ISSN: 2456 – 3137
International Journal of Computational Research and Development
Impact Factor 4.775, Special Issue, January - 2017
International Conference on Smart Approaches in Computer Science Research Arena
On 5th January 2017 Organized By
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

STATISTICAL DATA MINING – PREDICTION TASKS IN DATA MINING ALGORITHM

K. Samundeeswari* & Dr. K. Srinivasan**

* Guest Lecturer, Department of Computer Science, Govt. Arts College for Women, Krishnagiri, Tamilnadu
** Assistant Professor & Head, Department of Computer Science, Periyar University Constituent College of Arts & Science, Pennagaram, Dharmapuri, Taminnadu

Cite This Article: K. Samundeeswari & Dr. K. Srinivasan, "Statistical Data Mining – Prediction Tasks in Data Mining Algorithm", International Journal of Computational Research and Development, Special Issue, January, Page Number 11-15, 2017.

Abstract:

Data mining is a new discipline lying at the interface of statistics, database technology, pattern recognition, machine learning, and other areas. From a statistical perspective it can be viewed as computer automated exploratory data analysis of (usually) large complex data sets. Despite the obvious connections between data mining and statistical data analysis, most of the methodologies used in Data Mining have so far originated in fields other than Statistics. This paper explores some of the reasons for this, and why statisticians should have an interest in Data Mining. However, statistics provides the intellectual glue underlying the effort, it is important for statisticians to become involved. There are very real opportunities for statisticians to make significant contributions.

Key Words: Data Mining, Classification, Knowledge Discovery, Prediction Tasks, Statistics & Regression.

1. Definition and Objectives:

Data Mining (DM) is at best a vaguely defined field; its definition largely depends on the background and views of the definer. Here are some definitions taken from the DM literature:

Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. **- Fayyad.**

Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. **- Zekulin.**

Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data. **- Ferruzza.**

Data mining is the process of discovering advantageous patterns in data. **– John**

Data mining is a decision support process where we look in large data bases for unknown and unexpected patterns of information. **– Parsaye**

Data Mining is...

- ✓ Decision Trees
- ✓ Neural Networks
- ✓ Rule Induction
- ✓ Nearest Neighbors
- ✓ Genetic Algorithms

- Mehta

1.1 Objective: The term data mining is not new to statisticians. It is a term synonymous with data in the hope of identifying patterns. It has a derogatory connotation because a sufficiently exhaustive search will certainly throw up patterns of some kind by definition data that are not simply uniform have differences which can be interpreted as patterns. The trouble is that many of these "patterns" will simply be a product of random fluctuations, and will not represent any underlying structure. To statisticians, the term data mining conveys the sense of naive hope vainly struggling against the cold realities of chance. To other researchers, however, the term is seen in a much more positive light [10]. Statisticians have typically not concerned themselves with data sets containing many millions or even billions of records. Moreover, special storage and manipulation techniques are required to handle data collections of this size—and the database technology which has grown up to handle them has been developed by entirely different intellectual communities from statisticians. It is probably no exaggeration to say that most statisticians are concerned with primary data analysis. Data mining, on the other hand, is entirely concerned with secondary data analysis. In fact we might define data mining as the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners. There is urgency for statisticians to become involved with data mining exercises, to learn about the special problems of data mining, and to contribute in important ways to a discipline that is attracting increasing attention from a broad spectrum of concerns. The concept of DM is used in statistics are very large data bases must be stored and quickly accessed, and computationally intensive methodology applied to these data. This requires massive amounts of disk space and fast compute engines with large internal (RAM) memories. DM opens new markets for such hardware. Many organizations have large transaction oriented data bases used for inventory, billing, accounting, etc. These data bases were very expensive to create and are costly to maintain.

1.2 Statistical Analysis Procedure: The statistical analysis procedures provided by current DM packages nearly always include [1]:

- ✓ Decision tree induction (C4.5, CART, CHAID)
- ✓ Rule induction (AQ, CN2, Recon, etc.)

- ✓ Nearest neighbors (case based reasoning)
- ✓ Clustering methods (data segmentation)
- ✓ Association rules (market basket analysis)
- ✓ Feature extraction
- ✓ Visualization

In addition, some include:

- ✓ Neural networks
- ✓ Bayesian belief networks (graphical models)
- ✓ Genetic algorithms
- ✓ Self-organizing maps
- ✓ Neuro-fuzzy systems

Almost none of these DM packages offer:

- ✓ Hypothesis testing
- ✓ Experimental design
- ✓ Response surface modeling
- ✓ ANOVA, MANOVA, etc.
- ✓ Linear regression
- ✓ Discriminated analysis
- ✓ Logistic regression
- ✓ GLM
- ✓ Canonical correlation
- ✓ Principal components
- ✓ Factor analysis

2. Statistical Issues in Data Mining:

2.1 Size of the Data and Statistical Theory: Traditional statistics emphasizes the mathematical formulation and validation of a methodology, and views simulations and empirical or practical evidence as a less form of validation. The emphasis on rigor has required proof that a proposed method will work prior to its use [6]. In contrast, computer science and machine learning use experimental validation methods. In many cases mathematical analysis of the performance of a statistical algorithm is not feasible in a specific setting, but becomes so when analyzed asymptotically. At the same time, when size becomes extremely large, studying performance by simulations is also not feasible. It is therefore in settings typical of DM problems that asymptotic analysis becomes both feasible and appropriate. Interestingly, in classical asymptotic analysis the number of cases n tends to infinity.

2.2 The Curse of Dimensionality and Approaches to Address it: The curse of dimensionality is a well-documented and often cited fundamental problem. Not only do algorithms face more difficulties as the data increases in dimension, but the structure of the data itself changes. For example data uniformly distributed in a high dimensional ball. It turns out that (in some precise way, see Meilijson, 1991) most of the data points are very close to the surface of the ball. This phenomenon becomes very evident when looking for the k-Nearest Neighbors of a point in high-dimensional space. The points are so far away from each other that the radius of the neighborhood becomes extremely large. The main remedy offered for the curse of dimensionality is to use only part of the available variables per case, or to combine variables in the data set in a way that will summarize the relevant information with fewer variables [7]. This dimension reduction is the essence of what goes on in the data warehousing stage of the DM process, along with the cleansing of the data.

2.3 Automated Analysis: There are many examples where trivial non relevant variables, such as case number, turned out to be the best predictors in automated analysis. It is well known in statistics that having even a small proportion of outliers in the data can seriously distort its numerical summary. Such unreasonable values, deviating from the main structure of the data, can usually be identified by a careful human data analyst, and excluded from the analysis [14]. But once we have to warehouse information about millions of customers, summarizing the information about each customer by a few numbers has to be automated and the analysis should rather deal automatically with the possible impact of a few outliers. Statistical theory and methodology supply the framework and the tools for this endeavor.

2.4 Algorithms for Data Analysis in Statistics: Computing has always been a fundamental to statistic, and it remained so even in times when mathematical rigorousity was most highly valued quality of a data analytic tool. Some of the important computational tools for data analysis, rooted in classical statistics, can be found in the following list[3], efficient estimation by maximum likelihood, least squares and least absolute deviation estimation and the EM algorithm analysis of variance (ANOVA, MANOVA, ANCOVA), and the analysis of repeated measurements, nonparametric statistics, log linear analysis of categorical data, linear regression analysis, generalized additive and linear models, logistic regression, survival analysis, and discriminant analysis, frequency domain (spectrum) and time domain (ARIMA) methods for the analysis of time series; multivariate analysis tools such as factor analysis, principal component and later independent component analyses, and cluster analysis; density estimation, smoothing and denoising, and classification and regression trees (decision trees)[12], Bayesian networks and the Monte Carlo Markov Chain (MCMC) algorithm for Bayesian inference.

2.5 Visualization: Visualization of the data and its structure, as well as visualization of the conclusions drawn from the data, are another central theme in DM. Visualization of quantitative data as a major activity flourished in the statistics of the 19th 8 century,

ISSN: 2456 – 3137
International Journal of Computational Research and Development
Impact Factor 4.775, Special Issue, January - 2017
International Conference on Smart Approaches in Computer Science Research Arena
On 5th January 2017 Organized By
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

faded out of favor through most of the 20th century, and began to regain importance in the early 1980s [4]. This importance in reflected in the development of the Journal of Computational and Graphical Statistics of the American Statistical Association. Both the theory of visualizing quantitative data and the practice have dramatically changed in recent years. Spinning data to gain a 3-dimensional understanding of point clouds or the use of projection pursuit are just two examples of visualization technologies that emerged from statistics [4].

2.6 Scalability: In machine learning and data mining scalability relates to the ability of an algorithm to scale up with size, an essential condition being that the storage requirement and running time should not become infeasible as the size of the problem increases. Even simple problems like multivariate Statistical Methods for Data Mining histograms become a serious task, and may benefit from complex algorithms that scale up with size. Designing scalable algorithms for more complex tasks, such as decision tree modeling, optimization algorithms, and the mining of association rules, has been the most active research area in DM. Altogether, scalability is clearly a fundamental problem in DM mostly viewed with regard to its algorithmic aspects. We want to highlight the duality of the problem by suggesting that concepts should be scalable as well. In this respect, consider the general belief that hypothesis testing is a statistical concept that has nothing to offer in DM. The usual argument is that data sets are so large that every hypothesis tested will turn out to be statistically significant - even if differences or relationships are minuscule.

3. Prediction Tasks: Classification and Regression:

Classification and regression tasks are the most commonly encountered data mining tasks. These tasks, involve mapping an object to either one of a set of predefined classes (classification) or to a numerical value (regression) [2]. In this section we introduce the terminology required to describe these tasks and the framework for performing predictive modeling. We then describe several key characteristics of predictive data mining algorithms and finish up by describing the most popular predictive data mining algorithms in terms of these characteristics.

3.1 Terminology and Background: Most prediction tasks assume that the underlying data is represented as a collection of objects or records, which, in data mining, are often referred to as instances or examples. Each example is made up of a number of variables, commonly referred to as features or attributes. The attribute to be predicted is of special interest and may be referred to as the target or for classification tasks.

3.2 Characteristics of Predictive Data Mining Algorithms: It is useful to understand the characteristics that can be used to describe and compare them. These characteristics are described briefly in this section and then referred to in subsequent sections.

Table 1: Summary off Predictive Data Mining Algorithms

Learning Method	Tasks Handled	Expressive Power	Training Time	Testing Time	Model Comprehensibility
Decision Trees	Classification	Fair	Fast	Fast	Good
Rule-Based	Classification	Fair	Fast	Fast	Good
ANN	Classification, Regression	Good	Slow	Fast	Poor
Nearest-Neighbor	Classification, Regression	Good	No Time	Slow	Nomodel generated but predictions are explainable
Naïve Bayesian	Classification	Good	Fast	Fast	Poor

The first characteristic concerns the type of predictive tasks that can be handled by the algorithm. Predictive data mining algorithms may handle only classification tasks, only regression tasks, or may handle both types of tasks. The second characteristic concerns the expressive power of the data mining model. Algorithms with limited expressive power may not perform well on certain tasks, although it is difficult to determine in advance which algorithms will perform best for a given task [5]. In fact, it is not uncommon for those algorithms that generate less complex models to perform competitively with those with more expressive power. The format of the model impacts the third criterion, which is the comprehensibility, or explains ability, of the predictive model. Certain models are easy to comprehend or explain, while others are nearly impossible to comprehend and, due to their nature, must essentially be viewed as “black boxes” that given an input somehow produce a result. Whether comprehensibility is important depends on the goal of the predictive task. For example, if one can build an effective model for predicting manufacturing errors, then one may be able to use that model to determine how to reduce the number of future errors. The fourth criterion concerns the computation time of the data mining algorithm. This is especially important due to the enormous size of many data sets. With respect to computation time, we are interested in the training time, how long it will take to build the model, and the testing time, how long it will take to apply the model to new data in order to generate a prediction. The five listed methods are all in common use and are implemented by most major data mining packages.

4. Predictive Data Mining Algorithms:

We briefly describe some of the most common data mining algorithms. Because the purpose of this chapter is to provide a general description of data mining, its capabilities, and how it can be used to solve real-world problems, many of the technical details concerning the algorithms are omitted. A basic knowledge of the major data mining algorithms, however, is essential in order to know when each algorithm is relevant, what the advantages and disadvantages of each algorithm are, and how these algorithms can be used to solve real-world problems.

4.1 Decision Trees: Decision tree algorithms (Quinlan 1993; Breiman et al. 1984) are a very popular class of learning algorithms for classification tasks. The internal nodes of the decision tree each represent an attribute while the terminal nodes (i.e., leaf nodes displayed as rectangles) are labeled with a class value. Each branch is labeled with an attribute value and when presented with an

ISSN: 2456 – 3137
International Journal of Computational Research and Development
Impact Factor 4.775, Special Issue, January - 2017
International Conference on Smart Approaches in Computer Science Research Arena
On 5th January 2017 Organized By
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu
example, one follows the branches that match the attribute values for the example, until a leaf node is reached. The class value assigned to the leaf node is then used as the predicted value for the example. In this simple example the decision tree will predict that a customer will default on their automobile loan if their credit rating is "poor" or it is not "poor" (i.e., "fair" or "excellent") but the person is "middle aged" and their income level is "low".

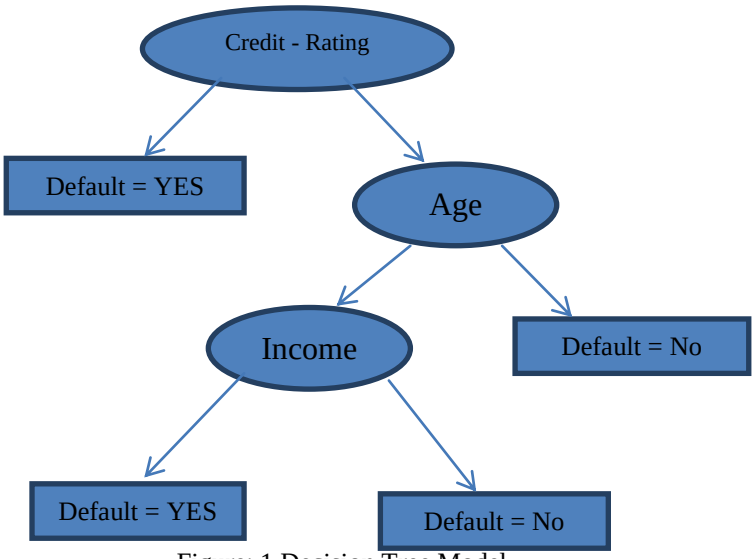

Figure: 1 Decision Tree Model

Decision tree algorithms are very popular. The main reason for this is that the induced decision tree model is easy to understand. Additional benefits include the fact that the decision tree model can be generated quickly and new examples can be classified quickly using the induced model. The primary disadvantage of a decision tree algorithm is that it has limited expressive power, namely because only one attribute can be considered at a time.

4.2 Rule-based Classifiers: Rule-based classifiers generate classification rules, for some rule-based systems the first rule to fire (i.e., have the left-hand side of the rule satisfied) determines the classification, whereas in other cases all rules are evaluated and the final classification is made based on a voting scheme. Rule-based classifiers are very similar to decision-tree learners and have similar expressive power, computation time, and comprehensibility. The connection between these two classification methods is even more direct since any decision tree can trivially be converted into a set of mutually exclusive rules, by creating one rule corresponding to the path from the root of the tree to each leaf.

4.3 Artificial Neural Networks: Artificial Neural Networks (ANNs) were originally inspired by attempts to simulate some of the functions of the brain and can be used for both classification and regression tasks (Gurney 1997). An ANN is composed of an interconnected set of nodes that includes an input layer, zero or more hidden layers, and an output layer. The ANN computes the output value from the input values as follows. First, the input values are taken from the attributes of the training example, as it is inputted to the ANN. These values are then weighted and fed into the next set of nodes, which in this example are H1 and H2. ANNs can naturally handle regression tasks, since numerical values are passed through the nodes and are ultimately passed through to the output layer. However, ANNs can also handle classification tasks by thresholding on the output values. ANNs have a great deal of expressive power and are not subject to the same limitations as decision trees. While the induced ANN can be used to quickly predict the values for unlabelled examples, training the model takes much more time than training a decision tree or rule-based learner and, perhaps most significantly, the ANN model is virtually incomprehensible and therefore cannot be used to explain or justify its predictions [12].

4.4 Nearest-Neighbor: Nearest-neighbor learners (Cover and Hart 1967) are very different from any of the learning methods just described in that no explicit model is ever built. That is, there is no training phase and instead all of the work associated with making the prediction is done at the time an example is presented [7]. Given an example the nearest-neighbor method first determines the k most similar examples in the training data and then determines the prediction based on the class values associated with these k examples, where k is a user specified parameter. The simplest scheme is to predict the class value that occurs most frequently in the k examples, while more sophisticated schemes might use weighted voting, where those examples most similar to the example to be classified are more heavily weighted. People naturally use this type of technique in everyday life. For example, realtors typically base the sales price of a new home on the sales price of similar homes that were recently sold in the area. Nearest-neighbor learning is sometimes referred to as instance-based learning. Nearest-neighbor algorithms are typically used for classification tasks, although they can also be used for regression tasks. These algorithms also have a great deal of expressive power. Nearest-neighbor algorithms generate no explicit model and hence have no training time. Instead, all of the computation is performed at testing time and this process may be relatively slow since all training examples may need to be examined. It is difficult to evaluate the comprehensibility of the model since none is produced. We can say that because no model is produced, one cannot gain any global (i.e., high-level) insight into the domain. However, individual predictions can easily be explained and justified in a very natural way, by referring to the nearest-neighbors.

4.5 Naïve Bayesian Classifiers: Most classification tasks are not completely deterministic. That is, even with complete knowledge about an example you may not be able to correctly classify it. Rather, the relationship between an example and the class it belongs to is often probabilistic [8]. Naïve Bayesian classifiers (Langley et al. 1992) are probabilistic classifiers that allow us to exploit statistical properties of the data in order to predict the most likely class for an example. More specifically, these methods use the training data and the prior probabilities associated with each class and with each attribute value and then utilize Bayes' theorem to determine the most likely class given a set of observed attribute values. This method is naïve in that it assumes that the values for each attribute are independent

5. Conclusion:

Statisticians have developed mathematical theories to support their methods and a mathematical formulation based on probability theory to quantify the uncertainty. Traditional statistics emphasizes a mathematical formulation and validation of its methodology rather than empirical or practical validation. How statistical data mining effectively work data mining algorithms have been explained, with the help of the statistical issues we can develop new concepts by using data mining techniques[7]. Data mining initially generated a great deal of excitement and press coverage and as is common with new "technologies", overblown expectations. All "knowledge workers" in our information society, particularly those who need to make informed decisions based on data, should have at least a basic familiarity with data mining. Apart from general issues arising from data set size and particular issues concerned with pattern search, most of these steps will be familiar to statisticians. It would be a great loss, for the reputation of statistics as a discipline as well as for individual statisticians.

6. References:

1. Breiman, L. 1996. Bagging predictors. Machine Learning, 24(2):123-140.
2. Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and Regression Trees. Wadsworth International Group.
3. Chakrabarti, S. (2002). Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data. Morgan Kaufmann.
4. Chen, Y., Zhang, G, Hu, D., and Wang, S. (2006). Customer segmentation in customer relationship management based on data mining. In Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management, 288-293. Boston: Springer.
5. Cohen, W. (1995). Fast effective rule induction. In Proceedings of the 12th International Conference on Machine Learning, 115-123, Tahoe City, CA.
6. Fayadd, U., Piatesky -Shapiro, G., and Smyth, P, From Data Mining To Knowledge Discovery in Databases", The MIT Press, ISBN 0–26256097–6, Fayap, 1996.
7. Gorunescu, F, Data Mining: Concepts, Models, and Techniques, Springer, 2011.
8. Han, J., and Kamber, M., Data mining: Concepts and techniques, Morgan-Kaufman Series of Data Management Systems San Diego:Academic Press, 2001.
9. Heikki, Mannila, Data mining: machine learning, statistics and databases, IEEE, 1996.
10. Piatetsky-Shapiro, Gregory, The Data-Mining Industry Coming of Age," IEEE Intelligent Systems, 2000.
11. Vapnik, V. N. (1998) Statistical Learning Theory. New York: John Wiley and Sons.
12. Vapnik, V. (1999) The Nature of Statistical Learning Theory, 2nd Ed. New York: Springer-Verlag.
13. Venables, W. N. and Ripley, B. D. (1999) Modern Applied Statistics with S-PLUS. Third Edition. New York: Springer-Verlag. [i]
14. Witten, I. H., and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition, Morgan Kaufmann, San Francisco.