A Novel Two Step Genetic Program for High-Dimensional Data

Mandava Mamatha [*1], Ms.J. Rama Devi [*2]

** PG Student, CSE Department, PVPSIT College/JNTUK University, Vijayawada, A.P, India*

**Sr. Asst. Professor, CSE Department, PVPSIT College/JNTUK University, Vijayawada, A.P, India*

Abstract— *The increasing volume of data to be analysed imposes new challenges to the data mining methodologies. Now a day's datasets have the problem curse of dimensionality. To solve these problems by the techniques PCA, 2SGP (Two Step Genetic Programming) are popular tools for linear dimensionality reduction and feature extraction. These traditional data mining methods are not integrate well with larger data sizes, miss the accuracy in terms of memory and time and not support for non-linear data. So, we propose a novel 2SKPCA algorithm .i.e; KPCA (Kernel Principal Component Analysis) and GP (Genetic Programming) to deal with high-dimensional non linear data also and reduce the dimensions. KPCA is the linear, nonlinear form of PCA, which better exploits the complicated spatial structure of high-dimensional features. GP produces the feature selections and derives the significant features from the original features.*

Keywords— *Feature extraction, high-dimensional data, kernel PCA, Genetic Programming, Classification.*

I. Introduction

In these day's we have seen the dramatic increase in the growth of information due to collection of data from various independent or connected applications and services. The most popular example is Internet. Internet usage is at high rate. The data which is stored on the net is very large. Every second data is added to the internet all over the world. The collective data and knowledge on the net is growing at a immense speed and even estimating the adding speed is also very tough task. This large amount of data is called as a Big Data. The term 'Big Data' means collections of huge, difficult or required data which become complex or impossible to process, analyse and store using current methodologies like database management. .Big data is a collection of large amount of data with different data types. Structured data are those data which are formatted in a database management system. By extracting meaningful associations, trends and patterns, from the large amount of data we make the quality of life and make our world put in a better place but it is unmanageable. Many IT companies to solve the big data challenges using a NOSQL database, such as Cassandra or HBase and HADOOP.

Data mining, the extraction of knowledge and useful information from large databases, is a powerful new technology to help companies for managing the most important information in their data warehouses. Traditional data mining methods are not integrate well with larger data sizes and are more expensive in terms of accuracy with respect to time. Most of the traditional algorithms require multiple times scanning the data and generate a large number of redundant dimensions. There is a demand for data mining algorithms in many big data application areas like biology, computer technologies, astrology and social networking.Big-data users demand speed of the data access, agility of the data, and constant iteration.

In the field of machine learning, to analyse the high dimensional data distributed data mining (DDM) algorithms received much more attention. But these data mining algorithms are not properly used to integrate the large amount of data. The purpose of this paper is to propose novel algorithms and procedures to solve the integration problem from the perspective of machine learning in classification problems.

II. LITERATURE SURVEY

ROBPCA: robust principal component analysis

Classical PCA is based on the covariance matrix of the data and is highly precious to outlying observations. Two robust approaches have been proposed. The first approach is the eigenvectors of robust scatter matrix such as the minimum covariance determinant and is limited to relatively low-dimensional data. The second approach is based on projection and can manage high-dimensional data. In these ROBPCA approach, which combines projection pursuit ideas with robust s-matrix evaluation. ROBPCA evaluates more accurate estimates at non contaminated dataset and more robust derived at contaminated data. ROBPCA can be examined rapidly, and is able to detect exact-fit situations. By by-product result, ROBPCA generates a diagnostic plot to provide and classifies the outliers. We apply the algorithm to several datasets from chemo metrics and engineering.

We have constructed a fast and robust algorithm for PCA of large dimensional data. The algorithm first applies PP techniques in the original data space. These results are then used to apply the observations into a subspace of small to big dimensions. By this subspace, ideas of robust covariance estimation are the applied. Throughout, we have the ability to detect exact fit situations and to reduce the dimensions. Simulations and applications to real data demonstrate that this ROBPCA algorithm yields very robust estimates when the data have outliers. The associated diagnostic plot is a useful graphical tool that allows one to visualize and classify the outliers.

Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure

Security and privacy issues are noticed by the volume, variety, and velocity of Big Data. The diversity of data sources, data flows, are grouped with streams of data acquisition and high volume create unique security issues. This paper describes the security challenges when organizations start moving sensitive data to a big repository is called hadoop. It generates the different threat models and the security control framework to address and eliminate security risks due to the derived threat conditions and usage models. The framework outlined in this paper is also meant to be distribution agnostic.

Hadoop and big data are no longer buzz words in large enterprises. If the reasons are correct or not, enterprise data warehouses are deploying into Hadoop and along with it come peta bytes of data. In this paper we have laid the groundwork for future security reasons on the Big Data ecosystem and providing security to this. However, there are many challenges when it comes to providing security in Big Data environment. This paper also provides the Big Data threat model to the reader for further expanding and analysing to their organizational environment. It also provides target reference architecture for Big Data security and covers the stack. There is a chance to get a head of the curve, test and deploy your tools, and techniques before big data becomes a big problem.

Distributed data mining: overview on Newsletter of the Technical Committee on Distributed Processing

Data mining means collecting the data from large databases and applies mining techniques on collected data and extract patterns for further decision making process. Data mining importance is its vast applicability. It is being used progressively more in business applications for understanding and extract valuable data, like consumer buying actions and buying tendency, profiles of customers, in industry and financial services. From this paper an overview of general architecture of centralized warehouse and distributed data warehouse is provided. It also presents various issues related to both the approaches. The paper ends with our comparision on comparative analysis of centralized warehouse and distributed warehouse.

Distributed data mining algorithms are faster, efficient and cost effective than centralized data mining. Of course, the techniques in distributed data mining are more complex than centralised data mining. Careful design is required for a distributed data mining task.

III.EXISTING SYSTEM

Big data refer to data sets which are so large and complex that is beyond the ability of typical software tools to capture, store, manage, and analyse it within a tolerable elapsed time. However, the massive volume of data makes it very difficult to perform effective analysis using the existing traditional techniques. In addition, other characteristics like 4v's and complexity put forward the big data issue more challenge. . Here we first consider ICA as the tools for reduces the features of local databases. Then, we will use GP to derive the local models from each distributed databases. Finally, we will integrate all local models into the global model by using genetic programming again. In the field of machine learning, high dimensional data analysis and distributed data mining (DDM) algorithms received much attention.

IV. PROPOSED SYSTEM

To propose algorithms and procedures and to solve the above problems from the perspective of machine learning in the classification problems. We follow the concept of DMM to first derive the local models from distributed databases and then calculate the global model. However, since high-dimension features may destroy the data structure of the problem and decrease the efficiency of algorithms, we should first reduce our features by some dimension reductionality techniques. In this 2SKPCA technique, we first consider KPCA as the tools for reduces the features of local databases. Then, we will use GP to derive the local models from each distributed databases. Finally, we will integrate all local models into
the global model by using genetic programming again. In addition, we will sample the validation sets as eigen graphical charts.

V. IMPLEMENTATION

A.KPCA and Genetic Programming

The common way to deal with the high-dimensional data for reducing the dimensions of a data set is feature extraction. Principal component analysis is well-known and recently used feature extraction method. For feature extraction in many applications like neural networks, neural computing & de-noising in nonlinear regression PCA technique is used. But some limitations are also present in using PCA. They have been reported as it's used only for two order statistics and a linear technique & neglects the aspects of non-Gaussian data.

Feature extraction or dimensionality reduction means to restrict the entire input space to a sub-space which is having lower dimensions. KPCA has been proposed to deal with many real-world applications like face recognition and active shape models. The concepts of KPCA can be described as follows:

ALGORITHM:

To derive kernel PCA

Step 1: Project the data to high-dimensional feature space k

Step 2: Compute covariance matrix of data in the feature space

Step 3: Compute principal components by solving eigen value problem

Step 4: The eigen vector can be expressed as linear combination of features.

A **genetic algorithm** (or **GA**) is a classification technique used in computing to find true or approximate solutions to optimization and classification problems. Genetic algorithms are divided as global search heuristics. Genetic algorithms are class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover. Genetic algorithms are developed as a computer simulation in which a population of abstract representations (like chromosomes or genotype) of candidate solutions to an optimization problem produces toward better solutions.

Normally, solutions are represented in binary format like 0s and 1s, but other codings are also possible. The evolution formally starts from a population of generated individuals and happens in generations. In each generation, from the fitness function individual population is evaluated, different individuals are choosen from the current population (based on their fitness), and modified (recombined) to form a new population. The new population is then used in the next iteration.

Finally, the algorithm terminates when either a maximum number of generations has been produced, or a fitness function has been reached for the population. If the algorithm has terminate after performing the maximum number of generations, a estimated solution may or may not have been reached.

ALGORITHM:

In genetic algorithm the following functions are also done.

1. **Start**: Generate random population of n chromosomes

2. **Fitness**: Evaluate the fitness f(y) of each chromosome y in the population

3. **New population**: Create a new population by iterating the following steps

i) Selection: Choose two parent chromosomes from a population according to their fitness

ii)Crossover: With this crossover the parents to form a new off spring. If no crossover was performed, offspring is copy of parents.

iii) Mutation: Using mutation probability mutate new offspring at each locus

iv) Accepting: Keep new offspring in new population

4 .**Replace**: Using new generated population for a future run of algorithms

5. **Test:** If the end condition is fully satisfied, stop, and return the best solution in current population

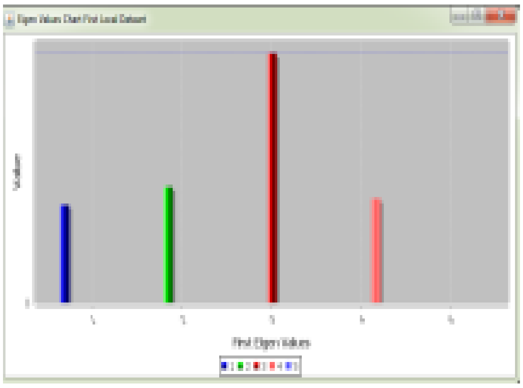
6. **Loop**: Go to step 2

B. 2SKPCA

First to derive the global models from distributed databases we use the concept of DMM. In this paper, we first consider KPCA technique to reduce the features of local databases. To obtain the global models from each data set we use GP. The concept of KPCA and GP is different in the feature selection. KPCA is used to replace the original features by generating the new features.GP is used to obtain the significant features from original features.

VI. RESULT AND DISCUSSION

Here, the input dataset is SUSY dataset. It has 18 features. In this 18 features first 8 features are kinematic properties. The remaining 10 features are functions of first 8 features. To differentiate these two classes this high-level features are used. In this proposed algorithm, the entire dataset is divided into three global records. Each global record contains some million random records. Here, the test dataset is derived. It contains the remaining million records. First, we use the KPCA to each global record to obtain and reduce the features. The eigen values of global datasets can be as shown In fig 1.



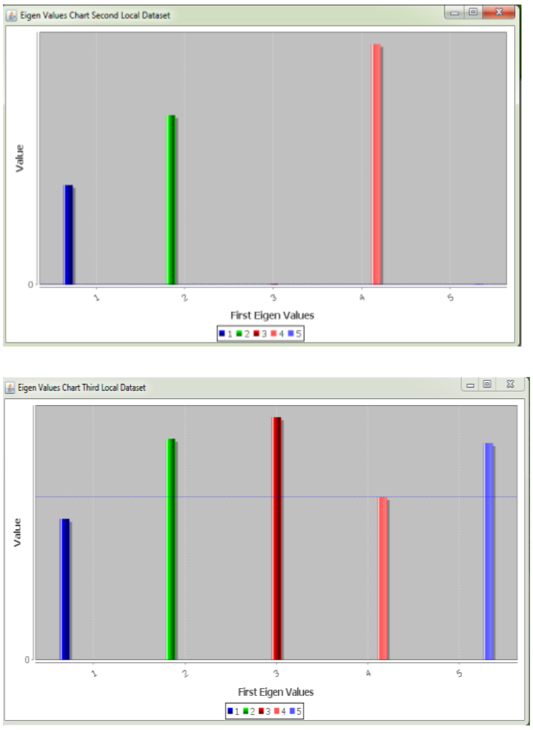


Figure 1.Eihen values of three global data sets

Here, we obtain the first four eigen values for first global dataset, three eigen values for second global dataset, five eigen values for third global dataset and detailed information of eigenvalues is 68.9%, 85.8%, and 80.7%,respectively. Here,we run genetic programming five times to calculate the accuracy rates of data, as shown in table 1.

From Table 1, we can observe the accuracy ratio of proposed algorithm is higher than the existing methods. In this proposed method deals with six features. Finally GP uses 18 features to generate the final model. So, our proposed method is suitable to handle the high-dimensional data.

TABLE 1. COMPARISON RESULTS

Accuracy ratio		Run1	Run2	Run3	Run4	Run5	Average
2SKPCA	Test	71.7382%	74.867%	71.3872%	71.8756%	70.8456%	72.8765%
	Training	71.681%	74.786%	71.2872%	71.6542%	69.3584%	72.7862%
2SGP	Test	70.8766%	72.043%	70.8047%	70.8318%	69.3050%	70.7723%
	Training	70.9277%	72.0702%	70.8342%	70.8934%	69.3677%	70.8186%
GP	Test	68.6981%	67.5822%	70.7864%	70.6515%	69.3182%	69.4072%
	Training	68.7535%	67.5216%	70.8327%	70.7242%	69.2915%	69.4247%

VII. CONCLUSION

The related information regarding the project is being referred and the requirements related to the project are analysed and problems are addressed in various approaches in the existing systems. First, we use KPCA to reduce dimensionality of global data sets and extract the knowledge and features. Next, we run GP to generate genetic trees as the local model and initial population of global model. Again we run GP to get the final global model. The problems such as accuracy and processing speed are being detected through analysing the Literature .The solution is to be analysed in detail way to ensure those problems addressed in previous approaches.

REFERENCES

[1] Aronis, J ., Kolluri, V., Provost, F .,&Buchanan, B. The WORLD: Knowledge discovery from multiple distributed databases. In proceedings of 10th international Florida AI Research symposium.

[2] Hubert, M., Rousseeuw, P. J., & Vanden Branden,K.ROBPCA: a new approach to robust principal component analysis.Technometrics,

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

[4] Zhang, Y., & Bhattacharyya, S. (2004). Genetic programming in classifying large-scale data: an ensemble method. Information Sciences,

[5] Chan, P. K., & Stolfo, S. J. (1998). Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. In KDD (Vol. 1998, 164-168).

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

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

[8] Koldovsky, Z., Tichavsky, P., & 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.

[9] Du, Q., & Kopriva, I. (2008). Automated target detection and discrimination using constrained kurtosis maximization. Geoscience and Remote Sensing Letters, IEEE, 5(1), 38-42.

[10] Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and Good practices. In Contemporary Computing (IC3), 2013 Sixth International Conference on, 404-409.