Data Mining and Optimization Techniques

Sunil Kawale

Department of Statistics, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad - 431 004, Maharashtra, INDIA.

Corresponding Address:

ksunila@rediffmail.com

Research Article

Abstract: Data mining is a modern area of science of extracting useful information from large data sets or databases. Applications of Data mining can be found in various areas. This paper introduces new contributions by optimization as a key technology in data mining. The methods suggested for solution of such important problems as where it deals with large data.

Keywords: Data Mining, Optimization, Support Vector Machine, Time series Data Mining, Logical Analysis.

Introduction

Data Mining:

Data Mining is the process of automatic discovery of useful information in large data repositories. Generally, Data Mining can be divided into two categories according to the objective of algorithms: 1) Classification Analysis 2) Association Analysis.

Many Data Mining methods involve with mathematical programming techniques. Optimization can be a component of a larger Data Mining process and New Data Mining techniques can be built using entirely optimization-based method. These optimization-based Data Mining techniques are applied mainly in Classification Analysis, where as very few algorithms in Association Analysis based as optimization.

Data Mining Process:

Data mining is an iterative process that typically involves the following phases:

a) Problem definition:

A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition.

In the problem definition phase, data mining tools are not yet required.

b) Data exploration:

Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital.

In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

c) Data preparation:

Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value.

In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

d) Modeling:

Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model.

In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required.

The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

e) Evaluation:

Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions:

i) Does the model achieve the business objective?

ii) Have all business issues been considered?

At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

f) Deployment:

Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets.

The following figure shows the phases of the Cross Industry Standard Process for data mining (CRISP DM) process model.

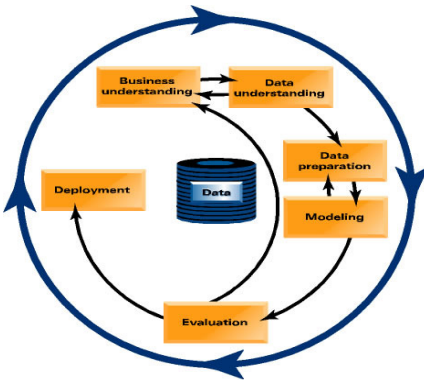

Figure 1: *The CRISP DM process model*

Intelligent Miner Modeling helps you to select the input data, explore the data, transform the data, and mine the data. With IM Visualization you can display the data mining results to analyze and interpret them. With IM Scoring, you can apply the model that you have created with IM Modeling.

Data Mining Models

1) Support Vector Machine Method:

A liner SVM searches for a linear classifier $w \cdot x + b = 1$ based on training data to label unknown data. This classifier is also known as a maximal margin classifier because it maximize the distance between data points is different classes.

$$\min_{w} \frac{\|w\|^2}{2}$$

such that $y_i (w \cdot x_i + b) \geq 1, \quad i = 1, 2 \ldots\ldots n$

It is a quadratic programming problem and w, x_i, y_i are vectors; b is scalar, which can be solved by the standard Lagrange multiplier method.

2) K-mean method:

K-mean is a cluster analysis algorithm and could be treated as an optimization problem, which minimize the sum of distance of each point to its nearest centroid. The clustering problem is then formulated

$$\min_{c^1 \ldots c^k} \sum_{i=1}^{m} \min_{l=1 \ldots k} \|x^i - c^l\|$$

where $x^i, i = 1, \ldots\ldots m$ are given data points

$c^l, l = 1, \ldots\ldots k$ are centroid of k clusters

$\|\cdot\|$ is same arbitrary norm on R^n.

Basic k-means and k-median algorithm for this case is as follows.

1) Select K points as initial centroid
2) Repeat
3) Form k clusters by assigning each point to its closed centroid
4) Compute the centroid of each cluster
5) Until centroid do not change.

3) Logical Analysis:

Logical Analysis of data is another Optimization Based Approach algorithm. It builds a classifier for a binary target variable based on learning a logical expression that can distinguish between positive and negative examples in a data set. If same data set is non-binary, cut-off value is applied to convert them into binary variable. And a table with all the binary attributes and target variables are obtained. The objective then becomes to explore a partially defined Boolean function (pdbf), with all the binary attributes as input and target variables as output.

4) Liner, Non-Liner and integer programming:

These are three powerful and well known optimization tools. Bradly et al. (1993) discussed all this in their research paper. The basic thing in such cases is that optimization of objective function subject to set of constraints. Bennetted and Mangasarien (1993) used linear programming problem for multi-category separation. Mangasarien (1996) described mathematical programming in machine learning. Mangasarien et al. tried to recognized patter through liner programming problem. Bradly et al. (1998) used mathematical programming for feature selection.

5) Network Analysis:

This includes the two approaches as link and affinity analysis. In link analysis, we construct pattern of behavior of the system entities and then search for similarities among them. Kimmo (2009) used network analyze decision making on different levels of network operation he identified the requirements decision making sets for knowledge discovery and data mining tools and methods. Peter Hoschka and willi (1991) discussed the use of network analysis in interpreting statistical data. Richard et al. (2002) said network that analysis can also be useful to fraud detection; they gave entire survey report on fraud detection.

6) Time Series Data Mining:

Time series data mining can be categorized in form:

a) Indexing:

Given a query time series Q and some similarity or dissimilarity measures D, find the nearest matching time series in database.

b) Clustering:

Find natural groupings of the time series in database under same similarity or dissimilarity measures D.

c) Classification:

Given an unlabeled time series Q, assign it to one of two or more predefined classes.

d) Segmentation:

Given a time series containing n data points, construct a model $\overline{Q}$ from K piecewise segments (k<<n) such that $\overline{Q}$ closely approximation Q.

7) Regression:

Regression is the oldest and most well-known statistical technique that the data mining community utilizes. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data.

8) Factor Analysis:

When the data composed of multiple attributes that are associated among them and we are interested in detection in such variables factor analysis will find an equivalent set of abstract factors combinations of the original attributes that describe the same problem availability, such equivalent set will be composed of uncorrelated variables, which can be stored in decreasing importance of their variance.

Conclusions

In this paper different data mining models are discussed which are very useful in data mining. These techniques are currently used by many persons, institutions and various commercial firms who give the solution of data related problems.

References

1. Bennett and Mangasarian. (1993). Multi category separation via linear programming, Optimization Methods and Software, 3: 27-39
2. Bradley, Fayyad, Mangasarian. (1999). Mathematical Programming for Data Mining: Formulations and Challenges. INFORMS Journal on Computing.
3. Bradley, Mangasarian and Street. (1998). Feature selection via Mathematical Programming, INFORMS Journal 10: 209-217.
4. Busygin, Prokopyev, Pardalos. (2007). An optimization-based approach for data classification. Optimization Methods and Software.
5. Keogh and Pazzani. (2002). On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration.
6. Kimmo. (2009). Data mining of telecommunications network log analysis Dept. of Computer Science Series of Publication A Report A-2009-1, University of Helsinki Finland.
7. Mangasarian. (1996). Mathematical programming in machine learning, Nonlinear optimization and Applications, 283-295.
8. Padmanabhan, Tuzhilin. (2003). On the use of optimization for data mining: theoretical interactions and eCRM opportunities. Management Science 2003 INFORMS, 1327-1343.
9. Peter and Willi. (1991). A Support system for interpreting statistical data, Knowledge Discovery in Databases, 325-345.
10. Richard and David. (2002). Statistical fraud detection: A review, Statistical Sci., 17 (3): 235-255.
11. Sunil Kawale. (2012). Statistical and Mathematical Models in data Mining, International Research Journal of Agricultural Economics and Statistics, 03: 359-362.
12. Tan, Steinbach, Kumar. (2005). Introduction to Data Mining.