Indian Journal of Science and Technology, Vol 9(28), DOI: 10.17485/ijst/2016/v9i28/87995, July 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

An Optimized Algorithm for Big Data Classification using Neuro Fuzzy Approach

Navneet* and Nasib Singh Gill

Department of Computer Science and Applications, MDU, Rohtak - 124001, Haryana, India;
navneet.khatri@gmail.com, nasibsgill@gmail.com

Abstract

Objectives: To optimize data mining technique for big data. **Methods/Analysis:** This paper designs a technique for the data mining of big data by modifying the existing data mining technique using fuzzy and neural network. The present technique firstly performs the dimension reduction. Then reduced dimension datasets are clustered, while the remaining attributes are used to classify such dataset by using automated fuzzy. **Findings:** The existing data mining techniques are not optimized on such data. The simulation using the fuzzy on various dataset shows the optimization of technique. The RNFCA algorithm is analyzed adding the RNFCA algorithm to the WEKA library on the Intel i5 @ 2.67 GHz using the eclipse IDE. The algorithm is analyzed on the datasets having 400 instances with 25 attributes and 32561 instances with 15 attributes. The detail description of these datasets is given in table 2. The performance of the RNFCA algorithm can be compared with existing CCSA algorithm and the decision tree i.e. J48. The figure 4 -7 shows the comparison graph of the J48, CCSA and RNFCA over various parameters. **Applications/Improvement:** The simulation using the fuzzy on various dataset shows the optimization of technique.

Keywords: BIG Data, Dimension Reduction, Fuzzy, K-Mean, Neural, Schwarz Criteria

1. Introduction

The growing culture and the increasing awareness of the world towards the internet usage lead to large volume of high velocity data from various fields. This results in the introduction of the BIG Data. The extraction of the information from BIG DATA seems a tough task[1,2]. Moreover, the profit enhancement from e-data needs optimized data mining techniques of BIG DATA. Due to the 3 V's i.e. velocity, volume and variety[3] and additional 2 V i.e. value, variability[4] of BIG DATA, the existing data mining techniques are not able to produce the efficient results. The data mining covers various operations like classification, clustering, regression, association etc., but this paper focus on the classification techniques. The classification technique is used to classify the data into various numbers of classes or categories on the basis of feature of the data. The classification process is divided in two steps; one is the training while other is the classification. The training step is used to build up a model for the classification that is used for the classification in next step. Theauthors[1,5] show that the first step i.e. training step doesn't get affected by the size of the data, while the second step; classification is not efficient on BIG DATA. Various researches had worked on the problem of classifying BIG DATA. The classification of the BIG DATA was done using the SVM. The author shows that the tuning of the parallel SVM resulting in optimized performance[6].Theexisting decision tree based classification algorithm was modified by cascading it with the clustering algorithm. The authors of paper firstly cluster the data by using the K-Mean clustering then classify the clusters by using J48 algorithm. The number of clusters has been decided by using the Schwarz criteria. The author also gives a classification algorithm by classifying the clustered data using association[7].The platfora method for high data delivery in large datasets focus on sky tree to analyze a machine learning language and data analytics platform focused on handling the Big Data[10].

*Author for correspondence

Another paper emphasizes the evolution of data processing adroitness to advanced data processing taxonomy from Mesolithic to recent years and a comparative study of prevailing tools/techniques which are useful for mainly the analysis of the bulky data[11]. Data Analytics for Rural Development gave conceptual framework for the application of data analytics in enhancing rural development by supporting different sectors such as agriculture, banking, governance and healthcare[12]. In another paper, the proposed methods have been applied to a Math teachers' selection problem of education in Iran. After determining the criteria that affect the Math teachers' decisions, fuzzy AHP and fuzzy TOPSIS methods are applied to the problem and results are presented[13].In folding of Fuzzy Hyperboloid introduced and study new connection between fuzzy retractions, fuzzy foldings and fuzzy deformation retracts of fuzzy open hyperboloid in fuzzy Minkowski space and fuzzy open ball in fuzzy Euclidean space[14]. In Survey on Clustering Techniques for Big Data Mining focuses on a keen study of different clustering algorithms highlighting the characteristics of big data[15]. In Improved Method for Handling and Extracting useful Information from Big Data the meaningful information is extracted from the large amount of data and provide the aggregated form of output to the users[16]. In Bigdata Platform Design and Implementation Model an optimized bigdata platform implementation model was proposed through S/W configuration based on open source[17].

This paper contributes by developing a novel technique to classify the BIG DATA on the basis of selected features i.e. attributes. The work is motivated by viewing the 3 V's property of the BIG DATA as the feature reduction on the BIG DATA solves the 3 V's issue.

2. Materials and Methods

2.1 Reduction, Neural and Fuzzy based Classification Algorithm (RNFCA)

The present technique i.e. RNCFA is completed in three phases. First phase is used to reduce the dimensions of a high dimension dataset. While the second phase perform the clustering on the reduced dimension dataset. The fuzzy generated rules through the remaining attributes are used in phase 3 for the classification using the ANFIS (Adaptive Neuro-Fuzzy Inference System). The architecture of the RNFCA is shown in figure 1.

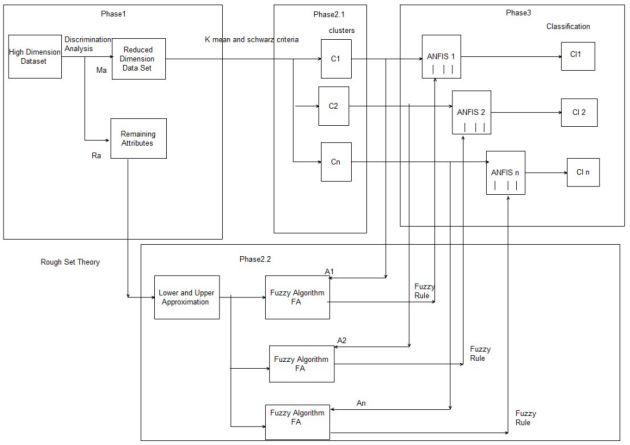

Figure 1. Architecture of the RNFCA.

The detailed description of all the three phases is given below.

2.1.1 Dimension Reduction Phase (Phase 1)

The Big data is a high dimension dataset due to various attributes available in the dataset. The discrimination analysis is used to discriminate the attributes on the basis of the features of the attributes. Suppose, input attributes has n_c classes for the a_n attributes of the dataset. The reduction process is described from equation no. (1) to (7). Some the symbols that are used are given in table 1:

Table 1. Various Symbols used

Symbol	Description
n_c	Number of Classes
a_n	Number of Attributes
a_{n_i}	Attribute of Class n_i
n_i	Number of Attributes in ith class
μ_i	ith Class Mean
μ	Mean of Complete Dataset
Sw_c	Scatter Matrix within class
Sb_c	Scatter Matrix Between Class

$$\mu_{i=\frac{1}{n_i}\sum a_{n_i}} \quad (1)$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i \quad (2)$$

The equation (1) calculates the mean of the particular class while the equation (2) is used to calculate the mean of complete dataset having n number of attributes.

$$S_{b_c} = \sum_{i=1}^{n_c} f_i (\mu - \mu_i)(\mu - \mu_i)^t \quad (3)$$

Where,

$$F_i = \frac{n_i}{n_c} \quad (4)$$

$$S_{w_c} = \frac{1}{a_n} \sum_{i=1}^{n_c} S_i \quad (5)$$

Where S_i can be calculated by using the equation (6).

$$S_i = \sum_{i=1}^{n_i} \left(a_{n_i} - \mu_i\right)\left(a_{n_i} - \mu_i\right)^t \quad (6)$$

The equation (3) calculates the scatter matrix between the classes while the equation (5) calculates the scatter matrix within the class. The total scatter matrix can be calculated by using the equation (7).

$$S_t = S_{w_c} + S_{b_c} \quad (7)$$

The Eigen vector of the S_t is used to select the attribute as Ma or Ra on the basis of equation (8).

$$IC = Min\left(E_v\left(S_t\right)\right) \quad (8)$$

The Attributes satisfying the criteria of equation (8) are marked as Ma while other as Ra. These attributes are given as input to phase2.

2.1.2 Phase 2

The phase 2 is divided into two sub modules, one sub-module uses the Schwarz criteria based k-mean clustering on Ma while the other module generates the fuzzy rules by using the rough set theory.

2.1.2.1 Schwarz based K-mean clustering

This phase is used to cluster the data by using the Schwarz based K-mean clustering of Ma attributes by using following algorithm:

1. Initiate K=smallest value(default k=2);
2. Apply K-means to generate number of clusters say Ma_{C0}, Ma_{C1}, Ma_{C2} ,……., Ma_{Cn}.
3. For each instance of dataset say **instance**
4. Calculate the Schwarz criterion for cluster Ma_{Ci} by using

$$\mathbf{SC = -2.In\hat{L} + k\,In\left(M_{a_n}\right)} \quad (1)$$

Where x = data within the cluster Ma_{Ci}

Ma_n = the number elements in Ma_{Ci}

k = the number of parameters to be estimated.

$\hat{L}$ = = The maximized value of the likelihood function of the model M i.e $\hat{L} = p(x \mid \hat{\theta}, M)$ where $\hat{\theta}$ are the parameter values that maximize the likelihood function.

5. Apply K-mean on Ma_{Ci} Clusters for k=2 say generated

Clusters are Ma_{Ci1} and Ma_{Ci2}

6. Calculate the SC for Clusters Ma_{Ci1} and Ma_{Ci2} by using

$$\mathbf{SC1 = -2.In\hat{L} + 2 * In(n)} \quad (2)$$

Here, the number of parameters gets doubled due to two clusters.

7. If SC>SC1 then n=n+1 i.e. new model preferred.
8. Ci=Ci1 and Cn=Ci2
9. i=i-1
10. End if
11. End

While the Ra attributes are used to generate the fuzzy rules by using phase 2.2.

2.1.2.2 Dynamic Fuzzy Rule Generation

The phase works on the rest dataset attributes i.e. Ra by following steps:

1. Scan the complete data.
2. Mark 1 to Ma and 0 to Ra.
3. Count frequency of the items with Ra.
4. Generate frequent item sets
5. Check the transaction of data is null
6. Put the value of support as the weight
7. Calculate the min and max by using the weight.
8. Calculate the distance with Euclidean distance formula
9. Generate distance vector value for the selection process
10. Initialized a population set ($t = 1$)
11. Compare the value of distance vector with population set
12. If value of support greater than vector value

13. Processed for encoding of data
14. Encoding format is binary
15. After encoding offspring are performed
16. A set of rules is generated.
17. Exit

These rules are used in phase 3 to classify each cluster element.

2.1.3 Phase 3

In this phase the classification process is done by using the ANFIS and by using the rules generated by phase 2.2. The architecture of the ANFIS is given in the figure 2.

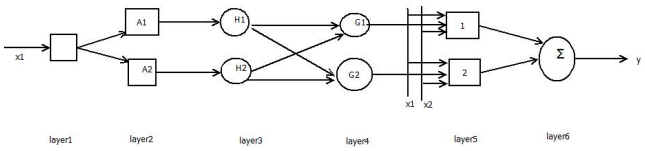

Figure 2. Architecture of the ANFIS.

The membership function for the fuzzy used in ANFIS architecture (shown in figure 2) is shown in the figure 3.

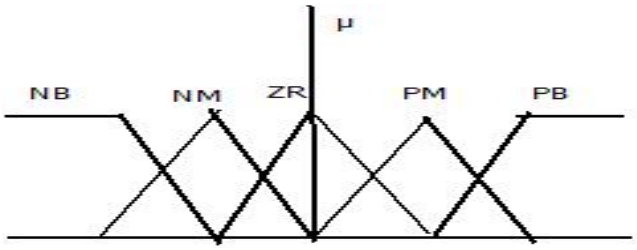

Figure 3. Membership functions for input and output.

The membership function can have lowest, lower, medium , high, highest values represented as NB, NM, ZR, PM and the PB. The result of the ANFIS is the class for each instance of the dataset. The next section describes the implementation with the results of the RNFCA on various dataset.

2. Results and Discussion

The RNFCA algorithm is analyzed adding the RNFCA algorithm to the WEKA library on the Intel i5 @ 2.67 GHz using the eclipse IDE. The algorithm is analyzed on the datasets having 400 instances with 25 attributes and 32561 instances with 15 attributes. The detail description of these datasets is given in table 2.

Table 2. Dataset Description

Data Type	Dataset 1 Medical	Dataset2 E-commerce
Number of attributes	25	15
Number of instances	400	32561
Number of classes	2	2
Attribute Type	Numeric, Nominal	Numeric, Nominal
Reference	UCI Repository[9]	UCI Repository[9]

The RNFCA algorithm is executed over two algorithms described in the table 2.Various parameters analyzed over dataset are accuracy, TP rate, FP rate, Precision and Recall. The description of these parameters can be found in [7]. The table 3 shows the performance of the RNFCA on dataset1 and dataset2.

Table 3. Performance of RNFCA algorithm

Parameters	Dataset1	Dataset2
Accuracy (%)	99.75	88.025
TP Rate	0.998	0.88
Fp Rate	.002	0.228
Precision	.998	0.88
Recall	0.998	0.873
F-Measure	0.998	0.871

The performance of the RNFCA algorithm can be compared with existing CCSA algorithm[8] and the decision tree i.e. J48. The figure 4 -7 shows the comparison graph of the J48, CCSA and RNFCA over various parameters.

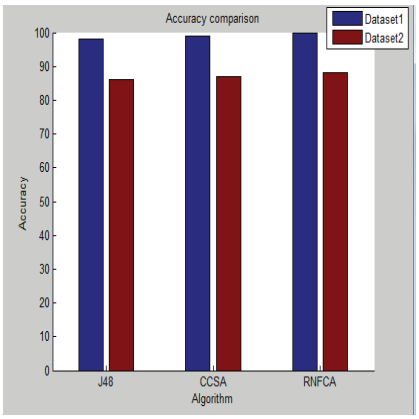

Figure 4. Accuracy comparison graph of the J48, CCSA and RNFCA.

The figure 4 shows the comparison of the accuracy for the techniques J48, CCSA and the RNFCA over dataset 1 and dataset 2 specified in table 2. The comparison clearly specifies that the RNFCA algorithm has improved the accuracy about 0.875% over the CCSA algorithm which has 0.9% enhancement over the J48 algorithm. The improvement in the accuracy results in enhanced mapping of elements with correct class.

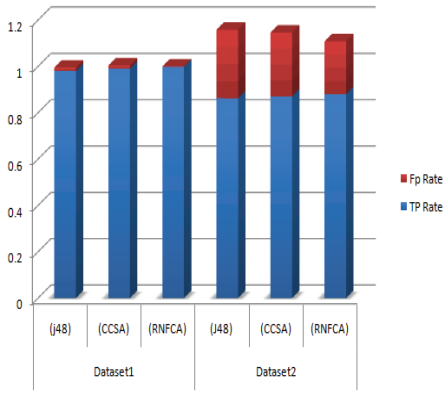

Figure 5. TP Rate vs FP Rate comparison graph of the J48, CCSA and RNFCA.

The figure 5 ROC graph for the specified algorithms over two datasets. The ROC graph shows the comparison of the TP rate vs FP rate. The enhancement in the TP with decrement in the FP rate shows the better receiver operating characteristics of the RNFCA algorithm as compared to CCSA & J48.

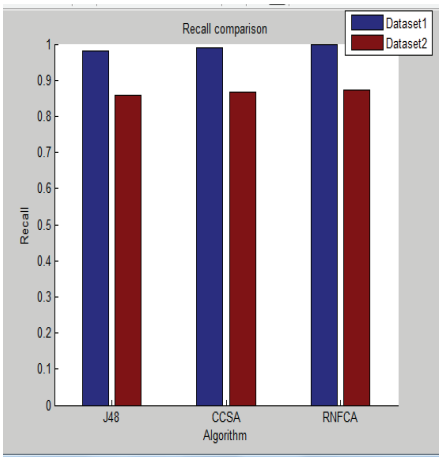

Figure 6. Recall comparison graph of the J48, CCSA and RNFCA.

The parameter recall specifies the relevancy of the output. The enhancement of the recall value of RNFCA algorithm over the CCSA and J48 represents more relevant data through RNFCA algorithm.

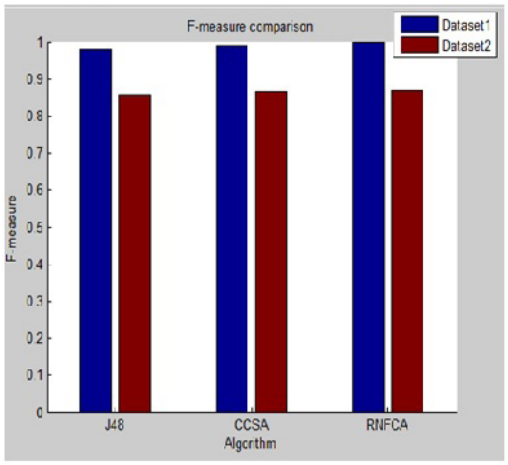

Figure 7. F-measure comparison graph of the J48, CCSA and RNFCA.

The F-measure is the harmonic mean of the precision and the recall which measures the effectiveness of the algorithm. The RNFCA algorithm exhibits better F-measure as compared to CCSA and J48 algorithm showing significance of the algorithm.

4. Conclusion

The paper implements the RNFCA algorithm and compares the performance with CCSA and J48 algorithm over two datasets. The performance of the algorithm is better when the numbers of attributes are large, while the performance enhancement can be recognized in other datasets also. The performance comparison depicts the same and shows the optimization by 0.9% using the RNFCA algorithm. The algorithm can also be analyzed on various other application areas'.

5. References

1. Li J, Xu Z, Jiang Y, Zhang R. The Overview of Big Data Storage and Management. 2014 13th International Conference on Cognitive Informatics and Cognitive Computing, London, ICCICC'14. 2014. p. 510 –13.
2. Pitre R, Kolekar V. A Survey Paper on Data Mining with Big Data. International Journal of Innovative Research in Advanced Engineering, (IJIRAE). 2014; 1(1):178–80.
3. Laney D. 3-D Data Management: Controlling Data Volume, Velocity and Variety, META Group Research Note. Available from: https://blogs.gartner.com/doug-laney/files/2012/01/

ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf , Date accessed: 06/02/2001.
4. Fan W, Bifet A. Mining Big data: Current Status and Forecast to the future, SIGKDD Explorations. 2013; 14(2):1–5.
5. Han J, Liu Y, Sun X. A Scalable Random Forest Algorithm Based on Map Reduce. 2013 4th IEEE International Conference on Software Engineering and Service Science, (ICSESS), Beijing. 2013. p. 849–52.
6. Cavallaro G, Ridel M, Richerzhagen M, et al. On understanding big data impacts in remotely sensed image classification using support vector machine methods. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2015; 8(10):4634–46.
7. Navneet Gill NS. Algorithm for producing compact decision trees for enhancing classification accuracy in fertilizer recommendation of soil. International Journal of Computer Applications. 2014; 98(2):8–14.
8. Navneet Gill NS. Classification using the compact rule generation. Oriental Journal of Computer Science and Technology. 2015; 8(1):49–58.
9. UCI Machine Learning Repository. Available from: http://archive.ics.uci.edu/ml. Date accessed: 1/12/2015.
10. Thiyagarajan VS. Platfora method for high data delivery in large datasets. Indian Journal of Science and Technology. 2015; 8(33):1–13.
11. Verma A, Kaur I, Arora N. Comparative Analysis of Information Extraction Techniques for Data Mining. Indian Journal of Science and Technology. 2016; 9(11):1–18.
12. Peisker A, Dalai S. Data Analytics for Rural Development. Indian Journal of Science and Technology. 2015; 8(4):50–60.
13. Moayeri M, Shahvarani A, Behzadi M H, Hosseinzadeh-Lotfi F. Comparison of Fuzzy AHP and Fuzzy TOPSIS Methods for Math Teachers Selection. Indian Journal of Science and Technology. 2015; 8(13):1–10.
14. El-Ahmady AE, Al-Rdade A. Folding of Fuzzy Hyperboloid. Indian Journal of Science and Technology. 2013; 6(8):1–6.
15. Sajana T, Sheela Rani CM, Narayana KV. A Survey on Clustering Techniques for Big Data Mining. Indian Journal of Science and Technology. 2016 Jan; 9(3). Doi no:10.17485/ijst/2016/v9i3/75971
16. Karthick N, Agnes Kalarani X. An Improved Method for Handling and Extracting useful Information from Big Data. Indian Journal of Science and Technology. 2015 Dec; 8(33) Doi no:10.17485/ijst/2015/v8i33/60744
17. Kyoo-sung N, Doo-sik L. Bigdata Platform Design and Implementation Model. Indian Journal of Science and Technology. 2015 Aug, 8(1 8). Do: 10.17485/ijst/2015/v8i18/75864