Statistics and Data Mining

Zhihua Xiao
Department of Information System and Computer Science
National University of Singapore
Lower Kent Ridge Road Singapore 119260
xiaozhih@iscs.nus.edu.sg

Abstract
From a statistical perspective Data Mining can be viewed as computer automated exploratory data analysis of 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 report will discuss the discrepancies of these two fields and give a survey of the work that has been done by current researchers and states what can be done next.

1. Introduction

Data Mining is used to discover patterns and relationships in data, with an emphasis on large observational data bases. It sits at the common frontiers of several fields including Data Base Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization. From a statistical perspective Data Mining can be viewed as computer automated exploratory data analysis of 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. Machine learning methods form the core of Data Mining Decision tree learning (g or rule induction) is one of the main Components of several Data mining algorithms, though there are also some difference from standard machine learning methods.

The field of data mining, like statistics, concerns itself with learning from data or turning data into information. The context encompasses statistics, but with a somewhat different emphasis. In particular, data mining involves retrospective analyses of data: thus, topics such as experimental design are outside the scope of data mining and fall within statistics proper. Data miners are often more interested in understandability than accuracy or predictability. Thus, there is a focus on relatively simple interpretable models involving rules, trees, graphs, and so forth. Applications involving very large numbers of variables and vast numbers of measurements are also common in data mining. Thus, computational efficiency and scalability are critically important, and issues of statistical consistency maybe a secondary consideration.

Furthermore, the current practice of data mining is often pattern-focused rather than model-focused, i.e., rather than building a coherent global model which includes all variables of interest; data mining algorithms (such as any of the many rule induction systems on the market) will produce sets of statements about local dependencies among variables (in rule form).

In this overall context, current data mining practice is very much driven by practical computational concerns. However, in focusing almost exclusively on computational issues, it is easy to forget that statistics is in fact a core component Data mining without proper consideration of the fundamental statistical nature of the inference problem is indeed to be avoided.

In this report, I will discuss from the perspective of Statistics, the main characteristics of Data Mining, its shortcomings and how statistics can, offered for the data mining process, what nowadays research have done, what still need to be Completed in future.
2. Data Mining

Data Mining (DM) is at best a vaguely defined field; its definition largely depends on the background and views of the definer. This also represents a main characteristic of it:

From Pattern Recognition *Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data;*

From Data Base, *Data Mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions;*

From machine learning, *Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data; data mining is: decision trees, neural networks, rule induction, nearest neighbors, genetic algorithms .......*

2.1 Primary goals of data mining

We can distinguish the primary goals of data mining by two types: verification and discovery. A data mining system of *verification* goal is designed to verifying the user’s hypothesis. A data mining system of *discovery* goal is designed to automatically finds new patterns for the user. We can further subdivide the discovery goal into *prediction*, where the system finds patterns for the purpose of predicting the future behavior of some entities; and *description*, where the system finds patterns for the purpose of presenting them to a user in a human-understandable form. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Description focuses on finding human-interpretable patterns describing the data.

Due to the complexity of the stored data, and of the data interrelations, verification-driven data mining is not sufficient for decision-making. It must be complemented with the discovery-driven data mining. Furthermore, in the context of Data Mining, description tends to be more important than prediction. This is contrast to pattern recognition and machine learning applications where prediction is often the primary goal.

2.2 Most Popular Data Mining Methods

There are many different techniques for data mining. Often which technique you choose to use is determined by the type of data you have and the type of information you are trying to determine from the data. The most popular data mining methods in current use are classification, clustering, neural networks, association, sequence-based analysis, estimation, and visualization. We elaborately explain each method in the below.

*Classification*

*Classification* employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The use of classification algorithms begins with a training set of pre-classified examples. The classification algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Once an effective classifier is developed, it is used in a predictive mode to classify new records into these same predefined classes. For example, a classifier capable of identifying risky loans could be used to aid in the decision of whether to grant a loan to an individual.
**Clustering**

Clustering approach addresses the segmentation problems. These approaches assign records with a large number of attributes into a relatively small set of groups or clusters. This assignment process is performed automatically by clustering algorithms that identify the distinguishing characteristics of the data set and then partition the n-dimensional space defined by the data set attributes along natural cleaving boundaries. There is no need to identify the groupings desired or the attributes that should be used to segment the data set.

Clustering is often one of the first steps in data mining analysis. It identifies groups of related data that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these clusters with respect to some desired outcome.

**Neural Network**

A neural network is a network of many very simple processors ("units"), each possibly having a (small amount of) local memory. The units are connected by unidirectional communication channels ("connections"), which carry numeric (as opposed to symbolic) data. The units operate only on their local data and on the inputs they receive via the connections. For a neural network algorithm, examples are provided one at a time. For each of these examples, the actual output vector is computed and compared to the desired output. Then weights and thresholds are adjusted, where weights that contributed to the correct output remain unchanged, and weights that contributed to an incorrect output are decreased if the actual value is lower than the desired value.

The algorithm terminates when all weights stabilize.

**Association**

Association approaches address a class of problems typified by a market-basket analysis. Usually we are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit to the market. The goal is to discover all associations such that the presence of one set of items in a transaction implies the presence of another set of items.

**Sequential-based System**

Given a database of transactions over a period of time, the goal of sequential-based system is to find inter-transaction patterns such that the presence of a set of items is followed by another set of items. Traditional market-basket analysis deals with a collection of items as part of a point-in-time transaction. A variant of this problem occurs when there is additional information to tie together a sequence of purchases (for example, an account number, a credit card, or a frequent buyer/flyer number) in a time series. In this situation, not only may the coexistence of items within a transaction be important, but also the order in which those items appear across ordered transactions and the amount of time between transactions.

Rules that capture these relationships can be used, for example, to identify a typical set of precursor purchases that might predict the subsequent purchase of a specific item. In health care, such methods can be used to identify both routine and exceptional courses of treatment.

**Estimation**

A variation on the classification problem involves the generation of scores along various dimensions in the data. Rather than employing a binary classifier to determine whether a loan applicant is a good or bad risk, this approach generates a credit-worthiness "score" based on a training set, which has been assigned the scores.
Visualization

Visualization provides analysts with visual summaries of data from a database. It can also be used as a method for understanding the information extracted using other data mining methods.

Features that are difficult to detect by scanning rows and columns of numbers in database, often become obvious when viewed graphically. Data mining necessitates the use of interactive visualization techniques that allow the user to quickly and easily change the type of information displayed, as well as the particular visualization method used. Visualizations are particularly useful for noticing phenomena that hold for a relatively small subset of the data, and thus are "drowned out" by the rest of the data when statistical tests are used since these tests generally check for global features.

Almost none of:

Hypothesis testing
Experimental design
Response surface modeling
Linear regression
Discriminant analysis
Logistic regression
Canonical correlation Principal components
Factor analysis

These latter procedures are of course the mainstay of our standard statistical package. So the statistics methodology has largely been ignored[2].

Although Data Mining appears to be a viable commercial enterprise, Those from academic community, who have accustomed to the strict theoretical structure of the logical formalism system, argue that whether Data Mining methodology is an intellectual discipline so far, Especially from the perspective of statistical data analysis, the answer is not yet. Data Mining packages implement well known are the procedures from the fields of Machine learning, pattern recognition, neural network and data visualization, they emphasize look and feel and the existence of functionality. The goal is to get to market quickly, most academic research in this area so far has focused on incremental modifications to current machine learning methods, speed-up or parallel implementation of existing algorithms. Data mining will need a sound theory base from Statistics, as well as this Statistics need to adapt to the new circumstance of the large real world data base.

The challenges for Data Mining also include[3]:

Large Data base: hundreds of fields, tables; millions of records;
High Dimensionality;
Over-fitting;
Assessing statistical significance: if search is over many models, some models will fit well based purely on chance.
Changing data and knowledge;
Missing and noisy data;
Complex relationships between fields;
Understandability of patterns;
User interaction and prior knowledge;
Integration with other systems.
3. What is Statistics?

Since its creation in the 18th century, statistics have served this purpose, providing the mathematical tools and analytic techniques for dealing with large amounts of data. Today, as we are confronted with increasingly large volumes of data, statistics are, more than ever, a critical component of the data mining and refining toolkit that facilitates making effective business decisions.

Statistics consists of two main parts, descriptive and inferential statistics. The methodology for organizing and summarizing the data for the sample is called descriptive statistics. When we attempt to use these summaries to draw conclusions about an entire population, we employ the methodology called as statistical inference[1].

This section briefly describes some of the central statistical ideas we think relevant to data mining, and will emphasis on statistical inference related contents.

1. Descriptive statistical

   **Probability distributions.**

   The statistical literature contains mathematical characterizations of a wealth of probability distributions, as well as properties of random variables functions defined on the events to which a probability measure assigns values. Important relations among probability distributions include marginalization (summing over a subset of values) and conditionalization (forming a conditional probability measure from a measure on a sample space and some event of positive measure). Essential relations among random variables include independence, conditional independence, and various measures of dependence, of which the most famous is the correlation coefficient. The statistical literature also characterizes families of distributions by properties that are useful in identifying any particular member of the family from data, or by closure properties useful in model construction or inference, for example conjugate families, closed under conditionalization, and the multi-normal family, closed under linear combination.

   The main purpose of Data Mining is to describe the data so a knowledge of the properties of distribution families can be invaluable in analyzing data and making appropriate inferences[].

2. Inferential statistics

   **Estimation.**

   An estimator is a function from sample data to some estimand, such as the value of a parameter. When the data comprise a sample from a larger actual or potential collection governed by some probability distribution, the family of estimators corresponding to all possible samples from that collection also has a probability distribution. Classical statistics investigates such distributions of estimators in order to establish basic properties such as reliability and uncertainty. Estimation almost always requires some set of assumptions. Such assumptions are typically false, but often useful. If a model (which we can think of as a set of assumptions) is incorrect, estimates based on it can be expected to be incorrect as well. Bayesian estimation emphasizes that alternative models and their competing assumptions are often plausible.

   **Hypothesis testing.**

   Since statistical tests are widely used, some of their important limitations should be noted.

   Viewed as a one-sided estimation method, hypothesis resting is inconsistent unless the alpha level of the testing rule is decreased appropriately as the sample size increases. Generally, an level test of one hypothesis and an level test of another hypothesis do not jointly provide an level test of the conjunction of the two hypotheses. An important corollary for data mining is that the level of a
test has nothing directly to do with the probability of error in a search procedure that involves testing a series of hypothesis. In data mining procedures that use a sequence of hypothesis tests, the alpha level of the tests cannot generally be taken as an estimate of any error probability related to the outcome of the search.

Model scoring.

The evidence provided by data should lead us to prefer some models or hypotheses to others, and to be indifferent between still other models. A score is any rule that maps models and data to numbers whose numerical ordering corresponds to a preference ordering over the space of models, given the data. For the reasons just considered, scoring rules are often an attractive alternative to tests. Typical rules assign models a value determined by the likelihood function associated with the model, the number of parameters, or dimension, of the model, and the data.

Given a prior probability distribution over models, the posterior probability on the data is itself a scoring function, arguably a privileged one. The Bayes Information Criterion approximates posterior probabilities in large samples. There is a notion of consistency appropriate to scoring rules; in the large sample limit, almost surely the true model should be among those receiving maximal scores. There are also uncertainties associated with scores, since two different samples of the same size from the same distribution may yield not only different numerical values for the same model, but even different orderings of models. For obvious combinatorial reasons, it is often impossible when searching a large model space to calculate scores for all models; it is, however, often feasible to describe and calculate scores for a few equivalence classes of models receiving the highest scores. In some contexts, inferences made using Bayesian scores can differ a great deal from inferences made with hypothesis tests.

Markov Chain Monte Carlo.

Historically, insurmountable computational difficulties forced data analysts to eschew exact analysis of elaborate hierarchical Bayesian models and complex likelihood calculations. Recent dramatic advances in Monte Carlo methods have, however, liberated analysts from some of these constraints. One particular class of simulation methods, dubbed Markov Chain Monte Carlo, originally developed in statistical mechanics, has revolutionized the practice of Bayesian statistics[13].

Generalized model classes.

A major achievement of statistical methodological research has been the development of very general and flexible model classes. Especially graphical models represent probabilistic and statistical models with planar graphs, where the vertices represent (possibly latent) random variables and the edges represent stochastic dependencies. This provides a powerful language for describing models and the graphs themselves make modeling assumptions explicit.

Graphical models provide important bridges between the vast statistical literature on multivariate analysis and such fields as artificial intelligence, causal analysis, and data mining.

One often used graphical model is Bayesian network,

Bayesian networks

A Bayesian network for a set of variables \( X = X_1, \ldots, X_n \) consists of a network structure \( S \) that encodes a set of conditional independence assertions about variables in \( X \), and; a set \( P \) of local probability distributions associated with each variable[6].

Together, these components define the joint probability distribution for \( X \). The network structure \( S \) is a directed acyclic graph. The nodes in \( S \) are in one-to-one correspondence with the variables \( X \). The probabilities encoded by a Bayesian network may be Bayesian or physical.
When building Bayesian networks from prior knowledge alone, the probabilities will be Bayesian. When learning these networks from data, the probabilities will be physical (and their values may be uncertain). In subsequent sections, we describe how we can learn the structure and probabilities of a Bayesian network from data, which is the process of using Bayesian network for data mining.

4. Using Statistics in Data Mining

Data Mining is a data-centralized process, so the property of the data determine how to design the algorithm. There are always problems with real world data, which data mining must face, they can be classified into five groups of: ultra large data, Noisy data, incomplete data, redundant data and dynamic data. Data driven techniques either rely on heuristics to guide their search through the large space of possible relations between combinations of attribute values or adopt some kind of data-reduction method to make the algorithm more efficient.

4.1 Ultra Large Data and data sampling

One of the important issues in data mining is related to the volume of data because many knowledge discovery techniques, involving exhaustive search over instance space, are highly sensitive to the size of data in terms of time complexity.

Current focus of statistics has gradually moved from model estimation to model selection.

Instead of looking for the parameter values that make a model fit the data well, also the model structure is part off the search process. This trend fits the goals of Data Mining nicely one does not want to fix the model structure in advance. The recent advances in, say Markov chain Monte Carlo (MCMC) methods[13], make it possible to consider far larger model spaces than previously. In addition to these techniques, the Data Mining community has lots to learn from statistics, e.g., in the handling of uncertainty. The main difference between Data Mining and statistics is perhaps in the extensive use of machine learning methods in Data Mining, in the volume of data, and in the role of computational complexity issues in Data Mining. For example, even MCMC method have difficulties in handling tens of thousands of parameter values; some sort of combinatorial preprocessing is needed to make the model selection task tractable. It seems that such combinations of methods can be useful: combinatorial techniques are used to prune the search space, and statistical methods are used to explore the remaining parts in great detail.

Data Sampling

The DM community may have to moderate its romance with big [4]. A prevailing attitude seems to be that unless an analysis involves gigabytes or terabytes of data, it cannot possibly be worthwhile. It seems to be a requirement that all of the data that has been collected must be used in every aspect of the analysis. Sophisticated procedures that cannot simultaneously handle data sets of such size are not considered relevant to DM. Most DM applications routinely require data sets that are considerably larger than those that have been addressed by our traditional statistical procedures (kilo-bytes). However, it is often the case that the questions being asked of the data can be answered to sufficient accuracy with less than the entire (giga or terabyte) data base. Sampling methodology, which has a long tradition in Statistics, can probably be used to improve accuracy while mitigating computational requirements. Also, a powerful computationally intense procedure operating on a sub-sample of the data may in fact provide superior accuracy than a less sophisticated one using the entire database. Of particular importance to data sampling are cases in which the algorithm used for data analysis requires a subset of the entire data, either for splitting the data for training/testing or evaluating the performance of the data analysis algorithm through an iterative process of varying the sample size, such as in neural networks applications. The
important issue here is the correct choice of samples in order to obtain and preserve the best possible performance for the algorithm in use.

For example, in neural networks applications, one usually has only a small set of correctly classified patterns, known as the training set. The remained main question is: does the given training set fairly represent the underlying class conditional probability density functions.

4.2 Incomplete data, Noisy data and Bayesian network

Bayesian network has many characteristics suited for Data Mining. It is a graphical model for probabilistic relationships among a set of variables. Besides Bayesian networks can readily handle incomplete data sets. For example, consider a classification or regression problem where two of the explanatory or input variables are strongly anti-correlated. This correlation is not a problem for standard supervised learning techniques, provided all inputs are measured in every case. When one of the inputs is not observed, however, many models will produce an inaccurate prediction, because they do not encode the correlation between the input variables. Bayesian networks offer a natural way to encode such dependencies.

Two, Bayesian networks allow one to learn about causal relationships. Learning about causal relationships are important for at least two reasons. The process is useful when we are trying to gain understanding about a problem domain, for example, during exploratory data analysis. In addition, knowledge of causal relationships allows us to make predictions in the presence of interventions. For example, a marketing analyst may want to know whether or not it is worthwhile to increase exposure of a particular advertisement in order to increase the sales of a product. To answer this question, the analyst can determine whether or not the advertisement is a cause for increased sales, and to what degree. The use of Bayesian networks helps to answer such questions even when no experiment about the effects of increased exposure is available.

Three, Bayesian networks in conjunction with Bayesian statistical techniques facilitate the combination of domain knowledge and data. Anyone who has performed a real-world modeling task knows the importance of prior or domain knowledge, especially when data is scarce or expensive. The fact that some commercial systems (i.e., expert systems) can be built from prior knowledge alone is a testament to the power of prior knowledge. Bayesian networks have a causal semantics that makes the encoding of causal prior knowledge particularly straightforward. In addition, Bayesian networks encode the strength of causal relationships with probabilities. Consequently, prior knowledge and data can be combined with well-studied techniques from Bayesian statistics.

Four, Bayesian methods in conjunction with Bayesian networks and other types of models offers an efficient and principled approach for avoiding the over fitting of data.

Using the Bayesian network in Data Mining is the process of building the Bayesian network for a given data set.

Reference