Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 8 (2017) pp. 2479-2490
© Research India Publications
http://www.ripublication.com

Multidimensional-RPCA Algorithm for Enhancing the Classification Accuracy

Dr. K.Mani[1] and N. Elavarasan[2]

[1]Associate Professor in Computer Science, Nehru Memorial College, Puthanampatti, Trichy.
[2]Research Scholar in Computer Science, Nehru Memorial College, Puthanampatti, Trichy.

Abstract

Multidimensional datasets is a significant problem across a wide variety of information processing fields including classification in data mining, data compression, machine learning and dataset navigation. Usual visualization techniques for multidimensional datasets such as root principle component analysis using pertinent dimensions detection do not scale well to higher dimensions. A common approach in solving this problem is dimensionality feature extraction. Existing dimensionality expectation maximization techniques usually used in selecting the feature extraction pertinent dimension subsets that are significant to the user without loss of information but classification accuracy is not improved because low dimension is used. A combined approach based on expectation–maximization algorithms, interactive ensemble model algorithms and the evolutionary algorithms have been used in this paper to obtain optimal dimension subsets which represents the original dataset without losing information for classification. For that NASA PROMISE2016 real dataset is considered. A comparative analysis is performed for high-dimensional datasets using the proposed method with existing EM's Root methods for classification accuracy.

Keywords: Pertinent dimension detection, Expectation Maximization , Root Principle Component Analysis , Classification Defect Prediction and Radial Basis Function complex.

I. INTRODUCTION

Data Mining sometimes called as knowledge discovery is the process of analyzing data from different perspectives and summarizing them into useful information. Feature extraction, extracts a subset of new features from the original feature set. Classification is the process of assigning label to the input data point given. There will be number of data point grouped and labeled into several categories. All given input data point is assigning any one of the label. In general, classification is performed over a clustered data points. It is based on certain similarity measures computed over the data points of different classes.

Machine learning is a subfield of computer science that evolved from the study of pattern recognition, computational learning in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data[5]. Feature extraction is the method of deciding on a subset of important and relevant features for building reliable learning models. It makes training and utilizing a classifier more efficient by reducing the size of the effective training set[7]. In this work, feature extraction and machine learning algorithms applied on publicly available PROMISE dataset PCA to analyze the predictive performance of the Classification. Feature Extraction (FE) selects the optimal subset of features by determining the worth of attributes. Root Principle Component Analysis (RPCA), feature extraction technique reduces the dimensions by merging of features based on eigenvalues. The compaction of feature set and boundary condition problems can be resolved by using EM-RBF Root. An Ensemble Model with expectation maximization learning approach implemented to train the sequence of RBF-EM component classifiers, whose model parameters are adaptively different manifesting in better generalization as compared to expectation–maximization approach [10].

The rest of the paper is organized as follows Section II describes the various work related with the classification accuracy. The proposed methodology with an example is discussed in section III. The experimental results obtained after implementing the proposed methodology is discussed in section IV. Finally section V ends with conclusion.

II. RELATED WORKS

The work carried out thus far by other researchers that related to defect prediction research using classification and machine learning concisely represented here.

A survey performed by Sadroddin Alavipanah et. al. [1] included feature extraction, data classification, and classifier evaluation. For that, the NASA datasets of PROMISE 2016 repository were selected to classify models for classification defect prediction. The datasets were collected and fed as input to the feature extraction process. Correlation-based feature subset extraction method gave the optimal feature sets. The selected features were used to classify into two classes namely defective and not-defective by using classification algorithms namely Bayes Net, Naive Bayes, Random Forest, Instance-based classifier and Random Tree. The result indicated that Random Forest algorithm would select the small subset of available attributes. In [2], Donald J et. al. discussed the concept of Bayesian networks to identify the influential set of metrics. They defined two new more metrics- Number of developers (NOD) and Lack of coding quality (LOCQ) in addition to metrics used in promise data repository.

In [3], Daniel Hausknost et. al. studied the empirical effect on predictive performance using different datasets with varying levels of imbalance of classification defect predication models. They proposed four evaluation techniques such as Mathews Correlation Coefficient (MCC), F-Measure, Precision and Recall used to measure the degree of predictiveness. The result indicated the predicative performance of Classification predication model get reduced when the data was imbalanced. Tong Tong et.al. [4] demonstrated various issues that affect the performance of defect predication models. They addressed the various aspects that remain unresolved such as a relationship between attributes and fault, no standard measures for performance assessment, issues with cross-project defect prediction, no general framework available, class imbalance problem, an economics of Classification defect prediction.

III. PROPOSED METHODOLOGY

The proposed methodology consists of three phases viz., identifying the concrete dimension, reducing the concrete dimension and improving the classification accuracy. In order to identify the concrete dimension, PDD is used. The concrete dimensions may have some irrelevant features. In order to eliminate them RPCA is used. Once the relevant features are obtained EM method is used in conjunction with Sushisen algorithm shown in Algorithm 3.1 to improve overall performance of Accuracy, Recall and Precision and they are visualized using the confusion matrix. The proposed methodology shown in the figure 1 and each phase is explained in the next subsections.

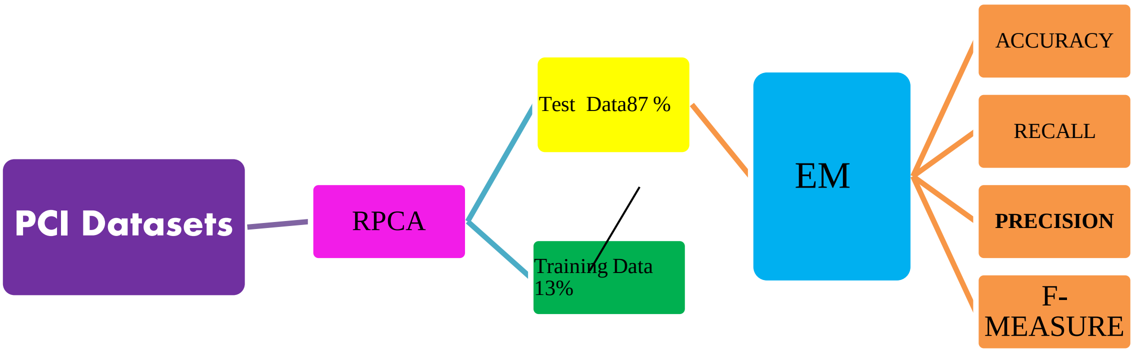

Fig 1: Proposed Methodology

3.1 PDD

In [9], N. Elavarasan has proposed a PDD algorithm. Based on identified attribute sets, this method computes multi attribute depthness similarity for each of the data points towards each class. The class which has more depthness value will be assigned as a class to the input data point. The classification accuracy is high compared to other algorithms by employing the PDD and classification algorithm. After identifying PD, then use Correlation feature Extraction (CFE) technique to select the optimal subset of features and to determine the degree of redundancy among them.

3.2 RPCA

It is an extension of PCA. The main difference between the conventional PCA and RPCA is that the PCA operates on zero-centred but RPCA operates by diagonalzing the covariance matrix which gives a polynomial decomposition of a covariance matrix. In RPCA first the desired root is chosen using K matrix and then center the feature space via the K matrix. RPCA feature extraction method is used to reduce the number of dimensions by merging features based on the theory of equations and then multiply the centered root matrix by the desired root corresponding to the largest root values. First, consider the root function K **[7]** which is a measure of closeness. If k=1 when the points coincide and equal to 0 at infinity.

3.3 EM

EM is a supervised learning method with associated learning algorithms used to solve classification and regression problems. It implements automatic complexity control to reduce over-fitting and uses a flexible representation of the class boundaries. It constructs a hyper-plane or set of hyper-planes in a high-or infinite-dimensional space, which can be used for classification, regression, or other tasks [6]. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a new data point will be in. The idea is to find the best hyper-plane that represents the largest separation between the two classes[11]. Therefore, we choose the hyper-plane so that the distance from it to the nearest data point on each side maximized. Various EM‟s Root includes linear root, non-linear root, polynomial root, radial basis function and multidimensional scaling.

3.3.1 Linear root

This classifiers consists of the set of training data points and finding the maximum-margin hyper-plane that divides the group of points so that that the distance between the hyper-plane and the nearest point root is maximized.

3.3.2 Non-linear root

This classifiers can be created by applying the Root trick to maximum-margin hyper-planes. The algorithm is similar to linear root(i=2), except that every dot product replaced by a nonlinear root function and allows the algorithm to fit the maximum-margin hyper-plane in a transformed feature space.

3.3.3 Polynomial root

In this classifier (i=3) kernels functions are used with EM and other models that represent the similarity of root in a feature space over polynomials of the original variables thus allowing learning of non-linear models.

3.3.4 Radial Basis Function (RBF) complex’s

This classier (i=4) is used to find a set of weights for curve fitting problem. The weights are in higher dimensional space than the original data. In this association learning means to find a surface in high dimensional space that provides the best fit to the training data. The complex includes three dimensions: input level, one hidden level and a linear output level. Hidden dimensions provide a set of functions that constitute an arbitrary basis for input patterns, these functions are called radial basis functions.

3.3.5 Multi-Dimensional Scaling (MDS)

This artificial neural root (i=5) model maps sets of input data into a set of appropriate outputs. The network consists of multiple dimensions of nodes like in the directed graph, with each level fully connected to the next one. Each node represents a neuron with a not- linear activation function, except for the input nodes. EM-RBF Root is a classifier model to solve boundary condition problem and to improve the overall performance of the system.

Once these classifiers are incorporated in EM's method, the performance of the classifier model is measured based on various predication parameters like accuracy, recall, precision, and F-measure and they are calculated using

Accuracy = TP+TN/ (TP+FP+FN+TN)

Precision or Positive Predictive Value (PPV) PPV = TP / (TP + FP)

Sensitivity or True Positive Rate (TPR) TPR = TP / (TP + FN)

F- Measure: F1 Score (F1) FDR = FP / (FP + TP)

3.4 Sushisen Algorithms

Sushisen algorithm is an effective feature extraction method and is widely used in PDD in "names" attributes. Information and classification algorithm concern the relation between a certain feature word and certain class, but treat all classes in training set as a whole and the importance of a certain word is measured by calculating the information amount that each class takes. Information classification of the feature word refers to the dataset value between the information amount of the whole training set without regard to feature word attribute and that of the training set with regard to feature word.

Input: Multi-Dimensional reduces Dimension Classification performance analysis

Init : EM Data and PROMISE2016 Dataset

Output: Ensemble Model Classification in Overall Accuracy

```
1. Initialization: Set $\boldsymbol{F} \leftarrow$ "initial set of $n$ features."
2. For $d(MD) = 1$ to $n$ do
3. RPCA Compute $I(xd, C)\ \forall xd \in \boldsymbol{F}$
4. end for
5. While $n >$ the desired number of RD features
6. Find a feature $xi$ that minimizes $I(xi, C)$;
7. Set $\boldsymbol{F} \leftarrow \boldsymbol{F} \setminus \{xi\}$;
```

```
For (PROMISE2016) $r = 1$ to $n$ $i$ _= $r$ do
    EM  Compute $I (xi , xi )$ $\forall xr \in \boldsymbol{F}$;
end for
Find a feature $x j$ that maximizes $I (xi , x j )$;
Set $\boldsymbol{F} \leftarrow \boldsymbol{F} \setminus$ _$x j$ _;
While the pair of features $(xi , x j ) > PDD$;
    Extraction of the feature: root equation
    Extract a feature $y$ from the pair of features odd degree
    $CFE (xi , x j)$  such that maximizes $I (y, C)$;
    Set $\boldsymbol{F} \leftarrow \boldsymbol{F} \cup \{y\}$;
    ROC  Compute $I (y,C)$;
    Confusion matrix $n \leftarrow n - 1$;
end while
The set $\boldsymbol{F}$ containing the created CM(confusion matrix) features .
```

IV RESULTS AND DISCUSSSION

The proposed methodology is implemented on NASA PROMISE2016 dataset using MATLAB. A comparative analysis is on different EM's with and without PCA based on accuracy, recall, precision, and F-measure and the results are tabulated. Table 1 shows the performance prediction measures.

Table 1: Performance Predication Measures

	Predicted 1	**Predicted 0**		**Predicted 1**	**Predicted 0**
True1	true positive	false negative	True1	TP	**FN**
True0	false positive	true negative	True0	FP	**TN**
	Predicted 1	Predicted 0		Predicted 1	**Predicted 0**
True1	Hits	Misses	True1	208	**0**
True0	false alarms	correct rejections	True0	68	**10**

The confusion matrix of different EM‟s on NASA PROMISE2016 dataset having 286 feature extraction without and with RPCA are shown in table 2 and table 3 respectively and their corresponding visual representations are shown in fig. 3 and fig. 4 respectively.

Table 2: Prediction performance of different EM's without RPCA

Non- Linear Root		Linear Root		Polynomial Root		
208	11	208	0	208	21	
57	10	68	10	49	8	
RBF complex's		Multi-Dimensional Scaling		Ensemble Model		
208	37	208	39	167	9	
30	19	18	23	10	100	

Table 3: Different EM's performance with RPCA

Non- Linear Root		Linear Root		Polynomial Root		
210	11	211	1	210	19	
55	10	65	9	47	10	
RBF complex's		Multi-Dimensional Scaling		Ensemble Model		
212	35	206	37	170	7	
26	13	20	23	13	96	

Table 4: Different EM's performance (with RPCA)

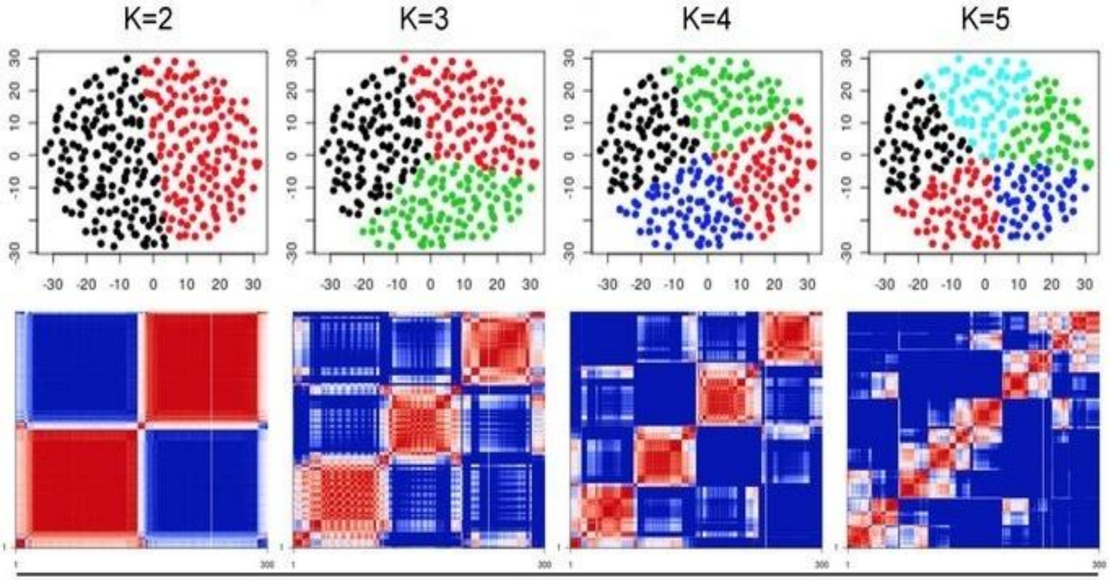

Fig 2: Graphical Comparison of different EM's under parameters (without RPCA)

4: Graphical Comparison of different EM's under parameters (with RPCA)

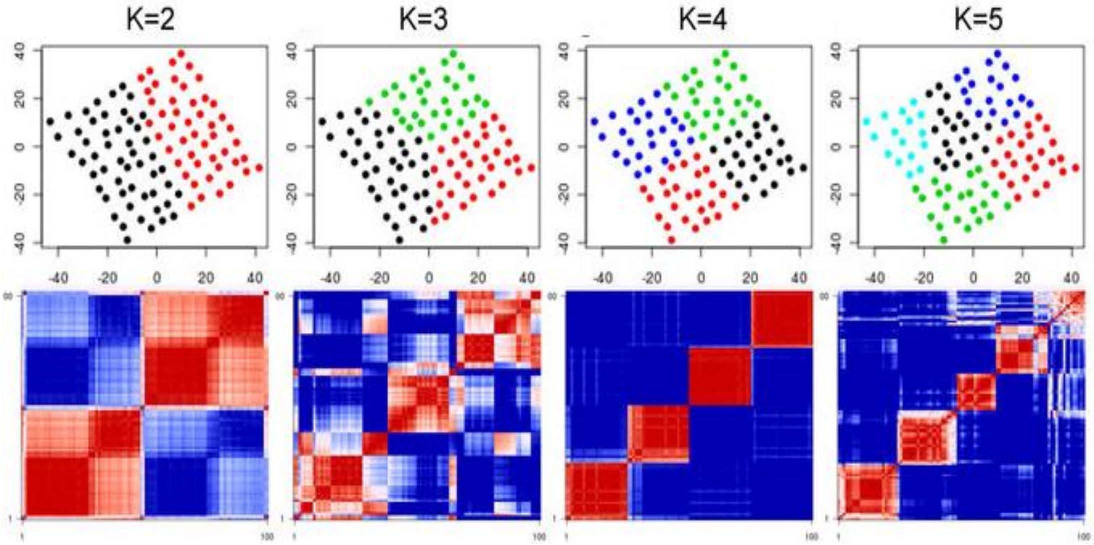

Fig 3: Graphical Comparison of different EM's under parameters (with RPCA)

The results indicate that even by reducing a number of features extracted from 286, the EM with RPCA outperforms than EM without RPCA provides better accuracy, precision, recall and F-measure

The third part represents prediction performance table of different EM"s after applying feature extraction RPCA technique by reducing the PCA dataset from 286 feature Extraction.

Table 4: Different EM's performance (with RPCA)

EM Type	Accuracy(%)	Precision(%)	Recall(%)	F- Measure(%)
Linear Root	76.22	98	75.36	0
Non- Linear Root	76.22	94.98	78.49	5.02
Polynomial Root	75.79	91.23	80.93	8.77
RBF complex's	77.21	84.9	87.39	15.1
MDS	80.21	84.21	92.04	15.79
Ensemble Model	93.36	94.89	94.35	5.11

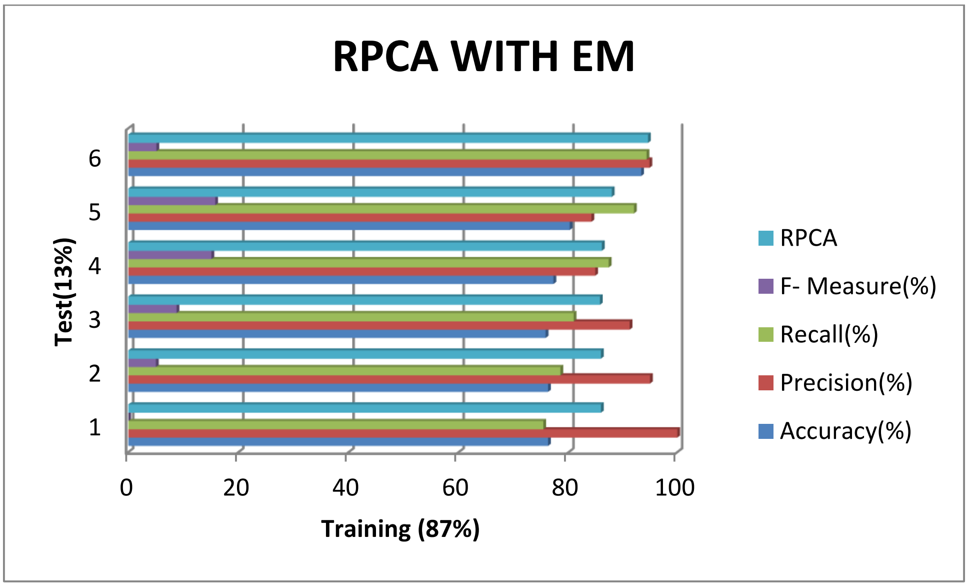

Fig 7: Overall EM's With PRCA Performance results

EM Type	Accuracy(%)		Precision(%)		Recall(%)		F- Measure(%)	
	Without RPCA	With RPCA	Without RPCA	With RPCA	Without RPCA	With RPCA	Without RPCA	With RPCA
Linear Root	76.22	76.92	75.36	76.44	75.36	99.52	24.63	23.55
Non- Linear Root	76.22	76.92	78.49	79.24	94.97	95.02	21.45	20.75
Polynomial Root	75.52	76.92	80.93	81.71	90.82	91.70	19.06	18.28
RBF complex’s	76.57	78.67	87.39	89.07	84.89	85.82	12.60	10.92
MDS	80.06	80.06	92.04	91.15	84.29	84.77	7.96	8.84
Ensemble Model	93.26	93.37	94.35	94.89	94.88	96.04	5.64	7.10

V. CONCLUSION AND FUTURE SCOPE

This work proposes an EM model for Classification defect prediction. EMs are used to provide component learning and to improve the overall performance of the system. There was significant difference in the performance of various EM‟s Root when feature extraction (RPCA) method was applied, which reduced the feature subset from 286 attributes. The experimental results clearly revealed that the predictive comparison of the proposed approach is better or at least comparable with other approaches. In order to improve the existing model performance by incorporating with the suitable EM networks as a part of expectation–maximization which gives better results. The performance of obtained Ensemble Model with EM from the result for PCA dataset with reduced number of features is accuracy 93.36%, precision is 94.89%, recall is 94.35% and F-measure is 5.11%.

REFERENCES

[1] Sadroddin Alavipanah, Dagmar Haase, Tobia Lakes, Salman Qureshi,”Integrating the third dimension into the concept of urban ecosystem services”, A Review Article Ecological Indicators, Volume 72, January 2017, Pages 374-398.

[2] Donald J. Lollar, Willi Horner-Johnson,”Public Health Dimensions of Disability”,International Encyclopedia of Public Health (Second Edition), 2017, Pages 190-199.

[3] Daniel Hausknost, Nelson Grima, Simron Jit Singh,The political dimensions

of Payments for Ecosystem Services (PES)",Ecological Economics, Volume 131, January 2017, Pages 109-118.

[4] Tong Tong, Katherine Gray, Qinquan Gao, Liang Chen, Daniel Rueckert, "The Alzheimer's Disease Neuro imaging InitiativeMulti-modal classification of Alzheimer's disease using nonlinear graph Pattern Recognition", Volume 63, March 2017, Pages 171-181.

[5] JianWen Tao, Dawei Song, Shiting Wen, Wenjun Hu,Robust," Multi-source adaptation visual classification using supervised low-rank representation, Pattern Recognition", Volume 61, January 2017, Pages 47-65.

[6] Yanxiong Li, Qin Wang, Xue Zhang, Wei Li, Xinchao Li, Jichen Yang, Xiaohui Feng, Qian Huang, Qianhua H,"Unsupervised classification of speaker roles in multi-participant conversational speech,Computer Speech & Language", Volume 42, March 2017, Pages 81-99.

[7] Wei Shen, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, Jie Tian,"Multi-crop Convolutional Neural Networks for lung nodule malignancy suspiciousness classification,Pattern Recognition", Volume 61, January 2017, Pages 663-673.

[8] R. Trigui, J. Mitéran, P.M. Walker, L. Sellami, A. Ben Hamida,"Automatic classification and localization of prostate cancer using multi-parametric MRI/MRS",Biomedical Signal Processing and Control, Volume 31, January 2017, Pages 189-198.

[9] N.Elavarasan,Dr.P.Mani," MADSE: An Enhanced Data Mining Algorithm to Improve Classification Accuracy" - WCCCT 2016, IEEE Xplore.

[10] F. Dornaika, Y. El Traboulsi,"Matrix exponential based semi-supervised discriminant embedding for image classification, Pattern Recognition", Volume 61, January 2017, Pages 92-103.

[11] Peng Zheng, Zhong-Qiu Zhao, Jun Gao, Xindong Wu,"Image set classification based on cooperative sparse representation, Pattern Recognition", Volume 63, March 2017, Pages 206-217.