www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 7 July 2017, Page No. 22041-22044
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i7.24

Big Data Feature Selection Data Stream Mining

Kapil A. Tupe[1], Prof. M. A. Wakchaure[2]
[1]AVCOE, Sangamner, Maharashtra, India
kapiltupe11@gmail.com
[2]AVCOE, Sangamner, Maharashtra, India
manoj13apr@gmail.com

Abstract: ***The Big Data has too many challenges that face each IT deployment and educational analysis communities, the huge amount of data coming from various sources which are based on data stream and curse of dimensionality. The Big Data depends on 3V challenges specifically, Volume, variety and velocity. It's typically illustrious that the data coming from various data sources in different format and gather together in very high speed and creating ancient batch based model which is infeasible for real time data processing. This can be the most important challenge with the Big Data. As velocity is one of the challenges in Big Data, the crucial issue is to mine most valuable or actual and relevant information. To perform data mining over such high speed information the Big Data technology obtaining importance currently a days. The Feature selection technique is employed for data stream mining on the fly in big data. Feature selection has been widely used to minimize the process load in causing the mining information model. To achieve the query accuracy within minimum processing time and to reduce the processing load the accelerated particle warm optimization (APSO) is employed.***

Keywords: Feature selection, classification, big data, particle swarm optimization

1. Introduction

Big data uses data mining technologies, there's vast improvement in varied field particularly on-line technologies and web. The Big data have three main problems specifically, Velocity, variety and Volume. Velocity problem is concerning huge quantity of information to be analyzed at associate high speed. As data is coming from various sources the data is in different format, therefore it is very difficult to handle and store that data. And Volume problem is concerning, huge amount of data want large volume of space for storing that data.

The traditional data mining technique is concern with loading of full set of data in other word batch based model, after loading of data into model the data is divided according to some divide and conquer strategy. The two algorithms about divide and conquer approach are Classification and Regression Tree and Rough set discrimination. In the data mining method whenever new data arrives, the method that produces the large dataset up to the bigger data. Every time the new data arrives the traditional model learning method needs to be re-run the model and built the model again with inclusion of new data.

Volume, variety and velocity are characteristics of the data stream, the new variety of algorithms called data stream mining methods are able to solve 3V issues of Big data. Data stream algorithm is capable to deal with classification model from bottom up approach; while not the requirement of reloading any previously seen data, for every time when new data arrives the incremental learning model update itself with inclusion of new data. The quantity of data streams handles by this kind of algorithm, and mining data stream on the fly. The classifier application like feature selection tries to select optimal subset so as to enhance the accuracy and improve model training time for the classifier.

The Accelerated Particle Swarm optimization (APSO) is suitable for making group of classification algorithm. Light-weight Feature selection algorithm is used for data stream mining on the fly in big data. Here the challenge for mining the data stream is about finding the appropriate model induction algorithm weather it can be traditional or incremental learning model. The classifier application like feature selection to select a set of most important and optimal features excluding non-relevant and redundant features so as to improve accuracy and speedup model training time for the classifier.

2. Problem Statement

The unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for Big Data.

3. Literature survey

P. F. Pai and T. C. chen [1] expressed that the ability to deal with numeric data, rough set theory, which might convey information in a very rule-based type, has been one in all the important technique in data mining. Applications of rough set theory for analyzing electricity load aren't extensively conferred. Thus, this exploration the employs rough set theory to judge electricity loads. In addition, Linear Discriminate Analysis is employed to get a scale back for rough set model; the time generating a scale back by rough set theory. Hence study designs a hybrid Discriminate Analysis and Rough Set Model to supply decision rules for representing relation in system of electrical load. During this to judge the feasibleness of the hybrid model the condition factors and variations of electricity load square measure are use. Experimental results show that the model will with efficiency and properly examine the relation between condition variables and variations of electricity load. As a result it shows potential for developing an electrical load system and bids decision rules base for the utility management moreover as operations employees.

M. M. Gaber, A. Zaslavsky and S. Krishnaswamy [2] present the advances in hardware and software system that permit to require totally different measurements of data in big selection of fields. These measurements are generated

continuous with high information rates. For instance, sensor network, web logs, and electronic network traffic. The storage, querying and mining of such data sets are extraordinarily machine stringent tasks. Mining data streams is concerning extract or mines the data structures described in models and patterns in unstoppable sequence of data. The analysis in data stream mining has achieved a high attraction it gives importance of its applications and also the increasing generation within the flow of data.

W. Fan and A. Bifet[3] this aims to find the datasets as a result of the datasets are complex and in huge size the large data could be used, and that we can't supervise them with data mining software system tool. Because of its volume, variety and velocity, it was non-acceptable to find the datasets. However Big data mining have ability of extracting or mining helpful data from these datasets or streaming nature of data. This velocity, variety and volume challenges in big data are changing into most necessary chance for succeeding years. During this paper author were present a quick summary of its current status, contention and forecast to the longer term.

A. Murdopo [4] presents the experiences of their user, huge data effectively analyze with internet corporations. The systems that are ready to deal with big data in terms of three dimensions: volume as data is continuously increasing, variety because the data is formatted differently and velocity because the data is coming terribly high speed into the system. Most of the prevailing system have addressed at two out of the three dimensions, a distributed machine learning framework that addresses the degree and variety dimensions, and large on-line analysis, a streaming learning machine framework that controls the variety and velocity dimensions. During this paper ascendable huge on-line Analysis, a distributed streaming machine learning framework is developed to handle the challenge. They place along scalable advanced massive on-line Analysis (SAMOA) with Storm, a progressive stream process engine, which permits SAMOA to inherit Storms quantifiable to handle velocity and volume.

S. Fong, X. S. Yang, and S. Deb [5], during this paper the documented drawback for building applicable classification model is to seek out an correct set of features from high dimensional data. However in data mining, some huge data don't seem to be solely huge in size however conjointly they're present with great amount of feature. During this paper author have developed new Feature selection algorithm known as Swarm search for distinguishing a best feature set by using meta-heuristics. For flexibility in incorporating any classifier as its fitness operate the swarm search is advantageous. Conjointly so as to facilitate heuristic search it install in any meta-heuristic algorithm. Some experiments they need done by testing the swarm search over data of high dimensionality.

L. Rokach and O. Maimon [6], during this paper, for illustration of classifier one amongst the foremost wide used approach is decision tree. From accessible data construction of decision tree is difficult task. During this paper the author has done survey on recent strategies of growing decision tree for classifier in a very top-down manner. The paper suggests the splitting criteria and pruning methodology. Within the unilabiate splitting criteria an enclosed node is split supported the worth of 1 attribute. The pruning methodology is developed to deal with the perplexity.

C. C. Aggarwal [7], the data are available in massive volume and to store massive volume of data is difficult task. What is more, the data is keep, the dimensions of incoming data is extremely immense thus it's unable to process the one explicit data over once. Therefore the operation of data mining like categorization, clustering, classification and frequent pattern mining is become more difficult. Owing to growing size of data it's not possible that the data is with efficiency propagated by multiple passes, rather than that the data is processed at the most once and this results in limit on the algorithm implementation.

P. Domingos and G. Hulten [8], nowadays several organizations have immense volume of databases. The database is updated sporadically in a corporation. The dimension of database grows while not limit and a number of other a lot of informational record is into the database per day. The streaming format of continuous data brings nice chance and challenges to mining. The Hoeffding bound accustomed guarantee that the data when mining has relevancy or similar to data. the data mining system that relies on Hoeffding trees i.e. VFDT has high performance.

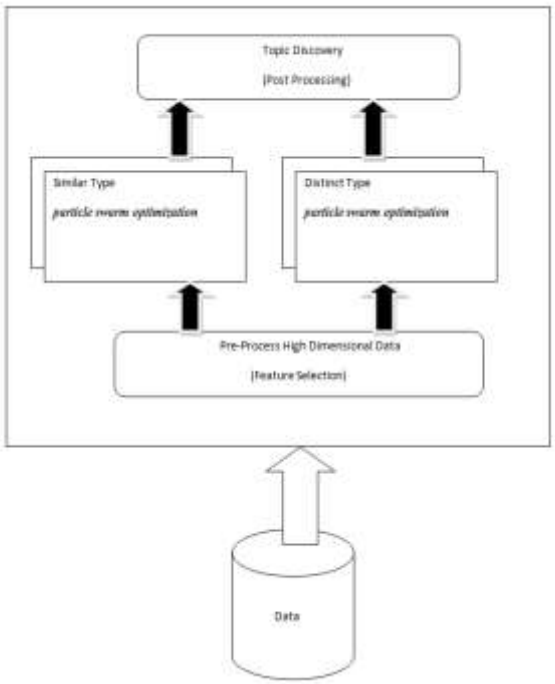
S. Fong, J. Liang, R. Wong and M. Ghanavati [9], to pick the identical features is one among the vital challenge for good prediction accuracy classification model. For optimum balance between generalization and over fitting the strategy of novel and economical feature clustering coefficient of variation (CCV) is projected during this paper. CCV search for optimum subset of attributes considerately of coefficient of variation of every attribute so as to enhance the correctness of classification. The operating of CCV is, it at first rank all the attributes based on the value of variations, then it split into two teams. At the top Hyper-pipe i.e. quick discrimination methodology is employed to look at that cluster generates higher accuracy in classification.

4. Proposed System

1. For data stream mining the feature selection technique by particle Swarm Search and Accelerated PSO is used.
2. The analysis results show that the incremental learning model technique obtained a better accuracy per second within the pre-processing as it update itself when new data is arrive..
3. Accelerated particle swarm optimization technique is used to improve the accuracy of query result and to minimize the processing load.
4. In incremental manner the combination of explosion is employed by applying swarm search approach. Whenever the high stream of data or sequence of data is arrive this swarm search approach is suitable with real-world applications.
5. Additionally, an incremental data model learning method is almost going to satisfy the challenges or demands of big data.

4.1 System architecture

DOI: 10.18535/ijecs/v6i7.24



4.2 System modules

Complex and evolving relationship:- To analysis and avoid deadlock and optimized to the complex query to the user and provide multiple service to user. It provides the relationship between the multiuser and multiple servers throughout the network.

Huge data with heterogeneous :- Anonymity data is to store and indexing with the database and provide the service to the user requirements.

Big data mining analysis:- To clustering data through out one client to another client. Extract the data relevant to query as a result.

Performance analysis:- Performance analysis is to provide the information about result of query by generating graph with comparision of start time and end time of query.

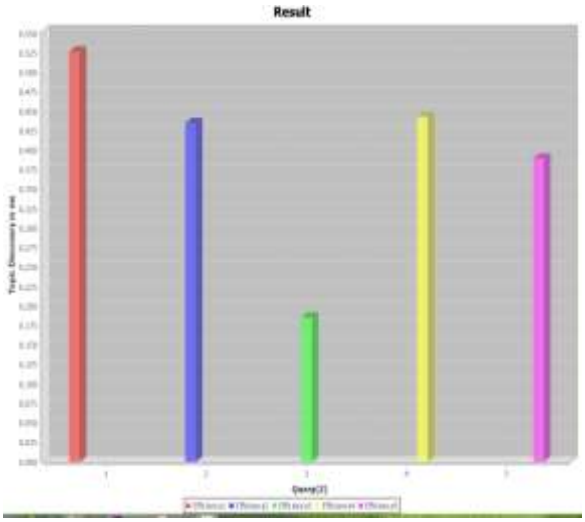
4.3 Proposed algorithm

Input:- Data streaming $(D_t) \rightarrow$ S
Output :-

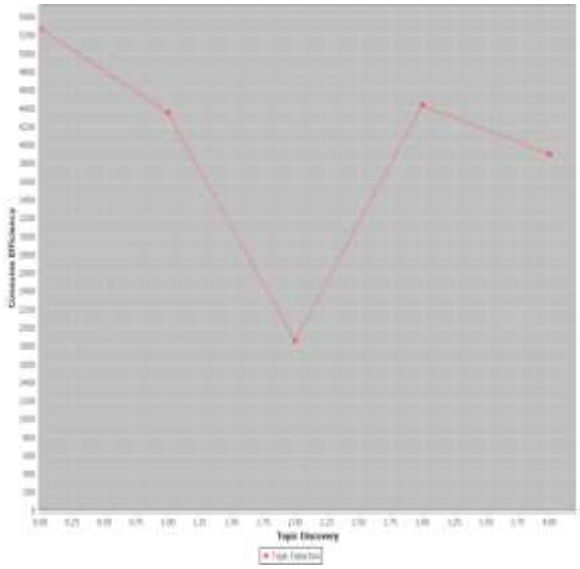
1. Start
2. Enter the Query
3. Stream the data on Google
4. S→ empty
5. Find total data on server

   Choose (d)= $\frac{L_{dis}-L_{sim}}{L_{total}}$
6. Data once again rank
7. Get result.

5. Experimental results



Above result show that the efficiency of topic discovery by particle swarm optimization in millisecond.



Above result show that efficiency level of query.

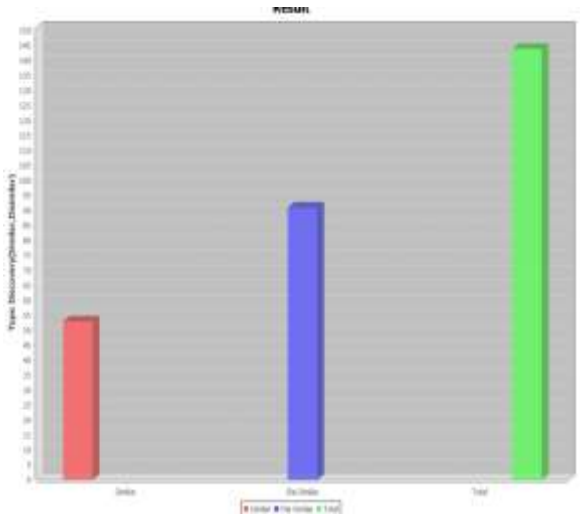


Fig : Accuracy comparision

$$Choose\ d^* = argmax_d \frac{l_1(d)-l_0(d)}{|l_0(d)|} \quad (1)$$

6. Conclusion

The Big data having various problems and challenges like volume, variety and velocity challenges. The data with different dimension and continuous sequence of data or in other word streaming nature of data worsen great computational challenges in mining of data. The experimental result show that the efficiency of each query with the help of topic discovery by particle swarm optimization in millisecond. The combinatorial explosion is self-addressed by used swarm search approach applied in incremental manner. This approach conjointly fits higher with real-world applications wherever their data arrive in streams. Additionally, associate incremental data mining approach is probably going to satisfy the demand of big data downside in computing.

Acknowledgment

It gives me an immense pleasure to express my sincere and heartiest gratitude towards my guide Prof. M.A.Wakchaure for his guidance, encouragement, moral support and affection during the course of my work. I am especially appreciative of his willingness to listen and guide me to find the best solution, regardless of the challenge. This work is also the outcome of the blessing guidance and support of my parents and family members and friends. I am also thankful to all who have contributed indirectly and materially in words and deeds for completion of this work.

References

[1] P.F. Pai and T.C. Chen, "Rough set theory with discriminant analysis in analyzing electricity loads," Expert Syst. Appl., vol. 36, pp. 8799–8806, 2009.

[2] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: A review," ACM SIGMOD Rec., vol. 34, no. 2, pp. 18–26,Jun. 2005.

[3] W. Fan and A. Bifet, "Mining big data: Current status, and forecast to the future," SIGKDD Explorations, vol. 14, no. 2, pp. 1–5,Dec.2012.

[4] A. Murdopo, "Distributed decision tree learning for mining big data streams," Master's of Science thesis, European Master Distrib. Comput., Jul.2013.

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

[6] L. Rokach, and O. Maimon, "Top-down induction of decision trees classifiers-a survey," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 35, no. 4, pp. 476–487,Nov.2005.

[7] C. C. Aggarwal Data Streams: Models and Algorithms, vol. 31. New York, NY, USA: Springer, 2007.

[8] P. Domingos, and G. Hulten "Mining high-speed data streams," in Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, 2000, pp. 71– 80.

[9] S. Fong, J. Liang, R. Wong, and M. Ghanavati, "A novel feature selection by clustering coefficients of variations," in Proc. 9th Int. Conf. Digital Inf. Manag., Sep. 29, 2014, pp. 205-213.

[10] Simon Fong, Raymond Wong, and Athanasios V. Vasilakos,, "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data" *in* IEEE Transactions On Service computing Jan/Feb 2016.

[11] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA, USA: Morgan Kaufmann, 1993.

Profile

Mr. K. A. Tupe is Pursuing Master in Engineering from Amrutwahini College of Engineering Sangamner. Received BE degree from University of Pune. His interested Areas are Data Mining, Big Data, Software Engineering.

Prof. M. A. Wakchaure is Assistant Professor in Amrutvahini College of Engineering, Sangamner. He is having 11 years of teaching experience. He is pursuing PHD from Savitribai Phule Pune University. His Research interests include Data Mining, Informational Retrieval, Software Engineering.