Chapter 2
Data Mining and Statistics: Tools for Decision Making in the Age of Big Data

Hirak Dasgupta
Symbiosis Institute of Management Studies, India

ABSTRACT

In the age of information, the world abounds with data. In order to obtain an intelligent appreciation of current developments, we need to absorb and interpret substantial amounts of data. The amount of data collected has grown at a phenomenal rate over the past few years. The computer age has given us both the power to rapidly process, summarize and analyse data and the encouragement to produce and store more data. The aim of data mining is to make sense of large amounts of mostly unsupervised data, in some domain. Data Mining is used to discover the patterns and relationships in data, with an emphasis on large observational data bases. This chapter aims to compare the approaches and conclude that Statisticians and Data miners can profit by studying each other's methods by using the combination of methods judiciously. The chapter also attempts to discuss data cleaning techniques involved in data mining.

INTRODUCTION

The fact that there has been a recent increase in the interest shown by many in the field of data mining or knowledge discovery or machine learning, has surprised many statisticians. Data mining attacks problems of descriptive data (i.e. effective summaries of data), identifies relationships among variables within a data set and uses a set of previously observed data to construct predictors of future observations. A well-established set of techniques for attacking all these problems have been developed by statisticians. Various algorithms and techniques such as: Statistics, Clustering, Regression, Decision trees, association rules, neural networks etc. are used for making predictions and also used in data mining.

Data mining, as it is practised at present, has evolved over nearly four decades, since the use of computers and accessories started being used for data collection and static data provision. Relational database management Systems (RDBMS) and Structured Query languages (SQL) were developed during the 80s

DOI: 10.4018/978-1-5225-2031-3.ch002

and 90s for providing dynamic data at the level of the record. Subsequently, online data processing and multi-dimensional databases and data warehouses came to be used (Cios et al., 2010).

The purpose of data mining is knowledge discovery. It extracts hidden information from large databases and hence is a powerful technology with a great potential for companies to focus on the analysis of the stored database (Adejuwon & Mosavi, 2010).

Both the techniques—Data mining and Statistic—use some common software packages by the software vendors (IBM, SAS, and many more). By strict definition "statistics" or statistical techniques are not data mining. They were being used long before the term data mining was coined to apply to business applications. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models. And from the users' perspective you will be faced with a conscious choice when solving a "data mining" problem as to whether you wish to attack it with statistical methods or other data mining techniques. Today people have to deal with up to terabytes of data and have to make sense of it and glean the important patterns from it. Statistics can greatly help in this process by helping to answer several important questions about their data: what patterns are there in the 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 some idea of what is contained in the database? For these reasons, it is important to have some idea of how statistical techniques work and how they can be applied. Data miners should have a foundation of knowledge in Statistics. Data mining is an interdisciplinary field with contributions from statistics, artificial intelligence, and decision theory and so on (Yahia & El-Mukashfi El-Taher, 2010).

Data mining is not just an "umbrella" term coined for the purpose of making sense of data. The major distinguishing characteristic of data mining is that it is data driven, as opposed to other methods that are often model driven. In statistics, researchers frequently deal with the problem of finding the smallest data size that gives sufficiently confident estimates. In data mining we deal with the opposite problem, namely, data size is large and we are interested in building a data model that is small (not too complex) but still describes the data well (Cios et al., 2010).

In other words, the essential difference between data mining and the traditional data analysis (i.e. statistics) is that data mining is to mine information and discover knowledge on the premise of no clear assumption.

Some definitions on data mining given by different authors are as follows:

- "Data mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules" (Linoff & Berry, 2014).
- "Statistics with Scale and Speed" (Darryl Pregibon).
- "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" (Hand, Mannila, & Smyth, 2001).
- "Statistics is at the core of data mining - helping to distinguish between random noise and significant findings, and providing a theory for estimating probabilities of predictions, etc. However Data Mining is more than Statistics. Data mining covers the entire process of data analysis, including data cleaning and preparation and visualization of the results, and how to produce predictions in real-time, etc" (Gregory Piatetsky-Shapiro).