International Journal of Pure and Applied Mathematics
Volume 118 No. 18 2018, 1857-1865
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

STREAMING CLASSIFICATION HOEFFDING TREE OF DIABETES MELLITUS USING BOOSTING

[1] J.Jeyalakshmi, [2]S.Poonkuzhali, [3]S.Sree Subha, [4]E.Mohana

Department of Information Technology
Rajalakshmi Engineering College
Chennai, Tamilnadu, India

ABSTRACT

Processing sequence of data of humongous length that may or may not vary over time in an efficient manner is the need of the hour with large amount of data available without inferences being made. Massive online analytics and Stream Analytics can be very helpful in this case. Healthcare data are more critical and may need immediate attention. One such challenge for healthcare related analytics is Diabetes which is a largely growing non-communicable disease. Prediction of diabetes can help in early intervention to control the disease. Stream Analytics is suitable for this area very much. The proposed work performs classification using massive online analytics and improvises the results of classification based on Hoeffding tree optimization using ozaboost techniques.

Keywords: Data Mining, Data Analytics, Machine Learning, Massive Online Analysis, Stream Analysis etc.,

[1] J.Jeyalakshmi, [2]S.Poonkuzhali, [3]S.Sree Subha, [4]E.Mohana

Department of Information Technology
Rajalakshmi Engineering College
Chennai, Tamilnadu, India
jeyalakshmi.j@rajalakshmi.edu.in; poonkuzhali.s@rajalakshmi.edu.in;
sreesubha.s@rajalakshmi.edu.in; mohana.e@rajalakshmi.edu.in

ABSTRACT

Processing sequence of data of humongous length that may or may not vary over time in an efficient manner is the need of the hour with large amount of data available without inferences being made. Massive online analytics and Stream Analytics can be very helpful in this case. Healthcare data are more critical and may need immediate attention. One such challenge for healthcare related analytics is Diabetes which is a largely growing non-communicable disease. Prediction of diabetes can help in early intervention to control the disease. Stream Analytics is suitable for this area very much. The proposed work performs classification using massive online analytics and improvises the results of classification based on Hoeffding tree optimization using ozaboost techniques.

Keywords: Data Mining, Data Analytics, Machine Learning, Massive Online Analysis, Stream Analysis etc.,

INTRODUCTION

In recent years, data streams have become an increasingly important area of research for the computer science, database and statistics communities. Data streams are ordered and potentially unbounded sequences of data points created by a typically non-stationary data generating process. Common data mining tasks associated with data streams include clustering, classification and frequent pattern mining.

The term diabetes mellitus describes a metabolic disorder of multiple etiology characterized by chronic hyperglyceamia with disturbances of carbohydrate,fat and protein metabolism resulting from defects in insulin secretion,insulin action, or both. The effects of DM include long-trem damage, dysfunction and failure of various organs. The dataset collected and generated are huge. Sometimes it is not possible to store and manage all data collected from the patients with the intention to intervene and probably reverse the state of the body. And the nature of the data is different, because with pervasive healthcare, data are pooled from various sources at different points of time with regard to a patient. Sometimes the streaming data have to be processed at a much faster rate where data analytics with the existing techniques is pretty cumbersome.

Hence, data stream real time analytics are needed to manage data currently generated huge. Streaming Data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously, and in small sizes. Streaming data includes a wide variety of data such as log files generated by customers using your mobile or web applications, ecommerce purchases, in-game player activity, information from social networks, financial trading floors, or geospatial services, and telemetry from connected devices or instrumentation in data centers.

This data needs to be processed sequentially and incrementally on a record-by-record basis or over sliding time windows, and used for a wide variety of analytics including correlations, aggregations, filtering, and sampling. Information derived from such analysis gives companies visibility into many aspects of their business and customer activity such as –service usage (for metering/billing), server activity, website clicks, and geo-location of devices, people, and physical goods –and enables them to respond promptly to emerging situations. For example, businesses can track changes in public sentiment on their brands and products by continuously analyzing social

media streams, and respond in a timely fashion as the necessity arises.

Stream learning algorithms are an important type of stream processing algorithms: In a repeated cycle, the learned model is constantly updated to reflect the incoming examples from the stream. They do so without exceeding their memory and time bounds. After processing an incoming example, the algorithms are always able to output a model. Typical learning tasks in stream scenarios are classification, outlier analysis, and clustering.

Since a multitude of algorithms exist for stream learning scenarios, a thorough comparison by experimental evaluation is crucial.For this purpose we introduce Massive Online Analysis (MOA) , for data stream mining. It includes tools for evaluation and collection of machine learning algorithms. It contains several classifier methods such as : Naïve Bayes, Decision Stump, Hoeffding tree, Hoeffding Option tree, Bagging, Boosting etc.[13]

It also has implementation of classification, regression, Clustering, frequent pattern mining and frequent graph mining. The goal of MOA framework for running experiments in data stream mining context by proving - Storable setting for data streams for repeatable experiments.. 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. [14] [15]

RELATED WORK

Recently, with the increase in usage of variety of content like audio, video, images, instantaneous typing with smart phones has become common place. And there is a strong need to process these data at a much faster rate. The following are systems that have been using the streaming and massive online analysis concepts.

Massive Online Analysis

Massive Online Analysis is a open source software to perform stream data mining in real time. MOA is initially started as a project by Machine Learning group in University of Waikato, New Zealand. Massive online analytics is a budding area which will in near future reach huge levels of growth and provide more insights in the future. The massive online analysis has been a much acclaimed technique for stream analysis. In recent times of Internet of Things, the streaming sensor data, video stream analysis and many other streams are used for analysis in pervasive analysis. The following are few related works that have been presented by various authors.

(Gao and Zhu, 2017) present a Integrated Analysis method using data stream analysis, and sort for Deep Learning over the streams. The authors suggest there has been better accuracy with the suggested algorithm. The drift risk parameters are identified in the paper and the temporarily deviating behaviour of the system may be reduced by identifying the drift parameters and controlling them. (Yaseen , Anjum & Antonopoulos, 2016) suggest a work in which the multimedia technology is concentrated for monitoring the video captured by devices like mobile smart phones and camera with less resolution. The content is pre-processed and the Stream classification is used in order to perform better object recognition. Reiz Transform is used to reduce the Gaussian Blur and the precision has been found to be accurate than existing techniques using stream classification.

(Song, He, Niu, & Gao.2016) present a framework for sensor data that is streaming in type. The classifier uses Hoeffding Tree over distributed content in Hadoop framework. Spatial Data identification is performed over this content. With Map Reduce and Cloud Storage the performance is further improved. They provide a simple framework for the data manipulation.

Diabetes

Diabetes is becoming a common condition due to lifestyle change, food consumption styles and

work habits adopted in current years. With the prevailing of the disease in large numbers, it is also mandatory to provide scientific aids powered by computing assistance to spread the awareness, faster screening of the disease, diagnosis and prediction, pervasive healthcare etc.

(Guevara.et.al.,2017) have presented a work on non invasive wearable devices to sense the glycemic levels on Type II Diabetes Mellitus patients. They have used neural networks for classification of Diabetes patients. The method is a screening technique which is used in automated fashion also. (Ünalir. M. O., Can. Ö., Sezer, Bursa, E. O., & H. Ak, 2017) present their work on Big Data application for healthcare system targeted on Diabetes Management. The author summarizes different traditional systems and also the advantages of the proposed work. The author suggests that Big Data supported screening methods are much better compared to the traditional work. The distributed processing of machine learning algorithms have always proven to be much better.

(MohammadRidha *et al*.,2018) provide a fully automated framework for Diabetes Screening for Type I patients. They use glucose levels at before and after food consumption to collect the data samples. They use simulators for this purpose. They simulate the pancreatic secretion of glucose and provide a much elaborate study on the description of mode of working of the pancreatic insulin secretion. (Nieto-Chaupis.,2017), present a paper on nano particles and their usage towards Diabetes. They conduct Monte Carlo Simulation to provide the viability of the device applicability and chemical composition based correction mechanism as to reverse the disease.

The literature discussed above, projects the uses and usages of Streaming Classification and Massive Online Analysis.

SYSEM DESIGN

The proposed system performs massive online analytics by preparing the needful data, taking this as input and converts to stream for processing. The stream is subjected to learning mechanisms and then it is optimized to improve the results.

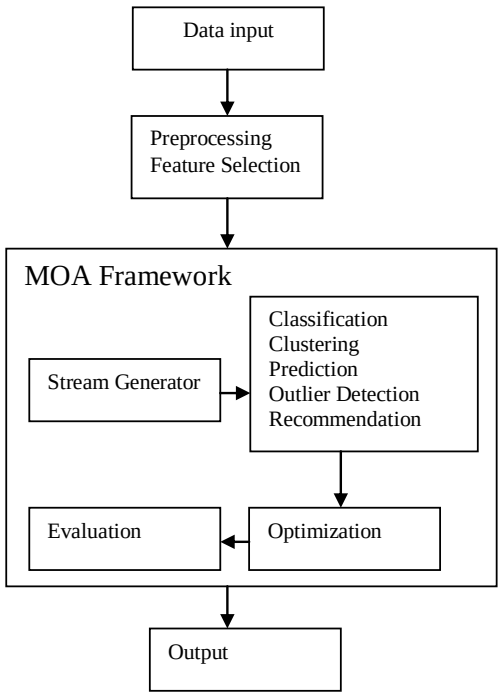

Figure 1. System Architecture

The proposed architecture involves collection and pre-processing of the dataset which is subjected to

needful feature selection. Then the stream generator collects and transforms the data to stream and then as per requirement of the problem, the classification, clustering or prediction platforms can be chosen.

The results of classification can be optimized by moving to boosting techniques and the results can be improvised. Later the evaluation and comparison reports are generated. The proposed system performs Hoeffding Tree Classification and then the results are boosted for optimization. The algorithms are briefed herein.

Algorithm : **Hoeffding Tree**

a. Assume a tree with single node as root

b. For each sample after sorting the observations and populating the leafs update the data to calculate information gain and add the number of independent observations for each sample

c. Compute Hoeffding Bound that helps to choose the splitting attribute with high information gain

$$\varepsilon = \sqrt{\frac{\left(R^2 \ln\left(\frac{1}{\delta}\right)\right)}{2N}}$$

the N = Number of independent observations whose range is R i.e R = range

d. Calculate the difference $\triangle \bar{G} = \bar{G}(x_a) - \bar{G}(x_b)$where the $\bar{G}(x_a)$ is node with highest Information Gain and $\bar{G}(x_b)$ is node with second largest information gain if this difference is greater than the hoeffding bound then it is assumed x_a is the attribute with highest value.

e. Then leaf node will be converted to the decision node.

The results of Hoeffding tree are boosted with ozaboost algorithm to improve the learning process which is mentioned below,

Procedure : Ozaboost

1. Initialize the number of learners to be boosted
2. Initialize the number of correct classification and incorrect classification as zero.
3. Choose a particular learner with particular example, calculate the poison value which is used as weight to be increase or decrease for the learner.
4. Update the learner with the current example with the weight calculated in the step 3.
5. If the learner correctly classified the example, then update the number of correct classification; else update the number of incorrect classification.
6. Repeat the steps 3, 4 and 5 for a learner with all the examples.
7. Repeat steps 3, 4, 5 and 6 for all the learner to be boosted

This algorithm assumes the initial weight as 1 and the Poisson distribution for the observations is utilized by the weak learner. If information is not properly classified, then the second weak learner focuses more on first weak learner and these chains ahead. Prediction is made using the weighted majority voting.

RESULTS AND PERFORMANCE EVALUATION

This section verifies the performance and optimization impact on accuracy and reliability of the proposed scheme through coding, output statements and visualization there from.

Dataset

The dataset is taken from UCI repository, known as PIMA Indians Diabetes Dataset of National Institute of Diabetes and Digestive and Kidney Diseases. The records were collected for 768 subjects. The attributes are explained below.

Table 1:Dataset Description

S.No	Attribute Name	Attribute Description
1	NPG	Occurrence of pregnancy
2	PGL	Concentration level noted in 2-hour OGT test
3	DIA	Diastolic Blood Pressure(in mmHg)
4	TSF	Triceps Skin Folds Thickness (in mm)
5	INS	Serum insulin noted at 2-hour interval
6	BMI	Body Mass Index (in Kg/mm2)
7	DPF	Diabetes Pedigree Function(Family History)
8	AGE	Age of person (in years)
9	Class	Has Diabetes or not

The implementation is performed in RStudio and the results are collected with MOA package in R. The results of the performance are collated for both stream based Hoeffding tree classification and boosted with ozaboost algorithm too. The summary of the results are shown below.

Table 2:Accuracy Comparison

Algorithm	Hoeffding Tree	Ozaboost on Hoeffding Tree
Accuracy	78.01	86.85

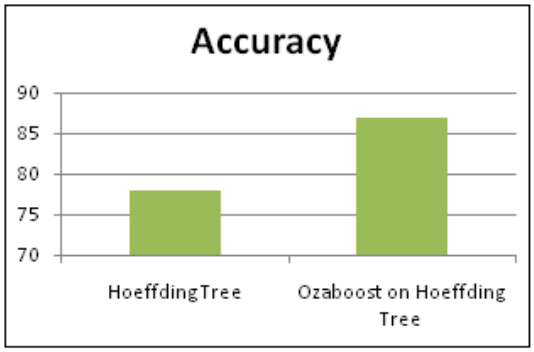

Figure 2: Comparison of Boosting Performance

As is shown in the table 1 and fig 2, the performance of the classification on Diabetes dataset, is implemented using streaming analysis of Hoeffding Tree and then it is optimized by ozaboost algorithm. The performance comparison shows that the boosting applied considerably increases the accuracy of the classifier.

CONCLUSION

Stream analytics is an emerging area. It is finding good scope for implementation these days with the advent of sensors being implemented for pervasive healthcare. India being a global healthcare destination, it is needful that new and innovating methods are being deployed for pervasive healthcare. The proposed work considers stream based hoeffding tree implementation over Diabetes dataset. The accuracy is boosted with ozaboost technique. The algorithm proves to improve well with boosting. The comparison is made in terms of accuracy and the comparison is focused only on stream algorithms. The batch learning classifiers are not considered herein since the context of usage is different.

ACKNOWLEGEMENTS

This research work is a part of the All India Council for Technical Education(AICTE), India funded Research Promotion Scheme project titled “Efficient Prediction and Monitoring Tool for Diabetes Patients Using Data Mining and Smart Phone System” with Reference No: 8- 169 /RIFD /RPS/POLICY-1/2014-15.

REFERENCES

Gao.K., and Zhu. Y.,(2017).Deep Data Stream Analysis Model and Algorithm With Memory Mechanism, *IEEE Access*, 5, 84-93.

Guevara. E., Torres-Galván. J. C., Elías. M. G. R., Luévano-Contreras. C., & González. F. J.,(2017) Non-invasive in vivo Raman spectroscopy of the skin for diabetes screening, 2017 Photonics North (PN), 1-2.

MohammadRidha T. et al.,(2018), Model Free iPID Control for Glycemia Regulation of Type-1 Diabetes, IEEE Transactions on Biomedical Engineering, 65(1),199-206.

Nieto-Chaupis. H.,(2017). Monte Carlo simulation for the very anticipated detection of charged giants proteins in type-2 diabetes patients based on the internet of bio-nano things, 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 1-4.

Song. X., He. H., Niu. S., & Gao. J.,(2016).A Data Streams Analysis Strategy Based on Hoeffding Tree with Concept Drift on Hadoop System, 2016 International Conference on Advanced Cloud and Big Data (CBD), Chengdu, 45-48.

Ünalir. M. O., Can. Ö., Sezer. E., Bursa. O. & Ak H.,(2017). Big data aware diabetes management: Requirements, solutions and reviews, 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), 1-6.

Yaseen. M. U., Anjum. A., & Antonopoulos. N.,(2016). Spatial Frequency Based Video Stream Analysis for Object Classification and Recognition in Clouds, 2016 IEEE/ACM 3rd International Conference on Big Data Computing Applications and Technologies (BDCAT), Shanghai, 18-26.