Abstract

Clustering is the process of grouping a set of objects into one based on their similarity metric. From past many years Clustering has been one of the exploratory research areas for many of the researchers. We have a wide number of clustering algorithms from past research. But none of the algorithms are proven to be work well when we use large set of data items. The greater part of existing clustering algorithms stumbles upon severe scalability and/or accuracy allied problems when used on databases with a huge number of records and/or attributes. In this paper, we propose a hybrid technique, which is based on two concepts namely ID3 decision tree and PFCM. The new technique is able to overcome many of these shortcomings. The primary intension of the research is clustering the data with efficient technique. The key idea is to create an unsupervised decision tree based on the data set which maximizes the difference between the feature-value means of the children nodes. We achieve this by employing ID3 decision tree algorithm. By using only best splits we can reduce the time. To make the clustering efficient we use possibility fuzzy c-means clustering on the results obtained from ID3. This method aims for large scale data.

1. Introduction

As the human population grew in number, so did the data about them. There are many fields such as Business, statistical data analysis, Fraud detection Medical, Engineering which require analysis of data to find the necessities of people and improve services.. But with the mounting amount of
data and arrival of computing in a variety of fields, extracting useful information from this data using a variety of complicated mathematical models and information became possible. This extraction of hidden and useful information from large amounts of high dimensional databases is said to be “Data Mining” [1]. Data mining aims at providing the knowledge regarding the data in terms of their organization, outlining, symbolic demonstration of data sets, concealed facts, prediction about some unknown values, pattern recognition, data distribution etc [2]. Predictably, the information that is mined is represented as a form of the semantic structure of the datasets. The model might be utilized for calculation and classification of new data. In current years the sizes of databases has amplified hastily. This is the main reason of development of several automatic knowledge extraction data tools. The word Data Mining or Knowledge Discovery in databases has been adopted for a field of research dealing with the automatic detection of inherent information or knowledge within databases [3].

The data mining functionalities mainly include association rule mining, classification, prediction & clustering. Association analysis is used for finding interesting relations between variables in big databases, which is specified in the form of rules to user. Knowledge discovery in number of domains have been made possible from the past several years, because of the wide range of data mining techniques and their significant contributions. These techniques are used in banking sector for loan payment prophecy, customer credit policy investigation, categorization of customers for beleaguered marketing, and discovery of money laundering schemes and other financial crimes. Similarly, in the retail industry, such techniques are used in customer retention and analysis of product sales. In the telecommunication industry these techniques help in identifying and comparing data traffic, system workload, resource usage, profit and fraudulent pattern analysis [8].

1.1 Classification Vs Clustering
Classification is the process of predicting the class labels from the targeted class which is also known as supervised learning as shown in Figure 1. In contrast Clustering which is unsupervised learning is the process of grouping the data into classes or clusters without any prior knowledge so that intra cluster similarity is very high and intra cluster similarity is very much less as shown in Figure 2. The objective of supervised learning is to find a set of characteristic descriptions of these classes [7]. Clustering is important for a variety of information needs and applications such as collection management to analyze collections of grant proposals for research portfolio analysis, clustering techniques use to require some of distance or similarity between pairs [9].
Businesses countenance challenges such as expansion, policy, globalization, rivalry, and financial changes, which oblige quick and excellent decisions rather than guess work. Taking excellent decisions requires perfect and lucid analysis such as forecast, assessment, segmentation and classification using data mining techniques. To discover interesting patterns, Decision tree induction and Clustering are two of the most important data mining techniques. There are several saleable data mining software in the souk, and most of them supply decision trees induction and clustering data mining techniques [5]. Decision tree classifiers are reasonably quick as compared to other classification techniques. A decision tree can be transformed into simple and easy to understand classification rules [10]. It is eminent that decision trees are probably the most popular classification model. Frequently used decision trees are usually constructed using a top-down approach, from common concepts to meticulous examples. Top-Down Induction of Decision Trees, is used to refer to this kind of algorithms. The ultimate goal of the decision tree learning process is to construct a decision tree which conveys interesting information in order to make predictions and classify previously unseen data [4].

2. Related Work

The majority of the clustering techniques are used for purely numerical or categorical data only, but not to both. In general, it was a nontrivial job to execute clustering on mixed data such as numerical and categorical attributes because there exists an awkward gap among the similarity metrics for categorical and numerical data. S.Sumathi and M.M.Gowthul Alam [11] have presented a general clustering framework based on the concept of object-cluster similarity and gives a unified similarity metric which could be simply applied to the data with categorical, numerical, and mixed attributes. They proposed a initialization method for mixed data which was implemented using K – Modes algorithm and further and iterative fuzzy K – Modes clustering algorithm.

As a result of the growing IT and producing methods and collecting data, it was admitted that today the data could be warehoused faster in comparison with the past. Therefore, knowledge discovery tools are required in order to make use of data mining. Data mining was typically employed as an advanced tool for analyzing the data and knowledge discovery. Indeed, the purpose of data mining was to establish models for decision. These models have the ability to predict the future treatments according to the past analysis and are of the exciting areas of machine learning and adaptive computation. Statistical analysis of the data uses a combination of techniques and artificial intelligence algorithms and data quality information. To utilize the data mining applications, including the commercial and open source applications, numerous programs are currently available. Ida Moghimipour and Malihe Ebrahimipour [12] have introduced a data mining and principal concepts of the decision tree method which are the most effective and widely used classification methods. In addition, a succinct description of the three data mining software, namely SPSS-Clementine, RapidMiner and Weka was also provided. Afterwards, a comparison was performed on 3515 real datasets in terms of classification accuracy between the three different decision tree algorithms.

Big data was the popular term used in the current era for extracting knowledge from large datasets. Big data was the collection of large and complex dataset. The challenge in big data was volume, variety and velocity (3V’s).variety could be classified into structured, unstructured data and semi structured. Structured data are the identifiable data, which was organized in some structure. Data
stored in the relational database are example of structured data. Unstructured data are the data without identifiable structure, audio, video and images are few examples. All web and bioinformatics data comes under semi structure data which does not have any regular structure, it was neither structured nor semi structured. Clustering the one of the best technique in knowledge extraction process. It was nothing but grouping of similar data to form a clusters. The distance between the data in one clusters and other should not be less. Many algorithms are practiced for clustering, in that k-mean clustering is the one of the popular term for cluster analysis. Nelson et al. [13] have aimed to design an algorithm to partition the dataset into k clusters based on some computational value. The limitation of k-mean clustering was that it could be applied to either structured or unstructured, not in combination of both. Here the limitation is overcome by proposing k –mean algorithm for extracting hidden knowledge by forming clusters from the combination of both structure and unstructured dataset.

Avinash et al. [14] have proposed a technique for analysis of high dimensional databases. Data mining was the method of discovering or fetching useful information from database tables. Many methods to sequential data mining have been proposed to extract useful information, such as time series analysis, temporal association rules mining, and sequential pattern discovery. Several basic techniques are used in data mining for describing the type of mining and data recovery operation. The rapid growth of the Internet could be largely attributed to the loose governance structure, which beyond some control over domain names, was open to be freely added to by anyone with access to a computer and an Internet connection.

Medical Data mining was the process of extracting hidden patterns from medical data. Asha et al. [15] have developed a hybrid model for classifying Pima Indian diabetic database (PIDD). The model consists of two stages. In the first stage, the K-means clustering was used to identify and eliminate incorrectly classified instances. The continuous data was converted to categorical form by approximate width of the desired intervals, based on the opinion of medical expert. In the second stage a fine tuned classification was done using Decision tree C4.5 by taking the correctly clustered instance of first stage. Further rules generated using cascaded C4.5 tree with categorical data are less in numbers and easy to interpret compared to rules generated with C4.5 alone with continuous data.

Dealing with data means to group information into a set of categories either in order to learn the artifacts or understand various domains. Data mining, having the tools of data classification and data clustering, is one of the most powerful techniques to deal with data in such a manner that it can help researchers identify the required information. As a step forward to address this challenge, experts have utilized clustering techniques as a mean of exploring hidden structure and patterns in underlying data. Improved stability, robustness and accuracy of unsupervised data classification in many fields including pattern recognition, machine learning, information retrieval, image analysis and bioinformatics, clustering has proven itself as a reliable tool. To identify the clusters in datasets algorithm are utilized to partition data set into several groups based on the similarity within a group. There was no specific clustering algorithm, but various algorithms are utilized based on domain of data that constitutes a cluster and the level of efficiency required. Clustering techniques are categorized based upon different approaches. Masood et al. [16] have presented a survey of few clustering techniques out of many in data mining. For the purpose five of the most common clustering techniques out of many have been discussed. The clustering techniques which have been
surveyed are: K-medoids, K-means, Fuzzy C-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Self-Organizing Map (SOM) clustering.

3. Problem Identification

A technique which is used to find useful information from database is said to be data mining. It is one of the fast growing research fields which are used in wide areas of applications. The data mining consists of many different algorithms such as classification, clustering, association analysis and searching algorithms. The common problem of most existing methods are given below,

- The majority of existing clustering algorithms encounter serious scalability and/or accuracy related problems when used on databases with a large number of records and/or attributes.
- When the size of a data set exceeds a certain limit, many existing algorithms will be no longer practical. Extremely large data sets in data mining challenge many existing algorithms which are effective for small data sets.
- One of the major problems in existing method is optimization via time complexity.

These are the problems of existing works and which motivate us to do this research in data mining.

4. Proposed Methodology

Data Mining is defined as mining of knowledge from vast quantity of data. It is one of the swift mounting research fields which are used in extensive areas of applications. Using mining we can envisage the nature and behavior of any variety of data. Clustering is an important task in data mining which is used for the purpose to make groups of the given data set based on the relationship between them. It is a main task in data mining, used in many fields, including statistical data analysis, machine learning, fraud detection, pattern recognition, image analysis, information retrieval, and bioinformatics. The primary intension of the research is clustering the data with efficient technique. At first create an decision tree based on the data set. To construct decision trees by splits that maximizes the difference between the feature-value means of the children nodes. This can be done by ID3 decision tree algorithm which gives faster and shorter tree. The key formulae is given as

Entropy is the measurement of uncertainty where the higher the entropy, then the higher the uncertainty.

\[ E(S) = - \sum_{i=1}^{m} p_i \log_2 (p_i) \]  
(1)

Information needed to identify the class of an attribute is given by

\[ \text{Info}(A) = \sum_{j=1}^{|D|} \frac{|D_j|}{|D|} \text{Info}(D_j) \]  
(2)

Information Gain or gain uses the entropy in order to determine what attribute is best used to create a split with. The Information Gain can be given as in (3) where A is the attribute and D is the data set.

\[ \text{Gain}(A) = \text{Info}(D) - \text{Info}(A) \]  
(3)

Gain ratio is the modification of the information gain that reduces its bias given as \( \text{Gain ratio} = \text{Gain} / \text{Split info} \), where

\[ \text{Split info} = - \sum_{i=1}^{n} \frac{m_i}{n} \log \frac{m_i}{n} \]  
(4)
This reflects the principle that the objects in different clusters should be as dissimilar as possible. Based on these we split the data from the dataset. The proposed method not uses all possible combination because it is time consuming so we only use the best split combination. Therefore our method is optimized which reduces the time complexity. Then the data sets are clustered based on the Possibility Fuzzy C-Means Clustering Algorithm (PFCM), which has the qualities of both the Fuzzy C-Means (FCM) and the Possibility C-Means (PCM). Based on these our method achieves the efficient clustering. The proposed method can be implemented in Matlab with various databases in UCI machine learning repository.

5. Conclusion

In this paper we tried to propose very rare combination of supervised learning mechanism which uses ID3 decision tree and unsupervised learning mechanism such as PFCM framework. It can be applied on very large data set which supports scalability requirement of clustering hence can be used to improve the system performance.

6. References


Authors Biography

1st. N. Deepika, Sr. Asst. Professor having 14 years of experience in Academics has pursued her M.Tech from JNTU; Hyderabad is currently working in NHCE, Dept of CSE, Bangalore. She has guided many UG & PG students for their Projects. Her Research areas include Clustering techniques, Data Mining, Web Mining and Big Data Analysis.

2nd. Dr. N. Guru Prasad is currently working in CSE department having more than 20 years of experience. He has more than 30 publications to his credit and has also authored 6 books. He is also a senior life member of various professional societies.