International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Knowledge Discovery in Data Mining and Massive Data Mining

Prof. Dr. P. K. Srimani[1] , Mrs. Malini M Patil[2]
[1]Former Chairman, Department of Computer Science and Maths, Director, R&D,
Bangalore University, Bangalore, Karnataka, India.
[2]Assistant Professor, Department of ISE, JSSATE, Bangalore, Karnataka, India.
Research Scholar, Bharthiyaar University, Coimbatore, Tamilnadu, India.

Abstract: *Knowledge discovery is a process of non trivial extraction of previously unknown and presently useful information. The rapid advancement of the technology resulted in the increasing rate of data distributions. The data generated from mobile applications, sensor applications, network monitoring, traffic management, weblogs etc. can be referred as a data stream. The data streams are massive in nature. The present work mainly aims at knowledge discovery using data mining and massive data mining techniques. The knowledge discovery process in both the techniques is compared by developing a classification model using Naive bayes classifier. The former case uses Edu-data, a data collected from technical education system and the latter case uses massive online analysis frame work to generate the data streams. Mining data stream is referred as Massive Data Mining. The data streams must be processed under very strict constraints of space and time using sophisticated techniques. The traditional data mining techniques are not advised on this massive data. Therefore the massive online analysis framework is used to mine the data streams. The present work happens to be unique in the literature.*

Keywords: *data mining; data streams; massive online analysis; massive data mining; classification;*

I. Introduction

In the recent decade it is observed that large volumes of data are collected in many organizations. The size of this large volumes of data ranges from some tera bytes to peta bytes[1]. Within these large masses of data which are oftenly referred to as very large data bases (VLDBs), there lies a hidden information of strategic importance. Data mining(DM) is a solution to find this hidden information from the VLDBs. The process of discovering the hidden information is called as knowledge discovery in data bases(KDD). Many business organizations worldwide are already using data mining techniques to find customer buying patterns, to reconfigure their products to increase sales, and to minimize losses. Data mining technology uses powerful technologies to quickly and thoroughly explore the data. Today's era of technology has resulted in the increasing rate of data generation. This is mainly because of different mobile applications, sensor applications, measurements in network traffic monitoring and management, log records and click streams in search engines, web logs, emails, blogs, twitter posts etc. This kind of data generated can be considered as a streaming data since it is obtained from an interval of time. Thus a data stream is defined as an ordered sequence of items that arrive in timely order[2,3]. Data streams are different from data in traditional databases. They are continuous, unbounded, usually come in high speed and have a data distribution which often changes with time[3]. Mining a stream data is referred as data stream mining or Massive Data Mining (MDM).

A. An overview of Data Mining and Massive Data Mining

This section is presented with the comparison of traditional and stream data mining techniques. In traditional data mining the data bases can range in gigabytes, terabytes or even peta bytes in size. Actually, DM is neither an empirical research nor a theoretical one, since no experiments are conducted with an initial start and no theory is proved by using the data. DM is concerned with the analysis of data and the use of the software techniques for finding patterns, regularities in the sets of data. The computational techniques are responsible for finding the patterns, which are previously unknown but, presently useful for future analysis. DM is an integral part of Knowledge Discovery in Databases (KDD), which is the overall process of converting raw data into useful and structured information. DM focuses on different ideas such as sampling, estimation, hypothesis testing from statistics, search algorithms, modelling techniques machine learning theories from artificial intelligence, pattern recognition , machine learning and hi-performance computing,. Thus, data mining is represented as a confluence of many disciplines. The advancement of technology has resulted in the evolution of different techniques in the area of DM. New research findings have resulted in new issues in each technique. To quote some; Association rule mining, Classification, Clustering. Etc

The data stream mining environment has many challenges. The challenges are: (i) Since the data collected is huge, multiple scans are not possible in data stream mining as compared with traditional data mining algorithms. (ii)The mining method of data streams should handle the change in data distribution. (iii) In case of online data streams mining methods should be more faster than the speed of incoming data. and (iv)Memory management issues related to data storage and CPU speed also matter more in data stream mining. The Important features of data streams are: (i) Data streams are huge in size, continuous in nature, fast changing and require fast response. (ii)Random access of data is not possible. (iii) Storage of data streams is limited. Only the summary of the data can be stored. Mining such data needs sophisticated techniques. The main requirements in mining data streams are summarized as follows: (i).The example has to be processed at a time, and inspected only once. (ii). Limited amount of memory can be used. (iii). Work in a limited amount of time. (iv) Any time prediction can be made. The different techniques available in MDM are classification, regression and clustering.

B. Types of Data in DM and MDM

DM and MDM can be applied on different kinds of data sets. Traditional DM techniques uses data available in spread sheets. Some of the software environments used for DM mainly require the data sets to be the format of .xlx, .csv, .arff etc. Many of the data sets taken for the analysis can be obtained from the UCI repository. Data streams can be classified into static (Offline) streams and evolving (Online) streams. Static streams are characterized by regular bulk arrivals. Web logs are considered as Static data streams because most of the reports are generated in a certain period of time. Another best example of Static (offline) data streams is queries on data warehouses. Evolving data streams[2] are characterized by real time updated data that come one by one in time. Examples for evolving (online) data streams are, frequency estimation of internet packet streams, stock market data, and sensor data. Such data should be processed online. Another very important feature of online data streams is that they should be processed online with the rapid speed with which they arrive and should be discarded immediately after being processed. Yet another important feature is bulk data processing is not possible in evolving data streams where as it is possible in static data streams. A quick review of the above discussion is presented in table 1.

Table 1: Comparision of DM and MDM environments

	Traditional	**Stream**
Number of passes	Multiple	Single
Processing Time	Unlimited	Restricted
Memory Used	Unlimited	Restricted
Type of result	Accurate	Approximate
Concept	Static	Evolving
Distributed	No	Yes

Rest of the paper is organized as follows: Section II focuses mainly on the related work in the area of DM and MDM. Section III discusses about the methodology used in DM and MDM. Section IV about the results and analysis in both DM and MDM respectively. Future enhancement of the work and conclusions are briefed at the end of the paper.

II. Related work

The techniques of data mining are exhaustively presented in [4,5,6,7]. Here, the authors have proposed a technique called Educational Mining (Edu-mining) is proposed using classification technique to discover the knowledge from Educational data (Edu-data). Edu-data is a large data repository consisting of data related to technical educational systems. Edu-data is evolved because of huge collection of data mainly from WWW, study material available in the internet, e-learning schemes, computerization of education system, online registration schemes for admission process in the universities, student information system, examination evaluation systems etc. Knowledge discovered helps the Technical Education System (TES) to take useful decisions for maintaining the quality of the education system. In an education system student, faculty and management are the three stake holders of the TES. Edu-mining is carried out on all the three stake holders of TES. The results of the exhaustive research work [6,7] are highly effective in taking optimal decisions at the managerial level. It has earned lot of scope in educational research. Literature survey reveals that lot of work is done in the area of data mining using different data sets and using different data mining techniques [4,5,6,7,8,9,10,11] . It is clear that mining Edu-data is a novel approach. The frame work used in Edu-mining is WEKA[12]. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a data set or called from the Java Code. Weka contains tools for Data-processing, Classification, Regression, Clustering, Association rules and Visualization. It is also well suited for developing new machine learning schemes. As discussed earlier massive collection of data from mobile applications, sensor applications, network monitoring, traffic management, etc., results in a new method of mining and such a type of data evolved is referred to as massive data mining. MOA[13] is the frame work used for massive data mining. Present work elaborates on the MOA framework for MDM. Authors have explored MOA in [14] by comparing the evaluation methods in MOA framework using Naive bayes classifier. Simple perceptron model is developed in [15] using perceptron algorithm.

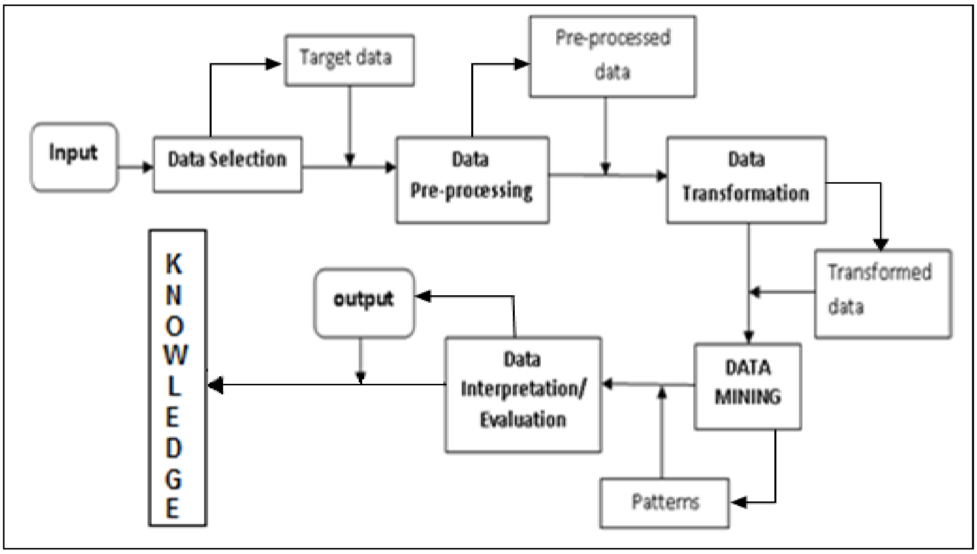
III Methodology

This section mainly discusses about the KDD process and the comparison of DM and MDM in both WEKA and MOA framework respectively. Evaluation process in MOA, Classification technique used in DM and MDM.

A. Steps of KDD

As an example, consider a transaction database maintained by a specialty consumer goods retailer. Suppose the client data includes a consumer name, zip code, and phone number, date of purchase, item code, price, quality and total amount. A variety of new knowledge can be discovered by KDD processing on this client database. During *data selection*, data about specific items or categories of items, or from stores in a specific region or area of the country, may be selected. The *Data cleansing* process then may correct invalid zip codes or eliminate records with incorrect phone prefixes. *Enrichment* typically enhances the data with additional sources of information. For example, given the client names and phone numbers, the store may purchase other data about age, income, and credit rating and append them to each record. *Data transformation and encoding* may be done to reduce the amount of data. For instance, item codes may be grouped in terms of product categories into audio, video, supplies, electronic gadgets, camera, accessories and so on. Zip codes may be aggregated into geographic regions; incomes may be divided into ten ranges, and so on. It is only after such preprocessing that data mining techniques are used to mine different rules and patterns. The results of data mining can be reported in a variety of formats, such as listings, graphic outputs, summary tables, or Visualizations. Representation of knowledge discovery process is shown in Fig 1.

Figure 1. Phases of KDD



B. WEKA frame work

The Weka frame work consists of four different modules: 1) Explorer: An environment for exploring data with 2)Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes. 3)Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning. 4)Simple CLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

C. Steps in MDM and MOA frame work

Massive data mining (MDM) is performed using Massive online analysis (MOA) framework[12]. It is a software environment for implementing algorithms and running experiments for online learning from evolving data streams. MOA is designed in such a way that it can handle the challenging problem of scaling up the implementation of the state of the art algorithms to real world data sets. It consists of offline and online algorithms for classification and clustering. It also consists of tools for evaluation and is an open source frame work to handle massive, potentially infinite, evolving data streams. MOA mainly permits the evaluation of data stream learning algorithms on large streams under explicit memory limits.

D. Evaluation process in MOA

There are two options in the case of evaluation process in MOA.Viz., *Holdout and Prequential*. The first case is suitable when the division between train and test sets is predefined so that the results from different studies could be directly compared. In the second case each individual example can be used to test the model before it is used for training and accordingly the accuracy can be incrementally updated.

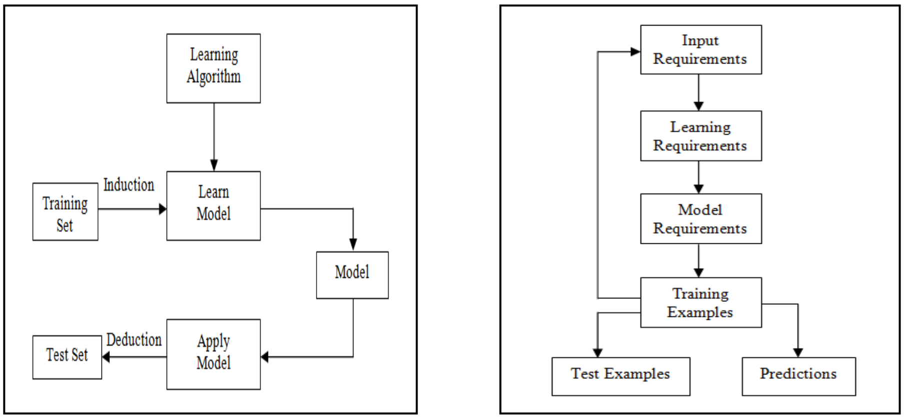
F. Data stream generator used in MDM

For the present analysis the data stream used is random tree generator. It generates a stream based on a randomly generated tree. It constructs a decision tree by choosing attributes at random to split, and assigning a random class label to each leaf. Once the tree is built, new examples are generated by assigning uniformly distributed random values to attributes which then determine the class label via the tree. The generator has parameters to control the number of classes, attributes, nominal attribute labels, and the depth of the tree. A degree of noise can be introduced to the examples after generation. In the case of discrete attributes and the class label, a probability of noise parameter determines the chance that any particular value is switched to something other than the original value. For numeric attributes, a degree of random noise is added to all values, drawn from a random Gaussian distribution with standard deviation equal to the standard deviation of the original values multiplied by noise probability.

G. Classification: A technique used in DM and MDM

The general approach for solving classification problems in DM and MDM are shown in Fig. 3(a) and (b). Firstly in DM, a training set which is used to build a classification model, consisting of records whose class labels are known must be provided. It is then subsequently applied to the test set which consists of records with unknown class labels using a learning algorithm. The objective in classification is to build a mapping function that assigns class labels to each new instance or to verify the appropriateness of class labels already assigned. For example Given a data base $X = \{ x_1, x_2, x_3 \ldots..x_n\}$ of tuples (items and records) and a set of classes $Y = \{y_1, y_2, y_3 \ldots.y_m\}$. Classification is the task of learning a target function f : x → y that maps each attribute set x to one of the predefined class labels y. Informally the target function is also known as a classification model. Secondly, The method MDM mainly consists of the four steps.viz., 1) Select the task. 2) Select the learner. 3)Select the stream generator. 4) select the evaluator. The model is configured with the above said steps. And the results are noted. The configuration of the modules is carried out using the above said steps.

Figure 3. General approach for solving classification problems in (a) DM and (b) MDM



IV Results and Discussions

Case 1: This case mainly focuses on the performance evaluation of the three types of classifiers viz, Rule Based, Decision Tree Based and Baysian networks in DM on Edudata which is a large repository consisting of data related to technical education system(TES) which is considered as a bench mark system for the study of Edumining.[3,4,5,6]. The three important stakeholders of the system are student faculty and management. The study comprises of three modules in student stake holder, three modules in faculty stake holder and finally an integrated module. Totally 3500 instances are taken for each module and the results of the present investigation predict :

1) The optimal classifiers for each module
2) The accuracy and time complexity for all the three classifiers
3) Facilitates to take the effective managerial decisions.

The details are not presented here. Several interesting results are found but only the final results are presented in table 2 for ready reference. In table 2, column 1 corresponds to correctly classified instances; column 2 corresponds to incorrectly classified instances; Column 3 corresponds to accuracy and column 4 to time complexity of the classifiers. The table is self-explanatory. The simulation is done using the WEKA tool.

Case 2: MOA framework basically uses two types of data mining techniques(MDM) namely, Classification and Clustering for data streams. The present investigation focuses on the classification aspect. (In particular Naïve Bayes classifier) The steps of our present analysis on MOA framework are as follows:

Step1: Configuring the model presents the following results:

Task selected: Measures Stream Speed.
Data stream selected: Random tree generator
Size of the Data Stream: 10,000,000

Results obtained from step1:

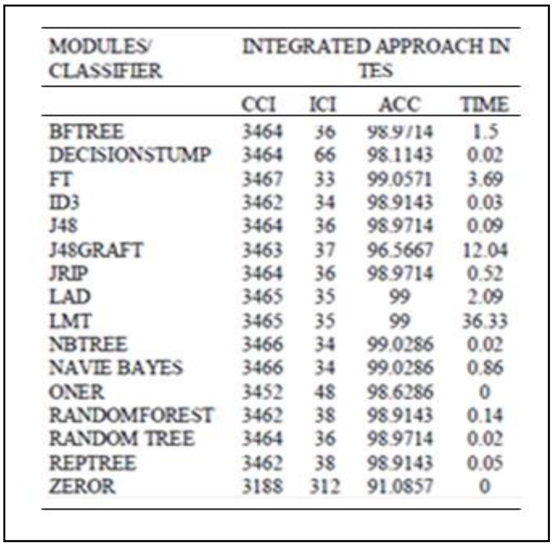
```
Number of instances generated = 10,000,000
Time elapsed = 12.964
Instances per second = 771,385.718
```

Step 2: Selecting the Learner:

Learner selected: Naïve Bayes (NB)　　Size of the data stream selected: 10,000,000

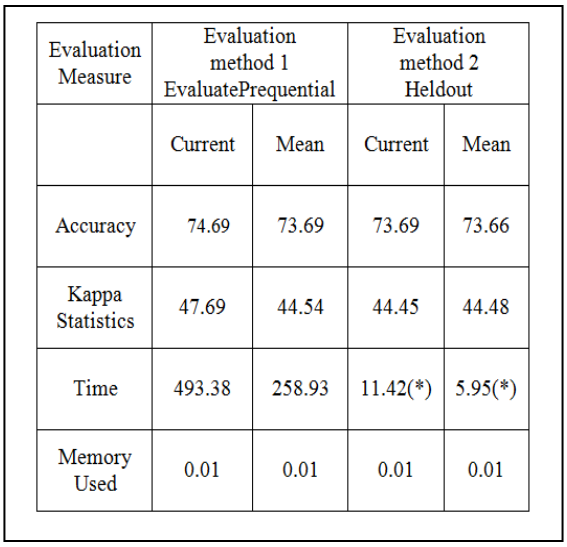
Step 3: Evaluation of the Model : The results of case 1 are presented in table 2 and table 3 gives the comparisons of two evaluation methods on the Naive Bayes learner.

Table 2. Results of the integrated approach in TES



MODULES/ CLASSIFIER	INTEGRATED APPROACH IN TES			
	CCI	ICI	ACC	TIME
BFTREE	3464	36	98.9714	1.5
DECISIONSTUMP	3464	66	98.1143	0.02
FT	3467	33	99.0571	3.69
ID3	3462	34	98.9143	0.03
J48	3464	36	98.9714	0.09
J48GRAFT	3463	37	96.5667	12.04
JRIP	3464	36	98.9714	0.52
LAD	3465	35	99	2.09
LMT	3465	35	99	36.33
NBTREE	3466	34	99.0286	0.02
NAVIE BAYES	3466	34	99.0286	0.86
ONER	3452	48	98.6286	0
RANDOMFOREST	3462	38	98.9143	0.14
RANDOM TREE	3464	36	98.9714	0.02
REPTREE	3462	38	98.9143	0.05
ZEROR	3188	312	91.0857	0

Table 3. Results of the evaluation methods of NB slearner



Evaluation Measure	Evaluation method 1 EvaluatePrequential		Evaluation method 2 Heldout	
	Current	Mean	Current	Mean
Accuracy	74.69	73.69	73.69	73.66
Kappa Statistics	47.69	44.54	44.45	44.48
Time	493.38	258.93	11.42(*)	5.95(*)
Memory Used	0.01	0.01	0.01	0.01

V Conclusions and Future Work

The present work focuses on the comparative study of DM and MDM techniques for Edudata and *random tree data generator* respectively. In the case of edu-data there are 3500 instances with 26 attributes and there are three stake holders mainly student, faculty and management. The student and faculty stake holders comprise of three modules each while the integrated module consists of management stakeholder. In the first case WEKA machine learning tool is used with 16 classifiers and the analysis is carried out in great detail. The final results are presented for ready reference. It is found that Functional Tree (FT) with Accuracy = 99.0571%, Time = 3.69 Sec and Naïve Bayes tree (NBtree) with Accuracy = 99.0286%, time = 0.02 Sec are found to be the best classifiers. In the case of *random tree data generator* the frame work of massive online analysis is used for the classification purpose. The data stream size is 10,000,000 and only Naïve Bayesian classifier is considered for the analysis. In this case the learning model is evaluated by using two evaluation methods Viz. prequential and held out methods respectively. Of the two methods prequential happens to be the best evaluation method with accuracy 74.69 and 493.98 sec. The results of the present study provide a strong platform for enhancing the accuracy of the method effectively. Further it is concluded that for massive data MDM technique is best suited and it has lot of scope for future research. The present work is first of its kind in literature.

References

[1] Ramesh Agarwal, Tomasz Imielinski and Arun Swami, " Mining Association Rules between Sets of items in large Data streams," Proceedings of ACM SIGMOD ,pp., 1-10, May 1993.

[2] C. Aggarwal, Book: "Data streams: Models and Algorithms," Springer, 2007.

[3] Sudipto Guha, Nick K Koudas and Kkyuseok Shim, "Data Streams and Histograms," ACM Symposium on Theory of Computing, 2001.

[4] Srimani P.K and Malini M Patil, " Edu-Mining : A Machine learning approach", ICM2ST-2011, Jaipur, Rajasthan, AIP Conference proceedings, 2011

[5] Srimani P.K and Malini M Patil, " A Classification Model for Edu-Mining," PSRC, ICICS-2012, Dubai, UAE, 2012.

[6] Srimani P.K and Malini M Patil," A Comparative Study of Classifiers for Student Module in Technical Education System(TES)," International Journal of Current Research, Vol 4, Issue, 01, pp., 249-254, January, 2012.

[7] Srimani P.K and Malini M Patil," Performance Evaluation of Classifiers for Edu-data: An integrated approach, " International Journal of Current Research, Vol 4, Issue, 02, pp.,183-190, February, 2012

[8] Cristobal Romero and Sebastian Ventura," Education Data mining A Review of the state of Art," IEEE Transactions on Systems, Man and Cybernetics November 2010, Vol 40, No. 6, pp., 601 -618.

[9] Ryan J.D. Baker and Kalina Yacef, "The State of Education Data mining : A Review and Future Visions,". Journal of Education Data mining. 2009, Vol ,1 ,Issue 1, pp., 3-17.

[10] R. Knauf, R. Boeck, Y. Sakurai, S. Dohi, and S. Tsuruta, "Knowledge mining for supporting learning processes," Proc. Of the 2008 IEEE International Conference on Systems, Man, and Cybernetics, Singapore.

[11] Y. Sakurai, S. Dohi, S. Tsuruta, and R. Knauf, R., "Modeling Academic Education Processes by Dynamic Story Boarding", Journal of Educational Technology & Society", vol. 12, 2009, pp., 307 -333.

[12] Ian H. Witten, Eibe Frank and Mark A Hall Book: " Data Mining : Practical Machine Learning Tools and Techniques," Morgan Kaufmann, 2011.

[13] Albert Bifet, Geoff Holmes, Richard Kirkby and Bernard Pfahringer " MOA: Massive Online Analysis", Journal of Machine learning Research, pp.,1601-1604, November, 2011.

[14] Srimani P.K. and Malini M Patil, "Simple Perceptron Model(SPM) for Evolving Streams in Massive Data Mining (MDM),"International Journal of Neural Networks, Vol 2, Issue 1, pp., 20-24 , Nov 2012.

[15] Srimani P.K. and Malini M Patil, " Massive Data Mining (MDM) on Data Streams Using Classification algorithm," International Journal of Engineering Science and Technology, Vol 6,pp., 2839-2848, June 2012.

Acknowledgements

One of the authors Mrs. Malini M Patil acknowledges J.S.S Academy of Technical Education, Bangalore, Karnataka and Bhartiyaar University, Coimbatore, Tamilnadu, India for providing the facilities to carry out the research work.