Indian Journal of Science and Technology, Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/107989, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Data Stream Classification using Random Forest and Very Fast Decision Tree

S. Divya* and N. Sairam

School of Computing, SASTRA University, Tirumalaisamudram, Thanjavur – 613401, Tamilnadu, India;
bosediv@gmail.com, sairam@cse.sastra.edu

Abstract

Objective: Data Stream Classification is a big problem. Ensemble based learning methods are used to tackle the data Stream classification problem. **Method:** Methods such as Random Forest and Very Fast Decision Tree (VFDT) for classification are used in the system. **Findings:** This hybrid approach is an Effective method for calculating hidden data and maintains accuracy in the large data set. It also calculates the neighbors between each pair of cases that can be used in clusters. It is used for finding unknown or (by scaling) gives informational views over the data. This hybrid approach achieves 85 % accuracy and result proves that the hybrid approach performs well when compared to other algorithms in terms of accuracy. This is the application where we can download data streams of any application at faster rate. Many methods are available to process data streams. But the proposed algorithm performs well when compared to other algorithms. **Applications:** Many real time data streams are downloaded and uploaded to test and train various Applications. Data streams are used in many applications such as medical applications and educational applications.

Keywords: Classification, Data Stream, Random Forest, Very Fast Decision Tree (VFDT)

1. Introduction

Data mining is used in many applications. Many tools are available to perform data mining tasks. Data mining is a very powerful tool for extracting useful information from a large data set. Data pre-processing is one of the data mining tasks. In data pre-processing, unstructured raw data are processed into structured data that is in understandable format[1]. Data pre-processing tasks are very essential in data mining because data mining is the process of mining useful information from a large data set. It contains many noisy data and outliers. So, data should be pre-processed to remove those noises and get useful information in a structured format[2]. Classification is the process of classifying the data based on similarity or any other criteria. Classification is very essential because if we classify the data means it is very easy for the learners to study and identify the data of a particular category. Classification plays a vital role in data mining tasks. For performing classification the above mentioned data mining tasks such as data pre-processing and clustering are needed. There are many classification techniques like K-nearest neighbor Classifier, Naive Bayes classifier; Random Forest, Similarity Based Classifier, Decision trees etc are available[3]. In Data Stream, the data flows continuously. The data flow in sequence, in a connection oriented environment. Data flows at a faster rate, so handling Data streams are very tedious process. It is necessary to classify the data streams because data are continuously flowing from various data streams. So, it is difficult to understand which data belongs to which category. So, it is important to classify the data streams. The input arrives, quickly in data streams. The algorithm should work fast enough to process data streams with less memory consumption. In the proposed system a hybrid approach that combines VFDT and Random forest is used for classification, which is efficient and manages abrupt concept changes that occur while handling data streams[4]. In[5] Ensemble learning framework is capable of handling errors in data and it's also able to handle concept drifts and they are capable of constructing Classification models with accuracy. The idea behind this is Decision tree pruning and selection of prototype. Maximum Variance margin

*Author for correspondence

Framework is used for filtering and cleansing the noise. In the Aggregate Ensemble first the base classifiers are trained using different algorithms and performed on different data instances. Model voting is performed to construct a Bayes classifier. The major drawback associated with this type of algorithm is that the accuracy of prediction is not optimal. In[6] the proposed algorithm for handling concept drifts in data streams, namely Ensemble tree decisions for Data streams with Concept drifts (EDTC). The Ensemble Decision tree algorithm uses the concept of random decision trees and double threshold to handle concept drifts. The main drawback of Hoeffding bound is that it is applicable only for application involving numerical data. In[7] used SLIQ—SLIQ expansion is a scalable decision tree, Rain forest and Boat for data stream classification.

2. Method

In the proposed system a Hybrid approach is used that combines the features of Random Forest and VFDT for classification. It was more powerful for making decisions during tie breaking while splitting the attributes. VFDT removes the nodes which are less promising. VFDT use the memory efficiently and handling drifts that arise in data streams. VFDT is efficient in memory and it is not suited for handling drifts that arise in data streams. VFDT is better than other algorithms in terms of accuracy, memory and time. Random forest is an effective method for maintaining accuracy in the large data set. In the System framework, the input data are pre-processed and the top k result is viewed, Random Forest and VFDT classifiers are used for classification shown in Figure 1. Hybrid approach combines the features of Random Forest and VFDT for classification.

3. Discussion and Results

The results of the Hybrid approach to perform data stream classification, achieves better performance when compared to other algorithms. The results obtained that this hybrid approach achieves accuracy of about 85 % is shown in Figure 2. The results are compared with the existing and the proposed model, the performance of the Data stream classifier in the proposed model is better than the existing model in terms of accuracy is shown in Figure 3. The proposed model consumes very less computation time and occupies less storage space when compared to the existing algorithms. Clearly the graph shows that the classifier used in this paper achieves the accuracy of about 85 % is shown in Figure 4.

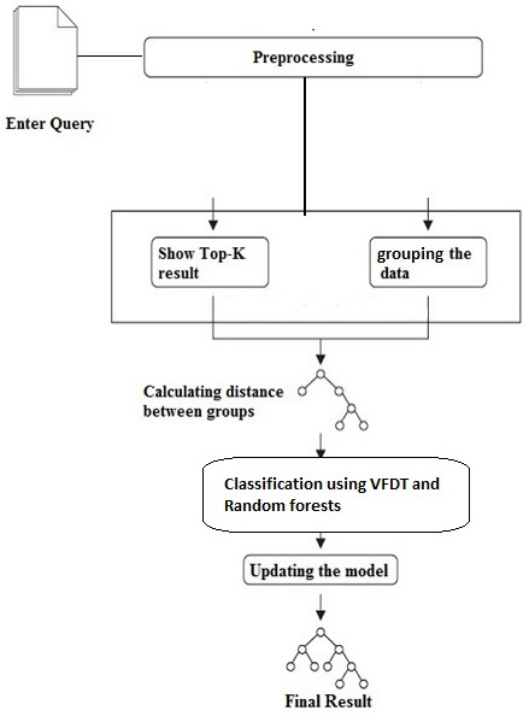

Figure 1. Block diagram of proposed method.

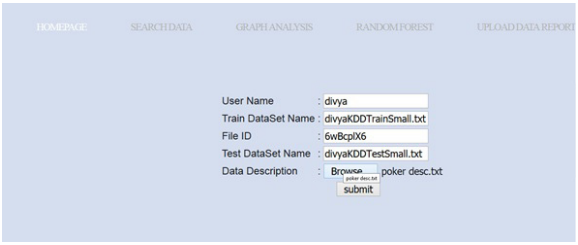

Figure 2. Data upload.

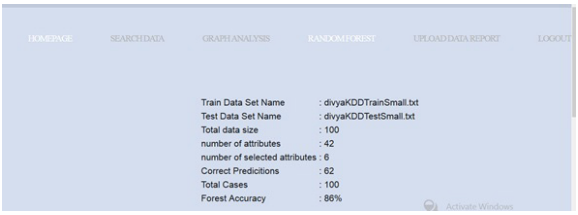

Figure 3. Result with forest accuracy.

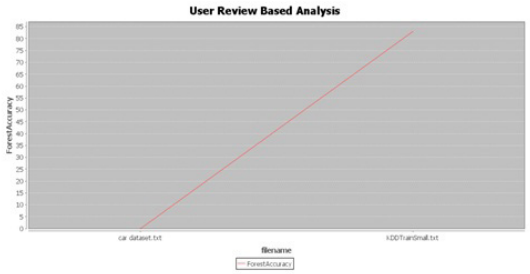

Figure 4. Graph analysis.

4. Conclusion

Data Stream Classification is a tedious process. Efficient algorithm should be used to handle data streams. A framework for Data stream classification using a hybrid approach is proposed. The best features of the Random Forest and VFDT are combined together to improve the accuracy of the data stream classification. By using this hybrid approach sudden concept change can be handled during data stream classification. The result shows that this hybrid approach performs well when compared to other algorithms. Hybrid approach is an Effective method for maintaining accuracy in the large data set.

5. References

1. Zhang P, Zhu X, Shi Y, Guo L, Wu X. Robust ensemble learning for mining noisy data streams. Decision Support System. 2011 Jan; 50(2):469–79.
2. Lia P, Wua X, Hua X, Wanga H. Learning concept-drifting data streams with random ensemble decision trees. Neurocomputing. 2015 Oct; 166:68–83.
3. Aggarwal CC. Data streams: models and algorithms. Springer-Verlag New York; 2006.
4. Brzezinski D, Stefanowski J. Reacting to different types of concept drift: The accuracy updated ensemble algorithm. IEEE Transactions on Neural Networks and Learning Systems. 2014 Jan; 25(1):81–94.
5. Karthika S, Sairam N. A naive bayesian classifier for educational qualification. Indian Journal of Science and Technology. 2015 Jul; 8(16):1–5.
6. Farid D, Zhang L, Hossain A, Rahman CM, Strachan R, Sexton G, Dahal K. An adaptive ensemble classifier for mining concept drifting data streams. Expert Systems with Applications. 2013 Nov; 40(15):5895–906.
7. Shaker A, Hullermeier E. Instance Based Learning (IBL) streams: a system for instance-based classification and regression on data streams. Evolving Systems. 2013 Dec; 3(4):235–49.