Analysis of Big Data Based On Ensemble Classification

Mrs. Marrynal .S. Eastaff, MSc., MPhil.,
Department of Information Technology & Computer Technology, Hindusthan College of Arts and Science, Coimbatore-28
Email: marrynalhindusthan@gmail.com
Mrs. Premalatha P, MSc., MPhil, (Ph.D)
Department of Information Technology & Computer Technology, Hindusthan College of Arts and Science, Coimbatore-28
Email: premalathap2000@gmail.com

---ABSTRACT--

Big Data describes a technology used to store and process the exponentially increasing dataset which contains structured, semi structured and unstructured data that has to be mined for valuable information. It deals with 3 V's: Volume, Variety and Velocity of processing data. It is associated with cloud computing for the analysis of large data sets in real time. 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 data does not mean just numbers, dates or strings but also geospatial data,3D data, audio, video, social files, etc. The main objective of the big data technology is to achieve high level quality of data and accessibility for business intelligence. It replaces the traditional warehousing, relational databases and complicates software techniques used to process these huge data's. The example of big data is petabytes and exabytes .With the rapid development in the networking technology which handles huge data at a time, the big data technology is gaining importance nowadays. It is now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. It is also used in government agencies, financial corporations, large enterprises, etc. Though big data technology has various advantages it also imposes challenges such as the data that is stored in big data is from different sources at different rates, hence the date from one source will be out of synchronization from the other source which is to be synchronized. Secondly, the greatest challenge lies in extracting useful data from the huge stored data at low cost which is known as data mining. This paper mainly aims at analyzing the big data technology and providing a detailed study of challenges and solution to these challenges.

---

1. INTRODUCTION

In the computer era, there is a massive improvement in all fields especially in internet and online technologies [1] by new and fast performing technologies. Since these fields use huge database such as data's generated by people, about things and a powerful data server is must. Cloud storage is needed to manage and reuse the data which are useful to people for security [2], hardware and software maintenances. These massive data has some problems [3] as they need high volume of storage space and it may perform the operations such as analytical, retrieval and process operations. Moreover these operations are very complex and purely time consuming one [4-6].

To overcome these difficulties introduction of big data mining [7-9] stores all these huge and complex data and the needed data can also be easily extracted from the large data base. This big data processing improves the speed of the data base transferring than simple data exchanges [10-12]. 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 and an efficient tool to be used in system detected by cyber-attacks. These big data helps the users to retrieve the data as per their wish. Hence feature selection and classification plays an important role in this big data to retrieve or search a data from a variety of big data sets. Also more efficient algorithms must be implemented when dealing with big data.

Big data depends on 3 V's such as Volume, Velocity and Variety. These 3 V's are main characteristic function in selecting clustering techniques. First V, Volume of data is very important in clustering process as they require storage space. Second V, Velocity deals with the processing speed according to data flow. Third V, Variety depends on the data types. It may be image, text or a video generated from various sources such as mobile phones, camera or sensors. The feature selection and clustering are the two steps very important in systems which use different domains such as pattern recognition, machine learning, bio-informatics, data mining, semantic ontology and in image retrieval.

There are many algorithms available for feature selection [13, 14] and clustering process [15, 16] . Hence an appropriate algorithm must be chosen properly for this big data. In paper[1] the authors proposed a measure for assessing structural correlations in heterogeneous graph data sets in social networks with events and analyzed the hitting time to aggregate the proximity among nodes which has same proximity among nodes which has same event to discover the highly correlated with graph structure. In paper[16] the authors evaluated a Phoenix, implementation of map reduce method for shared memory system created by Google to manage thread creation, dynamic task scheduling, fault tolerance and data partitioning across processor nodes and performance and error recovery features were evaluated.

In paper [14] an efficient OSFS algorithm was proposed for feature selection where huge volume of data's are need to transmit, to reduce the large memory space. And this proposed method improved the real time case study for better performance.

The objective of this paper is to analyze the big data in feature selection and classification by using proper algorithms. Genetic algorithm is proposed for feature selection and ensemble classifier is proposed for classification method in this paper. The speed of the proposed method is compared with a conventional method.

2. PROPOSED METHOD

The proposed method employs Genetic algorithm for feature selection from a large data set and ensemble classifier for classification to evaluate desirable output. The block diagram of proposed diagram is shown in Fig. 1 and the flow chart of the proposed diagram is shown in Fig. 2.

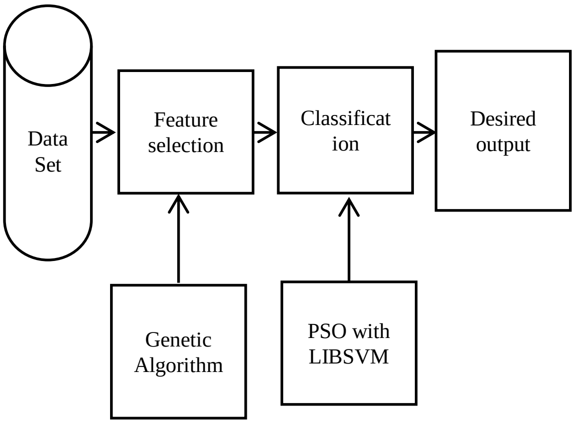

Fig. 1 Block diagram of proposed system

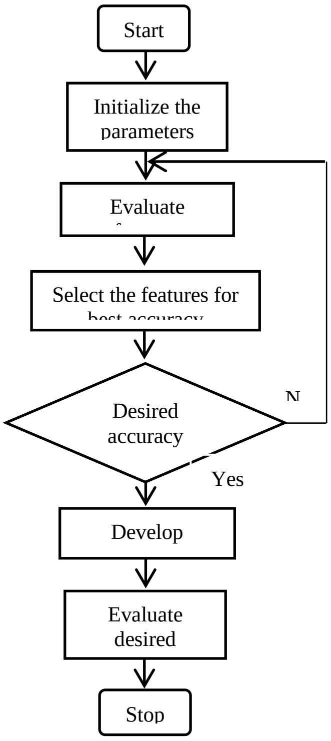

Fig. 2 Flow chart of the proposed system

2.1. FEATURE SELECTION

Feature subset of input variables is to find in feature selection from the large dataset. This feature selection will improve classification accuracy. The traditional approach has some limitation in discovery and in decision. The decision making includes multiple and complex objectives. In this paper Genetic algorithm is implemented. This evolutionary algorithm searches the appropriate feature subsets.

The genetic algorithm includes the following steps such as initialization, cross over, mutation and evaluate. The parameters are initialized first. In cross over, the two initial parents are recombined to form next generation which will become children. The mutation process takes randomly at any bit. The overall fitness which matches with mutation is evaluated.

2.2. CLASSIFICATION

Mapping a data item is an important task in classification into one of the predefined classes. A class or category must be developed consists of a data's of users which is very useful where large or huge data's are used for example web domain. In order to develop this class or category, extraction and selection is required. In this paper ensemble classifier using PSO with LSVM technique is implemented.

2.3. PARTICLE SWARM OPTIMIZATION

Particle swarm optimization (PSO) is an evolutionary algorithm. It is an optimization algorithm based on birds flocking and on the movement and intelligence of swarms. It uses a number of agents also known as particles that constitute a swarm moving around in the search space looking for the best solution. Each agent is treated as a point in N-dimensional space which adjusts its "flying" according to its own flying experience as well as the flying experience of other particles. Each particle keeps track of its coordinates in the solution space which are associated with the best solution (fitness) that has achieved so far by that particle and is treated as, pbest. The global best value is also obtained, denoted as gbest.

Let R be N dimensional search space, x_{ci} be the current position of i^{th} particle, p_{bi}, the personal best position of i^{th} particle andv_{ci}, the current velocity of i^{th} particle.

Let fn denotes the fitness function, then the personal best of i at a time step t is updated as:

$$p_{bi}(t+1) = \{\begin{matrix} p_{bi}(t) iffn(x_{ci}(t+1)) \geq f(p_{bi}(t)) \\ x_{ci}(t+1) iff\big(x_{ci}(t+1)\big) < f\big(p_{bi}(t)\big) \end{matrix}$$

The global best value is defined by

$$gbest \in \{p_{b1}(t), p_{b1}(t), \dots\dots, p_{bm}(t)\}$$
$$= \min\{f(p_{b1}(t)), f(p_{b2}(t)), \dots\dots f(p_{bm}(t))\}$$

The velocity and position of each particle is updated using the following equation

$$v_{ci}(t+1) = wi.v_{ci}(t) + a_1 r_1\big(p_{bi}(t) - x_{ci}(t)\big) + a_1 r_1\big(gbest - x_{ci}(t)\big)$$
$$x_{ci}(t+1) = x_i(t) + v_{ci}(t+1)$$

In the formula, wi is the inertia weight, a_1 and a_2 are the acceleration constants, r_1 and r_2 are random numbers in the range [0, 1] and VL_{i1} must be in the range $[-V_{Lmax}, VL_{max}]$, where $VL_{max}$the maximum velocity is.

2.4. LIBRARY SUPPORT VECTOR MACHINE

LIBSVM is a library for Support Vector Machines and developed in 2000 which gained more attention for regression and classification. LIBSVM easily apply SVM to other application. It involves two steps. Initially data set is trained to obtain model and using this model information is predicted.

2.5. KERNELS

Kernel methods are most used in SVM. SVMs are linear classifiers and regressors and through the Kernel trick it operates in space and even performs in non-linear classification and regression.

Radial Bias Function Kernel is expressed as

$$\mathrm{RBF} = \exp\left(\frac{1}{2\sigma^2\|\mathrm{x}-\mathrm{x_i}\|^2}\right)$$

Table 1. Comparison of Accuracy

Parameter	Existing method	Proposed method
Accuracy (%)	88.5	94.3

The PSO is initialized with features is obtained by using GA. Thus the feature selected from GA is then classified by this ensemble classifier (PSO using LIBSVM) to evaluate the desired data (gbest) from the big data mining which lowers the time consumption.

3. CONCLUSION

In this paper, feature selection and classification is proposed in big data mining and data storage, time consumption are analysed. Genetic algorithm is utilized for feature selection and PSO with LIBSVM technique is employed for classification. These techniques are implemented to learn more about the efficiency and accuracy which reduces the tedious, complexness and expensive. The proposed system is compared with the conventional method with the accuracy of 94.3% and proves that the proposed system is better than the existing method.

REFERENCES

[1] Jian Wu, Ziyu Guan, Qing Zhang, Ambuj K. Singh, and Xifeng Yan, "Static and Dynamic Structural Correlations in Graphs" IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 9, pp 2147-2160, September 2013

[2] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report UCB-EECS-2009-28, Univ. of California, Berkeley, Feb. 2009

[3] P. Mell and T. Grance, "Draft NIST Working Definition of Cloud Computing," http://csrc.nist.gov/groups/NS/cloudcomputing/index.html, June 2009.

[4] M. Arrington, "Gmail Disaster: Reports of Mass Emai Deletions," http://www.techcrunch.com/2006/12/28/gmail-disasterreportsof-mass-email-deletions/, 2006

[5] J. Kincaid, "MediaMax/TheLinkup Closes Its Doors," http://www.techcrunch.com/2008/07/10/ mediamaxthelink up –closesits -doors/, July 2008.

[6] Amazon.com, "Amazon s3 Availability Event: July 20, 2008,"http://status.aws.amazon.com/s3-20080720.html, July 2008.

[7] A. Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033, 2012

[8] J. Mervis, "U.S. Science Policy: Agencies Rally to Tackle Big Data," Science, vol. 336, no. 6077, p. 22, 2012.

[9] Nature Editorial, "Community Cleverness Required," Nature, vol. 455, no. 7209, p. 1, Sept. 2008

[10] G. Duncan, "Privacy by Design," Science, vol. 317, pp. 1178-1179, 2007

[11] B. Huberman, "Sociology of Science: Big Data Deserve a Bigger Audience," Nature, vol. 482, p. 308, 2012.

[12] E. Schadt, "The Changing Privacy Landscape in the Era of Big Data," Molecular Systems, vol. 8, article 612, 2012.

[13] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, "PrivacyPreserving Public Auditing for Secure Cloud Storage" IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[14] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu, "Online Feature Selection with Streaming Features," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 35, no. 5, pp. 1178-1192, May 2013.

[15] D. Gillick, A. Faria, and J. DeNero, MapReduce: Distributed Computing for Machine Learning, Berkley, 2006.

[16] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "Evaluating MapReduce for Multi-Core and Multiprocessor Systems," Proc. IEEE 13th Int'l Symp. High Performance Computer Architecture (HPCA '07), pp. 13-24, 2007.