ISSN 2278-8808
An International Peer Reviewed
SCHOLARLY RESEARCH JOURNAL
FOR INTERDISCIPLINARY STUDIES

AN INTERSECTING FUNCTIONALITY OF DATAMINING AND STATISTICS.

Adel Hedires Naathil Ghanem
,Islah High School
Ministry of Education
IRAQ

Abstract

Statistics is the collection tabulation analysis and interpretation of data selected for specific studies, Data mining is the technique to extract the hidden pattern from the large database for decision making purpose of any organization. Since both the fields are dealing with the large database and analyze and interpret it therefore both the fields are having some similarity. This paper examines the nature of both the disciplines and discusses its similarity and differences.

1. INTRODUCTION:

1.1.STATISTICS:

Statistics is the collection tabulation analysis and interpretation of data selected for specific studies, It deals with all aspects of data, including the planning of data collection in terms of the design of surveys and experiments [1]. The word statistics, when referring to the scientific discipline, is singular, as in "Statistics is an art." This should not be confused with the

word statistic, referring to a quantity (such as mean or median) calculated from a set of data [2], whose plural is statistics.

Some consider statistics a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data [3], while others consider it a branch of mathematics [6] concerned with collecting and interpreting data. Because of its empirical roots and its focus on applications, statistics is usually considered a distinct mathematical science rather than a branch of mathematics [4]. Much of statistics is non-mathematical: ensuring that data collection is undertaken in a way that produces valid conclusions; coding and archiving data so that information is retained and made useful for international comparisons of official statistics; reporting of results and summarized data (tables and graphs) in ways comprehensible to those who must use them; implementing procedures that ensure the privacy of census information.

Statisticians improve data quality by developing specific experiment designs and survey samples. Statistics itself also provides tools for prediction and forecasting the use of data and statistical models. Statistics is applicable to a wide variety of academic disciplines, including natural and social sciences, government, and business. Statistical consultants can help organizations and companies that don't have in-house expertise relevant to their particular questions.

1.2. DATA MINING:

Data mining is the technique to extract the hidden patterns from the large database for decision making for any organization. Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use [5]. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference

considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating[5].

Data mining involves six common classes of tasks [6].

- Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – Attempts to find a function which models the data with the least error.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.
- Sequential pattern mining – Sequential pattern mining finds sets of data items that occur together frequently in some sequences.

2. DATA MINING METHODOLOGY:

Data mining technique uses two different modeling tasks firstly it supervised the predictive model and second is unsupervised the predictive model. Predictive modeling tasks where the goal is to predict the value of other columns are called supervised tasks. The goal in descriptive modeling is to discover patterns and segments of the data. These are called unsupervised tasks. Unsupervised tasks provide insight into the data at a whole by showing patterns and segments that behave similarly.

3. WORKING OF DATA MINING:

How is data mining able to tell you important things that you didn't know or what is going to happen next? That technique that is used to perform these feats is called modeling. Modeling is simply the act of building a model (a set of examples or a mathematical relationship) based on data from situations where the answer is known and then applying the model to other situations where the answers aren't known. Modeling techniques have been around for centuries, of course, but it is only recently that data storage and communication capabilities required to collect and store huge amounts of data, and the computational power to automate modeling techniques to work directly on the data, have been available.

As a simple example of building a model, consider the director of marketing for a telecommunications company. He would like to focus his marketing and sales efforts on segments of the population most likely to become big users of long distance services. He knows a lot about his customers, but it is impossible to discern the common characteristics of his best customers because there are so many variables. From his existing database of customers, which contains information such as age, sex, credit history, income, zip code, occupation, etc., he can use data mining tools, such as neural networks, to identify the characteristics of those customers who make lots of long distance calls. For instance, he might learn that his best customers are unmarried females between the age of 34 and 42 who make in excess of $60,000 per year. This, then, is his model for high value customers, and he would budget his marketing efforts to accordingly.

4. STATISTICAL WORKING:

Statistics is the body of scientific methodology that deals with the logic of experiment and survey design, the efficient collection and presentation of quantitative information, and the formulations of valid and reliable inferences from sample data. Statisticians often work in conjunction with professionals from fields as biology, economics, engineering, medicine, public health, psychology, marketing, education, and sports.

Statistical procedures based on scientific sampling have become basic tools in diverse fields as weather forecasting, opinion polling, biological and agricultural estimation, and business trend prediction. Statisticians are in demand wherever quantitative studies are conducted.

5. STATISTICS AND DATA MINING:

Since Statists and data mining both fields deals with large database and used for analysis and prediction therefore there are some overlapping in both the fields, most of the techniques in data mining can be placed in a statistical framework. However data mining techniques are not the same as statistics. Statistical methods require great deal of user interaction in order to validate the correctness of a model. As a result statistical methods can be difficult to automate and statistical methods typically do not scale well to very large data sets. Statistical methods rely on testing hypothesis or finding correlations based on testing hypothesis or finding correlations based on smaller, representative samples of al larger population. However data mining methods are suitable for large data sets and can be more readily automated. In fact data mining algorithm often require large data sets for the creation of good model.

In case of business perspective it doesn't matter what is called Statistics, data mining or predictive analysis. Competitive advantage comes from decision making faster and more confidential. Here the question arises the difference between both the fields.

Let us see the definition of statistics "Statistics is the study of the collection, organization, analysis, interpretation and presentation of data[1]. It deals with all aspects of data, including the planning of data collection in terms of the design of surveys and experiments [1]". Its main goal is to extend knowledge about a subset of a collection to the entire collection. It draws valid conclusions and makes reasonable decisions on the basis of such analysis. Current and historical data in order to make predictions about the future events.

As far as data mining is concern it forms the predictive analysis that uses a variety of techniques to explore large amount of data to identify relationships between hundreds of data elements relationships that could not be uncovered through simple queries or reports. Data mining methodologies overlap with those in analysis disciplines such as statistics simulation, principle components and Bayesian methods. Forecasting and operations research.

Predicting customer behavior, identifying fraud and optimizing goods in supply chain often require a combination of analytical disciplines.

6. DATA MINING VS STATISTICS:

Data mining is categorized as either Descriptive or Predictive. Descriptive data mining is to search massive data sets and discover the locations of unexpected structures or relationships, patterns, trends, clusters, and outliers in the data. On the other hand, Predictive is to build models and procedures for regression, classification, pattern recognition, or machine learning tasks, and assess the predictive accuracy of those models and procedures when applied to fresh data.

The mechanism used to search for patterns or structure in high-dimensional data might be manual or automated; searching might require interactively querying a database management system, or it might entail using visualization software to spot anomalies in the data. In machine-learning terms, descriptive data mining is known as unsupervised learning, whereas predictive data mining is known as supervised learning.

Most of the methods used in data mining are related to methods developed in statistics and machine learning. Foremost among those methods are the general topics of regression, classification, clustering, and visualization. Because of the enormous sizes of the data sets, many applications of data mining focus on dimensionality-reduction techniques (e.g., variable selection) and situations in which high-dimensional data are suspected of lying on lower-dimensional hyperplanes. Recent attention has been directed to methods of identifying high-dimensional data lying on nonlinear surfaces or manifolds.

There are also situations in data mining when statistical inference — in its classical sense — either has no meaning or is of dubious validity: the former occurs when we have the entire population to search for answers, and the latter occurs when a data set is a “convenience” sample rather than being a random sample drawn from some large population. When data are collected through time (e.g., retail transactions, stock-market transactions, patient records, weather records), sampling also may not make sense; the time-ordering of the observations is crucial to understanding the phenomenon generating the data, and to treat the observations as independent when they may be highly correlated will provide biased results.

Data mining involves search architecture and require evaluation of hypothesis at a stage of search, evaluation of the search output, and appropriate use of the results. Statistics has little to offer in search architecture but helps a lot to evaluate hypothesis in course of a search for result evaluation.

Data mining does not replace traditional statistical techniques. It is an extension of statistical methods. There are number of statistical techniques that are being used in various tasks of data mining. These techniques include features of probability, distributions estimation, hypothesis testing model, scaring, Gibb's sampling, rational decision making, and casual inference and so on.

7. DIFFERENCE BETWEEN BOTH THE DISCIPLINES:

Statistics is concerned with probabilistic models, specifically inference on these models using data. Data Mining is (as I understand it) applied machine learning. It focuses more on the practical aspects of deploying machine learning algorithms on large datasets. It is very much similar to machine learning. Data mining exercise plays no role in data collection strategy, in this way it differs much of statistics therefore data mining some time referred to as secondary data analysis.

Statistics is just about the numbers, and quantifying the data. There are many tools for finding relevant properties of the data but this is pretty close to pure mathematics. Data Mining is about using Statistics as well as other programming methods to find patterns hidden in the data so that you can explain some phenomenon. Data Mining builds intuition about what is really happening in some data and is still little more towards math than programming, but uses both.

Statistics depends on assumptions as normal distribution, available characteristics, mean and standard deviation ability, statistics is described as being characterized by data sets which

are small and clean which permit straight forward answers via intensive analysis of single data sets.

Statistics has an emphasis that comes from mathematical statistics from computing with small data sets while knowledge discovery in data has a spin that comes from database methodology and from computing with large data sets.

8. CONCLUSION:

Data mining and statistics both the fields are working with data sets and used as the predictive analysis tools both the disciplines are having some similarity and differences. Hence there is some overlap in both the disciplines, But if we are strict on the definition “Statistics” or statistical techniques are not data mining. The challenge of collection, exploring and disentangling complicated interrelationships among various characteristics of data is what makes statistical analysis a rewarding activity. Data mining emphasize the new problems and opportunities that arise from data warehousing, and from the creation of new, often large database.

REFERENCES:

1. ^ [a] [b] Dodge, Y. (2006) *The Oxford Dictionary of Statistical Terms*, OUP. ISBN 0-19-920613-9
2. "Statistics". *Merriam-Webster Online Dictionary*.
3. ^ Moses, Lincoln E. (1986) *Think and Explain with Statistics*, Addison-Wesley, ISBN 978-0-201-15619-5 . pp. 1–3
4. ^ Chance, Beth L.; Rossman, Allan J. (2005). "Preface". *Investigating Statistical Concepts, Applications, and Methods*. Duxbury Press. ISBN 978-0-495-05064-3.
5. ^ [a] [b] [c] [d] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2011-10

6. ^ [a] [b] [c] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008.