IJIRSET

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, Octomber 2013

Data Mining & Data Stream Mining – Open Source Tools

Darshana Parikh, Priyanka Tirkha

Student M.Tech, Dept. of CSE, Sri Balaji College Of Engg. & Tech, Jaipur, Rajasthan, India

Assistant Professor, Dept. of CSE, Sri Balaji College Of Engg. & Tech, Jaipur, Rajasthan, India

Abstract : Term of data mining was available in mid of 1990's. Previously offered through shell script, command line argument, filtering , pipeline. Today its awkward. Now days no of tools are developed by research community in data mining. They provides GUI interface so users can easily communicate. Also they are provide free of cost using open source license. Because of open source users can extend any new method and also provides flexibility in handling various types of data. Here we describe open source tools for data mining and data stream mining. Here we introduce R, WEKA, ORANGE, KNIME and MOA. Our proposed structure use open source tool MOA. MOA contains collections of online and offline for both classification and clustering as well as tools for evaluation. And for data stream mining today MOA is best tool.

Keywords : MOA, WEKA, Data Stream, Big data, Novel class.

I. INTRODUCTION

Data Stream means continuous flow of data. Example of data stream include computer network traffic, phone conversation, ATM transaction, Web Searches and Sensor data. Data Stream Mining is a process of extracting knowledge structure from continuous, rapid data records. [2]Its can be considered as a subfield of data mining. Data Stream can be classified into online streams and offline streams. Online Data stream mining used in a number of real world applications, including network traffic monitoring, intrusion detection and credit card fraud detection. And offline data stream mining used in like generating report based on web log streams. Characteristics of data stream is continuous flow of data. Data size is extremely large and potentially infinite. It's not possible to store all data. But major problems related to data stream mining : Infinite length, concept evolution and concept drift. Infinite length means data stream have a infinite length so require infinite length storage and training time.[3] Concept evolution means developing novel class and concept drift means data changes over time. For our thesis main topic on concept evolution emergence of novel class. Novel class does not exist if we assume total no of classes are fixed. But some time data stream classification problem occur like intrusion detection, text classification and fault detection. So this assumption is not valid for real streaming environment. When new classes may be evolve at any time. Most existing data stream classification technique ignore this important aspect of stream data is the arrival of a Novel Class. Concept evolution solve the problem of infinite length and concept drift.[1]

Now days no of quantity of data generated. Sometimes it was not possible to store all data. Data stream real time analytics are needed to manage data currently generated. Increasing rate from such applications like sensor networks, measurements in network monitoring, traffic management, call detail records, blogging and twitter posts.
In data stream model data arrive at high speed and algorithm must process do under constraints of space and time. In data stream mining we are interested in three dimensions.

1. Accuracy
2. Amount of space(computer memory)
3. Time

This dimensions are typically independent. Issue of measurement of three dimensions simultaneously in data stream mining.[4]

II. BIG DATA

Big data is new term used to identify data set that due to large size. We can not manage them with data mining software tools. Big data mining is the capability of extracting useful information from largest data set or streams of the data. Big data analytics is prove as a important tool in improve efficiency and quality in organization. There are two methods for dealing with big data : sampling and distributed system. Sampling method is used when data set is too large. A good sampling method try to select best instances too have good performance using small quantity of time and memory. And distributed systems used now days are based on map reduced framework. Map reduced model divides algorithm in two steps: map and reduced. The input data is split into different data sets and each split is send to a mapper that will transform the data. The output of the mappers will be combined into reducers that will be output of final organization.[4]

Tools : Open Source Revolution

Early model inference and machine learning programs from the 1980's were most often invoked from a command prompt. Researchers mostly used scripting language as a Perl to separate implement ampling procedures and then execute programs. To compare different algorithms such scripts need to reformat data for each algorithm, parse textual outputs from each model and use them to compute corresponding performance scores. Implementation of this type require so much text processing and programming which is needless. Flexibility and extensibility in analysis software arise from being able to use existing code to develop or extend one's own algorithms.

A. **R**

R is a language and environment for statistical computing and graphics. Most of its computationally intensive methods are implemented in c, c++ , Fortan and then interfaced with R, a scripting language.[4] R includes extensive variety of techniques for statistical testing, predictive modeling and data visualization that become de facto standard open source library for statistics. Interface to R is command line and use through scripting. Extension of R is implemented as an R library and provides a graphical user interface to many of R's data analysis and modeling functions.[5]

B. **Weka**

Weka is best known machine learning and data mining environment. User can access components through JAVA programming or through a command line interfaces. Weka provides graphical user interface in an application called the Weka Knowledge Flow Environment featuring visual programming and Weka explorer. Weka is much weaker in classical testing than R but stronger in a machine learning.
There are mainly two ways to use weka t conduct your data mining tasks.[4]

- Use Weka GUI(graphical user interface)
    - GUI is
    - in KNIME with two data types – models and set of instances. But benefit of giving the user more control in setting up details of the experiment, such as separate preprocessing and training and testing example

sets. Here large straightforward and easy to use. But it is not flexible. It can not be called from you own application.

- Import Weka Java library to your own Java application
  - Developers can leverage on Weka java library to develop software or modify the source code to meet special requirements. Its more flexible and advanced. But it is not easy as GUI.

Overall goal of Weka to build a state-of-the-art facility for developing machine learning technique and allow people to apply them to real world data mining problems.[8]

C. **KNIME**

KNIME is nicely designed data mining tool that run inside the IBM's Eclipse development Environment. Its written in java and can extend its library of built in supervised and unsupervised data mining algorithm with provided by Weka. Each node performs a certain function, such as reading data, filtering, modeling, visualization. Nodes have input and output ports. Some node handle data model as classification trees.[4]

D. **Orange**

Orange is data mining suite built using the same principle as KNIME and Weka knowledgeFlow. In its graphical environment called Orange Canvas, the user places widgets on a canvas and connects them into a schema. Each widget performs some basic function, but unlike number of different visualization of data and models including intelligent search for good visualization. Orange is weak in classical statistics. It provides no widget for statistical testing. Computationally intensive parts of orange are written in C++ where as upper layers are developed in scripting language Python.[4]

E. **MOA**

MOA(Massive On-Line Analysis) is a framework for data stream mining. It includes tools for evalution and collection of machine learning algorithm. It has implementation of classification, regression, Clustering, frequent pattern mining and frequent graph mining. Related to the WEKA project it also implemented in JAVA. It includes a collection of offline and online as well as tools for evaluation: classification and clustering. Easy to design, extend and run experiments. The goal of MOA framework for running experiments in data stream mining context by proving

- Storable setting for data streams for repeatable experiments.
- A set of existing algorithm and measure from literature for comparison.
- An easily extendable framework for new streams, algorithms and evaluation methods.

Workflow in MOA: first a data stream is chosen and configured, second an algorithm is chosen and its parameters are set and third evaluation method or measure is chosen and finally results are obtained after running the task. To run an experiment using MOA , the user can choose between GUI or command line execution.[7]

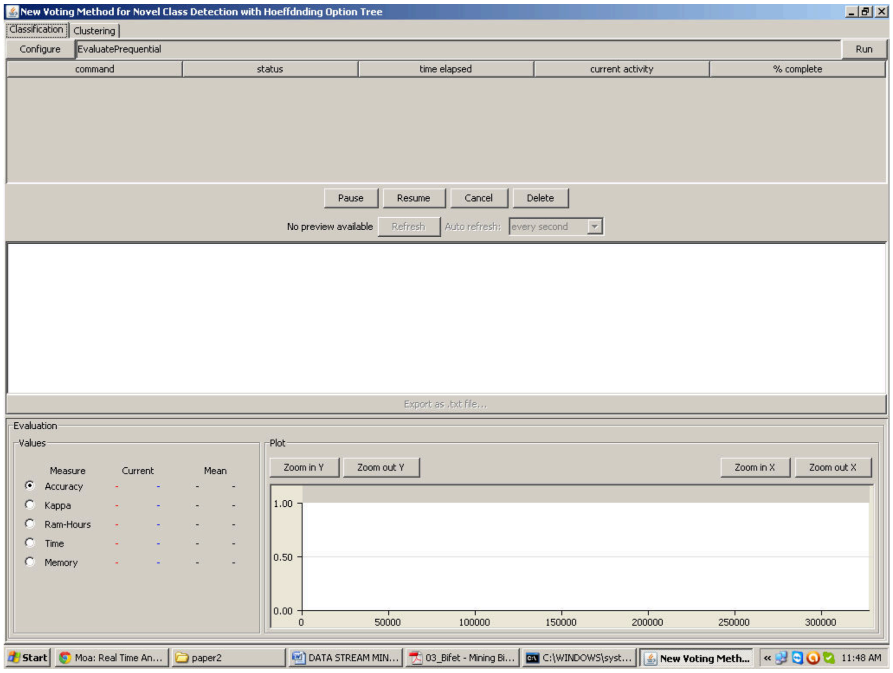
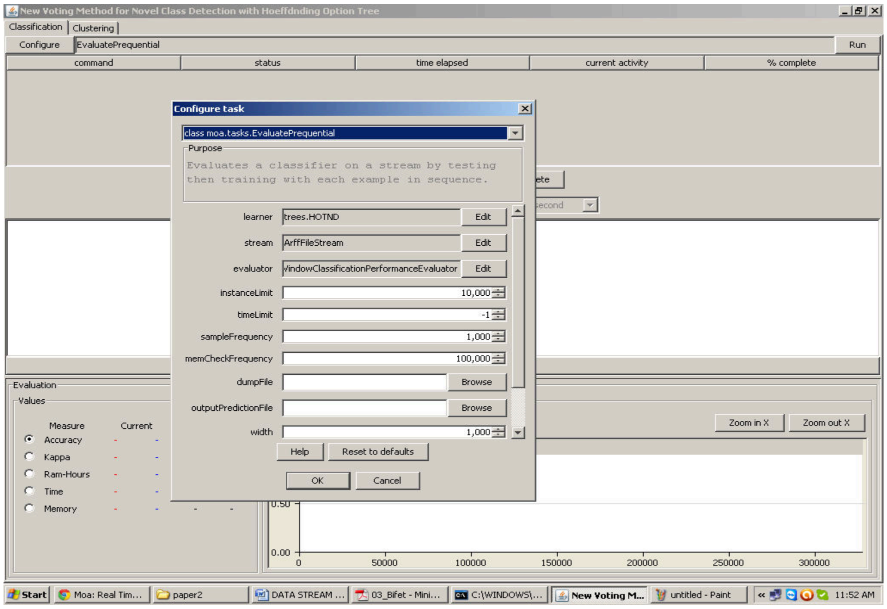


Fig-1 Classification Experimental Setting



class moa.tasks.EvaluatePrequential

Purpose

```
Evaluates a classifier on a stream by testing
then training with each example in sequence.
```

learner trees.HOTND

stream ArffFileStream

evaluator WindowClassificationPerformanceEvaluator

instanceLimit 10,000

timeLimit -1

sampleFrequency 1,000

memCheckFrequency 100,000

dumpFile

outputPredictionFile

width 1,000

Fig-2 Configure Task

III. PROPOSED WORK

MOA contains several classifier methods such as : Naïve Bayes, Decision Stump, Hoeffding tree, Hoeffding Option tree, Bagging, Boosting etc.

Hoeffding Tree:

A Hoeffding Tree is an incremental, any time decision tree induction algorithm that is capable of learning from massive data streams. Hoeffding Tree can often be enough to choose an optimal splitting attribute by Hoeffding Bound. The Hoeffding bound states that with probability 1 – δ the true mean of a random variable of range R will not differ from estimated mean after n independent observation by more than:

$$\sqrt{\varepsilon\ = R^2\ln(1/\delta) / 2n}$$

This bound is useful because it holds true regardless of the distribution generating values.[10]

Hoeffding Option Tree

Hoeffding option tree are regular hoeffding trees containing additional option nodes that allow several tests to be applied , leading to multiple Hoeffding trees as separate paths. They consists a single structure that efficiently represents multiple trees.[9]

IV. PROPOSED STRUCTURE

Here above we discuss about problem of novel class. When we classify data then some data are classified and some are misclassified. But misclassified instances class previously not defined then we consider its not outlier it's a novel class. It's a problem of data stream mining so we can use MOA tool for that and classifier method is Hoeffding Option Tree is used. But in proposed structure voting method of its changed. And also we get more accuracy because here we can use Hoeffding Option Tree.

V. CONCLUSION

State-of-the-art open source data mining provides GUI, focus on usability, interactivity and extendibility. Here we discuss some open source tools for data mining and data stream mining. Our proposed work on data stream mining. So most suitable tool is MOA. MOA build provides experimental framework for classification and clustering. Also this provides classifier Hoeffding tree and Hoeffding option tree. So using that we can modify method and improve accuracy to detect novel class.

REFERENCES

[1] Mohammad M Masud, Tahseen M, Al-khateeb, Latifur Khan, Charu Aggrawal, Jing Gao, Jiawei Han and Bhawani Thuraisinghum Detecting Recurring and Novel classes in Concept Drift Data Streams icdm, IEEE 11th International Conference On Data Mining, pp. 1176-1181, 2011.

[2] S.Thanngamani DYNAMIC FEATURE SET BASED CLASSIFICATION SCHEME UNDER DATA STREAMS *International Journal Of Communication And Engineering Volume 04 – No .04, Issue:01 March-201.*

[3] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani Thuraisingham Classification And Novel Class Detection In Data Stream With Active Mining M.J.Zaki etal.(Eds.): PAKDD 2010, Part II,LNAI 6119, pp.311-324 Springer- Verlag Berlin Heidelberg 2010.

[4] Albert Bifet Mining Big Data in Real Time Infomatica 37 (2013) pp. 15-20.

[5] Blaz Zupan, PhD[a,b,*] , Janez Demsar, PhD[a] Open Source Tools for Data Mining , Clin Lab Med 28 (2008) pp. 37-54.

[6] Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Philipp Kranen, Hardy Kremer, Timm Jansen and Thomas Seidl MOA: Massive Online Analysis, a framework for Stream Classification and Clustering.

[7] MOA , http://moa.cms.waikato.ac.nz .

[8] WEKA, http://www.cs.waikato.ac.nz/ml/weka .
[9] Geoffrey Holmes, Richard Kirkby, and Bernhard P Fahringer Mining Data Stream Using Option Trees(revised edition 2004).
[10] Pedro Domingos, Geoff Hulten Mining High-Speed Data Streams in proceeding of the 6th ACMSIGKDD International Conference On Knowledge Discovery and Data Mining, pp.71-80, ACM, August-2000