International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072

SVM CLASSIFIER ALGORITHM FOR DATA STREAM MINING USING HIVE AND R

Mrs.Pranamita Nanda[1],B.Sandhiya[2],R.Sandhiya[3],A.S.Vanaja[4]

[1]Assistant Professor,[2,3,4]Students

Department of Computer Science and Engineering

Velammal Institute Of Technology, Ponneri, Tiruvallur.

Abstract: *Big data is a challenging functionality for analyzing the large volume of data in the IT deployment in a different dimension. To make that analysis process in more efficient manner we use Hive tool for query processing and providing statistical report using RStudio. The processing load in data stream mining has been reduced by the technique know as Feature Selection. However, when it comes to mining over high dimensional data the search space from which an optimal feature subset is derived grows exponentially in size, leading to an intractable demand in computation. To reduce the complexity of using accelerated particle swarm optimization.(APSO), we connect the data by using Hadoop technology. Hadoop technology is easier to store and retrieve the data in a big data environment. With the dataset the data's are analysed and the statistical report is produced using SVM algorithm in R software where R language is used. This R-software environment is used to provide a statisical computing and graphics. This statistical report compares the accuracy between the linear and non linear grid where the higher accuracy dataset is efficient. The final graph provides combination of the linear and nonlinear with respect to cost and sigma which is the userdefined value. PSO with SVM algorithm increases the performance of analysing the data.*

INTRODUCTION:

The process of handling large volume of data, storing and retrieval of data is challenging factor. Data stream mining is the process of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many application of data stream mining can be read only once or a small number of times using limited computing and storage capabilities. Thus for retrieval of data we use data stream mining technique. To make the retrieval of data in efficient manner we use hadoop-hive tool for query processing. It takes less time to process. Process such as converting the unstructured data into structured data by creating schema. Then in hadoop environment there is a data storage place known as hadoop distributed file system where our database is imported from the external device or internal device such as server or system that we are working in to the HDFS using the hive query. The keyword inpath or externalpath is used for importing data from internal device and external device. Then the data is extracted from the database using test data and trained data. The trained data is already existing data's which is just a predicted one. With the trained data the testing is done for analyzing. Both the test data and trained data are used for classification algorithm known as Support Vector Machine. The SVM classifier is the classification algorithm. For a dataset consisting of feature s set and label set an classifier build a model to predict classes. The parameter used for this process is accuracy. The SVM classifier evaluate the predicted data and provides the accuracy. Thus the efficient accuracy is taken into consideration.

EXISTING SYSTEM:

The light weight feature selection technique known as swarm search is used for classfing the dataset. There are many feature selection technique like CCV, Improved PSO etc.,The amount of data feed is potentially infinite and the data delivery is continuous like a high speed train of information.The processing hence is expected to be real time and instantly responsive. The retrieval of data from large volume of data and maintaining them is difficult and the accuracy of the data is little lower which is been overcomed using best classifier algorithm. The complication on top of quantitatively computing the non-linear relations between the feature value and target classes is the temporal nature of such data stream, One must crunch on the data stream long enough for accurately modeling seasonal cycles or regular pattern if they ever exist. There are no straight-forward relations that can easily map the attribute data into a specific class without a long-term observation. This impacts considerately on the data mining algorithm design that should be capable of just reading and forgetting the data stream.

LITERATURE SURVEY:

Big Data though it is hype up-springing many technical challenges that confront both academic research communities and commercial IT deployment, the root sources of Big Data are founded on data streams and the curse of dimensionality. It is generally known that data which are sourced from data streams accumulate continuously making traditional batch-based model induction algorithms infeasible for real-time data mining. In order to tackle this problem which is mainly based on the high-dimensionality and streaming format of data feeds in Big Data, a novel lightweight feature selection is proposed. The feature selection is designed particularly for mining streaming data on the fly, by using accelerated particle swarm optimization (APSO) type of swarm search that achieves enhanced analytical accuracy within reasonable processing time. In this paper, a collection of Big Data with exceptionally large degree of dimensionality are put under test of our new feature selection algorithm for performance evaluation.[1]

The energy-saving research of virtualization of the cloud computing platform shows that there are problems in the management mode of the existing virtualization platform.This model is based on a single node managing the whole platform and the single model is responsible for migrating as well as scheduling all of the virtual machine.Therefore proposing a double management model of the virtual machine is used to solve the problem of single management node bottleneck and scope of the migration.A the same time,the improved PSO algorithm is used to make the plan for virtual machine migration.On the premise of meeting the service performance,the plan achieves energy saving by server booting to a minimum.Through the experiment,it proves that the proposed management mode not only solves the bottleneck problem of single management node, but also reduces the migration scope and the difficulty of the problem. The improved PSO algorithm obviously raises the speed of the migration and overall energy efficiency of scheme.[2]

The cloud storage problem is one of the interesting and important topics in the fields of cloud computing and big data. From the viewpoint of optimization, one discrete PSO algorithm is mainly utilized to handle with the cloud storage problem of the distributed data centers in China's railway and copy with the data between two data centers. In order to achieve the good performance considering the smallest transmitting distance, one discrete PSO algorithm essentially marries each other between two data center sets. Numerical results highlight that the discrete PSO algorithm can provide the guideline for the suboptimal cloud storage strategy of China's railway when the number of the distributed data centers is equal to 15, 17 and 18.[3]

One of the challenges in inferring a classification model with good prediction accuracy is to select the relevant features that contribute to maximum predictive power. Many feature selection techniques have been proposed and studied in the past, but none so far claimed to be the best. In this paper, a novel and efficient feature selection method called Clustering Coefficients of Variation (CCV) is proposed. CCV is based on a very simple principle of variance-basis which finds an optimal balance between generalization and overfitting.By the simplicity of design it is anticipated that CCV will be a useful alternative of pre-processing method for classification especially with those datasets that are characterized by many features.[4]

In a series of recent papers, Prof. Olariu and his co-workers have promoted the vision of vehicular clouds (VCs), a nontrivial extension, along several dimensions, of conventional cloud computing. The main contribution of this work is to identify and analyze a number of security challenges and potential privacy threats in VCs. Although security issues have received attention in cloud computing and vehicular networks, we identify security challenges that are specific to VCs, e.g., challenges of authentication of high-mobility vehicles, scalability and single interface, tangled identities and locations, and the complexity of establishing trust relationships among multiple players caused by intermittent short-range communications. Additionally, we provide a security scheme that addresses several of the challenges discussed.[5]

PROPOSED SYSTEM

We are proposing an approach called data stream mining using Hadoop – Hive technology. To implement the big data analytics in a huge scalability manner, big data needs hadoop for processing the data. The main research challenge here is about finding the most appropriate model induction algorithm for mining data streams. As an additional feature, pertaining to the possibility of embedding the data miner module into some small devices, the memory requirement is opt to be as little as possible for obvious reasons of energy saving and fitting into a tiny device size. In other words, the learned model, probably in form of generalized non-linear mappings between the values of the features to the predicted target classes, must be compact enough to execute in a small run-time memory. No room is wasted for storing the features and their relations that are neither significant nor contribute little to the model accuracy. To this end, without using feature selection is out of consideration, as the number of original features extracted from the data streams. Since these models are built based on a stationary dataset, model up-date needs to repeat the whole training process whenever new samples arrive, adding them to incorporate the changing underlying patterns. In dynamic stream processing environment, however, data classification model would have to be frequently updated accordingly.

ARCHITECTURAL DIAGRAM AND EXPLANATION:

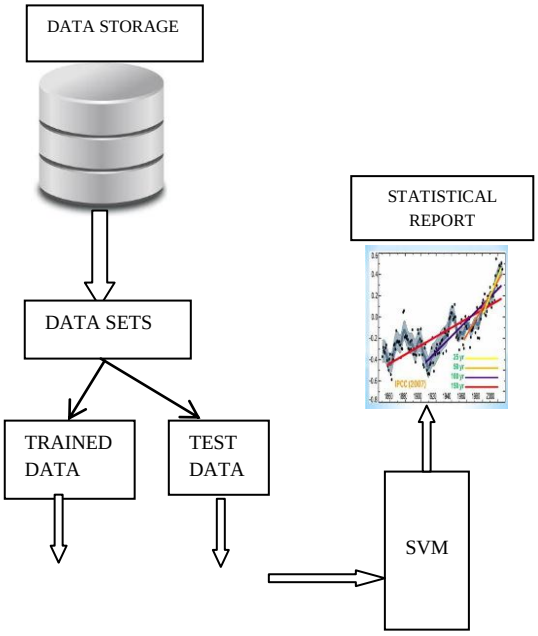


FIG: 2: PROPOSED ARCHITECTURE

From the database the datasets related to the user needs is retrieved using the hive query. The hive is the data warehouse used to analyse and retrieve the data. For this, first we need to continuously upload the data's in database. Then the datasets are retrieved for eg in database there will be the medical datasets, traffic light datasets, weather forcasting datasets etc., from these multiple datasets required one is retrieved using the hive query. The datasets have multiple fields here fields represents age, name, sex etc., The retrieval of data is based on these fields. With the retrieved datasets, analysis is done and divided into two segments known as trained dataset and test dataset. The trained dataset will be more than the test dataset. The trained dataset undergoes some filtering process. But the test dataset undergoes classification where the data's are sliced. And both sliced data and trained data enters into the SVM machine.

The SVM algorithm is used for binary, multi-class problem and anomalie detection. Using hyper planar the critical points are divided known as support vectors. The seperation is then perpendicular bisector of the line joining these two support vectors. These data's are entered into the R input frames. These R input frames is used to extract the data using statistical computing and graphics. It is used to provide statistical report. The statistical report is provided for linear and nonlinear. These report provides accuracy for both stream. Then linear accuracy and non linear accuracy is compared to see the efficiency. Then the grid analysis is done which combines both the accuracy and provides the graph. With that positive and negative data's are identified. The positive data is safe whereas the negative value is unsafe. It increases the efficiency and takes less time for anaysing and for retrieving the data. It improves the data processing speed. It can be able to analyse the large volume of data in a small time compare to another tools. It provides large scale integration of data.

MODULES:

- Create schema in data warehouse
- Importing the data to HDFS
- Extracting the data
- Performance evolution
- Statistical report

MODULE DESCRIPTION:

A) CREATE SCHEMA IN DATA WAREHOUSE:

In database the data's will be in the unstructured format which is unreadable. The database is uploaded in the system and to process the unstructured data in Hive, a schema is created. A schema is created using the attributes which is considered as field in Hive. These fields can be used to divide the data sets as test data and trained data where test data is a unpredicted data and trained data is a predicted data.

B) IMORTING THE DATA IN HDFS:

The Hadoop Distributed File System is designed to store very large dataset and to stream those data sets at high bandwidth to user application. The Database is converted from unstructured to structured format by creating the schema which is loaded into the HDFS. If the database is stored in the desktop then INPATH keyword is used where if it is stored in external devices then EXTERNALPATH keyword is used. The keyword OVERWRITE is used to replace old data with new data.

C) EXTRACTING THE DATA:

The hive query which is used for providing data summarization ,query and analysis. It gives an SQL like interface to query data data stored in various databases and file systems that integrate with Hadoop. Hive provides the necessary the necessary SQL abstraction to integrate HIVEQL into underlying java API without the need to implement queries in the low level API. Hive supports easy portability of SQL based application to Hadoop. It provides the sliced data from the datasets which is relevant to the user query. Using hive the data's are retrieved in faster manner and it can large volume of data. As the database is stored in the system and the processing also take place in the same system, the system act as both client and server.

D) PERFORMANCE EVOLUTION:

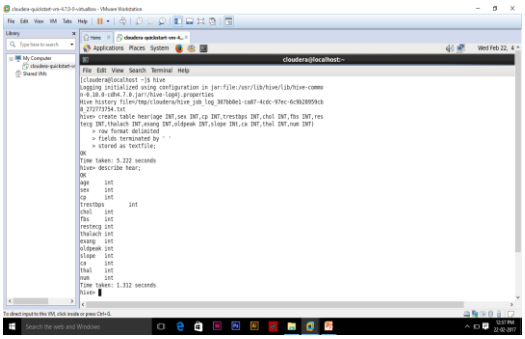
In this approach, the Support Vector Machine(SVM) algorithm is used for analysing and retrival of data. It is a linearized programming and supervised learning approach. It is processed on the basis of Machine Learning(ML) techniques. It accurately reduce the time complexity and code complexity. RStudio is adaptable with any type of data and produces the result with efficient improvement. The SVM algorithm is divided into two types they are linear and radial methods. Accuracy is the parameter which is determined using the SVM algorithm. The linear provides one accuracy and radial provides one accuracy. Comparing these two accuracy the highest accuracy is considered as efficient.

E) STATISTICAL REPORT:

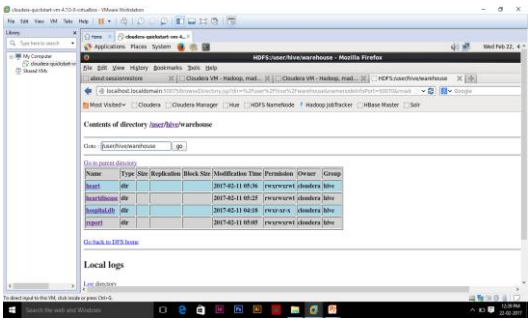
The Statistical report is determined using the Rstudio as per the user needs where R programming language is used for analysing the data. The Rstudio tool provides the graphical representation of the data for our input data. Both the linear and radial is combined to provide grid graph which helps to identify the highly positive and negative value

SCREENSHOTS:

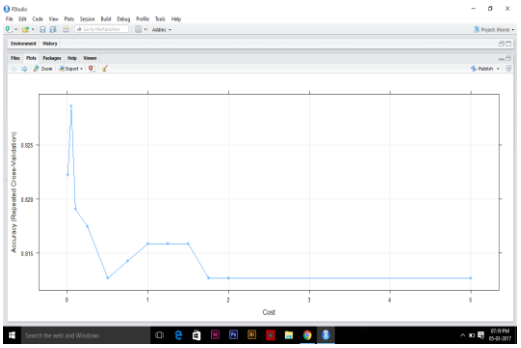
A) CREATING SCHEMA IN DATA WAREHOUSE:



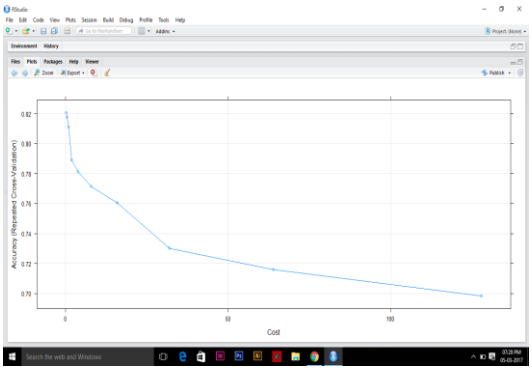
B) IMPORTING THE DATA INTO HDFS:



C) LINEAR KERNEL GRAPH:



E) RADIAL KERNEL GRAPH:



F) RADIAL GRID GRAPH:

CONCLUSION:

An approach known as Hive Tool which is used for storing and retrieving the data in large volume at higher speed. The Hive Tool can be used to process and store the exact data in a large database, compared to other data mining and cloud methodologies. The R-Studio is used o provide the statistical report by anlysing the data in the database as per the user requirement. The PSO with SVM algorithm improves the throughput efficiency.

FUTURE ENHANCEMENT:

In this paper the process of analysing is performed using Hive tool and statistical report is provided using R Software where R language is used. The statistical report provides

positive and negative value in the database. In future using these values prediction is done. This prediction says what will be the future problem with the help of past analysed data. Some new algorithm can be derived to increase the parameters efficiency ie.accuracy and also reduces the time consumption for the retrieval of data from the database.

REFERENCES:

[1] Simon Fong, Raymond wong, V.Vasilakos "Accelerated PSO swarm search feature selection for data stream mining bigdata", IEEE Transaction on Data engineering, VOL.10, NO.7, July 2016.

[2] Ge Rietai, Gao Jing "Improved PSO algorithm for energy saving research in the double layer management mode of the cloud platform", CloudComputing and Bigdata analysis(2016).

[3] Jun Liu, Tianyunshi, Ping Li "Optimal cloud storage problem in the distributed cloud data centres by the discrete PSO algorithm", Institute of computing technologies, china(2015).

[4] Fong.S, Liang.J, Wong.R, Ghanavati.M, "A novel feature selection by clustering coefficients of variations", 2014 Ninth International Conference on Digital Information Management (ICDIM), Sept. 29, 2014, pp.205-213.

[5] Gong Jun Yan, Ding Wen,Stephan dariu, Michael C Weigle "Security challenges in vehicular cloud computing", IEEE Transaction on Intelligent transportation systems, VOL.14, NO.1, March 2013.