Online Interactive Data Mining Tool

Mahesh Borhade\textsuperscript{a}, Preeti Mulay\textsuperscript{b}\textsuperscript{*}

\textsuperscript{a}\textsuperscript{*}MTech student, Symbiosis International University, Pune 412115, India
\textsuperscript{b}Research Guide, Symbiosis International University, Pune 412115, India

Abstract

Nowadays various fields of industries and studies require data mining tools to extract knowledge from variety of databases. Developing such data mining tool is nontrivial task, due to selections required from variety of available algorithms, professionally. 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.

\textcopyright{} 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC’15)

Keywords: interactive data mining; clustering; classification; association mining.

1. 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

\textsuperscript{*} Mahesh Borhade. Tel.: +91-9960909855.

E-mail address: mahesh.borhade@sitpune.edu.in
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. Following 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.

2. 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 behavior 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 tackled 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. One simple experiment is perform to evaluate
marking system as follows, wine dataset from uci repository with 100 instances and 10 attributes are provided as input to weka the generated result is as shown in following table:

<table>
<thead>
<tr>
<th>Algorithm/Parameter</th>
<th>Time to build</th>
<th>Space</th>
<th>No of cluster formed</th>
<th>Order independence</th>
<th>Cluster shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified Cobweb</td>
<td>80 ms</td>
<td>6 M</td>
<td>40</td>
<td>Yes</td>
<td>Contagious clusters</td>
</tr>
<tr>
<td>K-means</td>
<td>40 ms</td>
<td>5 M</td>
<td>39</td>
<td>Yes</td>
<td>Well Separated clusters</td>
</tr>
</tbody>
</table>

Observation in table 1 and graph shows that for given input amount of time to build and amount of space Cobweb is greater than K-means. In this if user expecting algorithm with less time complexity and space complexity which are basic algorithmic parameter then system must suggest as K-means is most suitable algorithm for given input data. If the considerable parameter varies like cluster shapes in clustering etc then the suggestion might be change.

3. OIIDM Design

![Fig 3.1. Flowchart of OIIDM](image-url)
Algorithm Selection Module is basically works as follows

![Algorithm Selection Module Diagram]

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.

3.1 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 an 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 \(O(nkt)\) [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].

3.2 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]

3.3 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].

4. Experiment

Traditional Cobweb is modified by using category utility function (CU) in OIIDM by using the formula

$$CU(C1, C2, ..., Ck) = \frac{1}{|C|} = \sum_{l=1}^{C} Pr(Cl) \left[ \frac{1}{\sqrt{\pi}} \sum_{i=1}^{A} \left( \frac{1}{\sigma_i} - 1 \right) \right]$$

| C | - Is the number of clusters in a dataset   | \( \sigma_i \) - is standard deviation of attribute \( a_i \)
| A | - Indicates the number of attributes of each instance

Pr | Cl | - Probability of Cluster Cl

Where Category utility is a measure of increase in predictability of attribute values given a clustering [7]. The value of category utility will be high when the clustering is good. Maximizing category utility achieves high predictability of a cluster for given variable values and vice versa. This modified Cobweb and original cobweb from weka tool is applied to the wine dataset which is downloaded from uci repository[10] the result is observable and as shown in fig (a) and (b).

![Fig 4.1 (a) Modified Cobweb result (b) Original Cobweb result](image)

If the user is unaware of his selection of data mining techniques then this proposed system will enquire about type of raw data set that user willing to provide as input. If the user raw data is completely numerical data set, then clustering will be most suitable option among available options. In case user have categorical or text data then the
priority should be given to classification or association mining. In case of mixed type data set with prior knowledge of classes, like in Wine and Wine Quality data set of UCI Repository [10], classification will be the best suitable options.

5. 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.

6. 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), 2012,3427-3431
10. archive.ics.uci.edu/ml/