IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Comparative Study of Different Classification Algorithms for Stream Data Mining Using MOA

Ashish P. Joshi[1*], Biraj V. Patel[2]

[1] Department of BCA,VP & RPTP Science College, Vallabh Vidyanagar, Gujarat, India
[2] G.H.Patel Department of Computer Science and Technology, Sardar Patel University Vallabh Vidyanagar, Gujarat, India

Available online at: www.ijcseonline.org

Accepted: 20/Nov/2018, Published: 30/Nov/2018

Abstarct: In the today's world, the data is much important and it is growing rapidly. It requires some intelligent analysis processing that helps to discover some knowledge from it. These massive data can be processed by the some framework like MOA (Massive Online Analysis). It has predefined data stream mining classification techniques which are used to distribute the data depends on its class. Some of these techniques like Hoeffding Tree, Decision Stump or Naive Bayes are well known. Comparative study of these techniques analyzes same type of data and compares the output. Which gives idea about different algorithms can be used for different purpose.

Keywords: Stream mining, Hoeffding Tree, Decision Stump, Naive Bayes, Classification, Massive Online Analysis (MOA).

I. INTRODUCTION

Stream data mining is very challenging task nowadays. The volume of data in today's world is increasing day by day. To manage or process these large volumes of data some special techniques required. Some of the online framework also provides the facility to analyze online stream data. Massive Online Analysis (MOA) is one of the online open source frameworks that provide predefined algorithms for the different types of stream data processing.

Stream Data Mining

Stream data are quickly changed in terms of memory (Volume), Speed (Velocity) and Form (Variety). Stream data mining is the process to find out the hidden knowledge from different kinds of stream data. MOA is well known framework which is used for process on stream data.

Massive Online Analysis

MOA is the most popular open source framework for data stream mining, with a very active growing community (blog). It includes a collection of machine learning algorithms like classification, regression, clustering, outlier detection, and concept drift detection and recommender system. Also it has variety of tools for evaluation. [1]

Massive Online Analysis (MOA) is also known for BigData Stream Analytics Framework is well suited for handling large volume of real-time data streams at a very soaring speed.

The different kinds of classification and clustering techniques of WEKA can be use with Massive Online Analysis. WEKA is for these techniques for offline data while MOA is use for online real-time stream data.

II. CLASSIFICATION TECHNIQUES

Every data set is associated with the label and this label may know as class. To classify the data, label plays important role. Based on the label associated with data, data can be classified. For example, student classified as slow learner and advance learner. If student get percentage below 50% then its class is slow learner and for above 50% then its class is advance learner. Here, in this paper, discussed about the three types of classification techniques like Hoefdding Tree, Decision Stump and Naive Bayes.

2.1 Hoefdding Tree

The hoeffding tree is algorithm which is use for classification in stream data mining. It is incremental model that can be handled and work on large amount of streams. As per the stream increment, a node is expanded in the decision tree. It is capable to expand tree at optimum splitting level. It is much useful for predict the decision based on the tree classification at the optimum level. Following Figure 1 shows the tree generated by the hoefdding tree for the production management system. It displays the production should increase if demand is more and production should decrease if demand is less.

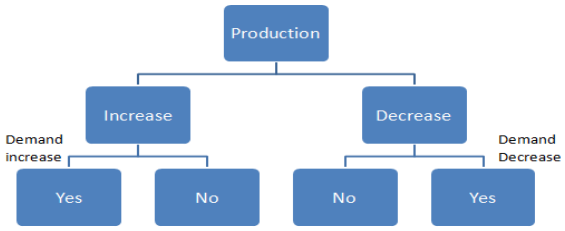

Figure 1: The tree generated by the hoefdding tree algorithm

2.2 Decision Stump

It is opposite to hoefdding tree, the decision stump model is enough capable to generating a decision tree with only single split. The leaf nodes of a decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute. Figure 2 shows tree for classification of student as slow or fast leaner. If percentage is more than 50 then they are in category of fast learner, slow learner otherwise.

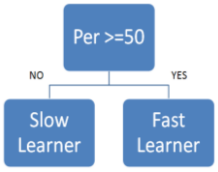

Figure 2: Tree for classification using decision stump algorithm Click on the Configure button, it display different kinds of algorithm for classification and other techniques.

Select specific classification technique from the “model” tab click by edit button.
Click OK button for complete the selection of algorithm.
Run the selected technique by clicking run button for the 10 iterations or 10 step period of time.
Save the result in csv file by selecting the taskResultFile.
Repeat the step from 2 to 6 for other algorithms and generate different csv files. It will create multiple different results in multiple different csv files for the different algorithms.
There is common column in each three file that is accuracy. Copy this column in separate sheet from multple files.
Compare these multiple columns (that means accuracy in different algorithm) by inserting the chart.

2.3 Naive Bayes

The Naive Bayes Classifier technique is based on Bayesian theorem and is particularly used when the dimensionality of the inputs is high. It is prediction based technique that classify the items based on its properties satisfies.

A naive Bayes classifier considers that the presence (or absence) of a particular feature (attribute) of a class is unrelated to the presence (or absence) of any other feature when the class variable is given. For example, a dry fruit is considered as cashew if it is white and curve. It is almond if it is red and rounded corner. The existence of features can predict the class (type) of the item.

III. 3. Comparison Of Different Classification Techniques

Step by step approach for analysis and comparison of similar data with the help of different algorithms.

3.1 Configure the MOA

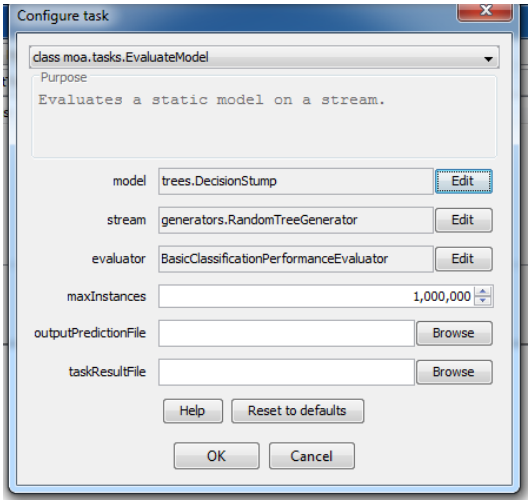

Figure 3: Configuration of MOA

Execute jar file of Massive Online Analysis framework (free download easily available).

3.2 Merge the data in Separate Excel Sheet

The final result looks like below. Figure 3 shows accuracy of three different algorithms.

	A	B	C	D
1		Accuracy		
2	evaluation instances	Hoeffding Tree	Decision Stump	*Naive Bayes*
3	100000	83.2713	57.4726	80.4461
4	200000	83.784	57.5312	80.4268
5	300000	84.2392	57.3596	80.5134
6	400000	82.0095	57.369	80.4112
7	500000	84.4589	57.6595	80.5242
8	600000	84.5056	57.617	80.4481
9	700000	84.5638	57.4336	80.3685
10	800000	83.8932	57.5054	80.4561
11	900000	84.5409	57.5713	80.4406
12	1000000	84.7091	57.4819	80.4432
13				

Figure 3: Accuracy analysis of three different algorithms

3.3 Generating chart from above data

After merging the data from different files in separate excel sheet, generate chart which will display the clear result of accuracy of different kind of algorithm as like figure 4.

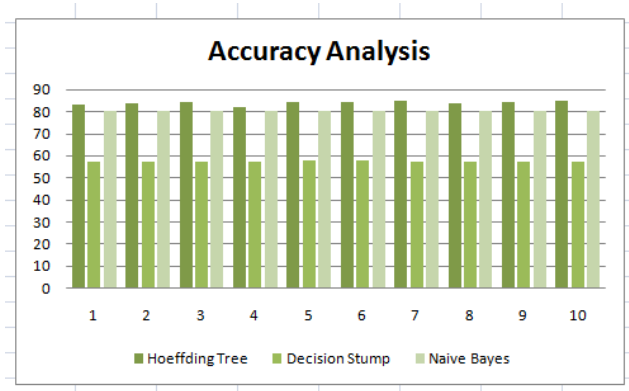

Figure 4: Accuracy analysis of three different algorithms by chart

IV. CONCLUSION

This paper describes the different classification algorithms used for analysis. Here, one can consider the Hoeffding Tree; it is best suited algorithm for accuracy purpose.

REFERENCES

[1]. https://moa.cms.waikato.ac.nz/
[2]. S.Muthukrishnan, ―Data streams: Algorithms and Applications‖.Proceeding of the fourteenth annual ACM-SIAM symposium on discrete algorithms, 2003
[3]. Tusharkumar Trambadiya, Praveen Bhanodia, ―A Comparative study of Stream Data mining Algorithms‖ in International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
[4]. https://www.cs.waikato.ac.nz/~abifet/MOA/ StreamMining.pdf
[5]. http://huawei-noah.github.io/stream DM /docs/HDT.html
[6]. Charu C. Aggarwal - A Survey of Stream Classification Algorithms, IBM T. J. Watson Research Center Yorktown Heights, NY 10598.
[7]. Prithvi Bisht, Neeraj Negi, Preeti Mishra, Pushpanjali Chauhan – A comparative study on various data mining tools for intrusion detection in International Journal of Scientific & Engineering Research Volume 9, Issue 5.
[8]. Dasrhana Parikh, Priyanka Tirkha – Data mining & Data stream Mining – open source tool in International journal of innovative research in science, engineering and technology vol.2 Issue 10.

Authors Profile

Mr. Ashish P. Joshi pursed Master of Computer Application from Saurashtra University, Gujarat in 2006.He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of BCA, VP & RPTP Science College, Sardar Patel University, Vallabh Vidyanagar, Gujarat, India. His main research work focuses on Data Mining, Real-Time Data Mining, Big Data Analytics, IoT and computational Intelligence. He has 12 Year of experience.

Dr. Biraj V. Patel pursued Ph. D. (Computer Science) from Sardar Patel Univetrsity, Gujarat, India in 2013. Working as a Lecturer since 2008 in Department of Computer Science, Sardar Patel University, Vallabh Vidyanagar, Gujarat, India. Research papers published in different International Journals 18. Published a book entitled "Meta Search Engine Optimization". Area of research includes, SEO and Data Mining.

