© 2017 IJSRST | Volume 3 | Issue 8 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

A Review on Various Approaches for data Preserving Clustering in Data Mining

Deep Kumar

Software Engineer, Igniva Solutions Private Limited, Mohali, Punjab, India

ABSTRACT

Data Mining is the process of extraction of valuable information from the raw data. Classification and clustering are the two main components that are used in data mining process.In the process of data mining various approaches have been used for clustering process so that data can be managed in the form of clusters. In this paper various approaches of clustering has been discussed. Various approaches have been used for clustering based on properties of the dataset instances. These approaches are based on centroid, rule based clustering and properties based clustering. On the basis of these approaches clustering approach that is suitable for large dataset has been selected.

Keywords: Clustering, K-Means, K-medioid, CobWeb, and DBSCAN

I. 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.

1.2 Types of Data Mining:

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

1.3Cluster: cluster is an ordered list of objects, which have some common objects. The objects belong to an interval.

1.4 Distance between Two Clusters: The distance between two clusters involves some or all elements of the two clusters.The clustering method determines how the distance should be computed. The distance between two points is taken as a common metric to assess the similarity among the components of a population.

1.5 Customer Data Clustering: Customer clustering is the most important data mining Methodologies used in marketing and customer relationship management (CRM). Customer clustering would use customerpurchase transaction data to track buying behavior and create strategic business initiatives. Companies want to keep high, profit, high, value, and low, risk customers. This cluster typically represents the 10 to 20 percent of customers who create 50 to 80 percent of a company's profits.A company would not want to lose these customers, and thestrategic initiative for the segment is obviously retention. A lowprofit, high, value, and low, risk customer segment is also an attractive one, and the obvious goal here would be to increase profitability for this segment.

Clusteringaims at discovering groups and patterns in data sets. In general, the output produced by a special clustering algorithm will be the assignment of data objects in dataset to different groups. In other words, it will be sufficient to identify data object with a unique cluster label.From the viewpoint of clustering, data objects with different cluster labels are considered to be in different clusters, if two objects are in the same cluster then they are considered to be fully similar, if not they are fully dissimilar. Thus, it is obvious that cluster labels are impossible to be given a natural ordering in a way similar to real numbers, that is to say, the output of clustering algorithm can be viewed as categorical. Since the output of individual clustering algorithm is categorical and so the cluster ensemble problem can be viewed as the categorical data clustering problem, in which runs of different clustering algorithm are combined into a new categorical dataset.

II. REVIEW OF LITERATURE

Mr. G. Nirmal Kumar.et. al [1] "A system learning of connection with humans by online social networking - A rapport by means of creating usable customer Intelligence from Social media Data"everything is just turning into online brand. The privacy and security are totally low in OSNS (Online Social Networking Sites). A data without user's interest is transferred; this can only be admitted to a particular level. To ensure secure data mining in OSNS sites a Privacy K mean algorithm is derived along with Check threshold algorithm for basic information transfer alone. Though very concern on OSNS it is seemed many number of people move out of online sites,which is that they just delete their account.

EssamShaabanet.al.[2] "A Proposed Churn Prediction Model"Churn prediction aims to detect customers intended to leave a service provider. Retaining one customer costs an organization from 5 to 10 times than gaining a new one. Predictive models can provide correct identification of possible churners in the near future in order to provide a retention solution. This paper presents a new prediction model based on Data Mining (DM) techniques. The proposed model is composed of six steps which are; identify problem domain, data selection, investigate data set, classification, and clustering and knowledge usage. A data set with 23 attributes and 5000 instances is used. 4000 instances used for training the model and 1000 instances used as a testing set. The predicted churners are clustered into 3 categories in case of using in a retention strategy. The data mining techniques used in this paper are Decision Tree, Support Vector Machine and Neural Network throughout an open source software name WEKA.

Xi Long et.al.[3] "Churn Analysis of Online Social Network Users Using Data Mining Techniques" A churn is defined as the loss of a user in an online social network (OSN). A set of 24 attributes is extracted from the data. A decision tree classifier is used to predict churn and non-churn users of the future month. In addition, k-means algorithm is employed to cluster the actual churn users into different groups with different online social networking behaviors. Results show that the churn and nonchurn prediction accuracies of 65% and 77% are achieved respectively. Furthermore, the actual churn users are grouped into five clusters with distinguished OSN activities and some suggestions of retaining these users are provided.

D. S. RAJPUT.et. al.[4] "Analysis of Social Networking Sites Using K- Mean Clustering Algorithm" Clustering is one of the very important technique used for classification of large dataset and widely applied to many applications including analysis of social networking sites, aircraft accidental, company performance etc. In recent days, Communication, advertising through social networking sites is most popular and interactive strategy among the users. This research attempts to find the large scale measurement study and analysis, effectiveness of communication strategy, analyzing the information about the usage, people's interest in social network sites in promoting and advertising their brand in social networking sites. The significance of the proposed work is determined with the help of various surveys, and from people who use these sites. Further a more specific pre-processing method is applied to clean data and perform the clustering method to generate patterns that will be work as heuristics for designing more effective social networking sites.

Chamatkar, A.J.et.al.[5] "Implementation of Different Data Mining Algorithms with Neural Network "With the huge amount of information available online, the World Wide Web is a fertile area for data mining research. Neural network is not suitable for data mining directly, because how the classifications were made is not explicitly stated as symbolic rules that are suitable for verification or interpretation by humans. Different concise symbolic rules with high accuracy can be extracted from a neural networkwith the proposed approach. The neural network is first trained to achieve the required accuracy in datamining. In this paper we are going to combine neural network with the three different algorithms which are commonly used in data mining to improve the data mining result. These three algorithms are CHARM Algorithm, Top K Rules mining and CM SPAM Algorithm. The different datasets of online e-commerce website flip-kart and Amazon are used to train the neural network and to use in data mining. The results of all three data mining algorithm with neural network techniques then tested on the available datasets and result are compared by computational complexity of the algorithm.

III. APPROACHES USED

Hierarchicalclustering: builds a cluster hierarchy or, in other words, a tree of clusters,also known as a dendrogram. Every cluster node contains child clusters; sibling clusterspartition the points covered by their common parent. Such an approach allows exploringdata on different levels of granularity. Hierarchical clustering methods are categorized intoagglomerative(bottom-up) and divisive (top-down). An agglomerativeclustering starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters. Adivisiveclustering starts with one cluster of all data points and recursively splits the most appropriate cluster.

Probabilistic Clustering: In the probabilisticapproach, data is considered to be a sample independently drawn from a mixture modelof several probability distributions. The main assumption is that data points

are generated by, first, randomly picking a model jwith probability ykjj:1,=τ, and, second, by drawing a point xfrom a corresponding distribution. The area around the mean of each (supposedly uni-modal) distribution constitutes a natural cluster.

K-Medoids: Methods: Ink-medoids methods a cluster is represented by one of its points. We have already mentioned that this is an easy solution since it covers any attribute types and that medoidshave embedded resistance against outliers since peripheral cluster points do not affectthem. When medoids are selected, clusters are defined as subsets of points close to respective medoids, and the objective function is defined as the averaged distance or another dissimilarity measure between a point and its medoid.

K-Means Methods: Thek-means algorithm is by far the mostpopular clustering tool used in scientific and industrial applications. The name comesfrom representing each of kclusters Cby the mean (or weighted average)cof its points, the so-called centroid. While this obviously does not work well with categoricalattributes, it has the good geometric and statistical sense for numerical attributes.

DBSCAN (Density Based Spatial Clustering of Applications with Noise): DBSCAN is designed to discover the clusters and the noise in a spatial database according to definitions and ideally, we would have to know the appropriate parameters Eps and Min Points Of each cluster and at least one point from the respective cluster. Then, we could retrieve all points that are density-reach-able from the given point using the correct parameters. But there is no easy way to get this information in advance for all clusters of the database. However, there is a simple and effective heuristic to determine the parameters Eps and MinPoints of the "thinnest", i.e.least dense, cluster in the database. Therefore, DB SCANuses global values for Eps and MinPoints, i.e. the same values for all clusters. The density parameters of the "thinnest"cluster are good candidates for these global parameter valuesspecifying the lowest density which is not considered to benoise.

COBWEB Algorithm: Whereas iterative distance-based clustering, such as K-means, iterate over the whole dataset until convergence in the clusters is reached. COBWEB works incrementally, updating the clustering instance by instance. The clustering COBWEB creates is expressed in the form of a tree, with leaves representing each instance in the tree, the root node representing the entire dataset, and branches representing all the clusters and sub clusters within the tree. It is worth noting that there is no limit to the total number of sub clusters except the limit imposed by the size of the dataset.

Farthest first traversal k-center algorithm: The farthest first traversal k-center algorithm (FFT) is a fast, greedy algorithm that minimizes the maximum cluster radius. This is also treated as an efficient algorithm which always returns the right answer. The pseudo code for the farthest first traversal algorithm is as follows:

Pick any Z € S and set T={z}
While
|T| < k
Z=arg max β(x,T)
T=TU{Z}

IV. CONCLUSION

Data mining is the process that has been used for selection of best dataset attributes and valuable information from raw information. In the process of clustering instances has been clustered on the basis of centroid that is based on k-means and K-medoids algorithm. DBSCAN algorithm has been used for the clustering process that is based on the density based. Various instances density has been computed. In the process of hierarchal clustering tree based structure has been developed that use root and leaves based process so that tree based clustering process can be achieved. On the basis of these approaches we can

conclude that tree based clustering provide better clustering rather than other approaches.

V. REFERENCES

[1]. Mr. G. Nirmal Kumar. "A system learning of connection with humans by online social networking - A rapport by means of creating usable customer Intelligence from Social media Data" IEEE International Conference on Science, Engineering and Management Research, 2014,pp. 1 – 6.

[2]. EssamShaaban."A Proposed Churn Prediction Model"International Journal,Research and Applications 2012, pp.693-697.

[3]. Xi Long."Churn Analysis of Online Social Network Users Using Data Mining Techniques"international Multi-conference of engineers and computer scientists,2012,pp-14-16.

[4]. D. S. RAJPUT."Analysis of Social Networking Sites Using K- Mean Clustering Algorithm" International Journal of Computer & Communication Technology, 2012, Vol. 3, pp. 975 –979.

[5]. Chamatkar, A.J."Implementation of Different Data Mining Algorithms with Neural Network"IEEE International Conference onComputing Communication Control and Automation,2015, pp. 374–378.

[6]. Huan Liu. "Some computational challenges in mining social media" IEEE International Conference on Advances in Social Networks Analysis and Mining, 2013, pp. 25-28

[7]. Au, W. H."A novel evolutionary data mining algorithm with applications to churn prediction" IEEE International Conference on Evolutionary Computation,2003,pp. 532 – 545.

[8]. Gok, Mehmet."A case study for the churn prediction in Turksat internet service subscription" IEEE International Conference on Advances in Social Networks Analysis and Mining,2015,pp. 1220–1224.

[9]. GuangliNie."Find Intelligent Knowledge by Second-Order Mining: Three Cases from China" IEEE International Conference on Data Mining Workshops,2010,pp. 1189–1195.

[10]. Chen Yu-bao."Study on Predictive Model of Customer Churn of Mobile Telecommunication Company" IEEE International Conference on Business Intelligence and Financial Engineering,2011,pp. 114 – 117.