Chapter 19
Predictive Analytics and Data Mining:
A Framework for Optimizing Decisions with *R* Tool

Ritu Chauhan
Amity University, India

Harleen Kaur
Hamdard University, India

ABSTRACT

High dimensional databases are proving to be a major concern among the researches to extract relevant information for futuristic decision making. Real world data is high dimensional in nature and comprises of irrelevant features, missing values, and redundancy, which requires serious concerns. Utilizing all such features can mislead the results for emergent prediction. Therefore, such databases are critical in nature to determine optimal solutions. To deal with such issues, the authors have developed and implemented a Cluster Analysis Study Behavior of School Children from Large Databases (CABS) framework to retrieve effective and efficient clusters from high dimensional human behavior datasets for school children in US. They have applied feature selection technique and hierarchical agglomerative clustering technique to discover clusters of vivid shape and size to retrieve knowledge from large databases. This study was conducted for Health Behavior in School-Aged Children (HBSC) using Correlation-Based Feature Selection (CFS) technique to reduce the inconsistent data records and select relevant features that will eventually extract the appropriate data to merge similar data and retrieve clusters. However, predictive analytics can facilitate a more thorough extraction of knowledge to facilitate better quality and faster decisions. The authors have implemented the current framework in R language where the clustering was emphasized using pvclust package. The proposed framework is highly efficient in discovering hidden and implicit knowledge from large databases due to its accessibility to handling and discovering clusters of variant shapes.

DOI: 10.4018/978-1-4666-9562-7.ch019

INTRODUCTION

In past decades there has being explosive growth in raw data from various sources. There are vibrant ranges of existing data resources from distinct background such as medical technology, business oriented organization, market based analysis, social analysis, science exploration, geographical information studies, and several other computerization, technologies for retrieval of information and storage of data. Numerous efforts are accomplished by researchers to retrieve effective and efficient patterns to discover knowledge from such large databases (Agrawal et al., 1998). But it's not possible by human capabilities all alone to retrieve patterns from such terabytes of data, to overcome such problem data analysis technique such as pattern detection, rules generation and decision making are widely appreciated around the globe for retrieval of relevant information from large databases. In recent review of literature we have found that statistical analysis tools involve series of flaws while handling large amount of complex and high dimensional databases. The method involves generation of several hypotheses testing on data sets for analysis and retrieval of knowledge; hence it was difficult to retrieve effective and efficient patterns for decision making. These techniques prove to be expensive, complex and time consuming for users. Therefore, Statistical tools were proven to be an irrelevant tool utilized for retrieval of information for knowledge discovery (Miller and Han, 2001; Mannila, 2002).

To overcome such flaws of statistical technique analyses among large databases, industry has developed several functionalities for databases such as: collection of data, creation of database, management of data, and finally analysis of data for better perceptive which tends to be relevant key factors for future discoveries of data (Han and Kamber, 2000). The last process involved in this development is defined as Knowledge Discovery in Database (KDD) (Branchmann and Anand, 1996). There are several definitions discussed in past for KDD, but most famous studied by Fayyad et al., 1996a. Several definitions for KDD are discussed below:

Definition 1: It is defined as a non-trivial process for retrieval of valid, novel, potentially useful and ultimately understandable patterns in data and describing them in a definite, concise and meaningful way (Fayyad et al., 1996b).

Data mining can be termed as most important step in KDD for retrieval of information or knowledge from raw data. Several KDD processes are followed by data mining techniques for extraction of patterns which are useful to discover hidden knowledge from large databases (Smyth, 2001). Therefore, without these valid additional steps there exists high risk of uninteresting and non valid patterns (Chen et al., 1996). Therefore, Data mining tasks can be defined as a process to extract hidden implicit interesting patterns from large databases for knowledge discovery process. It can be vitally defined as learning process to generate patterns from raw data automatically by building a computer based environment for complex datasets. There are several definitions discussed in past by researchers for data mining techniques such as:

Definition 2: The goal of Data mining focuses on retrieval of deep hidden information which can be utilized for knowledge discovery process for strategic decision making and equating fundamental research problems *(*Miller and Han, 2001).

Data mining techniques involves several of statistical techniques as well as sophisticated data analysis technique to retrieve effective and efficient patterns from large and complex databases (Afifi and Azen, 1972). The real world datasets