hardware Packet Classification System*, Journal of Applied Computer Science, Vol. 17, No. 2, 2009, pp. 97–111.

[12] Larose, D., *Metody i modele eksploracji danych*, PWN Warszawa, 2012.

[13] Nozer, S., *Reliability and risk : a Bayesian Perspective*, John WileySons, 2007.

[14] Pourret, O., Nam, P., and Marcot, B., *Bayesian networks : a practical guide to applications*, John WileySons, 2008.

[15] Duraj, A. and Krawczyk, A., *Finding outliers for large medical datasets*, Electrical Review, Vol. 86, No. 12/2010, 2010, pp. 188–191.

[16] Duraj, A. and Krawczyk, A., *Dobór miar odległości w hierarchicznych aglomeracyjnych metodach wykrywania wyjątków*, Electrical Review, Vol. 87, No. 12b, 2011, pp. 33–37.

[17] Duraj, A. and Krawczyk, A., *Outliers Detection of signals in biomedical information systems fusion*, Electrical Review, Vol. 20, No. 12b, 2012, pp. 128–131.