ORIGINAL ARTICLE ISSN:-2249-894X

Review Of Research

Abstract:-

The aim of this paper is to find the intersecting disciplines of statistics and data mining .Data mining (sometimes called data or knowledge discovery) Is a process to extract information from a dataset and transform it into an understandable structure for future use .It allows users to analyzes data from many different dimensions. Whereas Statistics is the study of collection, organization, analysis, interpretation and presentation of data, it deals with all aspects of data including planning and collection. In statistics not only we can find the patterns but we can improve the quality of data. In this paper we discuss the similarities and differences of statistics and data mining.

DATAMINING A STATISTICAL PERSPECTIVE

Meena Kumari

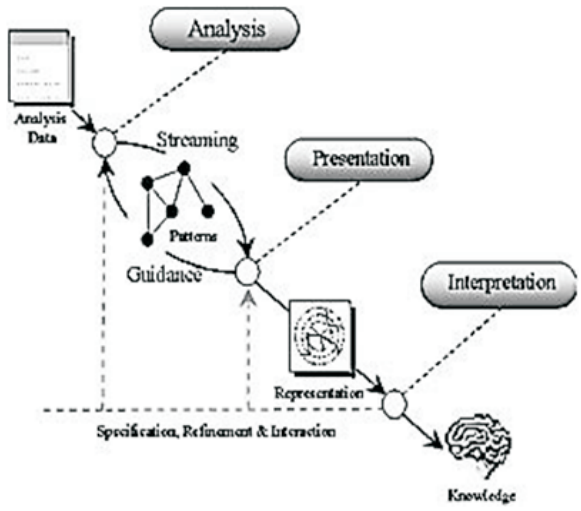
Assistant Professor, Hindu Girls College , SONEPAT,Haryana.

Keywords:

Statistical Method; Kdd; Data Dredging; Descriptive; Data Mining.

INTRODUCTION

DATA Mining is a term synonymous with dredging (data fishing, data snooping) that has been used to describe the process of uncover relationship in data in the hope of identifying patterns. It is also known as KDD.



This is a process of analyzing data from different perspectives and summarizing it into useful information but staved for knowledge. Early statisticians invented various techniques to handle any problem was at hand. Those techniques also proven their worth nowadays in all areas where data is being collected .For example, in the fields like psychological experiments, agricultural experiments, astrological experiments, medicinal experiments and even business etc., all requires data in their respective fields to arrive at some logical decision. After the development of computer technology and electronics data acquisition, recent decades have seen the growth of data bases in various areas as in banking sector, chemistry, medicines, astronomy and super market sales, to official and government statistics. These huge databases are viewed as a resource. Data mining is regarded as providing a set of tools by which information can be extracted. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational data bases.

There is some commonality between the statistics and data mining. Statistician have been 'manually' extracting patterns from the data for centuries long as 'Bays theorem' (1700s),regression analysis(1800s) and numerical analysis (extrapolation) for findings the patterns in the data but proliferation, ubiquity and increasing power of computer technology has increased data collection and storage. As data sets have increased in size and complexity, direct hands on data analysis has increasingly been augmented with indirect, automatic data processing. Data mining is used with the intention of uncovering hidden patterns. Data mining is the application of statistics as classification maps data into predefined groups or classes –Supervised learning, Pattern recognition; Prediction .Regression is used to map a data item to a real valued prediction variable .Clustering group's similar data together into cluster. Statistic, data mining is not only modeling and prediction but also a whole problem solving process. Indeed, it has even sometimes caused antipathy .There is a new discipline had an attractive name, almost calculated to arouse interest and curiosity that is Data mining .This new approach has particular relevance to commercial concerns for instance see E.Gantnar, Hastie, Fayyad usama et al.

2. Types of data-Statistic and Data mining both are dealing with data .Data is any facts, numbers or text. So one should know the types of data .In research mainly two types of data are studied-quantitative and qualitative.

Quantitative: These data have meaning as a measured such as a person height, weight IQ, or blood pressure or these can count, such as the number of stock shares a person own .this numerical measurement are called quantitative data. This can further dividing into two types: discrete and continuous.

Discrete data represent items that can be counted; they take on possible values that can be listed out. The possible value can be finite or infinite .for e.g. The number of heads in 100coin flips takes on value from zero to hundred (finite case), but the number of flips needed to100 heads takes on values from hundred on up to infinity .Its possible values may be listed as 101,102,103…..

Continuous data represent measurement; their possible values cannot be counted and can only be describe using intervals on the real number line. For e.g. A persons , height could be any value (with in the range of human heights) not just certain fixed heights. Time in a race, you could even measure it to fractions of a second, the length of a leaf.

Qualitative data: This type of data is a categorical measurement expressed not in terms of numbers, but

rather by means of natural description .Further it can be divided into categorical data -with order also known as nominal for e.g. male, female, literate, illiterate etc. Second category data with order also known as ordinal data for e.g. Veryhappy, happy, unhappy, and sad.

3. Methods for data mining: Data mining (sometimes called data or knowledge discovering) is the process of analyzing data from different perspective and summarizing it into useful information –that can be used to increase revenue, cuts costs or both .It allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified. Technically data mining is the process of finding correlations or patterns among dozens of fields in large relational data bases.

The simplest definition says that data mining uses statistical algorithms to discover patterns in data. There are many definitions as: "Data mining is finding interesting structure in database'' Fayyad and Bradley.

"Data mining is asset of methods used in the knowledge discovery process to distinguish previously unknown relationship and patterns within data''Ferruzza. "Data mining is the process of discovering advantageous patterns in data' John.

"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.

Most data mining methods are based on concepts from Machine learning, pattern recognition and statistics. Data mining uses methods that can sift through the data in search of frequently occurring patterns, can detect trends, produce generalization about the data, etc. These tools can discover this type of information with very little guidance from the user.

The main tasks well suited for data mining all of which involve extracting meaningful new information from the data. Knowledge discovery comes in two flavors: directed and undirected learning from data. The main activities of data mining are: classification, estimation, prediction, affinity grouping or association rules, clustering, description and visualization. Apart from these activities data mining has at least three major activities: classification, affinity grouping or association rule and sequence analysis.

In classification a database is analyzed and a set of rules which can be used to classify future data is generated. It allows finding rules that partition the data into several predefined classes.

An affinity grouping or association rule is a rule that implies certain association relationships among a set of objects in database. In this process association rules at multiple levels of abstraction from the relevant set(s) of data in a database are discovered. Mining association rules may require searching large relational database that is quite costly in processing.

In sequence analysis, patterns that occur in a sequence are discovered .This deals with data that appears in separate transactions. For example; ifa customer buys items X in the first week of the month, then he buys item Y the second week etc.

The most commonly used methods in data mining are:

1. Artificial neural network: non-linear predicative models.
2. Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression trees (CART) and Chi Square Automatic Interaction detection (CHAID) .
3. Genetic Algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts evolution.
4. Nearest neighbor method: A technique that classifies each record in a dataset on a combination of the classes of the K records most similar to it in a historical dataset (where k31). Sometimes called the k-nearest neighbor technique.
5. Rule induction: The extraction of useful if-then rules from data based on statistical significance.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data.

Commercial data-mining software and applications-

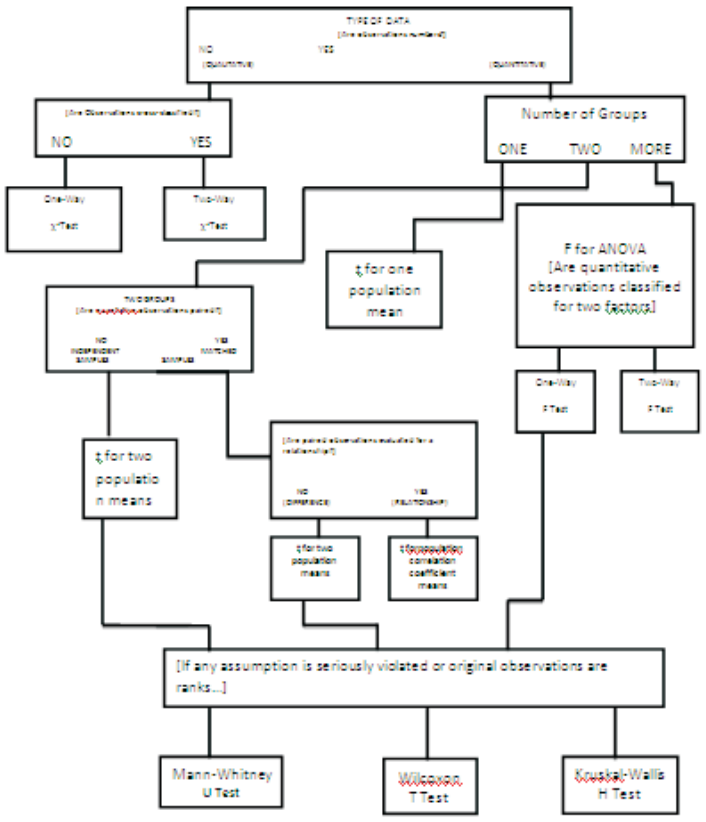
1 IBM SPSS Modeler: Provided by IBM.
2 Microsoft Analysis Services: Provided by Microsoft.
3 SAS Enterprise Miner: Provided by SAS Institute.
4 Statistica Data Miner: Provided by Stat Soft.
The statistical software is dramatically increasing the accuracy of analysis.

4.Methods for statistics-Statistics is the science off making effective use of numerical data relating to groups of individual or experiments. It deals with all the aspects of this, including not only analysis and interpretation of such data, but also the planning of the collection of the data, in terms of the design of surveys and experiments.

Statistical method include all those devices of analysis and synthesis by means of which statistics

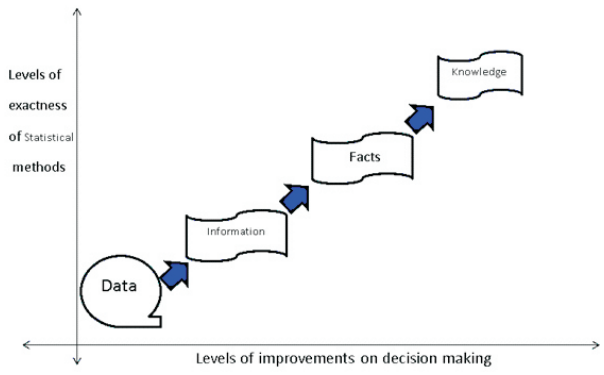
(data) are scientifically collected and used to explain or describe phenomena either in their individual or related capacities- Secrets Statistics is the branch of scientific method which deals with the data obtained by counting or measuring the properties of population of natural phenomena -Kendall Statistics are the numerical statement of facts capable of analysis and interpretation and the science of statistics is the study of principles and the methods applied in collecting,presenting, analysis and interpreting the numerical data in any field of inquiry- W.I King Decisions on how to evaluate data are crucial first steps in planning a study. Such decisions become part of the wider consideration of how to design study. Statistics can be a challenging subject for many investigators. There are several statistical tests to choose from, and choosing the right one for a particular set of data can be an overwhelming task, particularly such decisions are rendered after the data is collected. Even if an investigator does not perform his own analysis, he or she should have a working knowledge of the subject to communicate intelligently and effectively with the statistician and data miners.

Diagram depicting the statistical methods for data analysis:



Statistics and Data Mining:

As statistics and data mining both deal with the process of learning from data (pattern) so the main problem is to know how to get from data to information, from information to knowledge, from knowledge to decision and



decision to action.

Samplingmethodology, which has a long tradition in Statistics, can profitably be used to improve accuracy while mitigating computational requirements. Statistical learning problem can be considered as either supervised or unsupervised. Even if one were to grant the intellectual viability of data mining methodological development the issue remains as to whether Statistics as a discipline should be concerned with it.In the supervised learning the goal is to predict the value of the variable Y (outcome), based on number of other variables (predictors). As a result, the prediction model or learner is built. It will predict the outcome for new, unseen objects. In the unsupervised learning no outcome variable is available. Therefore its goal is to describe associations and patters among the data. Although Data Mining has its origins outside Statistics it uses many Statistical procedures, for example: classification and regression trees, rule induction, nearest neighbors, clustering methods, association rules, feature extraction, data visualization, etc. The sampling methodology is not used in Data Mining applications either, although it could improve accuracy while mitigating computational requirements .Computationally intense procedure operating on a subset of the data may in fact improve accuracy better than a less sophisticated one using the entire database. Extremely large data sets are usually quite complex, frequently containing scores of variables, many of which can be described by non –linear relationships. Numerous variables may also interact with each other. These issues all combine to make many Statistical procedures, such as Analysis of Variance or Regression analysis, difficult to use.

CONCLUSION:

Statistics and data mining both are doing some what same task of discovering structure in data. Some people regard data mining as a subset of Statistics. Butit is not a realistic assessment. Datamining also makes use of ideas, tools and methods from other areas- especially computational areas such as database technology and machine learning - and is not heavily concerned with some areas in which statisticians are interested. Most data mining techniques are statistical exploratory data analysis tools. Care must be taken to not “over analyze” the data, complete understanding of data and its collection methods are particularly important. Database sampling or cluster analysis may help in reducing the dimension and size of massive data sets .while large data sets introduce additional complications to their analysis, researchers should not disregard the basic statistical concepts that have served so well when analyzing smaller data sets. Data collection methods should reflect overall objectives and initial analysis should be composed of EDA and data visualization techniques. Once a complete understanding of the data has been gained more complicated methods, such as cluster analysis or data base sampling can be attempted.

REFERENCES

1.Clifton Christopher (2010) “Encyclopedia Britannica –Definition of Data Mining “, Retrieved 9/12/10
2.ABC Fayyad Usama, Piatesky-Gregory: Symth, Padhraic (1996), “From Data Mining to discovery in databases “, Retrieved 17 Dec 2008.
3.Miller , Harvey J and Han , Jiawei(eds)(2001) “ Geographic Data Mining and Knowledge discovery “ London, GB : Taylor and Francis.
4.T.Hastie, Trevor, Tibshirani, Robert, Friedmen2009 “The Elements of Statistical Learning: Data Mining inference and prediction “. Retrieved 2012-08-07.
5.E.Gatnor “Data Mining and Statistical data Analysis” Statistical revue 2, (1997) 309-316.
6.D.J. Hand, Data Mining: Statistics and more? –The American Statistician, 52(1998) , 112-118.
7.D. Kuonen, A statistical perspective of data mining CRM Zine, 48(2004),1-6.
8.Kantardzic’ Mehmed (2003): Data mining: Concepts Models ‘Methods and Algorithms. John Wiley and Sony-ISBNO-471-22852-4
9.Nibert,Robert,Elder,John,Miner,Gary (2009);Hand book of Statistical Analysis and Data mining Applications, Academic Press/ Elsevier IABN 978-0-12-374765-5
10.L.K.Grover and RajniMehra. The Lure of Statistics in Data mining. Journal of Statistics Education, 16, 1(2008), 26-33.