www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 1 Jan. 2017, Page No. 20112-20118
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.46

A Review in Statistical Aspects of Data Mining

K.Samundeeswari, Dr.K.Srinivasan
Guest Lecturer, Department of Computer Science,
Govt. Arts College for Women,
Krishnagiri - 635 001, Tamil Nadu, India
E-mail: samun.arun@gmail.com
Assistant Professor & Head,Department of Computer Science Periyar University Constituent College of Arts &Science,Pennagaram Dharmapuri – 636803,TamilNadu, India
E-mail: vasanmsc23@yahoo.co.in

Abstract

Statistics is defined as the science of collecting, analyzing and presenting data.Data mining is a new discipline lying at the interface of statistics, database technology, pattern recognition, machine learning, and other areas. KDD has a spin that comes from database methodology and from computing with large data sets, while statistics has an emphasis that comes from mathematical statistics, from computing with small data sets, and from practical statistical analysis with small data sets. Statistical techniques are driven by the data and are used to discover patterns and build predictive models. And from the users perspective view with a conscious choice when solving a "data mining" problem is attack it with statistical methods or other data mining techniques. However, since statistics provides the intellectual glue underlying the effort, it is important for statisticians to become involved. KDD is statistics and data mining is statistical analysis. "Knowledge Discovery in Databases" is not much different. The main statistical issues in Data mining (DM) and Knowledge Data Discovery (KDD) is to examine whether traditional statistics approach and methods substantially differ from the new trend of KDD and DM.

Keywords— Data mining, Knowledge Data Discovery, Multivariate Analysis, Statistics.

I.INTRODUCTION

A.Statistics in Data Mining

Statistics is a branch of mathematics concerning the collection and the description of data. Knowing statistics in your everyday life will help the average business person make better decisions by allowing them to figure out risk and uncertainty when all the facts either aren't known or can't be collected.

Today data mining has been defined independently of statistics though "mining data" for patterns and predictions is really what statistics is all about. Some of the techniques that are classified under data mining such as CHAID and CART really grew out of the statistical profession more than anywhere else, and the basic ideas of probability, independence and causality and over fitting are the foundation on which both data mining and statistics are built.

Statistics is defined as the science of collecting, analyzing and presenting data. KDD has a spin that comes from database methodology and from computing with large data sets, while statistics has an emphasis that comes from mathematical statistics, from computing with small data sets, and from practical statistical analysis with small data sets [1]. KDD is statistics and data mining is statistical analysis.

"Knowledge Discovery in Databases" is not much different. The components that seem needed are:

1. Computing skills required to manage the data and the analysis.
2. An understanding of design of data collection issues.
3. An understanding of statistical inferential issues.
4. Knowledge of relevant mathematics.
5. Insights from practical data analysis.
6. Application area insights.

7. Automation of data analysis.

Statistics is the traditional field that deals with the quantification, collection, analysis, interpretation, and drawing conclusions from data [4]. Data mining is an interdisciplinary field that draws on computer sciences (data base, artificial intelligence, machine learning, graphical and visualization models), statistics and engineering (pattern recognition, neural networks). DM involves the analysis of large existing data bases in order to discover patterns and relationships in the data, and other findings (unexpected, surprising, and useful). Typically, it differs from traditional statistics on two issues: the size of the data set and the fact that the data were initially collected for purpose of the DM analysis. Thus, experimental design, a very important topic in traditional statistics, is usually irrelevant to DM. On the other hand asymptotic analysis, sometimes criticized in statistics as being irrelevant, becomes very relevant in DM.

Statistics has been concerned with detecting structure in data under uncertainty for many years: that is what the design of experiments developed in the inter-war years had as its aims. Generally that gave a single outcome ('yield') on a hundred or so experimental points. Multivariate analysis was concerned with multiple (usually more than two and often fewer than twenty) measurements on different subjects.

Statistical pattern recognition where everything is 'learnt from examples'

Structural pattern recognition where most of the structure is imposed from a priori knowledge. This used to be called syntactic pattern recognition, in which the structure was imposed by a formal grammar, but that has proved to be pretty unsuccessful.

B.Multivariate Analysis

Multivariate analysis is concerned with datasets which have more than one response variable for each observational or experimental unit. The datasets can be summarized by data matrices X with n rows and p columns, the rows representing the observations or cases, and the columns the variables. The matrix can be viewed either way, depending whether the main interest is in the relationships between the cases or between the variables. Note that for consistency we represent the variables of a case by the row vector x. The main division in multivariate methods is between those methods which assume a given structure, for example dividing the cases into groups, and those which seek to discover structure from the evidence of the data matrix alone.

C.Statistics and its Need

Statistics is the science of learning from data. It includes everything from planning for the collection of data and subsequent data management to end-of-the-line activities such as drawing inferences from numerical facts called data and presentation of results [3]. Statistics is concerned with one of the most basic of human needs: the need to find out more about the world and how it operates in face of variation and uncertainty. Because of the increasing use of statistics, it has become very important to understand and practice statistical thinking.

D.Why Statistics Needed

Knowledge is what we know. Information is the communication of knowledge. Data are known to be crude information and not knowledge by themselves. The sequence from data to knowledge is as follows: from data to information (data become information when they become relevant to the decision problem); from information to facts (information becomes facts when the data can support it); and finally, from facts to knowledge (facts become knowledge when they are used in the successful completion of the decision process)[2]. That is why we need statistics. Statistics arose from the need to place knowledge on a systematic evidence base. This required a study of the laws of probability, the development of measures of data properties and relationships, and so on.

The statistical thinking process based on data in constructing statistical models for decision making under uncertainties.

II. WHY DATA MINING

Data mining got its start in what is now known as "customer relationship management" (CRM). It is widely recognized that companies of all sizes need to learn to emulate what small; service-oriented businesses have always done well – creating one-to-one relationships with their customers. In every industry, forward-looking companies are trying to move towards theone-to-one ideal of understanding each customer individually and to use that understanding to make it easier for the customer to do business with them

rather than with a competitor. These samecompanies are learning to look at the lifetime value of each customer so they know which ones are worth investing money and effort to hold on to and which ones to let drop. The term data mining was coined to apply to business applications.

"*Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules*." (M. J. A. Berry and G. S. Linoff)

"*Data mining is finding interesting structure (patterns, statistical models, relationships) in databases*." (U. Fayyad, S. Chaudhuri and P. Bradley)

"Data mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data using machine learning, statistical and visualization techniques"–(Frawley et al., 1992).

We think of data mining as the process of identifying valid, novel, potentially useful, and ultimately comprehensible understandable patterns or models in data to make crucial business decisions."Valid" means that the patterns hold in general, "novel" that we did not know the pattern beforehand, and "understandable" means that we can interpret andcomprehend the patterns. Hence, like statistics, data mining is not only modeling and prediction, nor a product that can be bought, but a whole problem solving cycle/process that must be masteredthrough team effort [4].

Data mining is one of the best ways to extract meaningful trends and patterns from huge amount of data. Data mining is concerned with the analysis of data and the use of software techniques for finding hidden and unexpected patterns and relationships in sets of data. The focus of data mining is to find the information that is hidden and unexpected.

The process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions is known as Data Mining [2].

Data are always dirty and are not readyfor data mining in the real world. For example,

- Data need to be integrated from different sources;
- Data contain missing values. i.e. incomplete data;
- Data are noisy, i.e. contain outliers or errors, and inconsistent values (i.e. contain discrepancies in codes or names);
- Data are not at the right level of aggregation.

The main part of data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It is the computer which is responsiblefor finding the patterns by identifying the underlying rules and features in the data. The choice of a particular combination of techniques to apply in a particular situation depends on both thenature of the data mining task to be accomplished and the nature of the available data. The idea is that it is possible to strike gold in unexpected places as the data mining software extractspatterns not previously discernible or so obvious that no-one has noticed them before. The analysis process starts with a set of data, uses a methodology to develop an optimal representation of thestructure of the data during which time knowledge is acquired. Once knowledge has been acquired this can be extended to larger sets of data working on the assumption that the larger data set has astructure similar to the sample data. This is analogous to a mining operation where large amounts of low grade materials are sifted through in order to find something of value.

III.DATA MINING TECHNIQUES

There are four main operations associated with data mining techniques which include [1]:

• Predictive modeling
• Database segmentation
• Link analysis
• Deviation detection.

Techniques are specific implementations of the· data mining operations. However, each operation has its own strengths and weaknesses. With this in mind, data mining tools sometimes offer a choice of operations to implement a technique. Many experts agree that data mining should not be automatic – human intervention and interpretation is essential.

IV.CHALLENGES OF DATA MINING

• Size of dataset

• High dimensionality
• Over-fitting
• Missing and noisy data
• Rapidly changing data
• Mixed dataset
• Human intervention and interpretation

V. DATA MINNG TASKS

Let us define the main tasks well-suited for data mining, all of which involve extracting meaningful new information from the data. Knowledge discovery (learning from data) comes in two flavours:directed (supervised) and undirected (unsupervised) learning from data. The six main activities of data mining are:

- *classification* (examining the feature of a newly presented object and assigning it to one of a predefined set of classes);
- *estimation* (given some input data, coming up with a value for some unknown continuous variable such as income, height, or credit-card balance);
- *prediction* (the same as classification and estimation except that the records are classified according to some predicted future behavior or estimated future value);
- *affinity grouping or association rules* (determine which things go together, also known as dependency modeling, e.g. in a shopping cart at the supermarket – market basket analysis);
- *clustering* (segmenting a population into a number of subgroups or clusters); and
- *Descriptionand visualization* (exploratory or visual data mining).

The first three tasks – classification, estimation and prediction – are all examples of directed knowledge discovery (supervised learning). In supervised learning the goal is to use theavailable data to build a model that describes one particular variable of interest[6], such as income or response, in terms of the rest of the available data ("class prediction"). The nextthree tasks – affinity grouping or association rules, clustering, and description and visualization – are examples of undirected knowledge discovery (unsupervised learning). Inunsupervised learning no variable is singled out as the target; the goal is to establish some relationship among all the variables ("class discovery"). Unsupervised learning attempts tofind patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes.

VI. KDD CONTRASTED WITH STATISTICS

Data mining and statistics have different intellectual traditions. Both tackle problems of data collection and analysis. Data mining has very recent origins. It is in the tradition of artificial intelligence, machine learning, management information systems and database methodology. It typically works with large data sets. Statistics has a much longer tradition. It has favoured probabilistic models, and has been accustomed to work with relatively small data sets. Both traditions use computing tools, but often different tools. Data mining may now be entering a less brash and more reflective phase of development, where it is more willing to draw from the statistical tradition of experience with data analysis. Efron's warning is apt:

> *"Statistics has been the most successful information science. Those who ignore statistics are condemned to re-invent it."*
> [Efron, quoted in Friedman 1997.]

There are many other points of common interest between data mining and mainstream statistical analysis, points that one would cover in a course on statistical regression and classification modeling. Variable selection is as much or more an issue in data mining as in mainstream statistical analysis. Depending on how results are to be used, the confounding of effects of variables may be a serious problem for interpretation.

Classical statistical methods do not scale up to these huge data sets. *Oftentimes the data should be scaled down.*Skills in the manipulation of large databases are necessary to do anything at all. It will take time to get widespread acknowledgement that the skills and tools needed to manipulate large data sets may not, on their own, be enough.

VII.DIFFERENCE BETWEEN STATISTICS AND DATA MINING

The techniques used in data mining, when successful, are successful for precisely the same reasons that statistical techniques are successful (e.g. clean data, a well-defined target to predict

and good validation to avoid over fitting). And for the most part the techniques are used in the same places for the same types of problems (prediction, classification discovery). In fact some of the techniques that are classical defined as "data mining" such as CART and CHAID arose from statisticians [6].

So what is the difference? Why aren't we as excited about "statistics" as we are about data mining? There are several reasons. The first is that the classical data mining techniques such as CART, neural networks and nearest neighbor techniques tend to be more robust to both messier real world data and also more robust to being used by less expert users. But that is not the only reason. The other reason is that the time is right. Because of the use of computers for closed loop business data storage and generation there now exists large quantities of data that is available to users. IF there were no data - there would be no interest in mining it.

Likewise the fact that computer hardware has dramatically upped the ante by several orders of magnitude in storing and processing the data makes some of the most powerful data mining techniques feasible today. The bottom line though, from an academic standpoint at least, is that there is little practical difference between a statistical technique and a classical data mining technique. Statistics can help greatly in this process by helping to answer several important questions about your data:

- What patterns are there in my database?
- What is the chance that an event will occur?
- Which patterns are significant?
- What is a high level summary of the data that gives me some idea of what is contained in my database?

Certainly statistics can do more than answer these questions but for most people today these are the questions that statistics can help answer. Consider for example that a large part of statistics is concerned with summarizing data, and more often than not, this summarization has to do with counting. One of the great values of statistics is in presenting a high level view of the database that provides some useful information without requiring every record to be understood in detail. This aspect of statistics is the part that people run into every day when they read the daily newspaper .Statistics at this level is used in the reporting of important information from which people may be able to make useful decisions. There are many different parts of statistics but the idea of collecting data and counting it is often at the base of even these more sophisticated techniques. The first step then in understanding statistics is to understand how the data is collected into a higher level form.

VIII.STATISTICAL ISSUES IN DATA MINING

A. 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 [1]. 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 [2]. Interestingly, in classical asymptotic analysis the number of cases n tends to infinity.

B.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. Take, for example, data uniformly distributed in a highdimensional 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.

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. It is an important and time-consuming stage of the DM operations, accounting for 80-90% of the time devoted to the analysis. The dimension reduction comprises two types of activities: the first is quantifying and summarizing information into a number of variables and the second is further reducing the variables thus constructed into a workable number of combined variables.

C.Automated analysis

The inherent dangers of the necessity to rely on automatic strategies for analyzing the data, another main theme in DM, have been demonstrated again and again. There are many examples where trivial nonrelevant 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 [5]. 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.

D.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 categorial 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); Bayesian networks and the Monte Carlo Markov Chain (MCMC) algorithm for Bayesian inference.

E.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 century, 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 pointclouds or the use of projection pursuit are just two examples of visualization technologies that emerged from statistics.

F. 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 9 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. Using* association rules as an example, one may

wonder whether an observed lift for a given rule is"really differ

IX.CONCLUSION

Data mining and statistics will inevitably grow toward each other in the near future because data mining will not become knowledge discovery without statisticalthinking, statistics will not be able to succeed on massive and complex datasets without data mining approaches. The field of data mining, like statistics, concerns itself with "learning from data" or "turning data into information.it is important to note that data mining can learn from statistics– that, to a large extent, statistics is fundamental to what data mining is really trying to achieve.

There is the opportunity for an immensely rewarding synergy between data miners andstatisticians. However, most data miners tend to be ignorant of statistics and client's domain; statisticians tend to be ignorant of data mining and client's domain; and clients tend tobe ignorant of data mining and statistics. 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. A question that has often been raised among statisticians is whether DM is not merely part of statistics.

REFERENCES

[1] 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.

[2] Gorunescu, F, Data Mining: Concepts, Models, and Techniques, Springer, 2011.

[3] Han, J., and Kamber, M. , Data mining: Concepts and techniques, Morgan-Kaufman Series of Data Management Systems San Diego:Academic Press, 2001.

[4] Heikki, Mannila, Data mining: machine learning, statistics and databases, IEEE, 1996.

[5] Piatetsky-Shapiro, Gregory, The Data-Mining Industry Coming of Age,"IEEE Intelligent Systems, 2000.

[6] Vapnik, V. N. (1998) Statistical Learning Theory. New York: John Wiley and Sons.

[7] Venables, W. N. and Ripley, B. D. (1999) Modern Applied Statistics with S-PLUS. Third Edition. New York: Springer-Verlag. [i]