Abstract

The Data Mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns. There are several major Data Mining techniques that have been developed and used in Data Mining projects. These techniques include association, classification, clustering, prediction and sequential patterns. Fuzzy logic is capable of supporting, to a reasonable extent, human type reasoning in natural form by allowing partial membership for data items in fuzzy subsets. Integration of fuzzy logic with Data Mining techniques has become one of the key constituents of soft computing in handling the challenges posed by the massive collection of natural data. This thesis is an extensive coverage of integration of fuzzy logic with various Data Mining techniques to handle vagueness and uncertainty at various levels to incorporate human type reasoning in modeling.

Since Data Mining is a computer based technology it is quite natural to build the principles of Data Mining using Boolean theory. But by using Data Mining techniques, what we aim is to model the behavior patterns of human interaction with various systems. The human behavior is always fuzzy in nature and it is very difficult to model human behavior patterns with crisp Boolean Data Mining systems. The modeling of imprecise and qualitative knowledge, as well as handling of uncertainty at various stages is possible through the use of fuzzy sets.

In Data Mining, Clustering is used for segmenting objects into similar groups. Using crisp clustering technique an object can become the member of only one cluster and the membership is either fully inclusive or exclusive. A review of crisp
clustering techniques and illustration of k-means algorithm is performed in this work to have clear idea about clustering principles. The humanistic approach to object segmentation is different that an object can become the member of more than one cluster at a time with varying degrees of association. Using fuzzy clustering techniques, we will be able to implement Data Mining segmentation exactly the way humans organize objects. The central idea in fuzzy clustering is the non-unique partitioning of the data in a collection of clusters with membership values between zero and one. The non-zero membership values with a maximum of one show the degree to which the data point belongs to a cluster. Fuzzy C-Means (FCM) algorithm incorporates the fuzzy concepts in clustering and this is used in vide range of application domains. In the thesis, the FCM algorithm is introduced and illustrated to bring out the advantages and limitations of the algorithm.

This research work includes an extensive analysis of the merits and demerits of fuzzy c-means clustering algorithm. The limitations of fuzzy c means algorithm including the problems associated with convergence and outlier points, lead to the development of many modified versions of fuzzy c means algorithm. The recent contributions reported in this area, which address the limitations of the c means algorithm, are reviewed. An efficient modification to the c means algorithm is recommended to address the projected limitations of c-means algorithm. In the modification, a novel, efficient and stable membership function for the fuzzy clustering algorithm is proposed.

Traditional clustering methods require the user to determine the number of clusters before we start any data exploration. In fuzzy clustering algorithms, the performance efficiency depends mainly on the initial estimation of clusters count and cluster seeds. This initial selection becomes a tedious task in clustering. The proposed fuzzy clustering membership function is used to devise a cluster count estimation algorithm. The new Fuzzy Threshold based Cluster Estimation
Algorithm (FTCEA) is presented and its performance is evaluated through experiments.

Other crisp Boolean Data Mining techniques like association rule mining, classification and prediction use intervals with crisp boundaries to divide the quantitative attributes into different groups. The intervals are applied with the conventional algorithms for finding patterns involving these attributes. Such techniques suffer from a problem called over estimation of boundary cases. In natural segmentation of quantitative values, the transition from one segment to another is not abrupt but gradual. These natural segments with overlapping boundaries can be modeled only by using fuzzy sets and linguistic variables. In Boolean Market Basket analysis also, we consider only the presence or absence of objects in a transaction to generate association rules. The factors like importance of items, the profit and the number of items in a transaction are not included in the Boolean association rule mining techniques. The importance of an item in a transaction could be represented using a fuzzy concept called weight of an item. Incorporating this idea, the unsupervised cluster estimation algorithm (FTCEA) is used to segment the quantitative attributes into natural linguistic segments. The resulting fuzzy linguistic segment values are used to form fuzzy association rules.

In web mining, the concept of a market basket can be extended as the pages visited by a user in one session. Association rule mining techniques are used here to find associations between web pages visited by users. The Boolean Apriori algorithm for association rule mining can be used to find the association between the web pages visited together by users. But by using Apriori algorithm, only the presence and absence of web pages in a browsing session can be considered. To get a better perspective of the underlying patterns we have to consider other important factors like the number of visits of a web category, the time spent on a web page etc. As an attempt to classify the web access sequences with respect to the importance of web categories, a novel Fuzzy Web Classification Algorithm (FWCA) using the
principles of fuzzy association rule mining is proposed. The efficiency of the algorithm is established through experiments.

Finally, the Fuzzy Data Mining concepts and methods presented in this thesis are applied on a real-time data set pertaining to the financial inclusion status of Idukki district of Kerala. This data is collected through a survey in the region and the information gathered is converted into a database table. The statistical analysis on this data provides insights about the actual situation in the region regarding financial inclusion. The data mining techniques in a fuzzy framework are applied on this data to reveal the patterns which lead to the present financial inclusion status of the area. The patterns evolved through this analysis reveal that education and exposure to modern technologies play the most significant role in financial inclusion.