ISSN:0975-9646

Classification and Analysis of High Dimensional Datasets using Clustering and Decision tree

Avinash Pal[1], Prof. Shraddha Pandit[2], Prof. JayPrakash Maurya[3]
[1,2]*Department of Computer Science & Engineering, R.G.P.V. Bhopal, India*
[3]*IES College of Technology, Bhopal, India*

Abstract— **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.**

Keywords— Data Mining, Clustering, decision tree, Synthesized data mining.

I. INTRODUCTION

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. 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. Data mining is the method 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. 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. 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 a way of evaluating the performance of any web site and its contents. It also helps to find out the success rate of web site among existing. 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 variety and the lack of structure of Web data, programmed discovery of targeted or unanticipated comprehension information still present many challenging research Problems. Most of the knowledge represented in HTML Web documents, there are numerous file formats that are publicly accessible on the Internet. Web mining can be decayed into the subtasks. First one is 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. Next is selection of information and preprocessing. It is a kind of modification processes of the original data retrieved in the IR process. Lastly 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 separation of data into small groups of related objects. Every grouping known as cluster consists of objects that are like amongst them and dissimilar compared to object of other groups. Clustering is the unsupervised categorization of patterns into groups identified as clusters. Clustering is a complex problem combinatorial, and divergence in assumptions and contexts in different communities have made the transfer of useful generic concepts and methodologies deliberate to occur. Clustering is considered an appealing approach for finding similarities in data and putting analogous data into disparate sets. Clustering partitions a data set into some kind of groups such that the similarity within a group is superior to that among groups. The basic idea of data grouping (clustering) is simple to use and in its nature and is very near to the human way of thinking; whenever they are offered with a large amount of data, humans are habitually tend to summarize this huge number of data into a small number of groups or categories in order to further facilitate its analysis. 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. 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].

Decision tree

Decision tree is one of the important analysis methods in classification. A decision tree is a classifier expressed as a recursive partition of the instance space. It builds its optimal tree model by selecting important association features. Even as selection of test attributes and separation of sample sets are two crucial parts in construction of trees. Unusual decision tree methods will implement different technologies to settle these problems. Traditional algorithms include ID3, C4.5, CART, SPRINT, SLIQ etc. Decision tree support tool that uses tree-like graph or models of decisions and their consequences [3][4], 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. 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.

II. RELATED WORK

A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree. This method was improved from traditional algorithms like CURE and C4.5 appropriately, and presents a new synthesized algorithm CA for mining large-scale high dimensional datasets. The basic idea of CA is 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. The experiments show the efficiency of CA is higher not only in clustering but also in decision tree. CA is receptive to a few parameters like the clustering number, shrink factors and the threshold. C4.5 only can covenant with the dataset that has the classification feature [1].

A survey on Web mining tasks and types was presented in [2]. There are two main affinities in Web Usage Mining determined by the applications of the discoveries; first is General Access Pattern Tracking and second is Customized Usage Tracking. The frequent access pattern tracking investigates the web logs to recognize access patterns and trends. These investigations can discard 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].

Comparisons among various clustering algorithms are presented in [5]. In this comparison six types of clustering techniques- k-Means Clustering, Optics, DBSCAN clustering, Hierarchical Clustering, Density Based Clustering and EM Algorithm are compared. Such clustering methods are implemented and analyzed using a clustering tool WEKA. 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 [5].

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 in [6]. 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 [6].

An Effective Algorithm for Web Mining Based on Topic Sensitive Link Analysis proposed in [7]. 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 Page Rank 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. This algorithm will improve the order of web pages in the result list so that user may get the relevant pages easily [7].

In year 2012, a survey on an Efficient Classification of Data Using Decision Tree was proposed by Bhaskar N. Patel et al [8]. K-means clustering algorithm was selected to improve the training phase of Classification. The Learning classification techniques in can be classified into three fundamental types; first is supervised second is unsupervised and finally third is reinforced. Training an average decision tree escorts to a quadratic optimization problem with bound constraints and linear equality constraints. Training support vector machines involves a huge optimization problem and many specially designed algorithms have been offered. This algorithm is called "Decision Tree Induction" that accelerates the training process by exploiting the distributional properties of the training data. The natural clustering of training data and the overall layout of these clusters are relative to the decision boundary of support vector machines. Decision tree is useful because construction of decision tree classifiers does not require any domain knowledge. The decision tree algorithm is a top-down induction algorithm. 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 [8].

In the same year; a study on Clustering Using Alternative Decision Tree were presented in [9]. In this met a Multi-Level Clustering mechanism via 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 offered MLC algorithm has been experimentally experienced on a set of data to find a cluster with intimately related object. This technique is used to defeat the existing system trouble, like manual intervention, misclassification and difficulties of finding a partition range and so on. MLC method forms a tree for the clustering process. In the tree structure, the height of each level of nodes represents the similar degree among clusters. MLC incorporate the vision of ADTree features and overcome the existing hierarchical clustering problem. This scheme offers more accurateness of cluster data without manual interference at the time of cluster formation. As compare to pre-defining clustering algorithm either partition or hierarchical, new technique is more robust and easy to reach the solution of real world complex business problem [9].

Classification algorithms discussed by Hanady Abdulsalam et al. [10], 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. [11] presented the problem of constructing decision tree classifiers on data with uncertain numerical attributes. The MacDiarmid's inequality, used in an 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 in [10]. 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 [12].

III. Proposed Algorithm

In this work, we are first applying Clustering algorithm on the original data set to form clustered data set. The clustered data set is then partitioned horizontally into two parties say P1 and P2. After this partition of the dataset into two sub sets horizontal Partition based ID3 Decision Tree Algorithm is applied and a decision tree is formed.

This Method has two phases.

1. Cluster the dataset using K-means Clustering Algorithm.
2. Classified the cluster data Using Enhanced ID3 algorithm with horizontal partitioning.

K-Means Clustering

Step 1: Initialize the cluster center, c_i, i=1,.., c.

This is typically done by randomly selecting *c* points from among all of the data points.

Step 2: Determine the membership matrix **U** by Equation (2).

Step 3: Compute the cost function according to Equation (1). Stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold.

Step 4: Update the cluster centers according to Equation (3). Go to step 2.

Horizontal Partitioning ID3 Decision Tree

```
Require: R, a set of attributes.
Require: C, the class attribute.
Require: S, data set of tuples.
1: if R is empty then
2: Return the leaf having the most frequent value in data set S.
3: else if all tuples in S have the same class value then
4: Return a leaf with that specific class value.
5: else
6: Determine attribute A with the highest information gain in S.
7: Partition S in m parts S(a1), ..., S(am) such that a1, ..., am are the different values of A.
8: Return a tree with root A and m branches labeled a1...am, such that branch i contains ID3(R − {A}, C, S (ai)).
9: end if
```

- Define P1, P2… Pn Parties. (horizontally partitioned).
- Each Party contains R set of attributes A1, A2, …., AR.
- C the class attributes contains c class values C1, C2, …., Cc.
- For party Pi where i = 1 to n do
- If R is Empty Then
- Return a leaf node with class value
- Else If all transaction in T(Pi) have the same class Then
- Return a leaf node with the class value
- Else
- Calculate Expected Information classify the given sample for each party Pi individually.
- Calculate Entropy for each attribute (A1, A2, …., AR) of each party Pi.
- Calculate Information Gain for each attribute (A1, A2,…., AR) of each party Pi.
- Calculate Total Information Gain for each attribute of all parties (TotalInformationGain( )).
- ABestAttribute ← MaxInformationGain( )
- Let V1, V2, …., Vm be the value of attributes. ABestAttribute partitioned P1, P2,…., Pn parties into m parties
- P1(V1), P1(V2), …., P1(Vm)
- P2(V1), P2(V2), …., P2(Vm)
- .
- .
- Pn(V1), Pn(V2), …., Pn(Vm)
- Return the Tree whose Root is labelled ABestAttribute and has m edges labelled V1, V2, …., Vm. Such that for every i the edge Vi goes to the Tree
- NPPID3(R – ABestAttribute, C, (P1(Vi), P2(Vi), …., Pn(Vi)))
- End.

IV. Simulation Result

As shown in the below Table is the time complexity comparison between existing id3 based decision tree and vertical partition based decision tree and was found that the proposed algorithm has less complexity when experimented on different values of dataset.

number_of_instances	id3_time(ms)	HP_time(ms)
20	80	17
25	97	18
50	115	19
100	135	33
200	160	37

Table 1: Time Comparison between existing id3 and horizontal id3

As shown in the below Table is the mean absolute error rate of the proposed rate which is less as compared to the existing id3 decision tree.

number_of_instances	ID3_Mean absolute error	HP_Mean absolute error
20	0.2860	0.1167
25	0.280	0.276
50	0.310	0.290
100	0.350	0.298
200	0.380	0.310

Table 2: Evaluation of Mean Absolute Error (MAE)

Dataset	Existing Work (MSE)	Proposed Work (MSE)
INI	563.055	483.75
STAD	568.129	492.281
OAK	39.3902	31.754
OAC	45.6978	39.594

Table 3. Comparison of Existing and Proposed Work

V. Conclusion

The proposed technique implemented here for the classification and analysis of high dimensional datasets using clustering and decision tree is efficient one in terms of error rate and time complexity. Here clustering is done using K-means clustering algorithm and then classification of data is done using horizontal partition id3 decision tree algorithm.

References

[1] 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.

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

[3] 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.

[4] M. Thottan and C. Ji, “Anomaly Detection in IP Networks,” IEEE Transaction on Signal Processing, vol. 51, no. 8, pp. 2191-2204, 2003.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.