Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2, Issue 11, November-2015

Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data using Genetic Algorithm

Himani Patel[1], Nilesh Mali[2]

[1]*Information technology, SCOE, Pune,*
[2] *Information technology, SCOE, Pune*

Abstract — *In the modern world there is brisk 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 minning of valuable information from such type of data the big data technology is gaining importance nowadays. Data minning application is used mostly in all fields right from science and engineering domains to social networking and biomedical science. It 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 minning 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 does 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 challenge in working with big data. Second challenge is related to minning 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 focus on analyzing the big data technology and provide detail study of accelerated PSO Swarm search feature selection.*

Keywords- *Data mining, Big data, Feature selection, Particle swarm optimization, Generic Algorithm*

I. 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 per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. In todays world data size is in zettabytes, and it 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. Moreover these operations are very complex and purely time consuming one.

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. These tools are very useful to end users in providing quality service. Big data helps the users to recover the data as per their desire.

As of late a ton of news in the media advocates the buildup of Big Data that are showed in three risky issues. They are the 3V difficulties 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 finds storing, handling, and investigation over them both computational and documenting testing.

With respect to these 3V difficulties, the conventional data mining methodologies are based on the full batch-mode learning may be unable in meeting the demand of systematic proficiency. That is just in light of the fact that the conventional data mining model development strategies require the full set of data, and after that 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, the traditional induction method needs to re-run and the model that was built should be re-built with the consideration of new data. Interestingly, the new type of algorithms known as data stream mining methods have the capacity to decrease these 3V issues of enormous information, since these 3V difficulties are mainly the attributes of data streams. Data stream calculation is not stemmed by the massive volume or fast data accumulation.

The algorithm is fit for instigating an arrangement or prediction model; each pass of data from the data streams triggers the model to incrementally update itself with no need of reloading any formerly seen data. This kind of algorithm can conceivably handle data streams that add up to infinity, and they can keep running in memory analyzing and mining data streams on the fly.

Data stream mining over Big Data is emerging and it demands for an efficient classification model that is capable of mining data streams and making a prediction for unseen samples. Traditional classification approach is referred to a method of top-down supervised learning, where a full set of data is used to construct a classification model, by recursively partitioning the data into mapping relations for modeling a concept. Since these models are built based on a static dataset, model update needs to repeat the whole training process whenever new samples arrive. The traditional

models might have a good performance on a full set of historical data, and the data are relatively static without expecting much new changes. In dynamic stream processing environment, the classification model would have to be frequently updated accordingly. Therefore a new generation of algorithms, generally known as incremental classification algorithms or simply, data stream mining algorithms has been proposed to solve this problem. This paper offers insights to developers who want to design a data stream mining applications over Big Data that may grow continually both in volumes and dimensions.

II. FEATURE SELECTION BY SWARM SEARCH AND APSO

"A contemporary type of feature selection algorithm, specially designed for choosing an optimal subset from a huge hyperspace is called Swarm Search-Feature Selection (SS-FS) Model". SS-FS is wrapper based feature selection model which retains the accuracy of each trial classifier built from a candidate feature subset, picks the highest possible fitness and deems the candidate feature subset as the choice output. The workflow of the SS-FS Model is shown in Figure 1.It can be seen that the operation iterates starting from a random selection of feature subset, continues to refine the accuracy of the classification model by searching for a better feature subset, in stochastic manner. The flow enables the classification model and the chosen feature subset finally converges.

The wrapped classifier is used as a fitness evaluator, advising how useful the candidate subset of features is; the optimization function searches for candidate subset of features in stochastic manner. This approach if run by brute-force testing out all the possible subsets, it will take an extremely long time. For there are 10,000 features in the "arcene" data, just for example, there are $2^{10000} \approx 1.9951 \times 10^{3010}$ possible trials of repeatedly building the wrapped classifier. While the increase in data features goes by O_2, the high computation costs intensify proportional to the amount of instances; in the case data stream mining, the data feed to the growth of BigData may amount to infinity!

In this regard, a stochastic-based search strategy called Swarm Search is used. Instead of testing on every possible feature subset, the Swarm Search which is enabled by multiple search agents who work in parallel would be able to find the most currently optimal feature subset at any time. In order to reduce the search time, a speedup is implemented in the initialization step in the Swarm Search, hence this approached is named as Accelerated ParticleSwarm Optimization (APSO). Further to improve the speed of feature selection process from big data, generic algorithm is proposed for feature selection in the present scope of study.

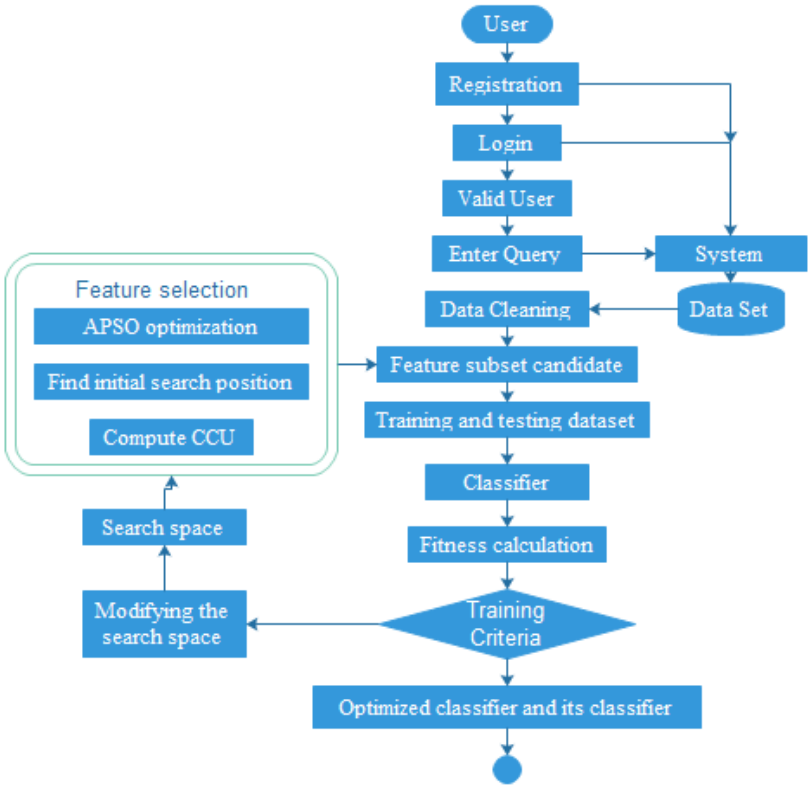

Figure 1: Work flow of SS-FS model

III. PROPOSED ALGORITHM

The proposed method employs Genetic algorithm for feature selection from a large data set and ensemble classifier for classification to evaluate desirable output. The work flow diagram of proposed method is shown in Figure 2.

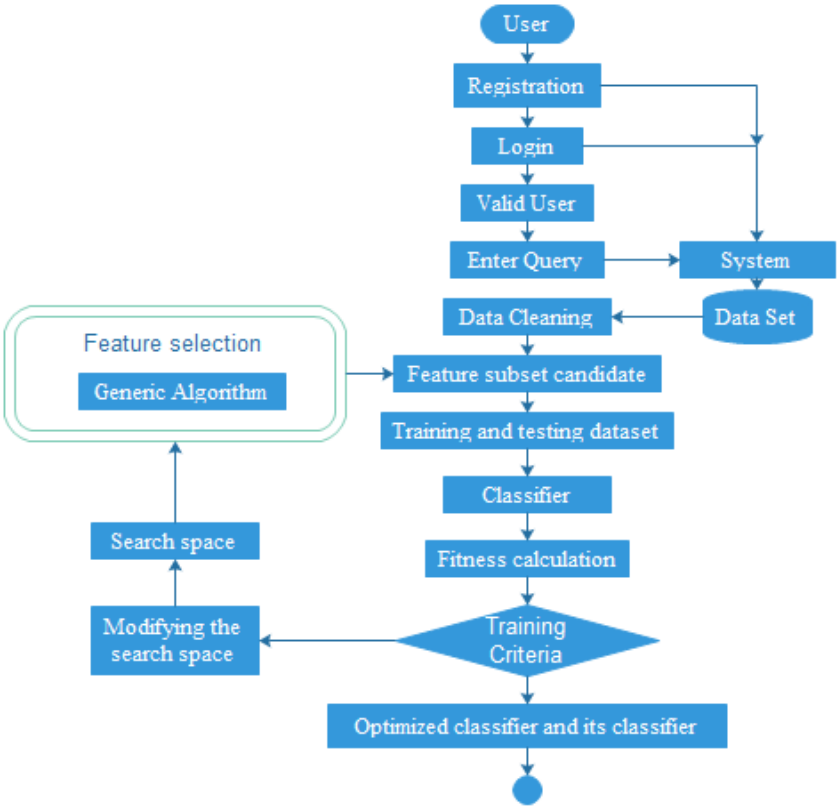

Figure 2: Work flow of SS-FS model with Genetic algorithm

3.1 Feature subset selection

Feature subset selection (FSS) is a process for identifying n most informative features V = {t1, t2, . . . , tn} from N known features S = {x1, x2, . . . , xN} (n < N), assuming that the C target classes W = {w1, w2, . . . , wC} are given and an observed dataset described by M samples (instances) is available. The selection of the most informative feature set leads to an improvement in classification accuracy, faster and more cost-effective classification performance, and a better understanding of the underlying process of the observed dataset. It would not be an overstatement to claim that FSS is the most important step when designing an engineering system because to its necessity for online classification of complex data. Many factors affect the computational costs (e.g., input dimensionality, and the complexity of a feature extraction and classification) but the size of a selected feature subset is of great importance mainly because the feature subset size dictates the memory allocation, which is the main bottleneck in mobile computing systems. The division of effective FSS into a filter method and a wrapper method is a key paradigm that provides a functional difference when evaluating the quality of a selected feature subset. In the filter method, the criterion function utilizes quantitative information such as the interclass distance of selected features. However, the criterion function used by the wrapper method relies on performance metrics for the classifier such as accuracy, specificity, and precision. The wrapper method often performs better at classification compared with the filter method, but it requires significantly higher computational costs because the fitness evaluation of a subset requires cross-validation or a bootstrapping procedure during the error estimation for each subset. Furthermore, the choice of the classifier inevitably biases the characteristics of the selected feature subset, which often leads to the loss of any generalization capability. Various criterion functions have been used for the filter method and they can be categorized into two groups, i.e., distance-based measures and relation-based measures.

In this paper Generic algorithm is implemented for feature selection. This evolutionary algorithm searches the appropriate feature subsets. The genetic algorithm includes the following steps:

(1) Initialization: The parameters are initialized first.
(2) Cross over: The two initial parents are recombined to form next generation which will become children.
(3) Mutation and evaluate: The mutation process takes randomly at any bit. The overall fitness which matches with mutation is evaluated.

3.2 Particle Swarm Optimization

Particle swarm optimization (PSO) is a population based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995, inspired by social behavior of bird flocking or fish schooling.

PSO shares many similarities with evolutionary computation techniques such as Genetic Algorithms (GA). The system is initialized with a population of random solutions and searches for optima by updating generations. However, unlike GA, PSO has no evolution operators such as crossover and mutation. In PSO, the potential solutions, called particles, fly through the problem space by following the current optimum particles.

Each particle keeps track of its coordinates in the problem space which are associated with the best solution (fitness) it has achieved so far. (The fitness value is also stored.) This value is called *pbest*. Another "best" value that is tracked by the particle swarm optimizer is the best value, obtained so far by any particle in the neighbors of the particle. This location is called *lbest*. When a particle takes all the population as its topological neighbors, the best value is a global best and is called *gbest*. The particle swarm optimization concept consists of, at each time step, changing the velocity of (accelerating) each particle toward its p*best* and *lbest* locations (local version of PSO). Acceleration is weighted by a random term, with separate random numbers being generated for acceleration toward *pbest* and *lbest* locations.

In past several years, PSO has been successfully applied in many research and application areas. It is demonstrated that PSO gets better results in a faster, cheaper way compared with other methods.

Another reason that PSO is attractive is that there are few parameters to adjust. One version, with slight variations, works well in a wide variety of applications. Particle swarm optimization has been used for approaches that can be used across a wide range of applications, as well as for specific applications focused on a specific requirement.

IV. ADVANTAGE AND DISADVANTAGE

4.1 Disadvantages of Existing System

1) 3V challenges of Big Data Technology:
    a) Velocity problem that gives rise to a huge amount of data to be handled at an escalating high speed.
    b) Variety problem that makes data processing and integration difficult because the data come from various sources and they are formatted differently.
    c) Volume problem that makes storing, processing, and analysis over them challenging.
2) Each time when fresh data arrive, which is typical in the data collection process that makes the big data inflate to bigger data.
3) The traditional induction method needs to re-run and the model that was built needs to be built again with the inclusion of new data.

4.2 Advantages of Proposed System

1) The new breed of algorithms known as data stream mining methods is able to subside 3V problems (Volume, Variety, and Velocity) of big data.
2) Each pass of data from the data streams triggers the model to incrementally update itself without the need of reloading any previously seen data.
3) Classification has been widely adopted for supporting inferring decisions from big data.

V. CONCLUSION

In Big Data investigation, the high dimensionality and the spilling way of the approaching information disturb awesome computational difficulties in information mining. Enormous Data becomes persistently with crisp information are being produced at all times; henceforth it requires an incremental calculation approach which has the capacity screen expansive size of information powerfully. Lightweight incremental calculations ought to be viewed as that is equipped for accomplishing vigor, high exactness and least preprocessing inactivity. In this paper, we explored the likelihood of utilizing a gathering of incremental grouping calculation for characterizing the gathered information streams relating to Big Data. As a contextual investigation experimental information streams were spoken to by five datasets of distinctive do-primary that have expansive measure of components, from UCI file. We analyzed the conventional grouping model prompting and their partner in incremental actuations.

REFERENCES

[1] Quinlan, J.R., C4.5: Programs for Machine Learning. Morgan Kauf-mann Publishers, 1993
[2] Ping-Feng Pai, Tai-Chi Chen, "Rough set theory with discriminant analysis in analyzing electricity loads", Expert Systems with Applications 36 (2009), pp.8799–8806
[3] Mohamed Medhat Gaber, Arkady Zaslavsky, Shonali Krishnaswamy, "Mining data streams: a review", ACM SIGMOD Record, Volume 34 Issue 2, June 2005, pp.18-26
[4] Wei Fan, Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2, pp.1-5
[5] Arinto Murdopo, "Distributed Decision Tree Learning for Mining Big Data Streams", Master of Science Thesis, European Master in Distributed Computing, July 2013
[6] S. Fong, X.S. Yang, S. Deb, Swarm Search for Feature Selection in Classification, The 2nd International Conference on Big Data Science and Engineering (BDSE 2013), 2013, 3-5 Dec. 2013.
[7] [Rokach, Lior, and OdedMaimon. "Top-down induction of decision trees classifiers-a survey." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 35, no. 4 (2005): 476-487.

