International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3, Issue No.1 pp : 51-55 1 Jan 2014

A Review on Comparative analysis of different clustering and Decision Tree for Synthesized Data Mining Algorithm

Avinash Pal, JayPrakash Maurya, Prashant Sharma

Department of Computer Science and Engineering
IES College of Technology, Bhopal
avinashcse.1010@gmail.com, jpeemaurya@gmail.com, prashsharma2003@gmail.com

Abstract— Web mining is the sub category or application of data mining techniques to extract knowledge from Web data. With the advent of new advancements in technology the rapid use of new algorithms has been increased in the market. A data mining is one of the fast growing research field which is used in a wide areas of applications. The data mining consists of classification algorithms, association algorithms and searching algorithms. Different classification and clustering algorithm are used for the synthetic datasets. In this paper various techniques that are based on clustering and decision tree for synthesized data mining are discussed.

Index Terms— Data Mining, decision tree, Clustering, Synthesized data mining.

I. INTRODUCTION

Data mining is 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, is open to be freely added to by anyone with access to a computer and an Internet connection. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to either enhance profits, cuts costs, or both.

Clustering and decision tree are two of the mostly used methods of data mining which provide us much more convenience in researching information data. This paper will select the optimal algorithms based on these two methods according to their different advantages and shortcomings in order to satisfy different application conditions. Classification is an important task in data mining. Its purpose is to set up a classifier model and map all the samples to a certain class which can provide much convenience for people to analyze data further more. Classification belongs to directed learning, and the most important methods take account of decision tree, genetic algorithm, neural network, Bayesian classification and rough set etc [1].

Web mining is used to understand customer performance, estimate the usefulness of a specific Web site, and help measure the success of a marketing crusade. Web mining can be decayed into the subtasks:

Resource finding

The task of retrieving intended Web credentials. By resource judgment means the procedure of retrieving the data that is either online or offline from the text sources available on the web such as electronic newsletters, electronic newswire, the text inside of HTML documents obtained by removing HTML tags, and also the physical selection of Web resources.

Selection of Information and pre-processing

Automatically selecting and pre-processing specific information from retrieved Web resources. It is a kind of modification processes of the original data retrieved in the IR process. These alteration could be either a kind of pre-processing that are mentioned above such as discontinue words, twigging, etc. or a pre-processing intended at obtaining the desired representation such as finding phrases in the training amount, transforming the illustration to relational.

Generalization

It automatically discovers general patterns at individual Web sites as well as across multiple sites. Machine knowledge or data mining techniques are typically used in the process of simplification. Humans play a vital role in the information or knowledge discovery process on the Web since the Web is an interactive medium [2].

CLUSTERING

Clustering is a division of data into groups of analogous objects. Every grouping known as cluster consists of objects that are like amongst them and dissimilar compared to object of other groups. Instead of data by fewer clusters essentially loses certain fine details, but achieves generalization. It represents many data objects by only some clusters, and therefore it models data by its clusters. Clustering is the unsupervised classification of patterns into groups known as clusters. Clustering is a difficult problem combinatorial, and dissimilarity in assumptions and contexts in different communities have made the transfer of useful generic concepts and methodologies slow to occur. Cluster analysis aims at identifying groups of related objects and, hence helps to discover distribution of patterns and interesting correlations in large data sets. So that it can be used in wide research since it arises in many application domains. Especially, in the last years the availability of huge transactional and experimental data sets and the arising requirements for data mining created needs for clustering algorithms that scale and can be applied in diverse domains.

Clustering is considered an interesting approach for finding similarities in data and putting similar data into dissimilar sets. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups. The idea of data grouping, or clustering, is simple to use and in its nature and is very near to the human way of thinking; whenever they are presented with a large amount of data, humans are usually tend to summarize this huge number of data into a small number of groups or categories in order to further facilitate its analysis. Most of the data collected in many problems seem to have some inherent properties that lend themselves to natural grouping. Clustering is a challenged research field which belongs to unsupervised learning. The number of clusters needed is unknown and the formation of clusters is data driven completely. Clustering can be the pretreatment part of other algorithms or an independent tool to obtain data distribution, and also can discover isolated points. Common clustering algorithms are K-MEANS, BIRCH, CURE, DBSCAN etc. Each of them has respective advantages: KMEANS is simple and easy to understand, DBSCAN can filter noises splendidly, CURE is insensitive to input etc. But now there still has no algorithm which can satisfy any condition especially the large-scale high dimensional datasets, so it is important for us to improve and develop clustering methods [1].

Various Clustering Techniques

K-Means Clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The algorithm is called k-means, where k represents the number of clusters required, since a case is allocated to the cluster for which its distance to the cluster mean is the negligible. The achievement in the algorithm centres on finding the k-means [3].

Hierarchical Clustering builds a cluster hierarchy or a tree of clusters, it is also known as a 'dendrogram'. All cluster nodes contains child clusters; sibling clusters partition the points covered by their common parent [3].

DBSCAN finds all clusters properly, independent of the shape, size, and location of clusters to everyone, and is greater to a widely used Clarans technique. The DBSCAN method is based on two key concepts: density reach ability and density connect ability. These both concepts depend on two input parameters of the DBSCAN clustering: the size of epsilon neighborhood e and the minimum points in a cluster m. The number of point's parameter impacts recognition of outliers. Points are confirmed to be outliers if there are few other points in the e-Euclidean neighborhood. Parameter 'e' manages the size of the neighborhood, as well as the size of the clusters. The Euclidean space has an open set that can be divided into a set of its connected components. The execution of this idea for partitioning of a finite set of points requires concepts of density, connectivity and boundary [3].

OPTICS ("Ordering Points to Identify the Clustering Structure") is an algorithm for finding density-based clusters in spatial data. Its fundamental idea is comparable to DBSCAN, but it addresses one of DBSCAN's main weaknesses: the problem of detecting significant clusters in data of changeable density. Consecutively the points of the database are linearly ordered such that points that are spatially closest become neighbours in the ordering. Furthermore, a special distance is stored for each point that corresponds to the density that needs to be accepted for a cluster in order to have both points belong to the same cluster [3].

Decision tree

Decision tree support tool that uses tree-like graph or models of decisions and their consequences [4][5], including event outcomes, resource utility and costs, frequently used in operations and research in decision analysis help to recognize a strategy most likely to reach a goal. In data mining and machine learning; decision tree is a analytical representation that is mapping from observations about an item to conclusions about its objective cost. The machine learning method for suggest a decision tree from data is called decision tree learning. A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree contains nodes that form a rooted tree; rooted tree is a directed tree with a node called "root" that has no arriving edges. All other nodes have accurately one arriving edge. A node with extrovert edges is called an internal or test node. All other nodes are known as leaves (also known as terminal or decision nodes). Inside a decision tree, each internal node divides the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most common case, each test believes a single attribute so that the instance space is separated according to the attribute's cost. In numeric attributes case the situation refers to a range.

II. BACKGROUND

The rapid growth of Web technology has made the World Wide Web an important and popular application platform for disseminating and searching information as well as for conducting business. This growth gave a way to the development of ever smarter approaches to extract patterns and build knowledge with the aid of artificial intelligence techniques. These techniques have been used, together with information technology in an extensive variety of applications. This is where semantics, social network analysis, content, usage, web structure, and further aspects have already been and will increasingly keep being included in many application domains. The Web provides rich medium for communication, which goes far beyond the conventional communication media. Several data mining methods can help achieve effective Web intelligence.

III. RELATED WORK

In year 2010, Dan et al [1] presented A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree. They improve the traditional algorithms like CURE and C4.5 appropriately, and present a new synthesized algorithm CA for mining large-scale high dimensional datasets. The basic idea of CA is shown as follows: first introduce PCA to analyze the relevancy between features and replace the whole dataset with several countable composite features; then improve CURE to part the set into several clusters which can be the pretreatment of other algorithms and achieve the reduction of sample scale; finally introduce parallel processing into C4.5 to enhance the efficiency of building decision tree. The decision tree classification part of CA algorithm is improved based on C4.5, and the improvements are mainly embodied in threshold

partition and selection of testing features. In traditional C4.5 algorithm, they will divide the datasets dynamically, and select the values with the biggest gain ratio to split continuous features. Introduce three different classifiers to ascertain the correctness of selecting features and avoid bios problems. This synthesized CA algorithm is improved based on traditional CURE and C4.5 methods. It introduces scale reduction, feature reduction and classification analysis to handle large and high dimensional datasets. By applying CA algorithm in maize seed breeding, they can find out the important features which will influence breeding tremendously and obtain the classification model of whole maize samples. The experiments show the efficiency of CA is higher not only in clustering but also in decision tree. There also exist some problems needed to research further more. CA is sensitive to some parameters like the clustering number, shrink factors and the threshold. C4.5 only can covenant with the dataset that has the classification feature. The dataset treated is a little small which will impact the final output of algorithms [1].

Chintandeep Kaur and Rinkle Rani Aggarwal presented a survey on Web mining tasks and types. They discussed about data mining and their classifications. The past few years have seen the emergence of Web mining as a rapidly increasing area, suitable to the efforts of the research society as well as various organizations that are practicing it Web data mining is a fast rising research area today. Web data is mainly semi-structured and unstructured. Due to the heterogeneity and the lack of structure of Web data, computerized discovery of targeted or unanticipated knowledge information still present many challenging research Problems. Most of the knowledge represented in HTML Web documents, there are numerous other file formats that are publicly accessible on the Internet. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking. The common access pattern tracking investigates the web logs to understand access patterns and trends. These investigations can shed light on improved structure and grouping of resource providers. Customized usage tracking analyzes individual trends. Its intention is to modify web sites to users. The information demonstrated the depth of the site structure and the format of the resources can all be dynamically customized for each user over time based on their access patterns. Web Data Mining is perhaps still in its infancy and much research is being carried out in the area [2].

Manish Verma et al [3] proposed A Comparative Study of Various Clustering Algorithms in Data Mining. They provide a comparative study among various clustering. They compared six types of clustering techniques- k-Means Clustering, Optics, DBSCAN clustering, Hierarchical Clustering, Density Based Clustering and EM Algorithm. Such clustering methods are implemented and analyzed using a clustering tool WEKA. Performances of the 6 techniques are presented and compared. Running the clustering algorithm using any software produces almost the same result even when changing any of the factors because most of the clustering software uses the same procedure in implementing any algorithm [3].

Andrew McCallum et al [6] offered Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching. have focused on reference matching, a particular class of problems that arise when one has many different descriptions for each of many different objects, and wishes to know (1) which descriptions refer to the similar object, and (2) what the best explanation of that object is. They present experimental results for the domain of bibliographic reference matching. Another significant illustration of this class is the merge-purge problem. Companies often purchase and merge multiple mailing lists. The resulting list then has multiple entries for each household. Even for a single person, the name and address in each version on the list may diverge slightly, with middle initials absent or present, words shortened or expanded, zip codes present or absent. This problem of merging large mailing lists and eliminating duplicates becomes even more complex for house holding, where one wishes to collapse the records of multiple people who live in the same household [6].

In recently 2012, Akerkar et al [7] proposed Desiderata for Research in Web Intelligence, Mining and Semantics. In order to offer Web data in suitable formats, Web logs, the Web-site contents, and the Hyperlink Structure of the Web, have been considered as the main source of information. Web log analysis can also help build customized Web services for individual users. Ever since Web log data presents information about specific pages' popularity and the methods used to access them, this information can be integrated with Web content and linkage structure mining to help rank Web pages, classify Web documents, and construct a multilayered Web information base. They also explain about semantic web data [7].

In 2012, S. N. Mishra et al proposed An Effective Algorithm for Web Mining Based on Topic Sensitive Link Analysis. They introduced a concept towards this direction; search based on ranking of some set of categories that comprise a user search profile. Some page ranking algorithms Weighted PageRank and Page Rank are generally used for web structure mining. Topic sensitive weighted page rank makes use of a subset of the ODP category structure that is associated with individual information needs. This subset is processed locally, aiming at enhancing generic results offered by search engines. Processing involves introducing significance weights to those parts of the ODP structure that correspond to user-specific preferences. The outcomes of local processing are consequently combined with global knowledge to derive an aggregate ranking of web results [8].

In this approach, the first step is to generate a biased weighted page rank vectors using a set of some basis topics. This step is the pre-processing step of the web crawler. This step is performed offline. The second step of approach will be performed at the time of query. User will provide a query q; let query q be the context of query. In other words, if the query was issued by highlighting the term q in some Web page, then query q' consists of the terms in web page. They chose to make it uniform, although they could personalize the uncertainty results for different users by varying this distribution. A new concept based on Topic-Sensitive PageRank and Weighted PageRank for web page ranking is based on PageRank algorithm, and provides a scalable approach for search rankings using Link analysis. For each Web page, compute an importance score per topic. At query time, these importance scores are combined based on the topics of the query and associated context to form a composite PageRank score for those pages equivalent the query. This

score can be employed in conjunction with other scoring schemes to produce a final rank for the result pages with respect to the query. This algorithm will improve the order of web pages in the result list so that user may get the relevant pages easily [8].

In year 2012, Bhaskar N. Patel et al [9] presented an Efficient Classification of Data Using Decision Tree. They survey many techniques related to data mining and data classification techniques. They also select clustering algorithm k-means to improve the training phase of Classification. The Learning classification techniques in data mining can be classified into three fundamental types; first one is supervised second one is unsupervised and finally third one is reinforced. There are at least three techniques which are used to calculate a classifier's accurateness. One method is to divide the training set by using two-thirds for training and the other third for approximation presentation. In another method called cross-validation, the training set is separated into reciprocally exclusive and equal-sized subsets and for each subset the classifier is trained on the union of all the other subsets. The usual rate of the error for each subset is therefore an estimate of the error rate of the classifier. Validation called Leave-one-out is an exceptional case of cross validation. All test subsets consist of a single instance. This validation type is more luxurious computationally, but functional when the most exact estimation of a classifier's error rates requisite. Training a average decision tree escorts to a quadratic optimization problem with bound constraints and one linear equality constraints. Training support vector machines involves a huge optimization problem and many specially designed algorithms have been offered. The algorithm that was used is called "Decision Tree Induction" that accelerates the training process by exploiting the distributional properties of the training data, i.e. the natural clustering of the training data and the overall layout of these clusters relative to the decision boundary of support vector machines [9].

The method called k-means since each of the K clusters is represented by the mean of the objects within it. It is also called the centroid method since at each step the centroid point of each cluster is assumed to be known and each of the remaining points are allocated to the cluster whose centroid is closest to it. The K-means algorithm keeps on as follows. First, it arbitrarily selects k of the objects; each object primarily represents a center. For each of the outstanding objects, an object is allocated to the cluster to which it is the most alike, based on the distance among the object and the cluster. Then it calculates the new mean for all clusters. This procedure repeats in anticipation of the criterion function converges [9].

Decision tree is useful because construction of decision tree classifiers does not require any domain knowledge. It can handle hi-dimensional data. The classification and learning steps of decision tree induction are simple and fast. Their representation of acquired knowledge in tree form is easy to assimilate by users. The decision tree algorithm is a top-down induction algorithm. The mean of this algorithm is to construct a tree that has leaves that are harmonized as potential. The most important step of this algorithm is to carry on dividing leaves that are not homogeneous into leaves that are as homogeneous as possible. Once the result obtained, it can be reused for next research. This research depicts on compares reformulated decision tree with standard decision tree for dataset. This comparison is from threshold (complexity) from low to high with reference to the testing accuracy. With this research a set of threshold taken to show that this method ives better accuracy with decision tree rather than K- Means [9].

In same year Gothai, E. and P. Balasubramanie proposed An Efficient Way for Clustering Using Alternative Decision Tree. Their study proposes a Multi-Level Clustering mechanism using alternative decision tree algorithm that combines the advantage of partition clustering, hierarchical clustering and incremental clustering technique for rearranging the most closely related object. The clustering initiation should happen based on the short name value, each short name pointing to the appropriate whole record object. The proposed MLC algorithm has been experimentally tested on a set of data to find a cluster with closely related object. This method is used to overcome the existing system problem, such as manual intervention, misclassification and difficulties of finding a partition range and so on. MLC forms a tree for the clustering process. In the tree structure, the height of each level of nodes represents the similar degree between clusters. MLC incorporate the futures of ADTree features and overcome the existing hierarchical clustering problem [10].

ADTree divide the data based on short name; if cluster is already available with the short name then insert a record into the same cluster else create a new cluster with the new name of short name then insert into a new cluster. In every cluster, sub-set diminutive name points to the whole record. The cluster formation method mainly focus on form a similarity value in single group, for this purpose they are using different method and result of each method is different cluster based on data and spread condition. The experimental results shows the proposed system has lower computational complexity, reduce time consumption, most optimize way for cluster formulation and better clustering quality compared with the existing hierarchical clustering algorithm. They ran extensive experiments with them to find time consumption and compared them with various versions of existing algorithms in order to show this new system reduces the time consumption. The new method offers more accuracy of cluster data without manual intervention at the time of cluster formation. Compare to existing clustering algorithm either partition or hierarchical, new method is more robust and easy to reach the solution of real world complex business problem [10].

In recently 2013, L. Rutkowski et al [11] offered Decision trees for mining data streams based on the McDiarmid's bound. The goal was to compute the heuristics procedures, e.g. Gini index or information gain, based on these N examples, and then to split the learning sequence according to this attribute. To resolve the difficulty hundreds of researches used the so-called 'Hoeffdings trees', resulting from the Hoeffding's bound, for pulling out data streams. The Hoeffding's inequality is not an adequate tool to solve the underlying problem and the previously obtained and presented in literature results, based on it, in a general case are not true. The MacDiarmid's inequality, used in a appropriate way is an efficient tool for solving the trouble. One of the most popular techniques in data mining is based on decision tree training. Traditionally, the first decision tree algorithm was ID3; later C4.5 and CART algorithms were developed [11].

Traditional techniques for data mining require multiple scans of data to extract the information that is not feasible for stream data. It is not probable to store complete data stream or to scan through it multiple times due to its wonderful quantity. The quantity of events in data streams that had previously happened is usually extremely large. To evaluate the performance of the McDiarmid Tree algorithm, numerous simulations were conducted. Because ϵ for the Gini gain tends to zero much faster than for the information gain, only the Gini gain is measured in the subsequent experiments. Synthetic data were used and generated on a basis of synthetic decision trees. They suggested using the term 'McDiarmid Trees' instead of 'Hoeffding Trees' in all algorithms previously developed and based on the Hoeffding's inequality [11].

Classification algorithms discussed by Hanady Abdulsalam et al. [12], holds of three phases; a training phase that contains of labeled records, a test phase using earlier unseen labeled records, and a consumption or deployment phase that classifies unlabeled records. In conventional decision tree classification, an attribute of a tuple is either unconditional or numerical. Smith Tsang et al. [13] presented the problem of constructing decision tree classifiers on data with uncertain numerical attributes.

IV. COMPARISON TABLE

Title	Author/ Publication	Technique used	Advantages	Limitations
A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree	Ji Dan, Qiu Jianlin, Gu Xiang, Chen Li, He Peng	A new synthesized data mining algorithm named CA which improves the original methods of CURE and C4.5. CA introduces principle component analysis (PCA), grid partition and parallel processing which can achieve feature reduction and scale reduction for large-scale datasets.	Tthe efficiency of CA is higher not only in clustering but also in decision tree.	Make full use of data mining technologies to direct agricultural industry is a good research topic. We can obtain some potential useful rules or models from mass agricultural datasets according to these methods.

V. CONCLUSION

Clustering is a technique of grouping simmilar elements so that on the basis of grouping of values classification can be done. There are various efficient techniques implemented for the clustering of data such as J48, ID3, Naïve Bayes Random Forest etc. Here in this paper a survey of all these techniques and their various application is discussed so that in the future an efficient technique is implemented.

REFERENCES

i. *Ji Dan, Qiu Jianlin, Gu Xiang, Chen Li, and He Peng "A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree", IEEE 10th International Conference on Computer and Information Technology (CIT), pp. 2722 – 2728, 2010.*

ii. *Chintandeep Kaur , Rinkle Rani Aggarwal " Web mining tasks and types", International Journal of Research in IT & Management (IJRIM),Volume 2, Issue 2 ,February 2012.*

iii. *Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta " A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, pp.1379-1384, Issue 3, May-Jun 2012.*

iv. *D. Mutz, F. Valeur, G. Vigna, and C. Kruegel, "Anomalous System Call Detection," ACM Trans. Information and System Security, vol. 9, no. 1, pp. 61-93, Feb. 2006.*

v. *M. Thottan and C. Ji, "Anomaly Detection in IP Networks," IEEE Transaction on Signal Processing, vol. 51, no. 8, pp. 2191-2204, 2003.*

vi. *Andrew McCallum, Kamal Nigam and Lyle H. Ungar "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining, pp. 169-178, 2000.*

vii. *Rajendra Akerkar, Costin Bădică, and Dumitru Dan Burdescu "Desiderata for Research in Web Intelligence, Mining and Semantics", Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, June 2012.*

viii. *Shesh Narayan Mishra, Alka Jaiswal and Asha Ambhaikar "An Effective Algorithm for Web Mining Based on Topic Sensitive Link Analysis", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 2, Issue 4, pp. 278 – 282, April 2012.*

ix. *Bhaskar N. Patel, Satish G. Prajapati and Dr. Kamaljit I. Lakhtaria "Efficient Classification of Data Using Decision Tree", Bonfring International Journal of Data Mining, ISSN 2277 – 5048, Vol. 2, No. 1, pp. 6-12, March 2012.*

x. *Gothai, E. and P. Balasubramanie "An Efficient Way for Clustering Using Alternative Decision Tree", American Journal of Applied Sciences, ISSN 1546-9239, vol. 9, no. 4, pp. 531-534, 2012.*

xi. *L. Rutkowski, L. Pietruczuk, P. Duda and M. Jaworski "Decision trees for mining data streams based on the McDiarmid's bound", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, Issue 6, pp. 1272 – 1279, June 2013.*

xii. *Hanady Abdulsalam, David B. Skillicorn, and Patrick Martin, "Classification Using Streaming Random Forests",IEEE Transactions on Knowledge And Data Engineering, Vol. 23, No. 1., pp.22-36, January 2011.*

xiii. *Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, and Sau Dan Lee, "Decision Trees for Uncertain Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 1, pp 63-78, January 2011.*