



Data Mining and Classification Techniques

Daniela Gotseva, Roumen Trifonov, Vasil Angelov
Computer Systems Department
Technical University of Sofia
Sofia, 1000, 8 Kl. Ohridski Blvd
Bulgaria

Abstract: The processed data in the world is increasing. Using different data mining techniques we can extract different patterns which can be used for discovering of tendencies in the data. This can be achieved with classification via methods from the artificial intelligence and machine learning. The result from the classification algorithm can be analyzed with different statistical methods and by performing this, the generated model can be evaluated.

Keywords: data mining, artificial intelligence, classification, algorithms, data analysis.

I. Introduction

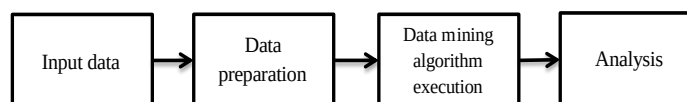
We are overwhelmed with constantly over-increasing volumes of electronically stored data. Regardless of the data origin (from science experiments, forecasts, statistics), there are patterns, describing certain trends. The job of the data scientist is to discover those patterns in order to make sense of them in particular filed.

Data mining is the automatic or semiautomatic computing process of identifying patterns in data. The researches in the field of Data Mining are based on achievements from artificial intelligence, machine learning, statistics and databases. The differentiation of Data Mining as individual and interdisciplinary field is based on both the integration on the aforementioned achievements and the development of the databases technologies for data management. The Data Mining analysis is based on unknown properties represented as “hidden meanings”. This leads to the idea that the raw data contains deeper layers of hidden knowledge which can only be identified via detailed in-depth study.

The input data consists of multiple examples in appropriate format. The output contains hypothesis, based on those examples using specific algorithm. There are fundamental classes of tasks that the algorithms of Data Mining can perform: association; chaining or sequencing; classification; clustering; forecasting. [1][2]

Data Mining is about solving problems by analyzing data already present in databases. The process can be summarized in Figure 1. It includes data preparation (in order to transform the data in the correct format), executing Data Mining algorithm (which identifies the useful information) and the analysis phase (reviews the statistical information after the completion of the algorithm).

Figure 1 Data mining process



The Data Mining methods can be applied in any of the following fields:

- Business – the already present data in the data warehouses can be used as input for Data Mining analysis. The goal is to find hidden patterns and tendencies. The online sales alone reach the amount of 100 million per day. The classification of the customers in categories will lead to optimal interaction between the companies and the clients and with this – the increasing of the company’s revenue;
- Science and Engineering – bioinformatics, genetics, medicine, education and electrical engineering are only a small portion of applicable fields of Data Mining. The human genetics is focused on the relationship between the DNA sequence and the susceptibility of the organism to diseases. The aim is to determine how the mutation of the DNA is correlating to the risk of specific disease. This will improve the ways of diagnosing, prevention and treatment of the disease. In the electrical engineering the Data Mining is applied at the observation of high voltage installations, in order to gather valuable information of the technical parameters of the equipment (the state of the insulators for example). In the education the goal is to determine the factors shaping the student’s behavior;
- Geographical Information Systems (GIS) – wireless sensory networks can be used for improvement of data gathering for GIS (air pollution for example). The gathered data has very closed values and there is excess

of data. By measuring the correlation between the data values from different sensors, can be defined new classes highly-specialized DM algorithms;

- Pattern recognition – includes already existing patterns in the data volumes. In that context the pattern is associative rule. Patterns with data anomalies can also be discovered (for terrorist activities).

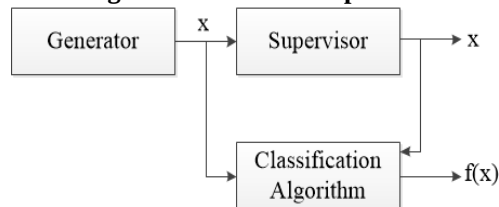
These are the main areas of applications of the Data Mining techniques.

II. Classification

Object classification is a form of data analysis used to extract important patterns for description of classes of data and prediction of future tendencies in data variation.

The classification is operation of dividing different objects in separate classes. Those classes can be defined by business rules, limits and mathematical dependencies. The classification process belongs to the Supervised Learning models. The aim of the classification is to be generated common model of the alternation of dependent attribute in correlation with one or several attribute-predictor.

Figure 2 Classification process

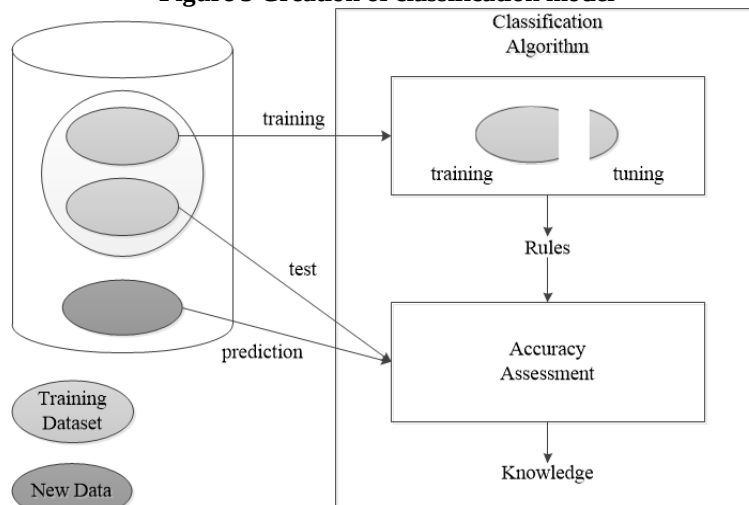


The classification is the most disseminated task in the Data Mining analysis. It includes allocation of studied objects, phenomena and processes in types and grouping of the data in order of degree of similarity.

The completion of classification task incorporates m examples, described as (x_i, y_i) , $i \in M$, where $x_i \in R^n$ is a vector of n predicting variables for the i th example and $y_i \in H = \{v_1, v_2, \dots, v_H\}$ is the class of the targeted variable. For example in classification with binary variable $H = 2$, we have two classes that can be referenced as $H = \{0, 1\}$ or $H = \{-1, 1\}$.

Let us assume that F is a class of functions $f(x): R^n \rightarrow H$, called hypotheses and representing the hypothetical relations between x_i and y_i . The classification task requires and includes the definition of appropriate hypothetical set F and algorithm A_F . This algorithm is used to discover the function $f^* \in F$, which describes the optimal relation between the predicting variables and the predictable variable (the class). The probability distribution $P_{x,y}(x, y)$ of the examples in the dataset D , defined in the set $R^n \times H$, is usually unknown and the most part of the classification models are not parametrized, because they do not make preliminary assumptions for the distribution $P_{x,y}(x, y)$. The process in Figure 2 includes three components – Generator, Supervisor and Classification algorithm.

Figure 3 Creation of classification model



The generator extracts random vectors x in correspondence with the probability distribution $P_x(x)$. For each vector x the supervisor returns the class according to the conditional probability $P_{y|x}(y | x)$, which is also unknown/undefined. The classification algorithm A_F , also called classifier, selects function $f^* \in F$ in the hypothetical space to minimize the loss function.

Part of the examples in the dataset D are used for learning – for extracting the functional connection between the targeted variable and the input variables, represented by the hypothesis $f^* \in F$. The rest of the data is used to

evaluate the accuracy of the generated model and to select the best model in the set of the models, generated by different classification algorithms.

The creation of classification model on Figure 3 includes three main phases:

- Training – the classification algorithm is applied on the examples of the subset T, which are part from the dataset D. The goal is to extract classification rules for determining the class y for each observation x;
- Test – the rules, generated during the training phase, are used for classifying the examples V in the dataset D, not included in the subset T ($V=D-T$). Each example from the test dataset is compared to the class, predicted from the classifier. The training and the test datasets have to be independent – examples from T should not be include in V, in order to achieve realistic evaluation;

Prediction – includes the usage of the classification model for determining the class of new examples added in the future. The prediction includes the applying training phase rules to the input variables, describing the new examples.

III. Classification Methods

The classification method can be divided into several separated groups: decision trees, generator of covering rules, K-nearest neighbours and neural networks.

A. Decision trees

The decision tree is a diagram with tree-like structure. Each internal node represents test procedure for the given attribute. Each branch represents the test result and the leaf nodes represent the searched class. The initial node is called root node.

For each node of the tree is chosen optimal rule for object division for new nodes. The rule includes selection of attribute for division and attributes values. There are different criteria for choosing appropriate division rule, differing in number of nodes and attributes. At each node are applied different stop criteria to determine if the tree should continue to grow or the node is a leaf node (this means that an appropriate classification class is found). At the end is applied criterion for tree pruning in order to avoid overfitting between the model and the training dataset. The pruning can be performed during the training phase (preliminary pruning) or after reaching the maximum size of the tree (subsequent or finishing pruning). The preliminary pruning can result in early tree growth which leads to not optimal decisions. This is the reason for the vast usage of finishing pruning. [3][4]

B. Generator of covering rules

These are rules that are valid only for a certain part of the objects. Most of the algorithms start with the discovery of most covering rule – this rule is valid for the highest number of objects. After that these objects are removed from the dataset and the algorithm is applied again in order to find the rule that includes the highest number of objects. This approach is applied until the dataset with objects for classification is empty. The generated rules divide the dataset on areas. The rules are in the form “If – Then”. [3]

C. K-nearest neighbours

This approach is also known as memory based reasoning or K-nearest clustering [3][5][6]. It uses conclusions, based on past experiences. Each object can be represented as a point in the multidimensional space via its vector. The algorithm accepts that the proximity in the objects can define the similarity in the vectors. This algorithm can be used for prediction by defining of the values of the variables, describing the objects. The steps of the algorithm are as follows:

- Defining the number K of the nearest objects, used for classification;
- Finding the nearest K nearby objects to the object of unknown class by using distance measuring function between the objects;
- Defining the class of the new object based on conclusion on the found K nearest neighbours.

The objects have attributes from different type – numeric and nominal. The distance between the objects is calculated using the following approach: first the distances between each one of the variables describing the objects (A and B) is calculated; then all the distances are combined in one summarized distance function from Table I, which represents the distance between the two objects

Table I Distance between two objects

<i>Absolute value of the difference</i>	<i>Square difference</i>	<i>Normalized absolute value</i>	<i>Absolute value of the difference between the standardized values*</i>
$ A - B $	$(A - B)^2$	$\frac{ A - B }{\max(A - B)}$	$\frac{(A-B)}{\sigma}$,  is standard deviation of the examined values

After the distances between the separated variables are calculated, they are united with one common distance function from Table II:

When the searched K nearest objects is found, the new class for the object is selected on “democratic principle”. Each of the neighbours “votes” for its class. The proportion of votes for each class corresponds to the probability of the object’s appurtenance to the class. The object is classified as the class with the most voted one.

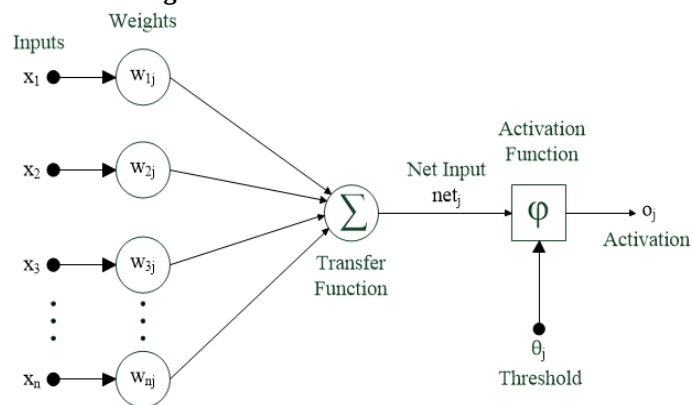
Table II Common Distance Function

Sum	Normalized sum	Euclidean distance
$d_{sum}(A, B) = \sum_{i=1}^n d_{var.i}(A, B)$	$d_{norm}(A, B) = \frac{d_{sum}(A, B)}{\max(d_{sum})}$	$d_{Euclid}(A, B) = \sqrt{\sum_{i=1}^n (d_{var.i}(A, B))^2}$

D. Neural networks

Originated from the field of machine learning, where they are used for mimicking the neural-physiological activity of the human brain, they consist of multiple elementary computing elements, called neurons, connected by weighted connections and organized in layers. The neuron j with threshold value θ_j receives input signals $x = [x_1, x_2, \dots, x_n]$ from the elements in the previous layer with which it is connected. Each signal is related/linked with weight $w_j = [w_{1j}, w_{2j}, \dots, w_{nj}]$, determining its importance. [3]

The neuron processes simultaneously the input signals, their weights and the threshold value via the combination or transfer function, which produces value, called potential or net input. Another function, called activation function, transforms the potential in output signal in correspondence with the threshold value θ_j .

Figure 4 Neural network architecture

The transfer function is usually linear, so therefore the potential is a weighted sum of the input signals, multiplied by the weights of the relevant connections. This sum is compared with the threshold value. The output signal y_j of neuron j is:

$$y_i = f(x, w_j) = f(P_j) = f\left(\sum_{i=0}^n (x_i w_{ij})\right) \quad (1)$$

One of the elements that have to be determined in the definition of the neural network is the activation function. Usually there are three types of used functions in Table III with additional constants α and β .

Table III Types of Used Functions.

Linear	Stepwise	Sigmoidal
$f(P_j) = \alpha + \beta P_j$	$f(P_j) = \begin{cases} \alpha, P_j \geq \theta_j \\ \beta, P_j < \theta_j \end{cases}$	$P_j = \frac{1}{1 + e^{-\alpha P_j}}$

The layers of the neural network are three types: input, output and hidden. The input layer receives information only from the external environment. This layer does not perform any calculations. It only transmits the information to the next layer. The output layer transmits the final results outside the system. Between these two layers are one or more hidden layers, used only for analytical purposes – to discover existing relations between the input and the output variables.

Under architecture of the neural network is understood its organization – number of layers, number of elements in each layer and the way the elements are connected

IV. Classification Approaches**A. Holdout method**

The holdout method includes separation of the dataset in two parts – one for training and another for testing the model. The objects in the training subset are chosen randomly and the remaining objects are placed in the testing subset. The volume of the training subset can vary from 1/2 to 2/3 of the total objects in the examined dataset. The accuracy of the classification algorithm depends on the test subset; therefore the accuracy can be overestimated or underestimated, related to the real prediction accuracy.

B. Repeated random sampling

In the repeated random sampling the procedure of dividing the general dataset in training and test subsets is executed multiple times. Each iteration includes two arbitrary chosen subsets with accuracy recalculation of the algorithm. At the end the total accuracy of the algorithm is the arithmetic mean value of the separated accuracy estimations

for the iterations. The number of the repetitions can be determined by preliminary application of different statistical techniques. This method presents more reliable accuracy in comparison with the holdout method, but it does not control the number of hits of the general dataset in the training or the test subsets. Therefore the variables can receive unusual values, called exceptions.

C. Cross-validation

The cross-validation is a variety of the repeated random sampling method. But this method guarantees uniform/equal participation of the objects from the examined dataset in the training subset and only one in the test subset. In the practice the most used method is the so called tenfold cross-validation. The general dataset is divided into ten subsets and the algorithm is executed ten times. Nine of the ten subsets are used for training and the remaining subset is used for testing. [3]

V. Classification Evaluation

A. Confusion matrix

The prediction of binary variables is the most common case of data mining tasks for classification. The presented explanations are for this case and the matrix is with two dimensions (2×2). When the predicted variable has N different values, the matrix is $N \times N$ dimensional. The classification model divides each object in one of two possible classes – for example Positive and Negative, with respectively True or False label. This leads to four possible classification classes for each object, represented in confusion matrix or contingency table.

Figure 5 Confusion matrix

		Predicted Class	
		Positive	Negative
Real Class	Positive	True Positive TP	False Negative FN
	Negative	False Positive FP	True Negative TN

The elements at the main diagonal contains the correctly predicted objects (TP and TN) and the remaining objects are the wrongly classified ones (FP and FN).

B. Classification accuracy and classification error

These measures are interdependent. The evaluation is executed by comparing the values from the test subset and the real value for each object. The accuracy is measured in percentage: $\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$ and the classification error, also measured in percentage, is calculated by: $\text{Error Rate} = \frac{FP+FN}{TP+FP+TN+FN}$. The TP rate shows what part of the objects in class Positive are recognized by the classification model. The FP rate determines what part of the incorrectly classified as Positive objects belongs to class Negative. All the rates are calculated by the formula in Table IV.

Table IV Calculation Formulas

TP Rate	FP Rate	TN Rate	FN Rate
$TP \text{ Rate} = \frac{TP}{TP+FN}$	$FP \text{ Rate} = \frac{FP}{FP+TN}$	$TN \text{ Rate} = \frac{TN}{TN+FP}$	$FN \text{ Rate} = \frac{FN}{FN+TP}$

The Precision measure is calculated by the formula: $\text{Precision} = \frac{TP}{TP+FP}$. The Recall measure is equivalent to TP rate: $\text{Recall} = \frac{TP}{TP+FN} = \text{TP Rate}$. The F-Measure combines Precision and Recall: $F - \text{Measure} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$.

C. Kappa Statistic

It measures the degree of consistency between the predicted and the real class. The value of this measure is in the interval [0; 1]. When the value is 1.0 there is full consistency between the predicted and the real classes.

D. Receiver operating characteristic (ROC) curve

Used for visual evaluation of the accuracy of the classification method. This is a two-dimensional diagram with the TP rate on the y-axis and the FP rate on the x-axis for several confusion matrixes for one classification model. The curve consists of all the points with coordinates (FP Rate, TP Rate) for the different values of the model's parameters. The area above the ROC curve is used for evaluation of the classification model. The values are between

0 and 1. Bigger area denotes better classification model. If ROC Area < 0.05 , then the evaluated model is worse than the random classifier (which is with ROC Area = 0.05) [3].

VI. Conclusion

The data mining is a rapidly evolving and broadly applicable field. Many business organizations use data mining techniques for data analyses and business intelligence, such as customer trends.

The represented groups of analytics methods evaluate the generated model. By applying different algorithms to specific datasets and by evaluating the received models we can choose the best algorithm for specific area of application.

VII. References

- [1] Fayyad, U., P. Smyth (1996). From Data Mining to Knowledge Discovery in Databases, AI Magazine, Fall Edition of 1996, USA, pp. 44-50.
- [2] Sumathi, S., S. Sivanandam (2008). Introduction to Data Mining and its Applications, Springer, USA, pp. 195-209.
- [3] Witten, L., E. Frank (2011). Data Mining – Practical Machine Learning Tools and Techniques, 3rd Edition, USA, pp. 86-250, 469-472.
- [4] Quinlan, R., (1993). C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, USA, pp. 71-78.
- [5] Duda, R., P. Hart, D. Stork (1996). Pattern Classification and Scene Analysis, New York, USA, pp. 15-16.
- [6] Smola, A., S. Vishwanathan, (2010). Introduction to Machine Learning, Cambridge University Press, Cambridge, UK, pp. 24-36.