Accelerated PSO Swarm Search Feature Selection with SVM for Data Stream Mining Big Data

Author(s): Himani Patel
Department of Information Technology, Siddhant College of Engineering, Sudumbare, Pune

ABSTRACT: **In the modern world there is huge development in the field of networking technology which handles huge data at a time. This data can be structured, semi structured or unstructured. To perform efficient mining of valuable information from such type of data the big data technology is gaining importance nowadays. Data mining application is been used in public and private sectors of industry because of its advantage over conventional networking technology to analyze large real time data. Data mining mainly relies on 3 V's namely, Volume, Varity and Velocity of processing data. Volume refers to the huge amount of data it collects, Velocity refers to the speed at which it process the data and Variety defines that multi-dimensional data which can be numbers, dates, strings, geospatial data, 3D data, audio files, video files, social files, etc. These data which is stored in big data will be from different source at different rate and of different type; hence it will not be synchronized. This is one of the biggest challenges in working with big data. Second challenge is related to mining the valuable and relevant information from such data adhering to 3rd V i.e. Velocity. Speed is highly important as it is associated with cost of processing. This paper focuses detailed study of accelerated PSO Swarm search feature selection and use of support vector machine.**

Keywords: Feature Selection, swarm intelligence, classification, big data, particle swarm optimization.

Introduction

In the computer era, there is a massive improvement in all fields especially in internet and online technologies by new and fast performing technologies. Google has more than 1 billion queries per day, Twitter has more than 250 million tweets daily, and YouTube has over 4 billion views on daily basis. The data size is growing around 40% every year. These massive data has some problems as they need high volume of storage space and it may perform various operations like analytical, retrieval and process operations which are very difficult and time consuming.

To overcome these difficulties, introduction of big data mining stores all these huge and complex data and the required data can also be extracted easily from the large database. This big data processing improves the speed of the data transferring than simple data exchanges. This big data mining is now kept on blooming in different online services and provides a best service to end users or customers. Big data helps the users to recover the data as per their desire. There are the 3V difficulties of Big Data known as: Velocity issue that offers ascend to a tremendous measure of information to be taken care of at a raising rapid; Variety issue that makes information preparing and reconciliation troublesome in light of the fact that the information originate from different sources and they are organized in an unexpected way; and Volume issue that makes storing, handling, and analysis over them difficult.For these 3V difficulties, the conventional data mining methodologies are based on the full batch-mode learning may be not appropriate to meet the systematic proficiencydemand. The conventional data mining model development strategies require the full data set and the data are apportioned by some divide and conquer methodology. Every time when fresh data arrive, the data collection process makes the big data increase to bigger data, it needs to re-execute and the model that was built should be re-created by consideringfresh data. The otherkind of algorithms recognized as data stream mining methods are capable to decrease these 3V issues of enormous information. The algorithm is suitable for instigating an arrangement or prediction model; each pass of data from data streams activates the model to renew itself in incremental manner withoutreloading any previously seen data. This kind of algorithm can conceivably handle data streams that add up to infinity by analyzing in memory.

Therefore a new generation of algorithms, commonlyfamous as incremental classification algorithms or data stream mining algorithms has been proposed to solve this problem. This paper offers insights to inventors who want to design a data stream mining applications over Big Data that may grow continually both in volumes and dimensions.

I. Related work

Two kinds of algorithms were considered for incremental learning: functional-based and decision tree-based. The functional based is building a model, possibly towork as a black-box with numeric weights and coefficients that illustrates the associations between the inputsand the predicted outputs. Two such most well-known functional-based incremental learning algorithms are KStar andUpdatable Naive Bayes.

KStar algorithm learns incrementally per instance by similarity function that calculates the entropic distance between the test instance and the other instances. KStar usually required longer processing time [2].

Updatable Naive Bayes is extended from the popular Naive Bayes classifiers whichassumestrong independence between the features. This assumption is advantageous that it needs little amount of training data to estimate the means and variancesof the features for computing the probabilities of all the possible outcomes for performing classification.The decision-tree based algorithms are second type of algorithms. Some preferred algorithms, includingKStarand Updatable Naïve Bayes is kept into experimental test. Such incremental decision tree algorithms using HB in node splitting test are called Hoeffding Tree.

James Kennedy and Russell Eberhart [5] in 1995 proposed Particle swarm optimization (PSO).It is an optimization technique inspired by natural behavior of bird flocking or fish schooling of finding optimal regions of complex search spaces by interactions of individual particles of population.PSO is initialized with a group of random particles that constitute a swarm, moving around in the search space.Each particle keeps track of its coordinates in the solution space which are affiliated with the finest fitness that has achieved so far by that particle. Chung-JuiTu, Li-Yeh Chuang, Jun-Yang Chang, and Cheng-Hong Yang [3] discussed feature selection using PSO and SVM. The feature selection method is required for sample classification to fasten the processing rate, predictive accuracy and to avoid incomprehensibility. PSO is used to apply a feature selection and SVM as a fitness function of PSO for classification.

Cheng-Lung Huang and Jian-Fan Dun[4] suggested a novel PSO-SVM model which combines the discrete PSO with the continuous-valued PSO to simultaneously improve the input feature subset selection and the SVM kernel parameter setting.Z.-H. Zhan, J. Zhang, Y. Li, and H. S.-H. Chung [6] extended PSO to Adaptive PSO that introduces two new parameters to the PSO, inertia weight to balance the global and local search capabilities in PSO and other one is control of the acceleration coefficients.

II. Proposed algorithm

Aim of the proposed algorithm is to minimize the time required to search for the given input. The APSO swarm search is proposed with SVM classifier to perform the search within the minimum time.

SVM performs a classification by finding the hyper-plane that differentiates the two classes very well. There may be many hyper planes, select the one which segregates the classes better.

(a) For locating data in 2D plane,

For the inputted keyword, a random number is generated.

If the number is positive, then that keyword may be assigned to one of the previously created class.

If the number is negative, then new class is created.

The value is calculated according to the equation

value = y_lower + (y_upper – y_lower) * (value – y_min) / (y_max – y_min)

The keyword is assigned to that class whose value matches approximately closer to the value of the existing class.

where, y_lower & y_upper are the margins of the classified classes.

x_min, x_max are minimum and maximum x co-ordinates.

y_min, y_max are minimum and maximum x co-ordinates.

(b) Termination Criteria

For the searching, the particles are moving continuously. They have their own velocity and position. Each particle has its own local best position. They move continuously to achieve the global best position.

The search is done for maximum 100 iterations. If the keyword is not found, then no result is displayed.

III. Pseudo code

```
Step 1: For the inputted keyword, a random number is generated.
Step 2: If the number is positive, then that keyword may be assigned to one of the previously created class.
Step 3: If the number is negative, then new class is created.
Step 4: Calculates a value that decides the keyword is assigned to which class.
Step 5: Go to the step 4.
Step 6: Obtain results.
```

Results

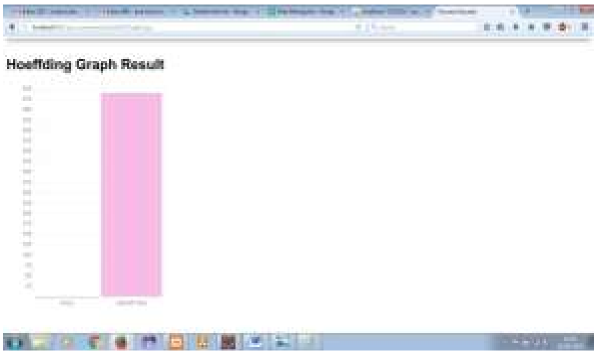

Fig. 1.Existing System Search Time

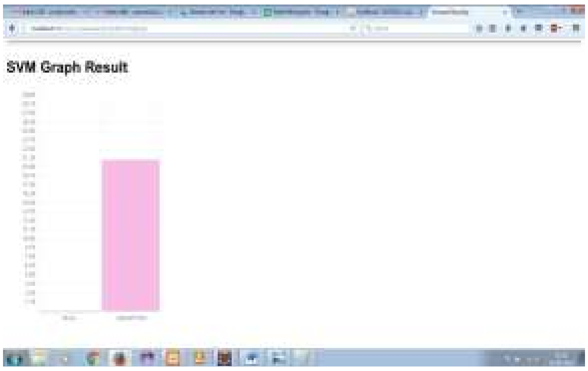

Fig 2. Proposed System Search Time

IV. Conclusion and Future Work

Big data technology has 3V challenges: Volume, Variety and Velocity. The data delivery is continuous. Hence, processing needs to be real time and quickly responsive. Moreover, the memory requirement must be compact enough to fit in a small run time memory. No space should be wasted for storing non-significant features. To shorten the search process and to improve the accuracy of classification model, APSO with SVM is used. The advantage of SVM is that it produces very accurate classifiers and it is robust to noise. The SVM is used with images also in place of text data for the future work.

References

(1) "Accelerated PSO swarm search feature selection for data stream mining big data" by Simon Fong, Raymond Wong, V.Vasilakos, IEEE transactions on services computing, vol. 9, no. 1,Feb. 2016

(2) "Mining big data: current status, and forecast to the future" by Wei Fan, Albert Bifet SIGKDD Explorations, vol. 14, no. 2, pp. 1–5,Dec. 2012.

(3) "Feature Selection in life science classification : Metaheuristic swarm search" by Simon Fong, Suash Deb ,Xin-She yang, Jinyan Lie, IEEE IT Prof. Mag., vol. 16, no. 4, pp. 24–29, Aug. 2014.

(4) "Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis" by Li-Fei Chen, Chao-Ton Su, Kun Huang Chen, Pa-Chun Wang, Springer, 3 May 2011

(5)"Swarm search for feature selection in classification," in Proc. 2nd Int. Conf. Big Data Sci. Eng., Dec. 2013, pp. 902–909.by S. Fong, X. S. Yang, and S. Deb, Proc. 2nd Int. Conf. Big Data Sci. Eng., Dec. 2013, pp. 902–909.

(6)"Feature Selection using PSO-SVM" by Chung-Jui Tu, Li-Yeh Chuang, Jun-Yang Chang, Cheng-Hong Yang, Proc. 3rd Int. Conf. Netw. Digital Technol., Macau, China, Jul. 11–13, 2011, pp. 53–66.