A Study of Online Interactive Data Mining Tool (OIIDM) used for clustering, Incremental Clustering, Classification, Association Mining

Rushikesh M. Shete[1], Palash M. Gourshettiwar[2]

Assistant Professor, Department of Information Technology, DMIETR, Sawangi (M), Wardha, Maharashtra, India[1]

Assistant Professor, Department of Computer Technology, KDKCE, Nagpur, Maharashtra, India[2]

ABSTRACT: These days various fields of industries and studies require data mining tools to extract knowledge from variety of databases based on different data. Developing such data mining tool is nontrivial task, due to selections required from variety of available algorithms. In this paper Online Interactive Incremental Data Mining tool (OIIDM) is presented. This tool provides variety of data mining tasks like clustering and incremental Clustering, classification, association mining. These tasks are achieved through interacting with user to provide satisfaction of performed task. OIIDM help user to get appropriate data mining algorithm among the available algorithms by performing analysis of algorithm based on input data by Considering Algorithmic parameter. This tool also support to the incremental approach of data mining to user as incremental data is one of the issues in data mining

KEYWORDS: Interactive Data Mining; Clustering; Classification; Association Mining.

I. INTRODUCTION

In study of history it is observed that humans are maintaining records for various purposes but their storing style varies from generation to generation, with those tradition recent developments in database techniques and data gathering techniques from various data sources such as social networking, remote sensing, business data generate huge amount of data. This data can be collection of text or multimedia posted by various users of social networking sites, in case of relational database data can be stored records about students or employee in tables or metadata of database, for business firms or retail shops data can be relation between different products and their sell, periodically added records of the network traffic can be the data, it can be images captured by geo-satellite such huge data can be transformed into meaningful information which will be further used for various purpose. Such extraction of meaningful information from huge amount of dataset is called as data mining. While performing data mining various technologies of data mining need to be consider those are machine learning, database systems, data visualization & statics, information theory. Issues need to be carefully handled in order to perform data mining task effectively and efficiently.

1. How user will decide particular algorithm is most appropriate for input dataset? As currently various data mining algorithm are upgraded and developed to deal with various problems of data mining. Results of data mining are varies with data mining algorithm and it is necessary that user should be satisfies with generated result.

2. How could the user be actively and interactively participating in the mining process until user's satisfaction? Since the background knowledge from user is crucial to the usefulness of mining result [6].

3. What happen if user tunes the parameters? In case of is not satisfied or partial satisfaction user may change the input parameters, which will further reflect in the generated output. This paper have proposed and implemented design of data mining tool which provide interactive and incremental clustering, classification and association rule mining, based on expert system.

- Tool can be used for different types of users these user may be beginner or expert in data mining.
- Interactive approach encourage user to communicate with system in data mining process.
- Incremental data mining approach allows user to add new batch of data in previous dataset.

II. RELATED WORK

Interactive data mining is one of the good techniques in data mining. The goal of interactive system design to integrate user's background knowledge into the entire data mining process. Interactive data mining can be considered under non deterministic computation system which is active system that implements context dependant and adaptive behaviour and dependant on user willingness.

There are various benefits of interactive data mining like

1. Mining different kinds of knowledge from database Need of different user is not same and different user may be interested in different kind of knowledge. Hence it is necessary to cover broad range of knowledge discovery.

2. Interactive mining of knowledge at multiple levels of abstraction- Mining process need to be interactive because it allows user to focus the search for pattern. In interactive system user are providing their feedback which is valuable to the system.

3. Adaptive and effective communication between user and system. User views, preferences, strategies play important role in user and system interactivity [6]. Data mining is iterative process and there should be scope for periodically added new dataset along with dataset which is processed and this issue can be handled by using incremental data mining techniques. Incremental data mining algorithm essentially reuse previously mined information and try to combine this information with fresh data to effectively compute new set of frequent item set. There are several advantages of this approach such as it save user time and efforts to go with new batch of data. Here in this paper algorithm Selection module is presented which is based on algorithm ranking system [3]. Which basically consider the various comparative parameters of algorithm like time and space complexity, efficiency of the algorithm after applying this algorithm to input dataset. It compares the values obtain for each parameter and accordingly assign points to it at the end perform summation of this points and provide ranking to each available algorithm in case of user is completely unaware of data mining process. This ranking system will be helpful to user to know which algorithm is most suitable according to system.
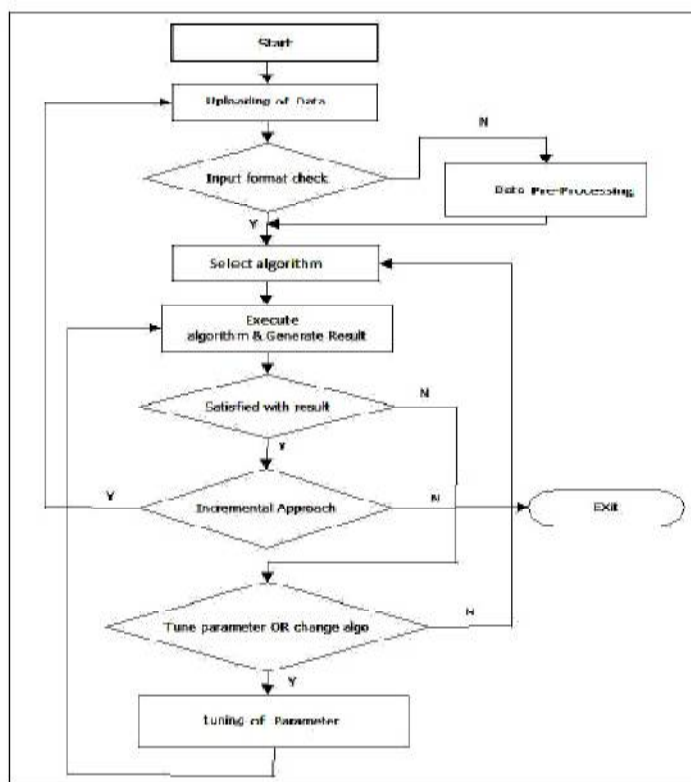
OIIDM Design

Fig.1. Flowchart of OIIDM

Algorithm Selection Module is basically works as follows:

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

A. Clustering technique includes

1. Incremental k-means Algorithm: Incremental k-means is a widely used clustering algorithm in various applications. K-means value algorithm is a efficient algorithm to resolve clustering issues, this algorithm is relatively simple and fast. For large data collection, this algorithm is relatively flexible and high efficient, because the Complexity is O (ntk) [2]. Among which, n is the times of iteration, k is the number of cluster, t is the times of iteration.

2. Cobweb: (COBWEB using the modified category utility) Cobweb is incremental system for hierarchical conceptual clustering, it generate hierarchical clustering where clusters are described by probabilistically. Cobweb uses heuristic evaluation measure called category utility to guide construction of tree in order to get the highest category utility [5]**.**

B. Classification techniques includes

1. C 5.0: The important task of classification process is to classify new and unseen sample correctly. C5.0 is a classifier which gives efficient classification in less time compare to other classifier. Memory usage is less in generating decision tree [1].

2. Bayesian algorithm: Bayesian networks are a powerful probabilistic representation, and their use for classification has received considerable attention. Bayesian algorithms predict the class depending on the probability of belonging to that class [9].

C. Association rule mining technique include

1. Predictive Apriori Association Rule mining algorithm: In predictive Apriori association rule algorithm, support & confidence is combined into a single measure called predictive "Accuracy". This predictive accuracy is used to generate the Apriori association rule. In Weka, this algorithm generates "n" best association rule based on "n" specified by the user [8].

2. Apriori Hybrid: this algorithm is combination of Apriori and AprioriTid. This combination is formed to remove disadvantages of mentioned algorithm so ultimately its performance is better than those [4].
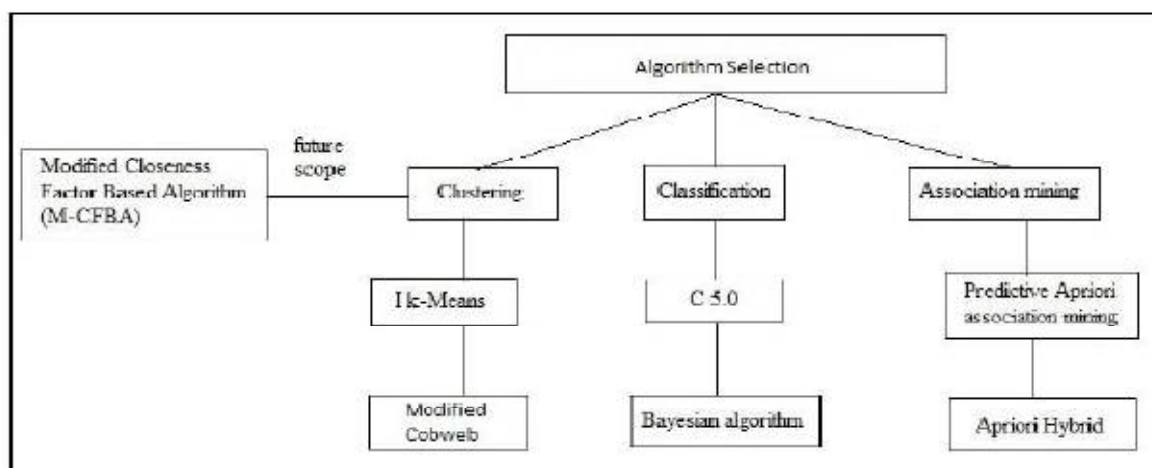

Fig 2: Algorithm Selection Module working

III. CONCLUSION

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

REFERENCES

1.   A. S. Galathiya#1, A. P. Ganatra#2 and C. K. Bhensdadia,. Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning. International Journal of Computer Science and Information Technologies, Vol. 3 (2), 3427-3431, 2012.
2.   Chunfei Zhang , Zhiyi Fang. An Improved K-means Clustering Algorithm. Journal of Information & Computational Science 10: 1 193–199, 2013
3.   Gholamreza Nakhaeizadeh, Alexander Schnabl. Development of Multi-Criteria Metrics for Evaluation of Data Mining Algorithms. American Association for Artificial Intelligence.: KDD 97.
4.   Jyoti Arora, Nidhi Bhalla, Sanjeev Rao. A review on association rule mining algorithms. International journal of innovative research in computer and communication engineering, vol. 1, issue 5, July 2013.
5.   Pyo Jae Kim 1 and Jin Young Choi. Incremental Conceptual Clustering Using a Modified Category Utility. Springer AI 2004, LNAI 3339, pp. 368–379, 2004.
6.   Qijun Chen, Xindong Wu, Xingquan Zhu, Online Interactive Data Mining. Supported by NASA EPSCoR grant, 2003.
7.   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
8.   Ms Shweta, Dr. Kanwal Garg, Mining Efficient Association Rules through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, ISSN: 2277 128X, June 2013
9.   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.
10.   archive.ics.uci.edu/ml/

BIOGRAPHY

Mr. Rushikesh Moreshwarrao Shete is an Assistant Professor in the Information Technology Department, Datta Meghe Institute of Engineering, technology & Research, Sawangi (M), RTM University Nagpur. I received Master of Engineering (M. E.) degree in 2014 from SGBAU, Amravati, MS, India. My research interests are Computer Networks (wireless Networks), Data Mining.

Mr. Palash Mohan Gourshettiwar is an Assistant Professor in the Computer Technology Department, KDK College of Engineering, Nagpur, RTM University Nagpur. I received Master of Engineering (M. E.) degree in 2013 from SGBAU, Amravati, MS, India. My research interests are Computer Networks (wireless Networks), Data Mining, Cloud Computing.