International Journal of Advanced Research in IT and Engineering ISSN: 2278-6244

AN OVERVIEW ON DATA MODELS FOR KNOWLEDGE DISCOVERY FROM DATABASES

Zohreh Mousavinasab*

Hedayat Bahadori*

Abstract: *In the process of data mining, the first step is to build models so that more time is allocated to the mining project. In this chapter, some of the algorithms and models used for analysis and data mining have been studied. Most models are varieties of search algorithms in computer science and statistical literature while they are used for a particular implementation.*

*Department of Computer Engineering, Omidiyeh Branch, Islamic Azad University, Omidiyeh, Iran

1- DATA MINING DEFINITION

Considering many cases not previously known, there is a need to discover methods that can figure out the essential parts of mining. The proposed methods were studied in this regard.

1-1- Summary of data image processing

Before collecting data for mining and developing an appropriate predictive model, the data must be well known. To start, some parameters such as mean, standard deviation, etc should be calculated.

Means of data imaging and graphing are useful for understanding the data. These tools are useful when preparing the data. For example, when using these tools, large distributions of data can be viewed in a graph. The Defected data rate can be approximately guessed. Analyses typically have many features that are closely associated. Multidimensional associations of these parameters should be displayed in two dimensions. Experts are needed to handle the operation, if feasible. This is the main problem with this tool.

1-2- Clustering

The goal of clustering is to divide the data into several groups. The data should be divided into different groups based on their major differences. Groups of data should be very similar.

Unlike categorizing in clustering, groups are not known in advance. The characteristic basis of each group is also unknown. Therefore after the groups have been divided, an expert should analyze these groups.

Sometimes it is necessary to delete some parameters which turn out irrelevant or less important from the clustering index when analyzing the clusters [1]. After the groups have been divided up in a logical and justifiable way, information regarding the clusters can be studied. New clusters can even be formed. The most important algorithms that can be used for clustering is called algorithm Kohnen and the k-mean.

Figure 1 is a client dataset including two attributes of age and income. The clustering algorithm divided the datasets into three groups based on these two characteristics. Cluster 1 included young clients with low income, cluster 2 contained middle-aged clients with an average income and the third cluster contained elderly clients with a rather low income.

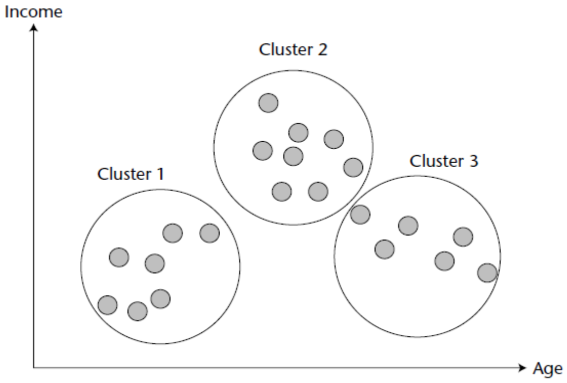

Figure 1: Clustering

1-3- Link Analysis

Data analysis is one method to describe the input and the relationships among the data values in the database. The most important method in link analysis is association discovery and order discovery.

In association discovery, rules that happen simultaneously are found, e. g. goods which are most likely purchased together at the same time. Order discovery is very similar, but the time parameter is also involved. Dependencies B are shown as A (A→B). “A” is prior and “B” comes subsequently, e. g. when a rule is as follows:

"If people buy a hammer, they will also buy a nail."

The first rule is to buy a hammer and a nail is also purchased as a result [2].

2- DATA PREDICTION MODELS

Previous data is used to construct a model of observed behavior. When the models consist of an existing input, the outcome of future behavior can be predicted. Some of the most popular data-mining prediction models are presented.

2-1- Grouping

The aim of classification is to identify different characteristics and include them in different groups. This model can be used on the existing data in order to predict new cases of behavior.

Data-mining can establish already-made classifications for the classification models. It can also make an inductive predictive model. These items may be obtained from a historical database [3].

2-2- Regression

Regression uses available values to predict other values. In the simplest form, standard statistical techniques such as linear regression are used. Unfortunately, many world problems are not based on simple linear regression values. Complex techniques (logic regression, decision trees and neural networks) may be needed to make accurate predictions.

The same models are used in both regression and classification. For example, regression decision tree algorithms and classifications (CART) are used when constructing classification trees and regression trees. Neural networks can be used for both as well [3].

2-3- Time series

Time series forecasting, predicts the unknown future values of a time series variable predictors and like in regression it uses its outcome to guide the prediction. Models must consider the characteristics of different time periods and their hierarchy.

2-4 models and algorithms for data mining

Many of these data mining products use algorithms. Usually, each of which has a particular strength. To use one of them professional guidance is needed to select the most reasonable product. These algorithms and models there are not necessarily better than each other and they should be chosen based on their efficiency and the kind of data used.

The neural network consists of an input layer. Each node in this layer is equal to one of the predictor variables. Nodes in the middle layer connect to nodes are connected to a number of nodes located in the hidden layer. Each input node is connected to all nodes located in the hidden layer. Nodes in the hidden layer can connect to nodes located in different hidden layers or to the output layer. Output layers consist of one or more output variables [4].

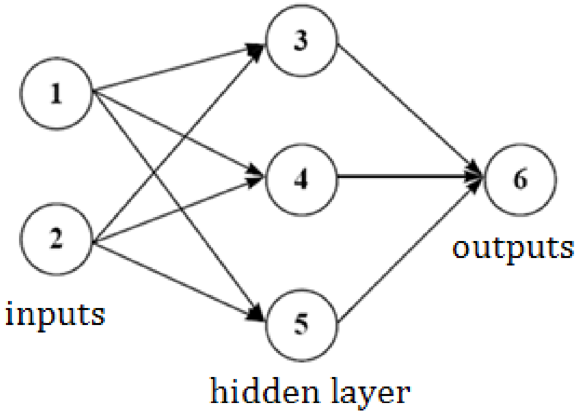

Figure 2: A neural network with one hidden layer

Each link between nodes x, y has weight which is shown by W_{xy}. Their weight is used in the calculation of the intermediate layers. Each node in the middle layers (except for the first layer) contains different input locations from different edges. Each has a specific weight. Each node in the input layer of each respective edge weight is multiplied and the results are added together. Then a predetermined function (activation function) is applied on the result and is considered as the output node of the next layer. The Weights of edges are unknown parameters of the system which are determined by a training method.

The Number of nodes, number of hidden layers and nodes and the way they are connected to each other, can be determined in neural network architecture (topology). Neural network software must determine the number of nodes, number of hidden layers, activation functions and constraints related to the weights of the edges.

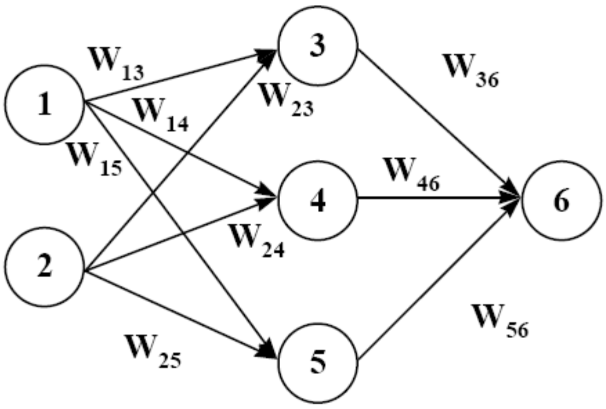

Figure 3: Weighted Neural Networks

Feed-Forward Back Propagation is one of the major types of neural networks described briefly as follows:

Feed-Forward means the amount of output parameter is determined based on input parameters and a number of primary weights. Inputs are combined and used in the hidden layers. Amounts of hidden layers are also mixed in order to calculate outputs.

In the back propagation, output error is computed by the comparison of the amount of output with the amount considered in the test data. These amounts are used for network correction and change of weights of edges. This trend starts from the output node and is calculated backward. This function is reiterated for each existing record in the databank. Each time this algorithm is performed for all the existing data in the bank is called a period. Periods continue until the amount of error does not change.

Since the number of parameters in neural networks is numerous, calculations of these networks could be time-consuming. However, if these networks are implemented for sufficient time, they will be usually successful. Another possible problem is excessive fitting. It means networks only work well on the training data while they are not suitable for other datasets. In order to overcome this problem, time of stop for network training must be determined. One of the ways is to operate the network on the test data besides on the training data repeatedly and assess the trend of error change in them. If the amount of error is increasing in the data, training is stopped even though error in the test data is constantly decreasing.

Since parameters of neural networks are numerous, a particular output can be created using different sets of parameters. Consequently, such parameters as weights of edges cannot be interpreted and have no special meaning. One of the most important benefits of networks is the capability of operating them on parallel computers.

2-5- Decision trees

Decision trees are known as a method for demonstrating a series of rules that lead to a class or amount. For example, an intention for dividing loan applicants into owners of credit risk is either good or bad. Figure 4-3 displays all the basic components of a decision tree solving this problem. These components include: decision node, branches and leaves [1].

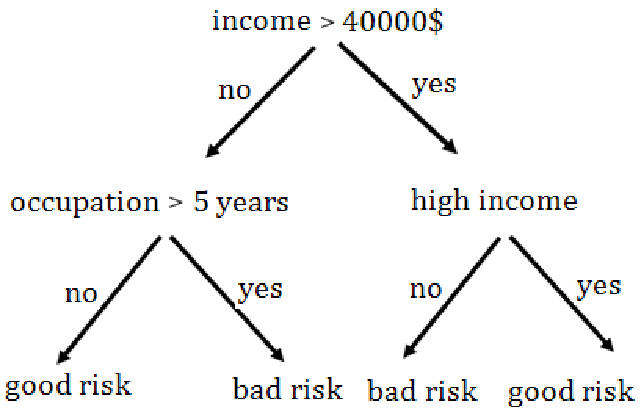

Figure 4: Decision tree

Based on algorithm, it is possible to have two or more branches. For example, CART creates trees with only two branches in each node. Each branch leads to another decision or leave node. By measuring a decision tree from the root to the leaves, we will reach a class as an amount attributed to a sample. Each node uses data of an item for making a decision about that branch.

Decision trees are built by sequential separating of data into distinct groups. The aim of this process is to increase the distance between nodes in each process of separation. One of the differences between various methods of making a decision tree is the way this distance is measured. Decision trees used for the prediction of clustering variables are called classification trees because samples are placed in the clusters or classes. Decision trees used for the prediction of continuous variables are called regression trees.

Each route towards the leaves in a decision tree is usually understandable. Hence, one of the main benefits of a decision tree is the ability to explain its own predictions. Nevertheless, this clarity might be rather misleading. For instance, hard separations in decision trees indicate a high precision which is less frequently manifested in reality. For example, why should a person with an income of 40,001 is good based on credit risk while a person with an income of 40,000 is bad? In addition, since several trees can show similar data with equal precision, what interpretation could be derived from the rules?

Decision trees pass through data for fewer times (once at most for each level of tree) and work very well with many predicting variables. As a result, models that are built rapidly make them fully suitable for datasets. If a tree is allowed to grow without any restriction, more time of building is spent which is not intelligently. Another important problem is that data is excessively fitting. Sizes of trees can be controlled by stopping rules. An ordinary rule for stopping is limiting the depth of tree growth.

Another method for stopping is truncating a tree. A tree can spread up to its ultimate size. Then, it grows to the smallest size without losing precision by the use of discovery methods or intervention of its user.

A usual flaw of a decision tree is that they perform division based on a greedy algorithm. Decision-making about which variable is used in the process of division disregards the effect of this division on the future divisions.

Furthermore, algorithms used for divisions are usually single variable i.e. they consider only one variable each time. Although this is one of the reasons for making these models, it makes recognition of the relation between predicting variables much harder.

2-6- Rule induction

Rule induction is a method for creating a set of rules that classify items. Although decision trees can make a series of rules, methods of rule induction create a set of independent rules that do not necessarily create a tree. Since the rule inducer has no compulsion for division on each level and can feed forward, it can find different than sometimes better patterns for classification. Despite the trees, the created rules may not cover all the possible items. Also, the rules may be involved in a contradictory prediction. A rule must be followed in each case. A method for solving these contractions is employing a degree of reliability for each rule and utilizing a rule with higher reliability [1].

2-7- K-Nearest Neighbor

Individuals usually refer to similar solutions previously taken when attempting to solve new problems. K-Nearest Neighbor is a clustering technique that uses a copy of this method. In this method, decision-making about within which cluster a new item is placed is conducted by the assessment of a number (k) of most identical items or neighbors. The number of items is counted for each class and the new item is attributed to a cluster with more neighbors to which it belongs [1].

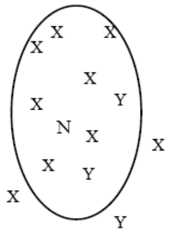

Figure 5: Neighborhood limits (most neighbors are placed in cluster x)

The first case for using K Nearest Neighbor is finding a criterion for the distance between attributes within data and calculating them. This is easy for numerical data but clustering variables need to be treated in a special way. When measuring the distance between different items, we can use a set of items previously clustered as a basis for clustering new items. The neighborhood distance and the method of counting them must be determined.

This method imposes a great computational load on a computer since the time of calculation increases in the form of factorial from all points (samples). Although using a decision tree or neural network for a new item is a fast process, K Nearest Neighbor requires time-consuming and new calculations for each new item. To enhance the speed of this method, usually all the data is kept within the memory.

Understanding models of K Nearest Neighbor is simple when the number of predicting variables is few. They are also useful for making various models of unstructured data like context. Different new data only require a suitable criterion.

2-8- Logistic Regression

Logistic Regression is a more generalized form of a linear regression. In the past, this method was used for the prediction of binary amounts or variables with multiple discontinuous classes. Since the amounts considered are applied for the prediction of discontinuous amounts, they cannot be modeled using linear regression. To this aim, discontinuous variables are converted to numerical and continuous variables by the use of a special method. The amount of probability logarithm considers the relevant variable and the outcome probability are calculated as follows:

"Probability of occurrence of division to the probability of lack of its occurrence"

Interpretation of this ratio is similar to what is used in most daily conversations about matches and bets or similar cases. For example, when we express that chance of winning for a team 3 to 1 in a match, we have used the same ratio and it means probability of winning for that team is 75%.

After obtaining suitable logarithm of probability, we are able to designate the favorable ratio in the proper class [1].

3- REFERENCES

[1] Tomoharu nakashima and hisao ishibuchi. 2005 : "*Using Boosting Techniques To Improve The Performance Of Fuzzy Classification Systems*" In "*Classification And Clustering For Knowledge Discovery*". Studies in Computational Intelligence. Vol 4. pp.146-157 . Springer, Netserlands.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996 : "*From Data Mining A Knowledge Discovery in Databases*".

[3] J.Han, and M.Kamber. 2001 : "*Data Mining: Concepts and Techniques*". San Diego Academic Press.

[4] David Hand, Heikki Mannila , Padhraic Smyth. 2001 : "Principles of Data Mining". The MIT Press.