Datastream Classification Algorithm For Multiple Target Classes

Shital Wale, Vrushali Bhor, Rohini Shirsath, Gaurav Nagarkar
Computer Department
Dr. D. Y. Patil Institute of Engineering and Technology, Pimpri, Pune, India

Abstract
At the present time, the data arriving is of continuous or non-continuous nature and is of infinitely long sequence, called data stream. The problems associated with the data stream are its classification in short time and to deal with changing class boundaries conditions, called “Concept drift”. In the proposed work, we introduce Data stream classification algorithm that is online, running in amortized time, adaptable, and able to handle multiple target classes deployed on Hadoop. The proposed algorithm is also able to handle irregular arrival of data records, and to adjust its parameters in order to respond to changing class boundaries. Map reduce implementation of C4.5 algorithm which is extension of ID3 algorithm is used for inducing classification rules in the form of decision tree.

Keywords: Data stream, decision tree, Hadoop
Abbreviation and acronyms: RF - Random Forests, CART - Classification and Regression Trees, HDFS - Hadoop Distributed File Systems

I. INTRODUCTION

Many of recent applications like data management, financial applications deals with endless sequences of data, called data streams. Analyzing and mining the data stream is crucial for extracting essential information from it. However, analyzing or mining data stream raises several issues when standard data mining algorithms are used. Standard data mining algorithm requires multiple passes of data to extract essential information, which is the biggest challenge in designing data stream classification algorithm. Hence, the data stream classification algorithm must be online as arrival rate of records is high.

On the other hand, it should be able to handle the changes in the boundaries of data stream. Hence, the classification algorithm should be incremental, so that changes required are updated rather than changing the model and it should be able to extract the essential information in relatively few numbers of passes. The purpose of introducing this algorithm is to overcome the drawbacks of the current existing data stream classification algorithms. The proposed work introduces a data stream classification algorithm that is incremental, able to update changes instead of creating new model deployed on Hadoop. The proposed algorithm handles ordinal or numerical values like flag field, source byte, destination byte, etc. It can also handle non-numerical values like protocol field. The algorithm achieves higher accuracy even for the multiple target classes. It combines the ideas of decision tree implementation using map reduce and Random forests. Random forest algorithm is used to select the attributes and to evaluate classification accuracy. Uses of map reduce technique for the formation of classification rules exhibit time efficiency and scalability. Map Reduce implementation of C4.5 algorithm which is extension of ID3 algorithm is used for inducing classification rules in the form of decision tree.

Following are the features of proposed data stream classification algorithm –

- Data stream classification algorithm has a capability in handling changing thresholds of newly arriving data called, concept drift.
- The algorithm deals with more than two classifiers that is the algorithm can have more than two target classes at its output.
- The training phase of the data stream classification algorithm requires limited amount of data, despite in case of current data stream classification algorithm where it has been observed that higher accuracy is achieved when that much data is used for training.
- Data stream classification algorithm has classification accuracy better than the current data stream classification algorithms.
- The proposed algorithm is incremental, able to update the changes rather than designing new model.

II. RELATED WORK

Random selection of attributes makes individual trees weak. Diverged base classifiers give good ensemble. The improvements suggested are such that individual base classifiers are strong as well as diverse [3]. Standard random forest algorithm gives the method to select the attributes randomly.

Changes are made in random forest algorithm based on accuracy. Comparable to standard version on real data sets, streaming version of random forest achieves
more classification accuracy [1]. But it is not able to handle the data other than numerical or nominal. Change in relation between the input data and target variable is survived online referred as concept drift. Existing traditional decision tree algorithms exhibit multiple limitations such as for a big dataset it becomes very time consuming to build a decision tree [2].

The data stream classification algorithm deals with only two target classes at its output. The training phase requires huge amount of data, as it has been observed that higher accuracy is achieved when extremely large amount of data is used for training in some conventional data stream classification algorithm [4].

Data specific random forest algorithm deals with the imbalanced data. Two approaches used to deal with the imbalanced data are based on sampling technique and cost sensitive learning [5].

III. RECOMMENDED FRAMEWORK

The proposed classification algorithm consists of three phases containing training phase, test phase and the deployment phase.

1. Training phase: The threshold for data classification is set initially on the basis of arrived labeled records.

2. Test phase: The unlabeled data is classified according to the parameters set in the first phase.

3. Deployment using phase: The changing boundary condition of newly arriving non-labeled data is handled. Minute changes in incoming data leads to no change in thresholds, but major changes leads to change in threshold. And on the basis of newly set threshold, the incoming data is classified into respective target classes.

Example:

An example of data stream classification is email spam detection system. Spam is defined as unsolicited commercial email or unsolicited bulk email it is becoming a grave problem on internet. Spam email may contain commercial advertisement, illegal services and products, viruses, cash prizes schemes. Spam increases load on the servers and increases the requirement of the ISP’s bandwidth. Also, it wastes the time of user in reading and deleting the emails. Hence, bifurcating of spam and ham is essential.

An incoming stream of emails is initially unlabeled but as user interacts with emails, their properties get changed. The user may report some emails as explicitly spam, whereas responding to an email is a strong indication that the email is a ham. The emails send by sender are stored on the e-mail server of the receiver. The data stream classification algorithm executes on the server of the receiver. At the server the arriving email is checked on the basis of the protocol, source byte, destination byte, flag field, etc. Based on the header field of the arriving message packet, the email is classified as spam or ham. The spam email is delivered in the spam folder of user, whereas the ham email is delivered to the user in their inbox.

Decision tree

Decision tree resembles flowchart, it consists of nodes. The internal nodes are denoted by rectangle and the leaf nodes are denoted by ovals. It is easy to understand and the most useful algorithm because of its lucid implementation. The outcome of decision tree predicts the classes of email received by the end user such as Spam, normal or other attacks.

Decision trees can be constructed relatively faster than other classification methods using map reduce method. The decisions trees can be easily translated into SQL statements to manipulate the database efficiently. A decision tree enables us to obtain accuracy that is better than conventional data stream classification algorithm methods. A random forest is an ensemble supervised machine learning technique that has emerged recently, which has application in data analysis and decision making.

A decision tree is drawn from left to right, and has burst nodes but no converging nodes. Hence, they can grow very big and are difficult to construct it manually.

A decision tree consists of 3 types of nodes:

1. Decision nodes - commonly represented by squares
2. Chance nodes - represented by circles
3. End nodes - represented by triangle.

Advantages

- Easy to use and understand - Trees are easy to create and visually simple to follow. A brief explanation enables us to understand the decision tree.
- Transparent - The diagrams for a decision clearly reflects out the choices and consequences so that all alternatives can be challenged. Results are clearly explained from the model using simple math.
- Provides an evaluation framework - The value and likelihood of outcomes can be quantified directly on the tree chart.
- Robust – Decision trees can accommodate new assumptions with probabilities when facts are not readily available.
Enables valuation of information - Decision trees enable a calculation of the value of perfect information or the value of knowing what will happen in the future. This can help to determine how much to spend on additional research to improve assumptions.

IV. MAPS REDUCE IMPLEMENTATION OF DECISION TREE ALGORITHM

Decision tree algorithm C4.5 which is extension of ID3 algorithm is used for inducing classification rules in the form of decision tree. The improvements of C4.5 include: (1) employ information gain ratio instead of information gain as measurement to select splitting attributes; (2) not only discrete attributes, but also continuous ones can be handled; (3) prune during the construction of trees to avoid over fitting[2].

V. CONCLUSION

In the proposed work, we will develop an algorithm that is able to classify data stream into multiple target classes, and handles concept drift. The algorithm handles dynamic data and is adaptive. Unlike most of the data stream classification algorithm, the proposed algorithm can deal with non-numerical values for classification of data into multiple target classes.

REFERENCES

[1] IEEE transactions on knowledge and data engineering, vol. 23, no. 1, January 2011 “Classification Using Streaming Random Forests” Hanady Abdulsalam, David B. Skillicorn, Member, IEEE, and Patrick Martin, Member, IEEE Computer Society
[2]. A Map Reduce Implementation of C4.5 Decision Tree Algorithm Wei Dai1 and Wei Ji, School of Economics and Management, Hubei Polytechnic University, Huangshi435003, Hubei, P.R. Chin