A REVIEW ON VARIOUS APPROACHES OF CLUSTERING IN DATA MINING

Abhinav Kathuria

Email - abhinav.kathuria90@gmail.com

Abstract: Data mining is the process of the extraction of the hidden pattern from the data available. Various classification approaches had been implemented in data mining process. These approaches have been used to divide the data into different sets so that easily relation between different attributes can be identified. Different data mining techniques have been used to help health care professionals in the diagnosis of Diabetes disease. Those most frequently used focus on classification: naïve bayes decision tree, and neural network. Other data mining techniques are also used including kernel density, automatically defined groups, bagging algorithm, and support vector machine. The problem of redundancy in is always occurred. In our work we will reduce this problem.

Keywords: Data Mining, clustering, KNN, Fuzzy-KNN, Naïve Bayes, Neural Network, Support Vector Machine.

1. INTRODUCTION:

1.1 Data Mining

Data mining process can be extremely useful for Medical practitioners for extracting hidden medical knowledge. It would otherwise be impossible for traditional pattern matching and mapping strategies to be so effective and precise in prognosis or diagnosis without application of data mining techniques. This work aims at correlating various diabetes input parameters for efficient classification of Diabetes dataset and onward to mining useful patterns. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare systems too. Data preprocessing and transformation is required before one can apply data mining to clinical data. Knowledge discovery and data mining is the core step, which results in discovery of hidden but useful knowledge from massive databases.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

1.2 Working of data mining

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

1.2.1 Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

1.2.2 Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

1.2.3 Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

1.2.4 Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

1.3 Issues in data mining

- Security and social issues
- User interface issues
- Mining methodology issues
- Performance issues
- Data source issues

1.4 Data Mining Clustering Techniques:

Apart from the two main categories of partitioned and hierarchical clustering algorithms, many other methods have emerged in cluster analysis, and are mainly focused on specific problems or specific data sets available. These methods include [HK01]:

Density-Based Clustering: These algorithms group objects according to specific density objective functions. Density is usually defined as the number of objects in a particular neighborhood of a data objects. In these approaches a given cluster continues growing as long as the number of objects in the neighborhood exceeds some parameter. This is considered to be different from the idea in partitioned algorithms that use iterative relocation of points given a certain number of clusters.

Grid-Based Clustering: The main focus of these algorithms is spatial data, i.e., data that model the geometric structure of objects in space, their relationships, properties and operations. The objective of these algorithms is to quantize the data set into a number of cells and then work with objects belonging to these cells. They do not relocate points but rather build several hierarchical levels of groups of objects. In this sense, they are closer to hierarchical algorithms but the merging of grids, and consequently clusters, does not depend on a distance measure but it is decided by a predefined parameter.

Model-Based Clustering: These algorithms find good approximations of model parameters that best fit the data. They can be either partitioned or hierarchical, depending on the structure or model they hypothesize about the data set and the way they refine this model to identify partitioning. They are closer to density-based algorithms, in that they grow particular clusters so that the preconceived model is improved. However, they sometimes start with a fixed number of clusters and they do not use the same concept of density.

Categorical Data Clustering: These algorithms are specifically developed for data where Euclidean, or other numerical-oriented, distance measures cannot be applied. In the literature, we find approaches close to both partitioned and hierarchical methods.

Scalability: The ability of the algorithm to perform well with large number of data objects (tuples). **Analyze mixture of attribute types:** The ability to analyze single as well as mixtures of attribute types. **Find arbitrary-shaped clusters:** The shape usually corresponds to the kinds of clusters an algorithm can find and we should consider this as a very important thing when choosing a method, since we want to be as general as possible. Different types of algorithms will be biased towards finding different types of cluster structures/shapes and it is not always an easy task to determine the shape or the corresponding bias. Especially when categorical attributes are present we may not be able to talk about cluster structures. **Minimum requirements for input parameters**: Many clustering algorithms require some user-defined parameters, such as the number of clusters, in order to analyze the data. However, with large data sets and higher dimensionalities, it is desirable that a method require only limited guidance from the user, in order to avoid bias over the result. Handling of noise: Clustering algorithms should be able to handle deviations, in order to improve cluster quality. Deviations are defined as data objects that depart from generally accepted norms of behavior and are also referred to as outliers. Deviation detection is considered as a separate problem. Sensitivity to the order of input records: The same data set, when presented to certain algorithms in different orders, may produce dramatically different results. The order of input mostly affects algorithms that require a single scan over the data set, leading to locally optimal solutions at every step. Thus, it is crucial that algorithms be insensitive to the order of input. High dimensionality of data: The number of attributes/dimensions in many data sets is large, and many clustering algorithms cannot handle more than a small number (eight to ten) of dimensions. It is a challenge to cluster high dimensional data sets, such as the U.S. census data set which contains attributes.

1.5 Classification of Clustering Algorithms:

Categorization of clustering algorithms is neither straightforward, nor canonical. In reality, groups below overlap. For reader's convenience we provide a classification closely followed by this survey. Corresponding terms are explained below.

- Hierarchical Methods

  Agglomerative Algorithms

  Divisive Algorithms

- Partitioning Methods
- Relocation Algorithms

f Probabilistic Clustering

f K-medoids Methods

f K-means Methods

f Density-Based Algorithms *f*

Density-Based Connectivity Clustering *f*

Density Functions Clustering

- Grid-Based Methods

f Methods Based on Co-Occurrence of Categorical Data

f Constraint-Based Clustering

f Clustering Algorithms Used in Machine Learning *f* Gradient Descent and Artificial

Neural Networks *f* Evolutionary Methods

- Scalable Clustering Algorithms *f*

Algorithms for High Dimensional Data

Subspace clustering *f* Projection

Techniques *f*

Co-Clustering Techniques 4

2. REVIEW OF LITERATURE:

Ravneet Jyot Singh, Williamjeet Singh, [1]"Data Mining in Healthcare for Diabetes Mellitus", purposed a n approach for data mining that is related to human health behavior. : Disease diagnosis is one of the applications where data mining tools are proving successful results. Diabetes disease is the leading cause of death all over the world in the past years. Several researchers are using statistical data. The availability of huge amounts of medical data leads to the need for powerful mining tools to help health care professionals in the diagnosis of diabetes disease. Using data mining technique in the diagnosis of diabetes disease has been comprehensively investigated, showing the acceptable levels of accuracy. Recently researchers have been investigating the effect of hybridizing more than one technique showing enhanced results in the diagnosis of diabetes disease.

S.Ummugulthum Natchiar et al [2] "Customer Relationship Management Classification Using Data Mining Techniques" CustomerRelationshipManagementpossesses Business Intelligence by incorporating information acquisition, information storage, and decision support functions to provide customized customer service. It enables customer representatives to analyze and classify data to address customer needs in order to promote greater customer satisfaction and retention. In this paper, a new feature selection method is proposed to resolve such CRM data set with relevant features by incorporating an efficient dataminingtechniques to improve data quality and feature relevancy after preprocessing. Finally it enhances the performance of classification.

Sankaranarayanan, S. et al [3] "Diabetic Prognosis through Data Mining Methods and Techniques" Datamining now-a-days plays an important role in prediction of diseases in health care industry. Datamining is the process of selecting, exploring, and modeling large amounts of data to discover unknown patterns or relationships useful to the data analyst. Medical datamining has emerged impeccable with potential for exploring hidden patterns from the data sets of medical domain. These patterns can be utilized for fast and better clinical decision making for preventive and suggestive medicine. In this paper, two major DataMiningtechniques v.i.z., FP-Growth and Apriori have been used for application to diabetes dataset and association rules are being generated by both of these algorithms.

C. M. Velu et al [4] "Visual Data Mining Techniques for Classification of Diabetic Patients", Clustering is a datamining technique for finding important patterns in unorganized and huge data collections. The likelihood approach of clustering technique is quite often used by many researchers for classifications due to its' being simple and easy to implement. It uses Expectation-Maximization (EM) algorithm for sampling. The study of classification of diabetic patients was main focus of this research work. Diabetic patients were classified by

dataminingtechniques for medical data obtained from Pima Indian Diabetes (PID) data set. This research was based on three techniques of EM Algorithm, h-means+ clustering and Genetic Algorithm (GA). These techniques were employed to form clusters with similar symptoms.

Deepti Mishra et al [5] "Analysis & Implementation of Item based Collaboration Filtering using K-Medoid" This thesis uses data mining classification algorithm classification algorithms to get useful information to decision-making out of customer ship transaction behaviors. Firstly, by business understanding, data understanding and data preparing, modeling and evaluating we get the results of the two algorithms and by comparing the results. This paper describes the use of classification trees and shows two methods of pruning them. An experiment has been set up using different kinds of classification tree algorithms with different pruning methods to test the performance of the algorithms and Pruning methods.

Wang, Guoyin et al [6] "Granular computing based data mining in the views of rough set and fuzzy set" datamining is considered as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. In our data-driven datamining model, knowledge is originally existed in data, but just not understandable for human. Datamining is taken as a process of transforming knowledge from data format into some other human understandable format like rule, formula, theorem, etc. In order to keep the knowledge unchanged in a datamining process, the knowledge properties should be kept unchanged during a knowledge transformation process. Many real world datamining tasks are highly constraint-based and domain-oriented. Thus, domain prior knowledge should also be a knowledge source for datamining.

3. APPROACHES USED:

3.1 Fuzzy KNN (k nearest neighbour)

A "Fuzzy KNN" algorithm utilizes strength of test sample into any class called fuzzy class membership and thus produces fuzzy classification rule. K-Nearest Neighbours algorithm is a non-parametric method used for classification and regression.In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

- In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

3.2 Fuzzy k-Nearest Neighbor Algorithm (FKNN):

The k-nearest neighbor algorithm (KNN) is one of the oldest and simplest non parametric pattern classification methods. In the KNN algorithm a class is assigned according to the most common class amongst its k nearest neighbors. According to his approach, rather than individual classes as in KNN, the fuzzy memberships of samples are assigned to different categories according to the following formulation.

3.3 Genetic Algorithm

A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic particle algorithms approximate the target probability distributions by a large cloud of random samples termed particles or individuals. During the mutation transition, the particles evolve randomly around the space independently and to each particle is associated a fitness weight function. During the selection transitions, such an algorithm duplicates particles with high fitness at the expense of particles with low fitness which dies. These genetic type particle samplers belong to the class of mean field particle methods.

3.4 SVM (Support Vector Machines)

Support Vector Machine (SVM) also called Support Vector Networks are supervised learning models that analyze data and recognize patterns. SVM models represent examples as point in space mapped in manner that separate

categories examples are divided by a gap thereby performing linear classification. Apart from this SVMs can also perform nonlinear classification using Kernel trick.

The main idea of SVM is that; it finds the optimal separating hyper plane such that error for unseen patterns is minimized. Consider the problem of separating the set of training vectors belonging to two separate classes.

3.5 Neural Network

Appliance learning algorithms make easy a lot in decision making and neural network has perform well in categorization function in medical field. Most accepted techniques in the middle of them are neural network. Neural networks are those networks that are the compilation of easy elements which function parallel. A NN can be skilled to do an exacting purpose by adjust the value of the weights connecting elements. Network function is resolute by the associations linking elements. There are several activation functions that are used to produce relevant output

4. CONCLUSION:

Data mining process can be extremely useful for Medical practitioners for extracting hidden medical knowledge. Various classification approaches had been implemented in data mining process. To applying data mining is beneficial to healthcare, disease diagnosis, and treatment; few researches have investigated producing treatment plans for patients. The main issue in the diabetes data classification is that due to in sufficient resources and data proper mining has not been done. To remove the issue of the data mining in healthcare proper data anomalies have to be pre-processed and redundancy must be removed from the dataset. We will reduce this naïve bayes.

REFRENCES:

1. Ravneet Jyot Singh, Williamjeet Singh**,** Data Mining in Healthcare for Diabetes Mellitus", International Journal of Science and Research (IJSR), Volume 3 Issue 7, July 2014
2. S. Ummugulthum Natchiar "Customer Relationship Management Classification Using Data Mining Techniques" International Conf. on Science Engineering and Management Research (ICSEMR), 2014, pp 1 – 5.
3. Sankaranarayanan, S. "Diabetic Prognosis through Data Mining Methods and Techniques", International Conf. on Intelligent Computing Applications (ICICA), 2014, pp. 162 – 166.
4. C. M. Velu "Visual Data Mining Techniques for Classification of Diabetic Patients", IEEE Conf. on Advance Computing Conference (IACC), 2013, pp. 1070 – 1075.
5. Deepti Mishra"Analysis & Implementation of Item based Collaboration Filtering using K-Medoid" International Conf. on Advances in Engineering and Technology Research (ICAETR), 2014, pp. 1 – 5.
6. Wang, Guoyin "Granular computing based data mining in the views of rough set and fuzzy set" IEEE Conf. on Granular Computing, 2008, PP 67.
7. Jagannathan, G. "Seventh IEEE International Conference on Data Mining Workshops", IEEE Conf. on Data Mining Workshops, 2007, pp. 1– 3