

Review of Applications of Statistics in Data Mining

Prof. Shriram Kargaonkar
MAEER's MIT Arts, Commerce & Science
College, Dehu Phata Road, Alandi - 412105.
E-mail: snkargaonkar@mitacsc.ac.in

Prof. Archana Kadam
MAEER's MIT Arts, Commerce & Science
College, Dehu Phata Road, Alandi - 412105.
E-mail: appawar@mitacsc.ac.in

Abstract

Data mining (DM) is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified into useful information. This information can be used to increase revenue, cuts costs, or both. From the point of view of Statistics, data mining can be viewed as the computer automated exploratory data analysis of large and complex data sets. The concepts in Statistics are widely used in data mining for analyzing large data sets and for finding correlations or patterns among various fields in large relational databases. Thus, Data mining is the application of Statistics in the form of exploratory data analysis and prediction models to reveal patterns and trends in large datasets. Data Mining is also called as Knowledge Discovery in Databases (KDD). The aim of this paper is to focus on applications of Statistics used in Data mining process.

Keywords: Data Mining, Statistics, Complex Data Sets, Data Mining Process and KDD.

INTRODUCTION

What is Data Mining?

Data mining is the extraction of hidden predictive information from large data and then transforms it into understandable structure. It predicts future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining is the core of Knowledge Discovery Process. Data mining process consists of an iterative sequence of several steps/process: data pre-processing; data management; data mining tasks and algorithms and post-processing. The applications of data mining and statistics cover a very wide spectrum. The most relevant fields are those where large volumes of data have to be analyzed. A survey on the www.kdnuggets.com portal in July 2005 revealed the main fields where data mining is used: banking (12%), customer relationship management (12%), direct marketing (8%), fraud detection (7%), insurance (6%), retail (6%), telecommunications (5%), scientific research (4%), and health (4%). Data mining is also called as Knowledge Discovery in Database (KDD).