IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 4, Ver. III (Jul – Aug. 2014), PP 93-99
www.iosrjournals.org

Performance Analysis of Hybrid (supervised and unsupervised) method for multiclass data set

Rahul R. Chakre[1], Dr. Radhakrishna Naik[2]
[1](PG Student, CSE, MIT, Aurangabad, Maharashtra, India)
[2](Prof. and Head, CSE, MIT, Aurangabad, Maharashtra, India)

Abstract: *Due to the increasing demand for multivariate data analysis from the various application the dimensionality reduction becomes an important task to represent the data in low dimensional space for the robust data representation. In this paper, multivariate data analyzed by using a new approach SVM and ICA to enhance the classification accuracy in a way that data can be present in more condensed form. Traditional methods are classified into two types namely standalone and hybrid method. Standalone method uses either supervised or unsupervised approach, whereas hybrid method uses both approaches. This paper consists of SVM (support vector machine) as supervised and ICA (Independent component analysis) as a unsupervised approach for the improvement of the classification on the basis of dimensionality reduction. SVM uses SRM (structural risk minimization) principle which is very effective over ERM (empirical risk minimization) which minimizes an upper bound on the expected risk, as opposed to ERM that minimizes the error on the training data, whereas ICA uses maximum independence maximization to improve performance. The perpendicular or right angel projection is used to avoid the redundancy and to improve the dimensionality reduction. At last step a classification algorithm is used to classify the data samples and classification accuracy is measured. Experiments are performed for various two classes as well as multiclass dataset and performance of hybrid, standalone approaches are compared.*

Keywords: *Dimensionality Reduction, Hybrid Methods, Supervised Learning, Unsupervised Learning, Support Vector Machine (SVM), Independent Component Analysis (ICA).*

I. Introduction

In the recent years data is increased not only in terms of number of samples but also in terms of number of dimensions in many applications such as gene expression data analysis, where the number of dimensions in the raw data ranges from hundreds to tens of thousands. From the theoretical point view more dimensions gives more understanding. This paper assumes the independency among the dimensions or features but in practical inclusion of more features or dimensions leads to undesirable performance i.e. known as curse of dimensionality. Multivariate data analysis brings great difficulty in pattern recognition, machine learning and data mining. Therefore the dimensionality reduction becomes very important task. Dimensionality reduction is a well-known data mining problem, which is usually considered as an improvement for pre-processing technique for subsequent machining. On completion of this task it will gives very desirable advantages such as reducing the measurement, storage and transmission, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance in terms of speed, accuracy and simplicity, facilitating data visualization and data understanding. A lot of dimensionality reduction methods were proposed to deal with these challenging tasks. Due to the computational complexity of data and classification in real-world applications, it seems not an easy task to build a general data reduction technique, so researches on dimensionality reduction have been conducted for last several decades and are still extracting much attention from pattern recognition and data mining society. Dimensionality reduction is nothing but the extracting the essential features in a way that high dimensional data can be presented in low dimensional space. Supervised and unsupervised are the two popular methods for the dimensionality reduction in first type class labels is known and in second type class labels are unknown. Hybrid dimensionality reduction method [1] is the combination of these two approaches which uses both the criteria. In supervised learning LDA [2] and Support Vector Machine [3] are the tow famous dimensionality reduction techniques, whereas in unsupervised learning PCA [4] and Independent Component Analysis [5] and last there is combination supervised and unsupervised approach which is known as Support Vector Machine and Independent Component Analysis.

II. Literature Survey

2.1 Linear Discriminant Analysis

Basically this is a supervised method in which class labels are known or in other words some prior information about the dataset is available. It is also called as fisher's discriminant analysis. There are two approaches for the LDA [2] namely class dependent and class independent transformation. This paper focuses

on the multiclass data set for the dimensionality reduction, therefore first approach i.e. class dependent transformation is used. This type of approach involves maximizing the ratio of between class variance to within class variance. The main objective is to maximize this ratio so that adequate class separability is obtained. The class-definite type approach involves using two optimizing criteria for transforming the data sets independently. The objective of LDA [2] is to perform dimensionality reduction while keeping as much of the class discriminatory information as possible.

Mathematical Formulation of LDA

$$S_B = \sum_{i=1}^{c} N_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{1}$$

Where,

S_B represents the between class scatter matrix.

$X_i \subset X$ Include many i th class data.

μ_i is the mean of data in X_i

Xi is the mean of the entire data X.

c Denotes the number of class in X.

$$S_w = \sum_{i=1}^{c} (\mathrm{x}_i - \mu)(\mathrm{x}_i - \mu)^T \tag{2}$$

Where S_W is the within scatter matrix.

$$W^* = \frac{\left\| W^T S_B W \right\|}{\left\| W^T S_W W \right\|} \tag{3}$$

Based on the S_B and S_W , the LDA a criterion is shown in above formula. W^* is the selected matrix in which maximizes ratio of between class scatter matrix and within class scatter matrix of the projects samples. For the above fisher again prove that for the maximization of this S_W should be non-singular so the effectiveness of class seperability is maximum.

2.1.1 Limitations of LDA

The major drawback of this method is it only focuses on the class discriminatory function it does not work for the feature discriminant functions. Again LDA have small sample size problem it does not work properly when there is small samples or instances are given , while working with this approach primary step is to calculate the class mean, again LDA have some problems with common mean problem, due to this classification accuracy gets reduced. LDA is a parametric in the nature so it assumes the unimodel of Gaussian likelihood therefore it does not work properly with the non Gaussian distribution. The robustness problem is also there in LDA .For avoiding some these problems there are methods namely null space LDA for small size small problem [6], discriminative common vector for the class mean problem, Orthogonal centroid method for robustness problem [7].The LDA [2] also extended for the kLDA for dealing with non-linearity, but all these method are inherited from the LDA so no one method avoid the all the above mentioned problems.

2.2 Principal Component Analysis

This is very popular unsupervised dimensionality reduction method in which class labels are unknown to us or In other words no prior information about the dataset is available to. It is also known for the the Karhunen-Loeve transform or Factor Analysis. The main objective of PCA [3] is to reduce the number of dimensions or features and transfer set of correlated variables into set of uncorrelated variables in a way that the original dataset is mapped into lower dimensional space. PCA projects the data in least square sense in which it captures the big variability in the data and ignores the small variability. Due to this features which are present in the various class will be separated in a way that dimensionality reduction takes place effectively, so it can place that feature from high dimensional space to low dimensional space. The classification accuracy is better in this type of method.

Mathematical Formulation for the PCA

$$W^* = \arg\max \left\| W^T S_T W \right\| \tag{4}$$

$$S_T = \sum_{i=1}^{c} (\mathrm{x}_i - \mu)(\mathrm{x}_i - \mu)^T \tag{5}$$

Where x_i is the i^{th} n dimensional data among N multidimensional dataset. When M is the dimension of the dimension vector s fulfills the $m << n$, $s_i = W^{*T} x_i \in R^m$ where W shows mapping of a good or optimal covariance of the dimensions described by the above total scatter matrix. Based on the predetermined the W . The projection is chosen in such way that the determinant of total scatter matrix will maximum.

2.2.1 Limitations for the PCA

It works very undesirable for the nonlinear dataset. Due to this classification accuracy is drastically reduced. But for that nonlinear component analysis is used i.e. KPCA [8], it gives a suitable kernel for transformation of low dimensional space into high dimensional but again it increases the computational complexity. PCA also suffers from the small instance problem. Due to which classification accuracy gets drastically reduced. PCA only focus on the feature or dimensions of the dataset it does not care about the class labels. This is also one of the major drawbacks while doing the dimensionality reduction. Regression model [9] is also another approach for the avoiding this problem but it fails at some point.

2.3 Support Vector Machine

SVM's precisely depend upon the theoretical model of learning with the surety of efficient performance. Basically SVM use the structure risk minimization principal by the use of linear function which is coming for the available datasets for the classification purpose. The main aim is to build a good classifier well for the unknown samples. The SVM is highest level of development of technique at present time classification method which is widely used in the statical learning environment. The objective of support vector machine is to obtain the most favorable hyper plane for linearly separable datasets and extends for the nonlinear datasets by the transformation of the original dataset into high dimensional space by the use of appropriate kernel function. SVM maximizes the margin around the separating hyper plane. The decision function is fully specified by subset of training samples, the support vectors i.e. the data points which are closest to the decision surface which are mostly difficult to classify. Due to the use of structural risk minimization principal the SVM does not suffers from the small sample size problem and common mean problem. It very effectively reduces the dimensions and gives better classification performance.

2.4 Independent Component Analysis

Independent component analysis is nothing but finding out the underlying factors or the dimensions from the multivariate statistical data. Basically it is famous blind source separation but it can be use for the dimensionality reduction. Independent Component Analysis [5] is different from other method because it looks for the dimensions or factors which are statistical independent and non-Gaussians. ICA is an unsupervised method in which prior knowledge about the class labels is unknown. It is a more advantageous than the PCA [4] because PCA seeks the projection which has maximum variance whereas ICA [5] seeks the projection which has maximum independence. That is the reason in this paper ICA is used for the maximization of independence due to which classification accuracy gets increased. For the implementation of the ICA there are two algorithms namely maximum likelihood source separations and Informix [10] but FastICA [11] algorithm due to its computational and conceptual simplicity for the multivariate data analysis. In ICA, the initial step is centering and whitening process then there is FastICA [11] algorithm.

2.5 Conclusion from Literature Survey

Dimensionality reduction is well known problem in real world application. In above there are only standalone approaches are used and every individual approach is suffering from some important issues due to that classification accuracy gets reduced. The existing hybrid methods are also suffers from the above problems because all the methods are inherited either from the supervised or unsupervised approach some of them are asymmetric principal and discriminant analysis(APCDA) [12], ICA augmented LDA [13] , discriminant non negative matrix factorization (DNNF) [14].Due to the above problems a new approach which is hybrid in the nature named SVM+ICA is used, it firstly utilizes supervised criteria that structural risk minimization and then it utilizes second unsupervised criteria independency among the feature maximization. So the steps for new proposed approach which is as follows

III. Proposed Work

3.1 SVM and ICA

As the name indicates it is the combination of two different approaches i.e. SVM as supervised and ICA as unsupervised. It is a hybrid method which is used for the dimensionality reduction. It uses both criteria's for the multivariate data analysis .As explained in the third point SVM uses SRM principal so it gives the best classification accuracy than other methods. For the implementation of SVM, LIBSVM software package [15] is used.

In more detail, SVM reduces the structural risk due to this the projection which are coming from the SVM gives the superior generalization ability to enhance the classification accuracy among the multivariate data set. The main objective of the SVM is to maximize the margin which gives a better data representation which results in the desirable classification performance and another advantage is that it gives projections which are well known for the construction of competent subspace for the dimensionality reduction. In this hybrid method SVM is treated as supervised part for dimensionality reduction. SVM provides optimum decision surface with the minimum structural risk by the use quadratic constrained optimization problem. The dual problem of multiclass classification is given by

$$\alpha^* = \arg\min_{\alpha}\{1/2\alpha^T Q\alpha - \alpha^T 1\}$$
$$\text{Subject to } \sum \alpha_i a_i = 0 \ , \ 0 < \alpha_i < b \ . \tag{6}$$

Where α_i is nothing but the langrage's multiplier for solving the optimization problem. y_i are the data samples and a_i will be the multiclass index. b is the some relaxation parameter to avoid the empirical risk, but X as a weight vector is used for the mapping in the linear way. so the it is given by

$$X = \sum \alpha_i^* a_i y_i \subset R^n \tag{7}$$

The most desirable set of mapping vectors derives from the SRM principle the starting process $X_{1,l}$ .there should be the pair wise orthogonally among the SVM and ICA, which is also denoted by the $X_{1,l} \perp X_{1+l,c}$ .The $X_{1,l}$ is calculated as a constrained optimization issue which is as follows

$$z^* = \arg\min \left\| y - z \right\|^2$$
$$\text{Subject to } X_{l,1}{}^T, z = 0 \tag{8}$$

Where Z represents the projected data onto the subspace orthogonal to $X_{1,l}$, and parallel to the decision hyper plane(s). Due to the orthogonality between $X_{1,l}$ and any components in the decision hyperplane, the structural risk minimization and independence maximization are isolated and performed one by one holding independence between any pair of x_i 's and x_j 's where $i = \{1.....l\}$ and $j = \{l+1.....c\}$

The second part of our hybrid method is ICA which plays the important role in this paper. ICA is mainly deal with the data which are statically independent and non Gaussian. ICA provides the projection which gives us the maximum independency among the features or dimensions results in a better data representation which have a key role in the improvement of the classification accuracy. In this system ICA is used as an unsupervised component. The FastICA algorithm involves two sequential processes, the one unit estimation and decorrealation among the weight vectors. The one unit estimates the weight vectors as follows,

$$x_i^+ = E\{zg(x_i^T, z) - E\{g'(x_i^T, z)\} \tag{9}$$

Where x_i^+ is the temporal approximation of the independent component with $j = \{l+1.....m\}$. g is the derivative of the non-quadratic function introduced in and $g(u) = \tanh(au)$,g' is the derivative of g, $g'(u) = \text{sech}^2(u).$

The purpose of the decorrelation process is to keep different weight vectors from converging to the same maximum. The deflation scheme based on symmetric decor relation helps remove dependency among x_i^+ 's as follows

$$X_{1+l,m} = X_{1+l,m}^+ \left[(X_{1+l,m}^+, X_{1+l,m}^+)^{-1/2} \ \right]^T \tag{10}$$

Where $X_{l+1,c}$ represents decorrelated mappings based on $x^{+}_{1+l,m} = [x^{+}_{1+l}, x^{+}_{m}]$ from independence maximization.

The third part of proposed hybrid system is nothing but the classification. For the classification purpose c nearest neighbor algorithm is used. The working principal of this algorithm is it stores the all available cases and creates new cases based on the distance function. This algorithm is also used in many areas of pattern classification. In the proposed system it plays very important role to avoid the dimensionality disaster.

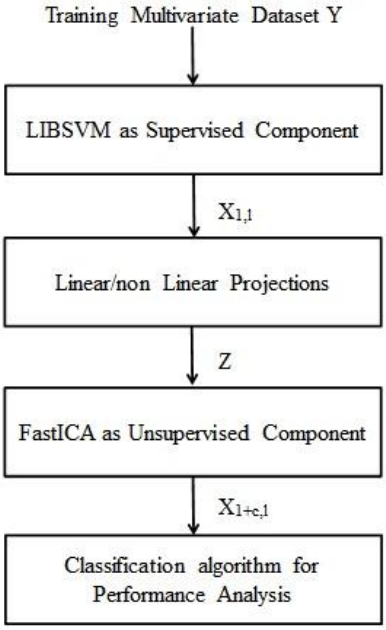

Figure. 1. Block diagram for the proposed hybrid method

Fig. 1 gives the complete information about the proposed hybrid dimensionality reduction method. LIBSVM software package is used for the implementation of SVM which is our first supervised component, then the second one is the projection both linear and nonlinear projections are used caused of linear and nonlinear SVM, third one is unsupervised component which is implemented by using FastICA algorithm, and the last one is one of the classification algorithm is used for the classification purpose which is a performance metric. Fig. 1 shows that multivariate data set named Y is given as input to the LIBSVM which has the dimension of n, after working on that it is reduced to the m dimension in a such way that $m << n$. Then by the use of projection matrix there is computation of a column vector c from the hybrid process SVM+ICA .The supervised component will generate the $X_{1,l}$ vector after that $X_{1,l+c}$ vector will be generated by the unsupervised component then the final weight vector will be given to the classification algorithm for the calculation of the classification accuracy.

IV. Experimental Results And Analysis

Natural dataset named cardiac arrhythmia which a multiclass in the nature is used for the dimensionality reduction purpose. Basically this hybrid method is analyzed over the accuracy of classification of arrhythmia dataset which contain list of 452 patients or samples which are to be classify among the 16 types of cancer due to some ambiguity in the dataset such as inconsistency in the samples, missing some samples or elements only 13 type of cancer are classified, for the missing elements some random variables are used. This dataset provides some different characteristics for samples per class and dimensionality. The number of feature or dimensions is reduced up to 95% due to the use of proposed hybrid method. The c nearest neighbor algorithm is used for the classification accuracy which is our performance metric. Some other metrics' are also used for this which are sensitivity and specificity. Above all the parameters are calculated by the use of confusion matrix. The experimental result shows that the proposed hybrid method performs extremely well with the classification accuracy up to which outperforms the other methods such as LDA PCA ICA.

Table 1: comparison with the standalone approaches

Approach	Classification accuracy (%)	No of projections	Sensitivity (%)	Specificity (%)
LDA	52.16	38	75.13	65.87
PCA	59.80	8	73.16	62.89
ICA	59.70	18	92.15	65.89
SVM+ICA	65.30	12+15	93.00	75

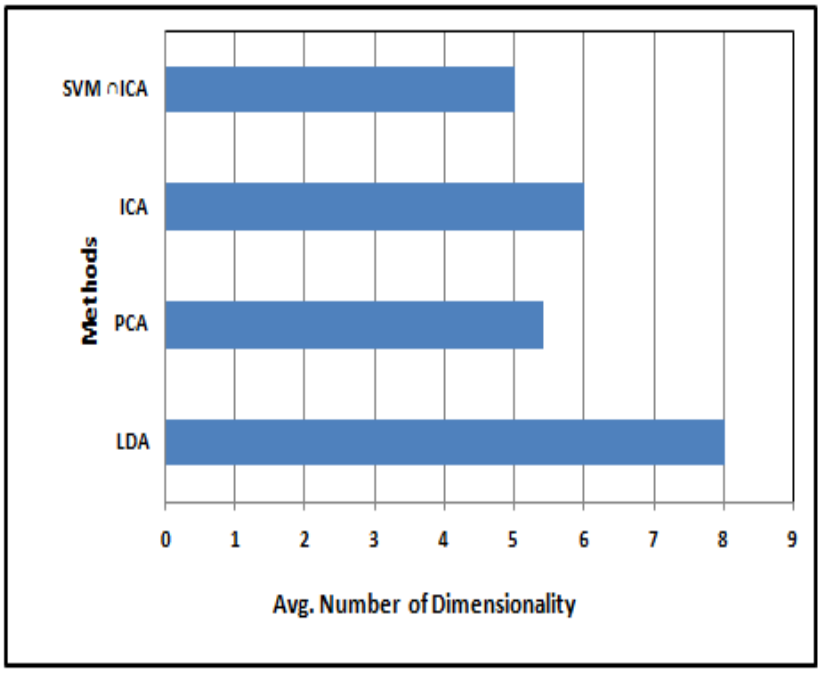

Figure 2: comparison with respect to reduced dimensionality

The Fig. 2 represents the comparison of other methods with respect to SVM and ICA. As shown in the chart the SVM and ICA work very well for dimensionality reduction.

V. Future Scope

As talking about the future scope, there are some important issues which are as mentioned, the first one is the constrained optimization techniques for SVM the second on is the construction of the subspace and the last one is the integration of ICA into the single formula for enhancement of the working speed. In another way single formulation does not gives the clear understanding about the dimensionality reduction but it is capable of avoiding the possible error came during the stepwise evolution of the process. Another major issue will be dealing with the non-linearity. There is a clear solution for this is use kernel because it is very user-friendly method but again there is problem that which type of kernel is to use for dealing with nonlinearity for transformation of low dimensional data in high dimensional space it is very hard to determine which kernel has to be trusted for the free dimensions.

VI. Conclusion

New hybrid method SVM and ICA which performs effectively for high dimensional data analysis, as it uses both supervised and unsupervised learning method and gives better classification results compare to the traditional method which are LDA, PCA and ICA. The traditional methods uses only one criterion either supervised or unsupervised. The proposed algorithm provides projections, such that SVM which is used as supervised component for minimization of the structural risk and ICA which is used as unsupervised component for maximization of independency among the features .The combination of both approaches gives the advantage of both methods, so it performs with the high degree of accuracy. This approach is used for the classification of

the multivariate data analysis .The experimental results and analysis shows dimensionality is reduced, for which "one against all" strategy is used (One against one for two class data set).

References

[1] Sangwoo Moon and Hairong Qi ,"Hybrid Dimensionality Reduction Method Based on Support Vector Machines and Independent Component Analysis", IEEE Transactions on Neural networks and Learning Systems,vol. 23, no. 5, may 2012.

[2] A.M.Martinez and A.C.Kak, "PCA versus LDA,"IEEE Trans.Pattern Anal.Mach.Intell. vol.23, no. 2, pp. 228-233, Feb. 2001.

[3] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):121–167, 1998.

[4] L. Cao, K. Chua, W. Chong, H. Lee, and Q. Gu. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. Nerocomp, 55:321–336, 2003.

[5] Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis and Its Application John Wiley & Sons,Inc. ,2001

[6] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.J. Yu, "Anew LDA-based face recognition system which can solve the small sample size problem," Pattern Recognit., vol. 33, no. 10, pp.1713–1726, 2000.

[7] H. Park, M. Jeon, and J. B. Rosen, "Lower dimensional representation of text data based on centroids and least squares," BIT Numerical Math vol. 43, no. 2, pp. 427–448, 2003.

[8] B. Scholkopf, A. Smola, and K. Muller, "Nonlinear component analysis as a Kernel Eigenvalue problem," Neural Comput., vol. 10, no. 5, pp, 1299–1319, 1998.

[9] H. Wold, "Estimation of principal components and related models by iterative least squares," in Multivariate Analysis. New York: Academic, pp. 391–420, 1966.

[10] Jean-Franc¸ois Cardoso,"Infomax and maximum likelihood for source separation", IEEE,Letters On Signal Processing Vol. 4, No. 4, pp. 112-114, 1997.

[11] Hurri, Gavert, Sarela, and Hyvarinen, A.,\The FastICA Package for MATLAB,"http://isp.imm.dtu.dk/toolbox/,1998.

[12] X.Jiang, "Asymmetric Principal component and discriminant analyses for pattern classification," IEEE Trans.Pattern Anal. Mach. Intell., vol. 31, no. 5, pp.931-937, May 2009.

[13] K. Kwak and W. Pedrycz, "Face recognition using an enhanced independent component analysis approach," IEEE Trans. Neural Network, vol. 18, no. 2, pp. 530–541, Mar. 2007.

[14] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," IEEE Trans. Neural Netw. vol. 17, no. 3, May 2006.

[15] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.