© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Data mining and soft computing

Ö. Ciftcioglu
Faculty of Architecture, Delft University of Technology
Delft, The Netherlands

Abstract

The term data mining refers to information elicitation. On the other hand, soft computing deals with information processing. If these two key properties can be combined in a constructive way, then this formation can effectively be used for knowledge discovery in large databases. Referring to this synergetic combination, the basic merits of data mining and soft computing paradigms are pointed out and novel data mining implementation coupled to a soft computing approach for knowledge discovery is presented. Knowledge modeling by machine learning together with the computer experiments is described and the effectiveness of the machine learning approach employed is demonstrated.

1 Introduction

Although the concept of data mining can be defined in several ways, it is clear enough to understand that it is related to information extraction especially in large databases. The methods of data mining are essentially borrowed from exact sciences among which statistics play the essential role. By contrast, soft computing is essentially used for information processing by employing methods, which are capable to deal with imprecision and uncertainty especially needed in ill-defined problem areas. In the former case, the outcomes are exact within the error bounds estimated. In the latter, the outcomes are approximate and in some cases they may be interpreted as outcomes from an intelligent behavior. From these basic properties, it may be concluded that, both paradigms have their own merits and by observing these merits synergistically these paradigms can be used in a complementary way for knowledge discovery in databases. In this context, in the following two sections the properties of data mining and machine learning paradigms are pointed out. In section four the methods of soft computing are given in perspective. A novel data mining-soft computing application for

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

knowledge discovery together with the experimental details of a case study are presented in section five, that is followed by conclusions.

2 Data mining

Data mining deals with search for description especially in large databases. Basically, this search is meant for certain information elicitation with the presence of others. More explicitly, data mining refers to variety of techniques to extract information by identifying relationships and global patterns that exist in large databases. Such information is mostly obscured among the vast amount of data. Since the data around the information of interest is considered to be irrelevant, they are effectively termed to be noise, which are practically desired to be absent. In this noisy environment, there are many diverse methods, which explain relationships or identify patterns in the data set. Among these methods reference may be made to cluster analysis, discriminant analysis, regression analysis, principal component analysis, hypothesis testing, modeling and so forth.

The majority of the data mining methods fall in the category of statistics. What makes the data mining methods different is the source and amount of the data. Namely, the source is a database, which supposedly has a big amount of relevant and irrelevant data in suitable and/or unsuitable form for the intended information to be extracted. Referring to the large size of the data set, the conventional statistical approaches may not be conclusive. This is because of the complexity of the data where only few is known about the properties so that it can neither be treated in a statistical framework nor in a deterministic modeling framework, for instance. A large data set from a stock market is a good example where apparently there is no established physical process behind so that the properties and behavior can only be modeled in a non-parametric form. The validation of such a model is subject to elaborated investigations.

The main feature of a data set subject to mining is complexity and the characteristic feature of the data mining methods distinguishes themselves by "learning" as to the conventional statistical methods. That is, even the well-known statistical or non-statistical or alternatively parametric or non-parametric methods are used, the final model parameters or feature vectors of concern are established partly or fully by learning. Note that, in conventional statistical modeling for relationship or pattern identification, model or pattern parameters are established by statistical computation in contrast with learning in data mining exercise. Although statistical techniques are apparently ubiquitous in data mining, data mining should not be carried out with statistics unless this is justified. Statistical methods assist the user in preparation the data for mining. This assistance might be in the form of data reduction and hypothesis forming. Such a preparation is especially beneficial for knowledge discovery by soft computing following the information extraction by data mining. In this work, learning in soft computing is accomplished by machine learning methods.

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

3 Machine learning

In data mining, the functional relationships present in a data set are essentially determined without any mathematical model and therefore they are computational models, which are non-parametric. Both study and computer modeling of the learning processes is the subject matter of the area called machine learning. Generally, a machine-learning system uses entire observations of its environment in the form of a finite set at once. This set is called training set and contains examples, observations coded in some form suitable for machine reading.

Data mining to establish the suitable training set should precede machine learning. If the data from the database is directly used for machine learning this becomes data mining which may be quite ineffective since the data in databases are not accumulated with the consideration of probable use for machine learning. Rather, they are chosen to meet the needs of applications. Therefore, properties or attributes that would ease the learning task are not necessarily present in the databases.

4 Soft computing

Soft computing possesses a variety of special methodologies that work synergistically and provides "intelligent" information processing capability, which leads to knowledge discovery. Here the involvement of the methodologies with learning is the key feature deemed to be essential for intelligence. In the traditional computing the prime attention is on precision accuracy and certainty. By contrast, in soft computing imprecision, uncertainty, approximate reasoning, and partial truth are essential concepts in computation.

These features make the computation soft, which is similar to the neuronal computation in human brain with remarkable ability so that it yields similar outcomes but in much simpler form for decision-making in a pragmatic way. Among the soft computing methodologies, the outstanding paradigms are neural networks, fuzzy logic and genetic algorithms. In general, both neural networks and fuzzy systems are dynamic, parallel processing systems that establish the input output relationships or identify patterns as prime desiderata which are searched for knowledge discovery in databases. By contrast, the GA paradigm uses search methodology in a multidimensional space for some optimization tasks. This search may be motivated by a functional relationship or pattern identification, as is the case with neural network and fuzzy logic systems.

4.1 Neural networks

Artificial Neural networks (ANN) consist of a fixed topology of artificial neurons connected by links in a predefined manner. They can perform information processing in a manner similar to a biological brain where the neuronal network, like biological network, is inherently nonlinear, highly parallel, robust and fault tolerant. The links represent synaptic weights and they

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

are adapted during learning. ANN can easily handle imprecise, fuzzy, noisy and probabilistic information and can generalize it to unknown tasks.

Due to powerful learning algorithms available and multi-layered structure a neural network is flexible enough to handle any data-mining task for all practical purposes. From this viewpoint, ANN can be seen as the most convenient soft-computing paradigm for data mining. The relationships present in the data set are represented as multivariable functional approximation where neurons play the role of nonlinear models with global or local nonlinearity. With the global nonlinearity, a neural network is widely regarded as a black box that it reveals little about its predictions but it provides ease in applications to data-mining.

4.2 Fuzzy logic

From the data-mining viewpoint fuzzy logic refers to numerical computations based on fuzzy rules, for the purpose of modeling functional relationships in a data set. It also encompasses fuzzy set-based methods for approximate reasoning. With this property fuzzy logic by contrast to neural networks, is a transparent box with its predictions. Also it allows linguistic and qualitative interpretations of the outputs.

For data-mining especially Sugeno-type fuzzy systems [1] is concern where the system approximates any functional relationship with a combination of several linear systems, by decomposing the whole input space into several partial fuzzy spaces known as fuzzy sets and representing each output space with a linear equation as local models. The fuzzy sets are described by functions called fuzzy membership functions as local models, which allow convenient application of powerful learning techniques for their identification from data.

To obtain fuzzy models from the data set, different approaches can be used, like fuzzy relational modeling [2], and product space clustering [3,4]. In a multidimensional variable space, fuzzy modeling is accomplished by means of one-dimensional fuzzy sets obtained from the multidimensional fuzzy sets by projections onto space of each input variable [5]. The complexity of fuzzy modeling exponentially increases with the dimension of the fuzzy product space, which is known as curse of dimensionality. This may be considered as the price paid in return to its desirable properties as being transparent, i.e., tracing possibility for reasoning on fuzzy model predictions.

4.3 Genetic algorithms

Genetic algorithms (GAs) are stochastic combinatorial optimization techniques inspired by the mechanism of natural evolution and genetics [6,7] involving the mutation and cross-over of chromosomes. The feature of GA is the ability of parallel search of the state space for optimization in contrast with the point-by-point search by the conventional optimization. GAs belong to a large class of meta-heuristics that are used to avoid local minima in heuristic search methods. They try to improve several solutions to a problem in parallel, mixing random jumps and crossing. With these peculiarities, it provides a paradigm for data-

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

mining with a consortium of appealing related methodologies like genetic search, tabu search, simulated annealing and so forth that are collectively referred to as *evolutionary algorithms*.

4.4 Neuro-fuzzy systems

Referring to different merits of both ANN and fuzzy logic systems for data mining, a judicious integration of the merits provides more effectiveness in applications. This incorporates the generic advantages of ANN like extensive parallel processing, robustness, and learning in a massive data environment and the generic advantages of fuzzy logic systems like modeling of imprecise, qualitative knowledge and transparent modeling of data. For example, learning fuzzy rules from data by means of data samples [8] is a good example. From data mining viewpoint, both ANN and fuzzy logic systems establish relationships that are determined as multivariable functional approximation. They learn from experience with sample data. From the neuro-fuzzy viewpoint, ANN and fuzzy logic systems can be seen as equivalent [9] by appropriate interpretation of the elements of both systems.

An important mathematical modeling for this equivalence is the approximation by means of radial basis functions. The model can be seen as an ANN where neuronal non-linear function is a gaussian, which is the most popular radial basis function in use. Such a network is known as radial basis functions network (RBFN) and it plays an important role in joint neural and fuzzy (neuro-fuzzy)computing. This is briefly considered below.

4.5 Radial basis functions network

An approximation to a function f(x) by radial basis function network (RBFN) is carried out by

$$f(x) = \sum_{i=1}^{M} w_i \phi(\| x - c_i \|) \tag{1}$$

where w_i are weights; M is the number of basis functions; $\phi(.)$ is the basis function based on the Euclidean distance metric

$$\| x - c_j \|^2 = (x - c_j)^T (x - c_j) \tag{2}$$

so that

$$\phi = \exp\{- \| x - c_j \|^2 / \sigma_j^2\} \tag{3}$$

where σ_j is the j-th width parameter that determines the effective support of the j-th basis function. After normalisation of the weights, the gaussian basis function play the role of membership functions in fuzzy systems. Therefore it is not difficult to see that RBFN exhibits key properties for neuro-fuzzy computing enjoying the key advantages of both systems. Generally, multilayer perceptron neural networks use sigmoid function causing global approximation features in function approximation. By contrast, gaussian functions in RBFN provide local approximation features. Therefore, an RBFN does need to use relatively more

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

artificial neurons for the same approximation task, while it maintains its local approximation properties.

The selection of the centers of the basis functions can be determined by means of advanced clustering algorithms available. From the data-mining viewpoint, such clustering forms the basic data mining procedures of knowledge discovery in databases. After the clustering task has been accomplished, the determination of the final network parameters establishes the knowledge model i.e., explicit relationships among the model variables as a part of the knowledge discovery. For data mining, the selection of the cluster centers among the data samples is important for the following reason.

In general a cluster center could give enough confidence about its validity of being representative enough only if such a center ever exists among the data samples. This is in contrast with a mathematical abstraction where such a center might appear due to mathematical clustering convenience and it may never belongs to the reality consistent with the actual data set. For clusters, which are to select among the data samples, first the clusters should be ranked according to their importance in the model. Thereafter appropriate number of high-rank clusters can be retained. The ranking process is done through a [NxN] dimensional matrix, which contains the outputs of the basis functions; that is N basis functions for N samples. Such a matrix is conventionally known as RBF matrix. However, for data mining applications, N is expectedly large so that conventional methods, orthogonal least squares [10] for instance, used for the ranking clusters from the RBF matrix are not efficient enough due to curse of dimensionality.

5 Soft computing for knowledge discovery: Case study

A case study using soft computing for knowledge discovery is carried out. In this study, an orthogonal transformation method, namely, principal component analysis (PCA) is applied to RBF matrix denoted here by G to eliminate complexity problem of RBFN for data mining applications. The elimination is due to the availability of efficient PCA algorithms, which can deal with large amount of data. The PCA decomposition of G, which is given by

$$G = Q \Lambda Q^T \tag{4}$$

where Q is a matrix ($Q^T=Q^{-1}$) whose columns are the eigenvectors of RBF matrix G and Q; Λ is the eigenvalue matrix. From the preceding matrix equation, we obtain

$$G[Q \Lambda^{-1}] = Q \tag{5}$$

This implies that the data set used for clustering is effectively the scaled projections of the RBF matrix G on the principal axes. The gradation (or ranking according to the influence) of the clusters where the clusters are among the data samples is accomplished as follows [11]. Due to the orthogonalisation process the energy delivery from each artificial neuron (or node) to the output is independent of the rest of the nodes. Initially energy delivery of each node to the

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

output is computed for each node using N samples and the node with the most energy delivery is selected as the first node of this initial network. The same procedure applied again for this network whose first node is already determined.

In this new computation among the remaining nodes, the node providing most energy delivery to the output with the presence of the first node is selected as second node. This procedure is repeated until all the nodes are considered and graded according to their influence on the function approximation where the influence is determined by their relative energy contribution to form the function. During this influence gradation scheme the columns of Q are interchanged so that the columns of eigenvectors follows the ranked sequence of the clusters, i.e., data samples.

In this method, the PCA forms the data mining process identifying the principal components of the data and the graded sequence of the influential variables. We consider this as information extraction. The knowledge discovery is accomplished by processing this information using soft computing methods. Note that, in this application the machine learning method is different than those conventionally used for ANN training. For data mining applications the training algorithms based on gradient computation can be trapped at local minima and therefore the convergence, that is training, is not guaranteed. However, in the PCA based training above, the training is obtained by a straightforward computation together with local approximation properties.

For this research, use is made of an actual data set from a design research. The actual design data set, which is also used in another data mining research [12] comprises 45 design variables and the total data samples are 196 [13]. However, since the amount of data 196 might be considered as small for a data mining exercise, a synthetic data set of 1000 samples with the same statistical properties as the actual data set is prepared by mathematical treatment as described by Fukunaga [14]

$$X = (\Lambda^{1/2} Q^T)^{-1} Y = Q \Lambda^{1/2} Y \tag{6}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & .... & 0 \\ & \lambda_2 & \\ ..... & .... & .... \\ 0 & & \lambda_N \end{bmatrix} \tag{7}$$

is a diagonal matrix of eigenvalues of the actual data; Y is the data set created which consist of independent and identically distributed variables each with the variance 1 where all samples are normally distributed; Q is the eigenvector matrix of the actual data. The number of variables is 45 as is the case with the actual design data and number of samples created is 1000 so that Y has the dimensions of [45x1000]. With the transformation (7) above, the data set X with the dimensions of [45x1000] has the same covariance matrix as the actual data set indicating the same correlation among the variables. To train an RBF network with 43 inputs and 2 outputs with 1000 training samples while the local approximation features are maintained is a heavy task for conventional machine learning methods.

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

The synthetic data set is subjected to data mining and knowledge discovery by the orthogonal transformation method presented above. The total execution time for modeling (knowledge discovery) is approximately 10 min. The conservative estimated computation time for well known OLS method which uses Gram-Schmidt orthogonal transformation for orthogonal decomposition of RBF matrix G in place of its principal components, is more than 200 times than that of the present execution time. However, due to probable numerical instabilities expected in the computations, even the due training result is not guaranteed in this case. It should be pointed that, since the complexity of the network increase exponentially by the number of data samples, the estimation of the execution time given above cannot be determined precisely. The training results are shown in figures 1-3, where figure 3 is obtained by zooming in figure 1.

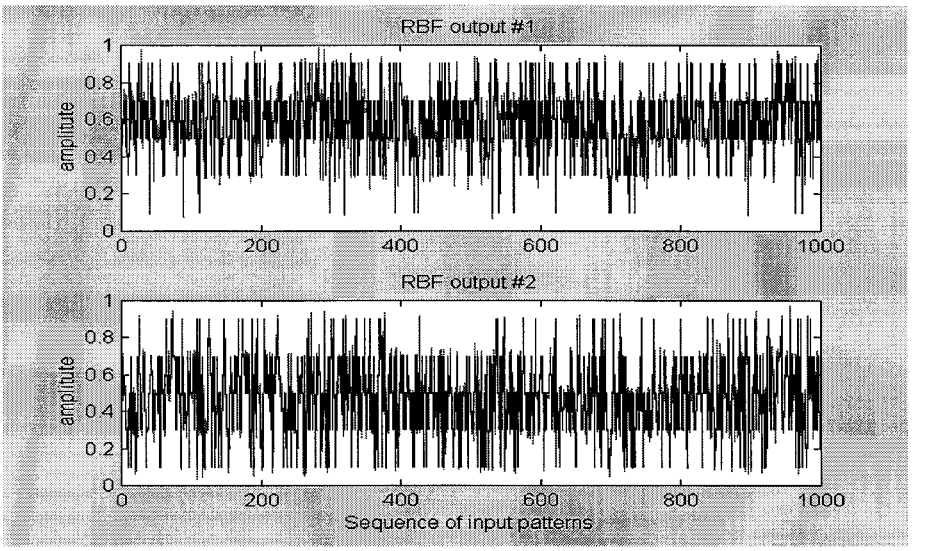


Figure 1: Machine learning results from the data set with 1000 data samples: The model predictions and the data set learned together. They almost overlap.

Since the data set originally stems from normally distributed identical and independent samples, more intensity of the samples around 0.5 is rather noticeable. Here it is important to note that, the learning process is accomplished efficiently as well as effectively for all knowledge discovery purposes, at large. The knowledge model having been obtained, this model can be used for meta-knowledge acquisition by means of advanced methods like sensitivity analysis [12], for instance.

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

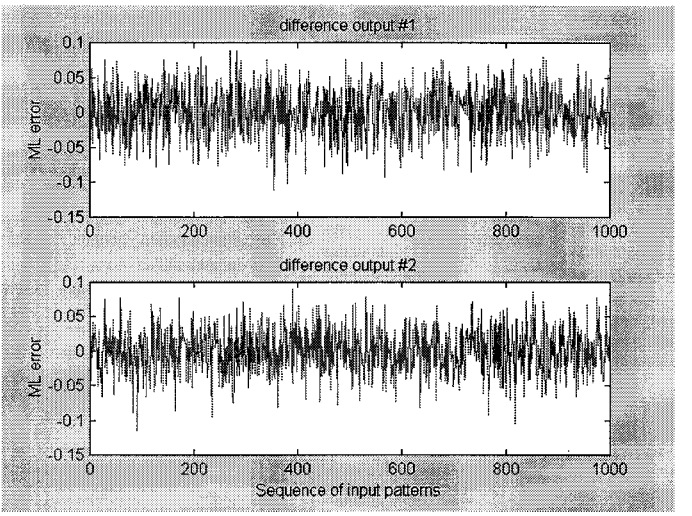


Figure 2: Machine learning results from the data set with 1000 data samples: The differences between the model prediction and the data set learned for outputs 1 and 2.

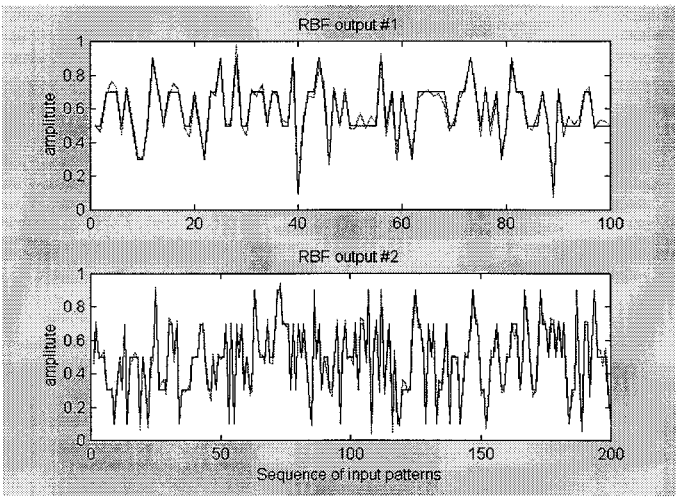


Figure 3: Machine learning results from the data set as to model predictions and the data set learned: First 100 samples for model output #1 and the data together (upper); the first 200 samples for model output #2 and the data together (lower) (s. Fig.1).

6 Conclusions

The aim of this presentation is to demonstrate the effectiveness of a novel application of an orthogonal transformation method using principal components

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

for data mining and knowledge discovery. This is accomplished by machine learning in a framework of soft computing methodology with a neuro-fuzzy system approach described.

References

[1] Takagi T. and Sugeno M., Fuzzy identification of systems and its application to modeling and control, *IEEE Trans. Sys., Man and Cybern.* vol. SMC-15, pp.116-132, 1985
[2] Pedrycz W., *Fuzzy Sets Engineering*, Boca Raton, FL: CRC, 1995
[3] Nie J. H. and Lee T. H., Rule based modeling: Fast construction and optimal manipulation, *IEEE Trans. Syst., Man, Cybern. A,* vol.26, pp.728-738, 1996
[4] Wolkenhauer O., *Data Engineering*, John Wiley & Sons, New York, 2001
[5] R. Kruse, J. G and Klawonn F., *Foundation of Fuzzy Systems*, Wiley, New York, 1994
[6] Holland J. H., *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI, 1975
[7] Goldberg D. E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Publishing Co. Inc., Reading, MA, 1989
[8] Wang L. X. and Mendel J. M., Generating fuzzy rules by learning from examples, *IEEE Trans. Syst., Man and, Cybern.*, vol. 22, pp. 1414-1427, 1992
[9] Jang, J.S.R. and Sun C. T., Fuctional equivalence between radial basis function networks and fuzzy inference systems, *IEEE Trans. Neural Networks*, vol.4, pp.156-159,1994
[10] Chen S., Billings S. A. and Luo W., "Orthogonal least squares methods and their application to non-linear system identification", *Int. J. Control*, vol.50, no.5, pp. 1873-1896, 1989
[11] Ciftcioglu O., Studies on the complexity reduction with orthogonal transformation, *Proc. World Congress on Computational Intelligence WCCI-2002, IEEE International Conference on Fuzzy Systems May 12-17,* Honolulu, pp.1476-1481, 2002
[12] Ciftcioglu, O, Model validation by sensitivity analysis for data mining, *Data Mining-2002*, Brebbia C. A. (Ed.), WIT Press, Southampton, UK, 2002
[13] Durmisevic S., *Perception Aspects in Underground Space by Means of Intelligent Knowledge Modeling*, PhD Thesis, Delft University of Technology, May 2002
[14] Fukunaga K., *Introduction to Statistical Pattern Recognition*, Academic Press, Inc, Boston, 2nd ed. 1990