SOLVING DISTRIBUTED AND HIGH-DIMENSIONAL PROBLEMS OF BIG DATA USING GENETIC PROGRAMMING

Anugu Sushma[1], P.Srilatha[2], Dr.G.Vishnu Murthy[3]
[1]M. Tech Student, [2]Assistant Professor, [3]Head of the Department, Department of CSE,
CVSR College of Engineering, Village Venkatapur, Mandal Ghatkesar, District Ranga Reddy,
Telangana, India

Abstract: This process can be express the potential of GP (Genetic Programming) as a base classifier algorithm in building ensembles within the context of large-scale information classification. Associate ensemble designed upon base classifiers that were trained with genetic programming was found to considerably shell its counterparts designed upon base classifiers that were trained with decision tree and logistic regression. The prevalence of genetic programming ensembles is attributed to the higher diversity, each in terms of the functional form of as well as with respect to the variables processes the models, among the base classifiers.

Index Terms: Big-Data; Distributed Problems; Data Extraction; Genetic Programming; Fast-ICA;

I. INTRODUCTION

The term massive data has been a popular topic recently in practice, academe, and government to reflect the needs of using the large data. Massive data refer to data sets that are thus massive and complex that is on the far side the flexibility of typical package tools to capture, store, manage, and analyze it among a tolerable period of time. The purpose of collecting massive data is similar to tradition data processing to resolve the key issues of society, business and science. However, the large volume of data makes it terribly difficult to perform effective analysis using the existing traditional techniques. Additionally, different characteristics like velocity, variety, variability, value and complexity suggests the large data issue additional challenge. To deal with the complexity of massive data, many information technologies and package are projected, for instance NoSQL, Hadoop and cloud computing. These solutions are sometimes technological orientation rather from the angle of theory. The problem of data integration comes from the property of variety in massive data. With such variety, a challenge is a way to combine the distributed and massive significant features for analysis. Although it is convenient to combine all options across tables, it should suffer the curse of dimensionality and issues of feature selection. On the opposite hand, the matter of skills accessibility is that the fact that the standard data processing strategies cannot deal with massive data thanks to these data is stored distributed. In the field of machine learning, high dimensional data analysis and distributed data mining (DDM) algorithms are recently developing topics and received a lot of attention recently. Though these problems are clearly related to massive data, they are not well-integrated and will be overcome appropriately. The value of massive data is unquestionable. However, a way to transform massive data to massive value is that the main issue. Though there are many tools and architectures, like Map-Reduce, Hadoop, No-SQL database etc, Map-Reduce programming paradigm involves distributed process of enormous data over the cluster. Under this paradigm the input data is spitted according to the block size. The data split is performed by the input format. These splits are assigned a specific key by the record reader and so a key, value combine is generated. Key, value pairs are then subjected to a two section process. This two section process comprises of a map section and a reduce phase. To search, manage, store, and management immense volume of data, the analysis of massive data will actually derive the nugget of massive data. Therefore, the aim of this paper is to propose algorithms and procedures to resolve the on top of issues from the perspective of machine learning in classification issues.

II. RELATED WORK

There are a number of papers are projected recently to handle the distributed high-dimensional data. as an example, projected a quick outlier detection strategy for distributed high-dimensional data sets with mixed features; used appropriate distance operate between the feature vector approximations to handle distributed high-dimensional issues. However, the previous have to limit the mixed options and therefore the latter will only suit in unvaried distributed databases. An additional flexible and integrated strategy or algorithm ought to be projected during this file. During this proposal, we tend to hope to deal with this issue appropriately projected collective data processing (CDM) for prophetic data modeling in heterogeneous environments. CDM may be foundations during which any operate are often represented in very distributed fashion exploitation an acceptable set of basic functions. Once the essential functions are the native analysis manufacture correct and helpful results which will be directly used as a part of the world model without loss of accuracy typically, the performance of CDM depends on the standard of estimated cross-terms. Several algorithms are projected based on the concept of CDM. As an example, collective principal part analysis distributed clustering algorithm collective variable regression and distributed call tree construction. Massively distributed storage may be a fundamental change in dealing with the property massive volume in big data. A distributed database system means that the analytic tables are located in several databases. However, we tend to cannot simply merge all tables for analysis because of the restrictions of the

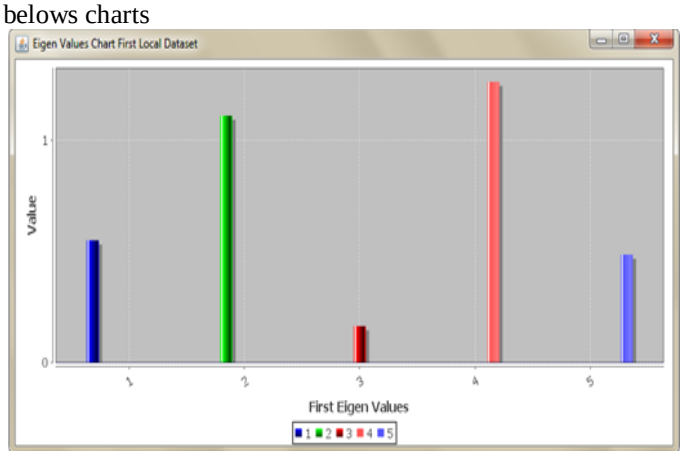
computing power and memory of a computer. Classical strategies are not designed to deal with this sort of drawback hence, massive data arise additional issues of high dimensional data analysis to understand heterogeneity and commonality across different.

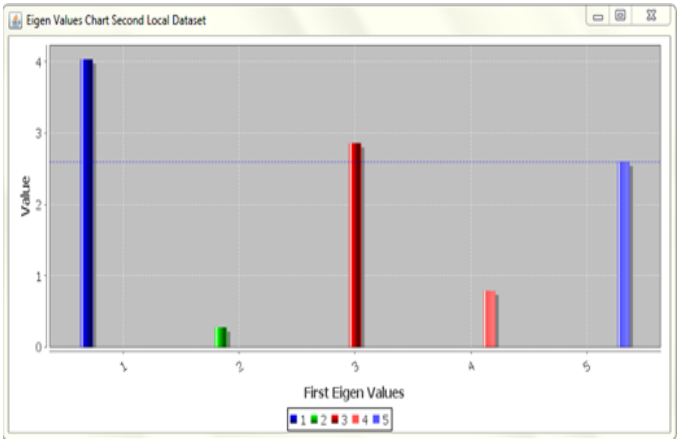
III. FRAME WORK

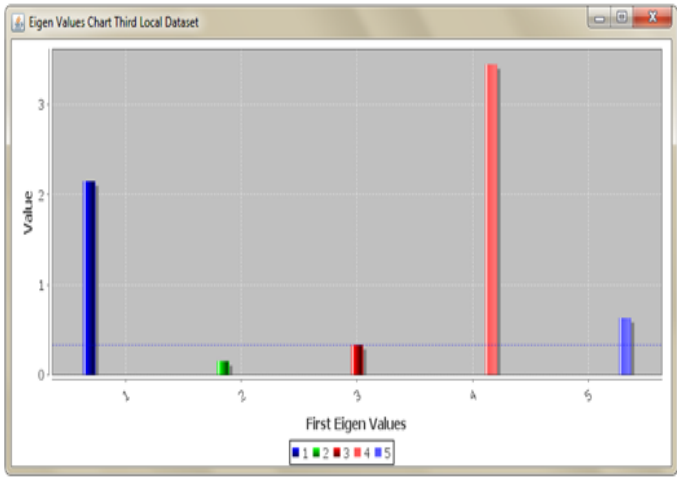
In this paper, we tend to follow the concept of DMM to first derive the native models from distributed databases then calculate the global model. However, since high-dimension options could destroy the information structure of the problem and decrease the efficiency of algorithms, we should first reduce our options by some dimension reduction techniques. During this paper, we tend to initial consider independent component analysis as the tools for reduces the options of local databases. Then, we will use genetic programming to derive the native models from every distributed database. Finally, we will integrate all local models into the global model by using genetic programming again. The genetic trees of the local models are the initial population of the global genetic programming. In addition, we will sample the validation set from the distributed database to determine the ultimate world model by the cross-validation technique. Feature extraction involves reducing the number of the dimensions of a data set and could be a common way to deal with the high-dimensional data. The most well-known and standard feature extraction technique is principle component analysis (PCA). Though many papers successfully used principle component analysis for feature extraction in several applications, the limitations of principle component analysis has been reported as only using two order statistics and a linear technique. In observe the transform defined by second-order strategies like principle component analysis is not useful for many functions wherever best reduction of dimension within the mean-square sense is not required. This is as a result of principle component analysis neglects such aspects of non-Gaussian information and independence of the elements (which, for non-Gaussian information, is not a similar as uncorrelated-ness). Hence, high-order technique could also be made for non-Gaussian information. Independent component analysis (ICA) is one in every of the applied math tools to extract the independent component (IC) from a determined variable series according to high-order statistics. The main distinguish of Independent component analysis Independent component analysis from alternative strategies is that it looks for components that are each statistically independent and non-Gaussian. Independent component analysis has been projected to deal with several real-world applications like signal process, (MEG) magneto encephalography, and image analysis. During this paper, the Fast- Independent component analysis algorithm is used to reduce the dimensionality of the distributed data tables. The algorithm minimizes the mutual data to derive ICs supported the fixed-point iteration schema. The Fast- Independent component analysis has been widely and with success used for several applications to reduce the dimensionality of the issues. Additionally, the nonlinear Fast-ICA algorithm has been proposed to consider nonlinear relationships between ICs. It will be additionally used here to obtain the subset of every data table. On the other hand, genetic programming (GP) was proposed by to automatically extract intangible relationships during a system and has been utilized in several applications, like symbolic regression and classification. The preparative steps of genetic programming contain to determine the terminals, functions, fitness function, parameters, and termination criterion. Of these preparatory steps is drawback dependent. The illustration of GP will be viewed as a tree-based structure composed of the operate set and terminal set. Once we tend to initialize a population of the genetic programming tree, the following procedures are the same as genetic algorithms (GAs) the initial GP-trees are randomly generated and not all features should appear at intervals a GP-tree. Hence, from the description of genetic programming, it can be seen that one in the distinctive characteristic of genetic programming is its built-in mechanism to pick the options that are related to the matter via the operators of evolution. During this approach, a nonlinear variable choice will be used for dimensional reduction. It should be highlighted that the thought of feature choice between independent component analysis and genetic programming are totally different. In Independent component analysis, new options are generated to replace the original features. On the other hand, genetic programming derives the significant features from the original features.

IV. EXPERIMENTAL RESULTS

In our experiments user click on read SUSY dataset after reading the dataset to generate model and Fast ICA algorithm means Fast ICA algorithm is used to reduce the dimensionality of the distributed tables. The algorithm minimizes the mutual information derive IC's (independent components) based on the fixed point iteration schema and generate model and Fast ICA algorithm can ask enter the model size to enter like 100, 200 it will divided in to three models like model1, model2 and model3 these three models also called as local datasets the training dataset will be generated and then we use ICA for each local dataset to extract and reduce features after that test-set will be generate after generating the test-set apply the genetic algorithm the genetic algorithm is used for to compare the both training sets and test-sets data and then we can display the final global model after that Eigen values charts will be generated it will be generate the three models charts to shown in belows charts







The above three charts are also called as models charts the first model chart will generate the five Eigen values are 0.6, 1.1, 0.2, 1.3, 0.5 and the model chart will be generated with using first five values like 4.4, 0.3, 2.9, 0.8, 2.6 the model chart will be generated with using first five values like 2.2, 0.2, 0.3, 3.4, 0.6 by using this approach we can avoid distributed and high dimensional problems.

V. CONCLUSION

In this paper, we tend to project an integrated algorithm to deal with the distributed classification drawback of massive data. First, we use ICA to reduce the spatial property of native data sets and retain the importance data of features. Then, we adopt genetic programming to generate genetic trees because the native model and initial population of the global model. Finally, we run genetic programming again to get the ultimate global model. As per the results of the empirical study, the projected technique outperforms than the standard genetic programming with respect to the accuracy ratio. Additionally, the projected techniques are often considered as the way to deal with the classification of massive data.

REFERENCES

[1] Koldovsky, Z., Tichavsky, & Oja, E. (2006). Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining CramÉr- Rao Lower Bound. Neural Networks, IEEE Transactions on, 17(5), 1265-1277.

[2] Koufakou, A., & Georgiopoulos, M. (2010). A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. Data Mining and Knowledge Discovery, 20(2), 259-289.

[3] Kriegel, H. P., Kunath, P., Pfeifle, M., & Renz, M. (2005). Approximated clustering of distributed high-dimensional data. In Advances in Knowledge Discovery and Data Mining (pp. 432-441). Springer Berlin Heidelberg.

[4] Kumar, P., & Pandey, K. (2013). Big Data and Distributed Data Mining: An Example of Future Networks. International Journal, 2, 36-39.

[5] Koza, J. R. (1992). Genetic programming: on the programming of computers by means of natural selection (Vol. 1). MIT press.

[6] Lange, K., Papp, J. C., Sinsheimer, J. S., & Sobel, E. M. (2014). Next-generation statistical genetics: Modeling, penalization, and optimization in high-dimensional data. Annual Review of Statistics and Its Application, 1, 279-300.

[7] Li, X. R., Jiang, T., & Zhang, K. (2006). Efficient and robust feature extraction by maximum margin criterion. Neural Networks, IEEE Transactions on, 17(1), 157-165.

[8] Li, G., Rosenthal, C., & Rabitz, H. (2001). High dimensional model representations. The Journal of Physical Chemistry A, 105(33), 7765- 7777.

[9] McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity, 2013.

[10] Park, B. H., & Kargupta, H. (2002). Distributed data mining: Algorithms, systems, and applications.

[11] Park, B., Kargupta, H., Johnson, E., Sanseverino, E., Hershberger, D., & Silvestre, L. (2001). Distributed, collaborative data analysis from heterogeneous sites using a scalable evolutionary technique. Applied Intelligence, 16(1), 19-42.

[12] Provost, F. J., & Buchanan, B. G. (1995). Inductive policy: The pragmatics of bias selection. Machine Learning, 20(1-2), 35-61.

[13] Rosipal, R., Girolami, M., Trejo, L. J., & Cichocki, A. (2001). Kernel PCA for feature extraction and de-noising in nonlinear regression. Neural Computing & Applications, 10(3), 231-243.

[14] Shorter, J. A., Ip, P. C., & Rabitz, H. A. (1999). An efficient chemical kinetics solver using high dimensional model representation. The Journal of Physical Chemistry A, 103(36), 7192-7198.

[15] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. Knowledge and Data Engineering, IEEE Transactions on, 26(1), 97-107.

