A Review on Feature Selection Data Stream Mining in Big Data

K.A.Tupe[1], Prof.M.A.Wakchaure[2]

[1] ME Student, Department of Computer Engineering, Amrutvahini COE, Maharashtra, India
[2] Assistant Professor, Department of Computer Engineering, Amrutvahini COE, Maharashtra, India

ABSTRACT

Though the Big Data has many challenges that face both commercial IT deployment and academic research communities, the Big Data have many root sources and that are founded on data stream and annoyance of dimensionality. In the field of Big data Network technology with respect to Agile development which handles huge amount of data at a time. The Big data relies on 3V challenges namely, Volume, Variety and Velocity of data. It is generally known that the data coming from various data sources in various formats accumulate continuously together making traditional batch based model infeasible for real time data mining. This is the biggest challenge with the Big Data. As Velocity is one challenge in Big data, the crucial thing is to mine valuable and relevant data. To perform efficient data mining over such high speed data the Big data technology getting importance now a days. The Feature selection technique is used for data stream mining on the fly in Big data. Feature selection has been widely used to minimize the processing load in inducing the mining data model.

Keyword: - *Big data, Feature selection, Classification, Swarm intelligence, metaheuristics, particle swarm optimization.*

1. Introduction

Big data uses new and fast performing data mining technologies, there is huge improvement in various field especially online technologies and Internet. The Big data have three problematic issues namely, Velocity, Variety and Volume. Velocity problem is about huge amount of data to be handled at an increasing high speed. As data coming from various sources the data is formatted differently, that means variety of data that makes data processing and data integration difficult. And Volume problem is about, huge amount of data need high volume of storage space to store that data. To meet the proficiency demand some conventional data mining methods based on full batch base mode learning are not suitable for the challenges in Big data.

The traditional data mining techniques require full data to load and loaded or integrated data is divided according to some strategy called Divide and Conquer strategy. In order to use divide and conquer strategy we have two classical algorithms namely, Rough Set Discrimination and Classification and Regression Tree algorithm. As the data come from various external and internal sources the different format of the data are integrated together and every time fresh data is coming. When fresh data come the traditional induction method needs to re-design the model with the inclusion of new fresh data.

The data stream mining algorithm is able to handle the challenges of Big data i.e. Volume, Variety, and Velocity. Data stream mining algorithm has capability to bring a classification model from bottom-up criteria; there is no need of loading any previous data, for each new pass whenever new fresh data come the model automatic increment and update itself. The Big data is huge not only because of its volume but they are very much large in size because they described by a large number of features. Swarm Search is the technique used for finding an optimal set of features over an high dimensional data. With the help of feature selection we improve the classification accuracy in representing the optimal features.

The Accelerated Particle Swarm Optimization is useful for finding a precise grouping of classification algorithm. For mining the Data stream on the fly the lightweight feature selection is useful. The data comes with high dimensionality and steaming form and feeds into a Big data in order to address this challenge the novel lightweight feature selection is use.

2. Literature Survey

P. F. Pai and T. C. Chen [1], stated that the ability to tackle with numeric and nominal information, rough set theory, which can put across knowledge in a rule-based form, has been one of the most significant technique in data analysis. On the other hand, Application of rough set theory for analyzing electricity load is not extensively conferred. Thus, this exploration the employs rough set theory to evaluate electricity load. Additionally, Linear Discriminate Analysis is used to generate a reduce for rough set model; the time generating a reduce by rough set theory. Therefore, this study designs a hybrid Discriminate Analysis and Rough Set Model to provide decision rules representing relation in an electric load information system. In this exploration nine condition factors and variations of electricity load are in use to evaluate the feasibility of the hybrid model. Experimental results disclose that the model can efficiently and correctly examine the relation between condition variables and variations of electricity load. As a result it shows potential for developing an electric load information system and bids decision rules base for the utility management as well as operations staff.

M. M. Gaber, A. Zaslavsky and S. Krishnaswamy [2], presents the current advances in software and hardware that allow to take different measurements of data in wide range of fields. These measurements are generated continuous with very high changing data rates. For example, sensor network, web logs, and computer network traffic. The storage , querying and mining of such data sets are extremely computational demanding tasks. Mining data streams is about extract the knowledge structures represented in models and patterns in non stopping flow of information. The research in data stream mining has achieved a high attraction due to the importance of its applications and the increasing generation in the flow of information. Applications of data stream analysis can differ from critical scientific and astronomical applications to important business and financial one.

W. Fan and A. Bifet [3], this aims to identify the datasets because the datasets are complex and in large size the Big Data is a new term used, and we can’t supervise them with data mining software tool or current methodologies. Due to its volume, variety and velocity, it was not possible to identify the datasets. But Big data mining have ability of extracting useful information from these large datasets or streaming nature of data. These 3v challenges in Big Data are becoming most exciting and important opportunity for the next years. In this paper author were present in this issue, a brief overview of its current status, controversy and forecast to the future.

A. Murdopo [4], presents in order to improve the experiences of their users, web companies needs to effectively analyze big data. They need to have systems that are able to handle big data in terms of three dimensions: volume as data keeps increasing, variety as the type of data is different and velocity as the data is continuously coming very rapidly into the system. On the other hand most of the existing system have addressed at most only two out of the three dimensions, a distributed machine learning framework that addresses the volume and variety dimensions, and Massive online analysis, a streaming machine learning framework that controls the variety and velocity dimensions. In this paper Scalable Advanced Massive Online Analysis, a distributed streaming machine learning framework is developed to address the afore-said challenge. Furthermore they put together Scalable Advanced Massive Online Analysis (SAMOA) with Storm, a state-of-the-art stream processing engine, which allow SAMOA to inherit Storms scalability to address velocity and volume.

S. Fong, X. S. Yang, and S. Deb [5], in this paper the well known problem for building appropriate classification model is to find an accurate set of features from high dimensional data. But in data mining, some big data are not only huge in size but also they are present with large amount of feature. In this paper author have developed new Feature Selection algorithm called Swarm Search for identifying an optimal feature set by using meta-heuristics. For flexibility in incorporating any classifier as its fitness function the swarm search is advantageous. Also in order to facilitate heuristic search it install in any meta-heuristic algorithm. Some experiments they have done by testing the swarm search over data of high dimensionality.

L. Rokach and O. Maimon [6], in this paper, for representation of classifier one of the most widely used approach is decision tree. From available data construction of decision tree is challenging task. In this paper the author has done survey on recent methods of growing decision tree for classifier in a top-down manner. The paper suggests the splitting criteria and pruning method. In the univariate splitting criteria an internal node is split based on the value of one attribute. The pruning method is developed to address the Dilemma.

C. C. Aggarwal [7], the data come in large volume and to store large volume of data is challenging task. Furthermore, the data is stored, the size of incoming data is very huge so it is unable to process the one particular data more than once. Hence the operation of data mining like indexing, clustering, classification and frequent pattern mining is become more challenging. Because of growing size of data it is impossible that the data is efficiently propagated by multiple passes, instead of that the data is processed at most once and this leads to limit on the algorithm implementation.

P. Domingos and G. Hulten [8], today many organizations have huge volume of databases. The database is updated periodically in an organization. The size of database grows without limit and several millions of informational record is into the database per day. The streaming format of continuous data brings great opportunity and challenges to mining. The Hoeffding bound used to guarantee that the data after mining is relevant or identical to data. The data mining system which is based on Hoeffding trees i.e. VFDT has high performance.

S. Fong, J. Liang, R. Wong and M. Ghanavati [9], to select the identical features is one of the important challenge for good prediction accuracy classification model. For optimal balance between generalization and over fitting the method of novel and efficient feature clustering coefficient of variation (CCV) is proposed in this paper. CCV search for optimal subset of attributes with consideration of coefficient of variation of each attribute in order to improve the correctness of classification. The working of CCV is, it initially rank all the attributes based on the value of variations, then it split into two groups. At the end Hyper-pipe i.e. fast discrimination method is used to examine which group generates better accuracy in classification.

3. Conclusion

The Big data is observed in various issues and challenges such as 3v challenges. The high dimensional data and streaming nature of data aggravate the great computational challenges in data mining.ss In this review paper we did study of how handle this Volume, Variety and Velocity issues. When data come from various source in different format the incremental computational technique is used to monitor the large volume of data dynamically. The data mining algorithms are used for achieving robustness, high accuracy and minimum pre-processing latency. The novel lightweight feature selection is used enlighten the processing load. The accelerated particle swarm search is used to find the accurate grouping of classification algorithm and data stream mining on the fly.

4. Acknowledgment

I am very much thankful of Prof. M.A. Wakchaure for his guidance and consistent encouragement in this paper work.

5. 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. 87998806, 2009.
[2]. M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, Mining data streams: A review, ACM SIGMOD Rec., vol. 34, no. 2, pp. 1826, 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. 15, Dec. 2012.
[4]. A. Murdopo, Distributed decision tree learning for mining big data streams, Masters 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. 902909.
[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. 476487, 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].Shitole, Manodnya A., and Manoj A. Wakchaure. "Patient-Centric and Privacy Preserving Clinical Decision Support System Using Naive Bayesian Classification." (2016).
[11]. Wakchaure, Mr Manoj Ashok. "Survey onDiscrimination Prevention in Data-Mining."

IJARIIE

BIOGRAPHIES

	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.

IJARIIE