Title of thesis : Clustering and Classification techniques in Health Care System

Name of Scholar : Parvesh Kumar

Name of Supervisor : Prof. S.K. Wasan

Department : Mathematics

Faculty : Natural Sciences

ABSTRACT:

Knowledge discovery in databases (KDD) is an independent research discipline, which discovers information from large amount of data. KDD is a series of processes including data collection, data preprocessing, data transformation, data mining and knowledge presentation. Among these processes data mining is a vital step.

Data mining is the process of extracting non-trivial, implicit, previously unknown and potentially useful information or patterns from large data repositories using applications of special algorithms built upon sound principles from numerous disciplines including statistics, artificial intelligence, machine learning, database science, and information retrieval. Generally speaking, there are two classes of data mining descriptive and predictive. A predictive mining makes a forecast about values of the data using known results found from different data. Predictive data mining tasks include classification, regression, time-series. Descriptive mining is to summarize or characterize general properties of data in data repository. Unlike predictive mining, a descriptive mining serves as a way to explore the properties of the data examined, not to predict new properties. Clustering, summarization, association rules and sequence discovery are usually considered as descriptive in nature. Clustering and classification are two important techniques among these various data mining techniques. In classification, we place objects into predefined classes using a classification algorithm whereas in clustering, we map objects into different groups, where objects in a group are similar to each other and dissimilar to the objects of another group. Classification algorithms are supervised (because class labels are predefined) while clustering algorithms are unsupervised (because class labels are not predefined).

There is a vast potential for data mining applications in healthcare. Data mining techniques, such as pattern association, classification and clustering, are

now frequently applied in cancer and gene expressions correlation studies. To classify the various types of cancer into its different subcategories, different data mining techniques have been used over gene expression data. There are many instances of reportedly successful applications of both hierarchical clustering and partitioning clustering in gene expression analyses. Yeung compared *k*-means clustering, CAST (Cluster Affinity Search Technique), single-, average- and complete-link hierarchical clustering, and totally random clustering for both simulated and real gene expression data. They favoured *k*-means and CAST. Gibbons compared *k*-means, SOM (Self-Organizing Map), and hierarchical clustering of real temporal and replicate microarray gene expression data.

In this thesis, different classification and clustering techniques are studied over different cancer datasets. Various new and old algorithms from classification and clustering techniques are used to classify the various types of cancer into its different subcategories. After giving a brief introduction of data mining, we introduce related concepts. The entire work is divided into six chapters. Chapter-1 deals with various concepts related with data mining and introduce various data mining techniques. It also describes related work done in literature. Chapter-2 includes the necessary theoretical foundations for the clustering and classification techniques. It describes various data types and data formats used in clustering. Dissimilarity and similarity measures are also discussed. We also describe different types of clustering and classification algorithms.

Chapter-3 discusses the variations of k-means algorithm which includes k-means, global k-means, x-means, k-means++, rough k-means, fuzzy c-means and efficient k-means algorithms. In order to explore the strength and weaknesses an attempt has been made to compare the variations of k-mean algorithms using high dimensional cancer datasets. Results of our study over different cancer datasets using these variations are shown in this chapter

Chapter-4 contains a comparative study of various clustering algorithms namely k-means, rough k-means, PAM, CLARA, EM and Accelerated EM to classify the cancer datasets. Comparison is made in respect of accuracy and ability to handle high dimensional data.

Chapter 5 includes a comparative analysis of various classification algorithms like C4.5, CART, Random Forest, LMT, ADT, Naïve Bayesian and Bayesian logistic Regression using cancer datasets. Results of cancer classification using these algorithms are also discussed in this chapter.

Chapter 6 presents the conclusion of the thesis and provides suggestions for further research