ONLINE DATA MINING TOOL

[1]Umang Kumar, [2]Kusum Sharma, [3]Sumrit Singh, [4]Indu Khatri
[1,2,3] B.Tech Student, [4]Assistant Professor
Department of Computer Science Engineering
Mahavir Swami Institute of Technology, Sonipat, New Delhi-NCR.

Abstract: - These days, in many fields of industries and studies, data mining tools extract information from a variety of databases. Developing such a data mining tool is a nontrivial task, due to selections required from a variety of available algorithms, professionally. In this paper, Online Interactive and Incremental Data Mining tool (OIIDM) is presented. This tool provides a variety of data mining engagements like clustering, association mining, and many more. These tasks are achieved through interacting with the user to provide the satisfaction of performed task. OIIDM helps the user at a time to get the most-correct data mining also among the available also presented at the time of analysis of algorithm root on the input data by considered algorithmic parameter. This tool supports the gradational approach of data mining to the user as graditional data is a major bug, error, or issue in data mining.

1. INTRODUCTION

In the study of history, it is observed that humans are maintaining records for various purposes but the storing style is very different from generation to generation and, with this tradition the recent development in the database method and techniques and the data gathering techniques from the many data sources such as a: - social networking, remote sensing, business data generated and huge amount of data. This data can be a collection of text or multimedia posted by various users from the social media networking sites and in the case of a relational database- the data can be stored in records about the student and the employees in tables or metadata of database, for business firms or retail shops data can be a connection between the many products and their marketing, periodically added to the write-down of the network traffic can be the data, it can be the image is clicked by the geo-satellite such an enormous data can be transmitted into the relevant info. Knowledge which will be used further for different kinds of motives or purposes. This type of extraction of information from the huge amount of dataset is known as "data mining".

While performing data mining various technologies of data mining need to be considered. Those are machine learning, database systems, data visualization & statics, information theory. The following issues need to be carefully handled to perform data mining tasks effectively and efficiently.

1. How will the user determine if a particular algorithm is most appropriate for the input dataset? As currently various data mining algorithms are upgraded and developed to deal with various problems of data mining. Results of data mining vary with data mining algorithms and it is necessary that users should be satisfied with the generated result.
2. How could the user be actively and interactively participating in the mining process until the user's satisfaction?
   Since the background knowledge from the user is crucial to the usefulness of mining results. And;
3. What happens if the user tunes the parameters? In the instance that it's not satisfactory or can be partial satisfaction, the user may change the input parameters, which will further reflect in the generated output.

This paper has proposed and implemented the design of a data mining tool that provides interactive and incremental clustering, classification, and association rule mining, based on the expert system.

= This tool is used by various or different types of users. These users may be beginners or experts in data mining.
= Interactive approach encourages the user to communicate with the system in the data mining process.
= Incremental data mining approach allows users to add a new batch of data to the previous dataset.

2. WORK RELATED TO D.M (DATA MINING)

Interactive data mining is one of the most useful techniques all-over in data mining. The goal of the interactive system is designed to integrate the user's background knowledge into the entire data mining process. Interactive data mining can be considered under a non-deterministic computation system which is an active system that implements context dependent and adaptive behavior and dependent on user willingness.
There are several different types of benefits for interactive data mining are:-

a). Mining different kinds of knowledge from the database.
The needs of different users are not the same and different users may be interested in different kinds of knowledge. Hence it is necessary to cover a broad range of knowledge discovery.

b). Interactive mining of knowledge at multiple levels of abstraction- The mining process needs to be interactive because it allows users to focus on the search for patterns. In an interactive system, users are providing their feedback which is valuable to the system.

c). Adaptive and effective communication between the user and the system. User views, preferences, and strategies play an important role in user and system interactivity.

Data mining is an iterative process and there should be scope for periodically adding new datasets along with the dataset which is processed and this issue can be tackled using incremental data mining techniques. Incremental data mining algorithms essentially reuse previously mined information and try to combine this information with fresh data to effectively compute a new set of frequent itemsets. There are several advantages of this approach such as it saves the user time and effort to go with a new batch of data.
Here in this paper, the algorithm Selection module is presented which is based on the algorithm ranking system. Which considers the various comparative parameters of the algorithm like time and space complexity, and the efficiency of the algorithm after applying this algorithm to the input dataset? It compares the values obtained for each parameter and accordingly assigns points to it at the end perform summation of these points and provide ranking to each available algorithm in case the user is completely unaware of the data mining process. This ranking system will be helpful to the user to know which algorithm is most suitable according to the system.

A simple experiment was performed to evaluate:-

the marking system is shown below in table format, the wine dataset from the UCI repository consists of the- 100 instances and 10 attributes are provided as input to weak the generated result is as shown in the following table:-

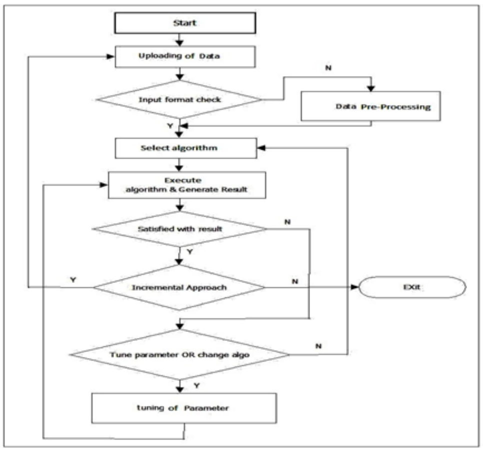
Algorithms and performance parameters:-

Algorithm / Parameter	Time to build	Space	No cluster formed	Order independence	Cluster shape
Modified Cobweb	80 ms	6 M	40	Yes	Contagious clusters
I K-means	40 ms	5M	39	Yes	Well Separated

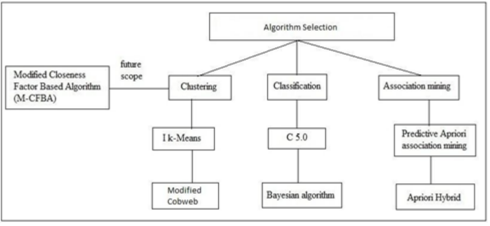
Observation in the above table that for given input the amount of time to build and amount of space Cobweb is greater than K-means. in this, if the user expects an algorithm with less time complexity and space complexity which are basic algorithmic parameters then the system must suggest as K-means is the most suitable algorithm for given input data. If the considerable parameters vary like cluster shapes in clustering etc. then the suggestion might change.

3. OIDM DESIGN

Flowchart for OIDM:-



The Algorithm Selection Module works as:-



Algorithm Selection module

After preferences from the user about the algorithm and execution of the algorithm, the user is again needed to answer queries for satisfaction. If the user is satisfied with the generated result then it's ok. Otherwise, the user can go and change the algorithm or tune the parameter. If the user is satisfied with the results and he/she has a new batch of data then an option is provided for incremental data mining. Modified Closeness Factor-Based Algorithm (M-CFBA) is the future scope of current research.

Clustering technique includes =

Increment k- Algorithm: Incremental - means that it's a mostly used clustering algorithm in different kinds of applications. K-means value algorithm is an efficient algorithm to resolve clustering errors or bugs, this also is little-bit simple and fast. For large data collection, this also is slightly- flexible and highly efficient generally, because of the difficulty or Complexity is O (ntk). And, among which the n is the time of iteration, k is the number of clusters, this is the time of iteration.

Cobweb: (COBWEB using the modified category utility) Cobweb is an incremental system for hierarchical conceptual clustering; it generates hierarchical clustering where clusters are described probabilistically. Generally, it uses a heuristic evaluation measure known as the category utility to guide the construction of a tree to get the highest category utility.

Classification techniques include =

1. C 5.0= the important task of the classification process is to classify new and unseen samples correctly. C5.0 is a classifier that gives efficient classification in less time compared to other classifiers. Memory usage is less in generating decision trees.

2. Bayesian algo = bayesian network is a very strong or powerful probabilistic characterization, and the uses of classification have received considerable attention. Bayesian algorithms predict the class is dependent on the chance of being into that class.

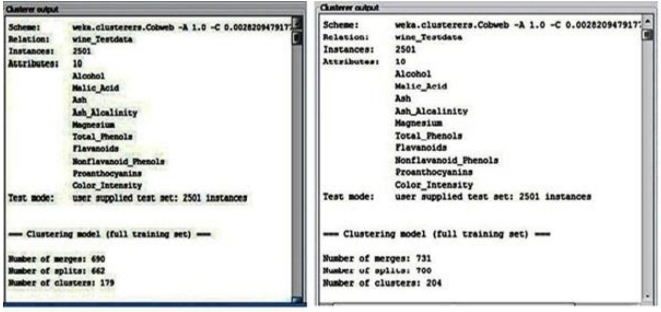
Association rule mining techniques include =

1. Predictive apriori association rule mining algorithm:- In predictive apriori association rule algorithm, support & confidence are combined into a single measure called predictive "Accuracy". And Predictive accuracy is generally used to generate the- apriori association rule. In Weka, this algo produces the “n” perfect or best association rule generally based upon the "n" state by the user. And;
2. Apriori Hybrid: this algorithm is a combination of Apriori and AprioriTid. This combination is formed to remove disadvantages of the mentioned algorithm so ultimately its performance is better than those.

4. EXPERIMENT

The long-established cobweb is improved by using this category utility function (CU) in OIIDM by using this formula.

Where Category utility is a measure of the increase in predictability of attribute values given clustering. The value of category utility will be high when the clustering is good. Maximizing the category achieves a high probability of a cluster for provided variable merit and it conversely. The modified Cobweb and original Cobweb from the Weka tool is applied to the wine dataset which is downloaded from the UCI repository, the result is observable and as shown in the following diagram:-
(a) and (b) ;



```
Scheme:       weka.clusterers.Cobweb -A 1.0 -C 0.00282094791
Relation:     wine_Testdata
Instances:    2501
Attributes:   10
              Alcohol
              Malic_Acid
              Ash
              Ash_Alcalinity
              Magnesium
              Total_Phenols
              Flavanoids
              Nonflavanoid_Phenols
              Proanthocyanins
              Color_Intensity
Test mode:    user supplied test set: 2501 instances


=== Clustering model (full training set) ===

Number of merges: 690
Number of splits: 662
Number of clusters: 179
```

```
Scheme:       weka.clusterers.Cobweb -A 1.0 -C 0.00282094791
Relation:     wine_Testdata
Instances:    2501
Attributes:   10
              Alcohol
              Malic_Acid
              Ash
              Ash_Alcalinity
              Magnesium
              Total_Phenols
              Flavanoids
              Nonflavanoid_Phenols
              Proanthocyanins
              Color_Intensity
Test mode:    user supplied test set: 2501 instances


=== Clustering model (full training set) ===

Number of merges: 731
Number of splits: 700
Number of clusters: 204
```

(a) Modified Cobweb result (b) Original Cobweb result

If the user is not known for his selection of any data mining techniques or methods then this system will be put forward and will enquire about the type of raw data set that the user is willing to provide as input. If the user raw data is a completely numerical data set, then clustering will be the most suitable option among available options. In case of user have categorical or text data then priority should be given to classification or association mining. In the case of mixed-type data set with prior knowledge of classes, like in the Wine and Wine Quality data set of UCI Repository, classification will be the most suitable option.

5. CONCLUSION

OIIDM is "one of its kind" in an effectual collection of various data mining tools. Such a collection "under one roof" was very essential to various categories of users including laymen, students, professionals, decision authorities, to name a few. OIIDM provides a platform for all types of data mining researchers, not only to decide which is the most suitable algorithm for their data but also to validate results given by their manual implementations or by other tools. The user achieves the most optimal and desired result without implementing a single line of code. The user only needs to answer the required expert system queries. where various tools are available in the market, though this tool is available the main intention behind the tool is to provide a powerful tool that allows users to apply different data mining techniques on target data as well as keep working until the user is satisfied and many more. ETC.

REFERENCES

1. Chunfei Zhang, Zhiyi Fang. An Improved K-means Clustering Algorithm. Journal of Gholamreza Nakhaeizadeh, Alexander Schnabl. Development of Multi-Criteria Metrics for Evaluation of Data Mining Algorithms. American association or Artificial Intelligence.: KDD-97
2. Jyoti Arora, Nidhi Bhalla, Sanjeev Rao. A review of association rule mining algorithms. Incremental Conceptual Clustering Using a Modified Category Utility. Springer AI 2004, LNAI 3339, pp. 368–379, 2004
3. Qijun Chen, Xindong Wu, Xingquan Zhu,. Online Interactive Data Mining. Supported by NASA's EPSCoR grant in 2003.
4. Mi Li, Geoffrey Holmes, and Bernhard Pfahringer. Clustering Large Datasets Using Cobweb and K-Means in Tandem. Springer AI 2004, LNAI 3339, pp. 368–379, 2004
5. Ms. Shweta, Algorithm Using Attributes and Comparative Analysis of Various Association Rules
6. Ms S. Vijayarani1, Ms M. Muthulakshmi. Comparative Analysis of Bayes and Lazy Classification Algorithms. International journal of advance research in computer and communication engineering, Vol. 2, Issue 8, August 2013. 5.archive.ics.uci.edu/ml/