A Survey of Distributed Data Classification using Multiple Support Vector Machine Based on ACO

Geetanjali Patel[1], Avinash Sharma[2], Madhuvan Dixit[3]

[1]*M.Tech (CSE, Student), MITS, Bhopal*

[2]*Professor (HOD, Department of CSE), MITS, Bhopal*

[3]*Associate Professor (Department of CSE), MITS, Bhopal*

Abstract- Big Data perform more-volume, complex, increasing information sets with different and original sources in concern distributed form. With the more improvement of internetworking, information storage, and data repository capacity, Big Data are now speedy improving in all science and engineering work areas, consisting physical, biochemical and biomedical sciences. This paper represents a MSVM-ACO (Multiple Support Vector Machine-Ant Colony Optimization) algorithm which differentiate the information of the distributed Big Data concern and supports a distributed Big Data implementation prototype, from the information mining concern. This information-driven prototype contains requirement and driven assembling of data sources, mining and analysis, programmer interest modeling, security and privacy issues. We examine the disputing matter in the information driven prototype and also in the distributed Big Data revolution. As distributed Big Data diligences are characterized with self directed sources and distributed controls, combining spread information sources to a one point site for mining is pre-plan prohibitive due to the potential forward cost and privacy policy. Although we can always carry out mining activities at each different location, the biased seen of the information organized at each location often leads to predetermine decisions or prototypes, same as the elephant and blind men situation. Big Data mining system has to act a data interchange and merge algorithm to confirm that all different sites can work together to accomplish a global optimization target. Prototype mining and correlations are main steps to confirm that prototypes or patterns discovered from different data sources can be consolidated to match the complete mining objective.

Keywords: Big Data Model, Distributed Processing System, Data Mining and Ant Colony Optimization.

I. INTRODUCTION

Author Dr. Yan Mo won 2012 Nobel Prize in Literature field. This is perfectly the most disputation Nobel Prize in this category. Find on Google search engine with "Yan Mo Nobel Prize," conclusions in 1,050,000 web results on the Internet. "For all praises and criticisms," said Mo speedy, "I am grateful." What kinds of praises and criticisms has Mo basically received over his 35-year writing career? As comments secure coming on the Internet and in different news media, can we optimize all kinds of opinions in multiple media in a real-time scenario, consisting modified, cross-referenced issues by critics? This kind of summarization code is an efficient example for Big Data processing model, as the data comes from different, heterogeneous, self directed sources with complex and evolving bonding and maintains growing. From the above example, the era of Big Data has arrived. Each day, 2.5 quintillion bytes of information are initiate and 90 percent of information in the world today where produced within the past three years. Our limitation for information generation has never been so powerful and tremendous ever since the innovation of the information technology in the early 19th century decade. As another example, on 4 October 2013, the first presidential debate in between of President Barack Obama and Governor Mitt Romney executed more than 10 million tweets within every 2 hours. Among all these tweet statements, the desired events that created the important discourse generally exposed the public involve, such as the discourse about Medicare and vouchers. Such online discussions support a new explanation to sense the public opinions and perform feedback in real- time and are importantly assuring compared to traditional media, like as radio or TV distribution. Another instance is Flickr, a public image sharing website, which received 2.8 million photos per day, on average, from March to February 2013. Considering the size of each image is 2.5 MB, this demand 3.6 TB storage every day. As an old saying states: "an image is worth a thousand words," the billions of images on Flicker are a broad tank for us to enhance the society, events, affairs, disasters and so on, only if we have the immune to harness the large amount of data.

II. SYSTEM MODEL

Big Data starts with large-volume, **h**eterogeneous, **a**utonomous sources with distributed and decentralized **c**ontrol, and seeks to **e**xplore complex and evolving bonding among information (HACE Theorem).

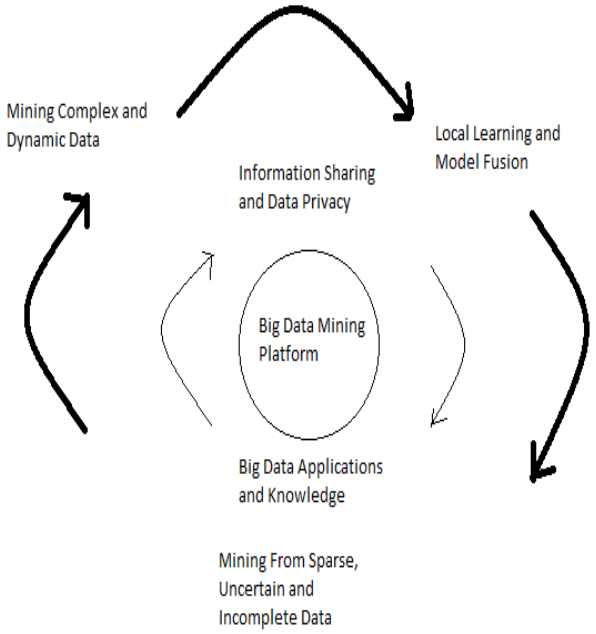

Figure 1: A Big Data processing framework

Raw data is useful only when it is transformed into knowledge or useful information. This consist information analysis and convert to extract interesting outlines and relationships among the problem variables. In practical applications, such type of transformation accepts efficient information access, analysis and representation of the result in a predefine manner. The opposite of more distributed environments in both medical and commercial domains introduces a new dimension to this process — a large number of distributed sources of data that can be used for discovering knowledge. Cost of data transmission between the distributed databases is a desired part in an increasingly handheld device and associated world with a more number of distributed information resources. This cost consists of several components like (a) Limited network bandwidth, (b) data security, and (c) existing organizational structure of the applications environment. The area of Distributed Knowledge Discovery and Data Mining (DDM) explains procedures, systems and human-computer attachment issues for knowledge discovery applications in distributed environments for minimizing this cost. In this dissertation work, we consider a Multiple Support Vector Machine (MSVM) model to represent classify data from distributed uncertain knowledge. Specifically, we address the problem of learning a MSVM from heterogeneous distributed data. The classify data is optimize through ACO technique.

III. PREVIOUS WORK

In this section, discuss the review of big data and data mining using MSVM based on ACO and different data classification techniques. The MSVM (Multiple Support Vector Machines) basically are capable of delivering maximum performance in words of classification accuracy than the other information classification methods.

Big Data support more-volume, complex, improving information sets with different, self driven sources [1]. With the fast growth of networking, information storage and information collection limit, Big Data are now freely expanding in all directional science and technical work areas consist physical, chemical and biomedical sciences [1]. As the website has improved as distributed information storage, individuals and organizations have been eligible to use the minimum-value knowledge and information on internet when taking business conclusions [2]. Various large organizations have different databases maintained in multiple sections and hence multi-database mining is an real task for information mining [3]. To reduce the find cost in the information from all databases, we need to signify which repositories are most suitable relevant to a information mining application [3]. We consider an interesting and challenging problem, online streaming feature selection, in which the size of the feature set is unknown, and not all features are available for learning while quit the number of evaluations constant [4]. The ACO system consist two protocols: (1) Local pheromone update protocol, which performed whilst constructing solutions. (2) Global pheromone updating perform, which performed after all ants construct a solution [5]. Furthermore, an ACO procedure consist two more methodology: trail expedition and desired, daemon actions [5]. Trail expedition decrement all trail costs over duration, in order to neglect unlimited accumulation of trails over some element [5]. We represent a cumulative method to learning a Bayesian network from distributed heterogeneous information [6]. We first implement a local Bayesian network at local site using local information [7]. Then each location represents the evaluations which are most desired to be evidence of associating between local and global variables and communicate a subset of these observations to a central site [8]. Rough set theory performs a needed mathematical issue to draw essential decisions from exact life information consisting vagueness, inaccurate and incomplete and hence performed successfully in the area of pattern identification, machine implementation and knowledge discovery [9]. We apply a fast repetitive procedure for justifying the Support Vectors of a mentioned set of nodes [10]. Our procedure tasks by evaluating a candidate support vector set. It needs a greedy method to select nodes for involve in the candidate set [10]. When the sum of a point to the desired candidate set is suspended by

reason of other points already exist in the set we use a backtracking method to prune away such nodes [10].

IV. PROPOSED METHODOLOGY

The Algorithm of proposed method is explained below:
Algorithm: [classifyDataset1, classifyDataset2] = MSVM_ACO (database1, database2)

Step 1: We take two dataset as multiple databases which consists complex, sparse and uncertain data.

Step 2: SVM technique is apply to input database in parallel form. The pseudo code of PSVM as follows.

```
candidatePSV = { nearest pair from different labels }
 while there are violating nodes do
       Find a nodeviolator
       Candidate_PSV = candidate_SV S violator
          if any αp < 0 , addition of c to S then
            candidate_PSV = candidate_PSV \ p
            continue till all nodess are pruned
          endif
 end while
```

Step 3: Above step are perform in repeated form, then classified data through PSVM from respective multiple data sources store.

Step 4: Now apply ACO method on image dataset for retrieving images. The sequence of steps for finds the best optimize descriptor point using ACO as follows.

4.1 The source facts comes from evaluating the exploitation of food sources among many ants, in that ants particularly constant cognitive abilities have securely been able to search the shortest path between a food resource and the nest food resource.

4.1.1 The first ant searches the food source (F), from whatever way (a) then coming back to the nest food source (N), exiting back a trail pheromone (b)

4.1.2 Ants every which way follow four potential routes, but the fortifying of the runway makes it more fines as the shortest route.

4.1.3 Ants take the little way long sections of other routes suffer their trail pheromones.

4.2 in a sequence of implementations on a colony of ants with a selection between two undesirable length ways leads to a source of food, biologists have calculated that ants inclined to use the shortly route. A model explicates this behavior is as bellows:

4.2.1 An ant tests more or less at random around the colony;
4.2.2 If it searches a food source, it exists more or less remotely to the nest food source, exiting in its way a trail of pheromone value;

4.2.3 These pheromones costs are suitable nearby ants will be depend to below, more or less organize, the track;

4.2.4 Come back to the settlement, these ants will strengthen the way;

4.2.5 If two ways are potential to scope the same food source, the shorter one will be, in the same time, travelled by more ants than the yearn way will;

4.2.6 The short way will be growth enhanced, and become more admirable;

4.2.7 The long way will eventually evaporate, pheromones are volatile;

4.2.8 Since, all the ants have designed and hence "select" the minimum way.

4.3 Eventually, if the quantity of pheromone cost remained the same over time on all sides, no way would be select. Therefore, because of feedback, a minimum variation on side will be simplified and thus permit the select of a side. The procedure will shift from an unstable situation in which no side is stronger than one another, to a stable situation where the way is composed of the strongest sides.

Step 5: As per optimize result of classified data, considering as classified optimize dataset from Multi-information sources.

V. IMPLEMENTATION TOOLS

MATLAB is a large-performance, effective and interactive language for essential computing approach. It combines computation, visualization, graphical and programming in an easy environment where questions and solutions are solved in familiar mathematical notation and graphical way. It include mathematical matrix form of image and other computational procedure development data acquisition modeling, picture processing, information processing, simulation and prototyping data processing, visual scientific, engineering drawing and graphics application development consisting graphical programmer interface area building

MATLAB (Matrix Laboratory) is an effective programming way whose basic information node is an array in multiple dimensional plan, which does not consider to mention dimensioning. This permits you to prove various technical problems in various formats, especially those with matrix form and vector calculations. In a few durations it would assume to write a code in a desired scalar language such as C or FORTRAN. MATLAB is stands for matrix laboratory. MATLAB was generally written to perform easy access to matrix software implemented by the LINPACK, EISPACK and various technical projects. Today, MATLAB compiler supports to incorporate the LAPACK packages, embedding the suitable in software for matrix evaluation and programming issues. MATLAB is used in every facet of computational mathematics. Following are some commonly used mathematical calculations where it is used most commonly:

- Dealing with Matrices and Arrays
- 2-D and 3-D Plotting and graphics
- Linear Algebra
- Algebraic Equations
- Non-linear Functions
- Statistics
- Data Analysis
- Calculus and Differential Equations
- Numerical Calculations
- Integration
- Transforms
- Curve Fitting
- Various other special functions

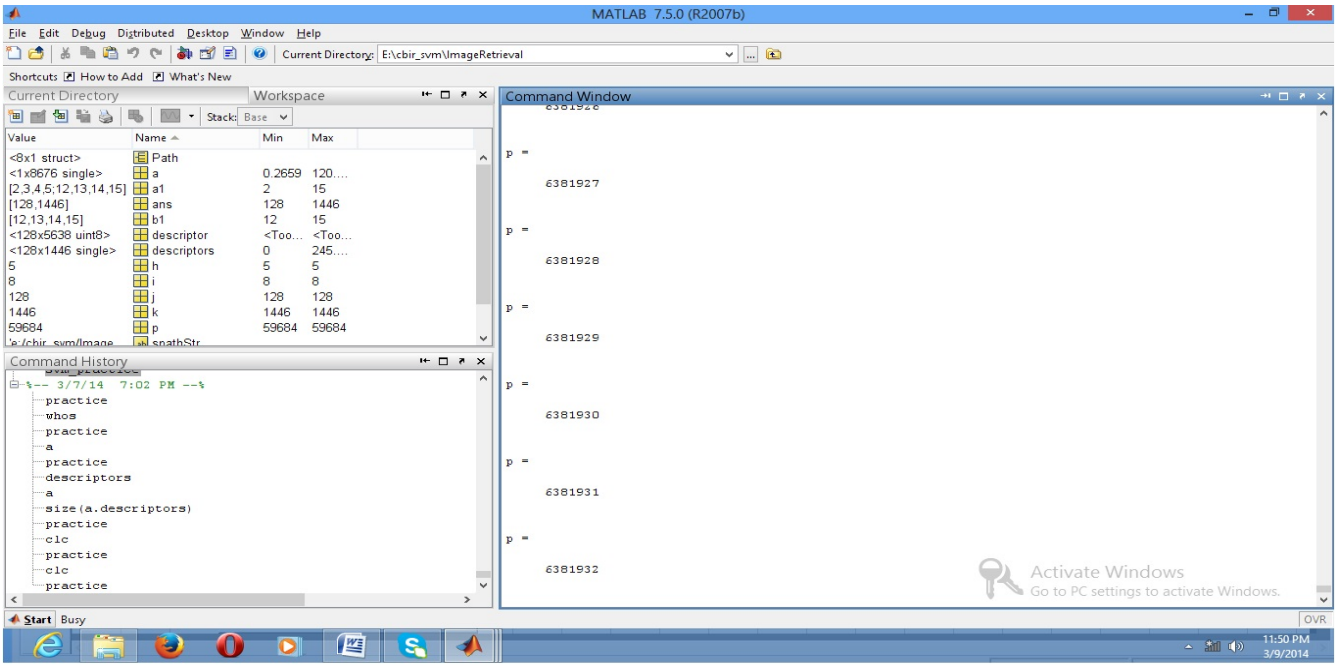

Figure 1: MATLAB Command Window

MATLAB has performed over duration of years with multiple inputs from various programmers. In university research area, it is the benchmark and effective instructional method for introductory and new generation courses in engineering and medical science field. In engineering field, MATLAB is the tool of select for best large-productivity research concern, development and analysis. MATLAB support basic key points a family of application depend solutions called toolboxes. Following are the basic features of MATLAB:

- It is a high-level language for mathematical computation, computer visualization and application development.
- It support an effective environment for iterative exploration, design and problem solving.
- It provides library of numerical functions for algebra, statistics, Fourier analysis, filtering, optimization, integration and solving ordinary differential equations.
- It provides built-in graphics function for visualizing information and tools for creating custom plots.

- MATLAB interface provide development tools, GUI used for enhancing program quality and maintainability and modify maximum performance.

- It support techniques for making applications with graphical interfaces.
- It support mapping for MATLAB algorithms with external applications.

VI. REFERENCES

[1] Xindong Wu and Xingquan Zhu, "Data Mining with Big Data," IEEE Tran. on knowledge and data engineering, VOL 26, NO. 1, JANUARY 2014

[2] Nada M. A. Al Salami, "Ant Colony Optimization Algorithm," UbiCC Journal, Volume 4, Number 3, August 2009.

[3] S.V.N. Vishwanathan, M. Narasimha Murty, "SSVM : A Simple SVM Algorithm," Science, vol. 337, pp. 337-341, 2012.

[4] Durgesh K. Srivastava, Lekha Bhambhu, "Data Classification Using Support Vector Machine," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[5] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012. [6] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012.

[7] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[8] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.

[9] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.

[10] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.

[11] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multi- media, (MM '09,) pp. 917-918, 2009.

[12] R. Chen, K. Sivakumar, and H. Kargupta, "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data," Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.

[13] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Efficient Algorithms for Influence Maximization in Social Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

[14] C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, and K. Olukotun, "Map-Reduce for Machine Learning on Multicore," Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS '06), pp. 281-288, 2006.

[15] G. Cormode and D. Srivastava, "Anonymized Data: Generation, Models, Usage," Proc. ACM SIGMOD Int'l Conf. Management Data, pp. 1015-1018, 2009.

[16] S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J. McPherson, "Ricardo: Integrating R and Hadoop," Proc. ACM SIGMOD Int'l Conf. Management Data (SIGMOD '10), pp. 987-998. 2010.

[17] P. Dewdney, P. Hall, R. Schilizzi, and J. Lazio, "The Square Kilometre Array," Proc. IEEE, vol. 97, no. 8, pp. 1482-1496, Aug. 2009.

[18] P. Domingos and G. Hulten, "Mining High-Speed Data Streams," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '00), pp. 71-80, 2000.

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