ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st&22ndMarch, Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Fisher Score Dimensionality Reduction for Svm Classification

Arunasakthi. K , KamatchiPriya.L, Askerunisa.A

PG Scholar, Dept of Computer Science and Engineering, Vickram College of Engineering, Enathi, Tamil Nadu, India.

Assistant Professor, Dept of Computer Science and Engineering, Vickram College of Engineering, Enathi, Tamil Nadu, India.

Head, Dept of Computer Science and Engineering, Vickram College of Engineering, Enathi, Tamil Nadu, India.

Abstract - The Support Vector Machine is a discriminative classifier which has achieved impressive results in several tasks. Classification accuracy is one of the metric to evaluate the performance of the method. However, the SVM training and testing times increases with increasing the amounts of data in the dataset. One well known approach to reduce computational expenses of SVM is the dimensionality reduction. Most of the real time data are non- linear. In this paper, F- score analysis is used for performing dimensionality reduction for non – linear data efficiently. F- score analysis is done for datasets of insurance Bench Mark Dataset, Spam dataset, and cancer dataset. The classification Accuracy is evaluated by using confusion matrix. The result shows the improvement in the performance by increasing the accuracy of the classification.

Key Terms – Support Vector Machine, Dimensionality Reduction, F- score Analysis, Confusion Matrix.

I.INTRODUCTION

Now days, real world data such as electrocardiogram signals, speech signals, digital photographs has high dimensionality. In order to handle these high dimensional data in the analysis makes difficulty and complexity. To get the efficient access with these data, the high dimensional data should be transformed into meaningful representation of the low dimensional data.

A. Dimensionality Reduction

Dimensionality reduction is a process of extracting theessential information from the data. The high-dimensional data can be represented in a more condensed form with much lower ,Dimensionality to both improve classification accuracy and reduce computational complexity. Dimensionality reduction becomes a viable process to provide robust data representation in relatively low-dimensional space in many applications like electrocardiogram signal analysis and content based image retrieval. Dimensionality reduction is an important preprocessing step in many applications of data mining, machine learning, and pattern recognition, due to the so-called curse of dimensionality.In mathematical terms, the problem we investigate can be stated as follows: D-dimensional data $X = (x_1 \ldots . x_D)$ is transformed into d dimensional data $Y = (y_1 \ldots . y_d)$. Dimensionality reduction captures the related content from the original data, according to some criteria. Feature extraction reduces the number of variables so that it can reduce the complexity which can improve overall performance of the system.

Data reduction can be applied on various applications like classification, regression, etc. In this paper, data reduction is applied on the classification problem and Support Vector Machine is used as the classifier. Accuracy is taken as a metric to evaluate the performance of the Support Vector Machine.

B. Dimensionality Reduction Techniques

Dimensionality reduction reduces the number of variables to improve the performance of the classification. High dimensional data is the major problem in many

applications which increase the complexity by taking the more execution time.

There are number of techniques available for reducing the dimensionality of the data. Each and every technique reduces the dimensions of the data based on particular criteria. In recent years, Principal Component Analysis (PCA) , Linear Discriminant Analysis (LDA), and Independent Component Analysis(ICA) are regarded as the most fundamental and powerful tools of dimensionality reduction for extracting effective features from high-dimensional vectors of input data.

In this paper, the feature selection is done by F-score Analysis. F-score analysis is a simple and effective technique, which produce the new low dimensional subset of features by measuring the discrimination of two sets of real numbers. Minimizing the distance between the same classes and maximizing the difference between the different classes makes this feature selection effectively. Though many techniques available for classification problem most of the methods support only for linear data. But in the case of Support Vector Machine classifier, it can handle both linear and Non - linear data. The experiments give better performance with low dimensional data rather than the high dimensional data.

C. Objective

The main objective of this paper is to transform the high dimensional data into low dimensional data by reducing the number of variables on the dataset. In this paper, Dimensionality reduction improves the performance of the classification problem with the F-score analysis. Classification is the process of analysing the data that which belongs to which one of the class. There are number of techniques for the classification. Among these techniques, Support Vector machine handles both the linear and non-linear data. On the other side, F-score is the simple and effective technique to select the meaningful information from the high dimensional data.

Dimensionality reduction reduces the dimension of the original data that will automatically increase the performance of the classifier by decreasing the execution time & space complexity. This paper mainly focuses on to improve the accuracy of the classifier by reducing the dimension of the original data.

II.RELATED WORK

In this section, the various techniques which are already used in several applications are discussed. Linear Discriminant Analysis is one of the techniques which reduce the data by finding the linear discriminants. Zizhu [1] uses Linear Discriminant Analysis (LDA) to reduce the dimensions on linear data. It is found that, the major problems of LDA are Small Sample Size (SSS) Problem, Singularity and Common Mean (CM) Problem. LDA is extended Joint Global and Local Linear Discriminant analysis (JGLDA) [2] to represent both local and global structure of the data. It is found that, the major problems of LDA are singularity problem and Small Sample Size (SSS) Problem. LDA/QR composition method solves the problem of singularity [3]. Jing Peng [4] finds the linear discriminants using regularized least squares and Yuxi Hou [5] used null based LDA (NLDA) to solve the Small Sample Size problem. Fisher Linear Discriminants (FLD) [6] and Generalized Discriminant Analysis (GDA)[7] are some other techniques to handle linear data.

Principal Component Analysis (PCA) is an unsupervised technique projects the uncorrelated data. The major problem of PCA is sensitive to outliers. Two dimensional PCA (2D PCA), Robust Principal Component Analysis (RPCA) are used to overcome the problem of outliers [8][9]. PCA based on L1-norm is less sensitive to outliers rather than the PCA based on L2-norm [10].

F-score analysis is a simple and effective technique to select the most relevant feature from the dataset. It finds the subset by analysing all the features and maximizing the distance between the different classes and minimizing the distance within classes. It can be used to handle the non-linear data and removes the irrelevant and redundant data from the high dimensional space and gives the relevant data in the form of low dimensional data[11]-[14].

Support Vector Machine (SVM) is an effective classifier, which is used to handle linear and non – linear data. By comparing with other techniques, SVM works very well in the presence of few data samples and exploits a margin-based geometrical approach rather than the statistical methods [15]-[19]. Though it works well, it is not suitable for the high dimensional data. The Performance of the SVM is degraded when the dimensions of the data is increased. The effetiveness of the feature reduction is shown on speaker verification and the accuracy is improved with the low dimensional data [20].

In real world, most of the data is in the form of non-linear and high dimensionality. Taking all these data for the analysis cause to increase the complexity and it consumes more time for execution. To reduce the complexity of the system the dimensions of the data should be reduced into low dimesional data. In this paper, F-score analysis is chosed as a technique to reduce the dimensions of the data, which can handle the non-linear data. To show the effectiveness of the dimensionality reduction, it is applied on the Support Vector Machine Classifier.

III.METHODOLOGY

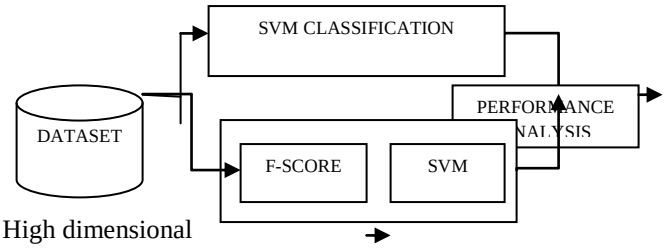

Fig 1 Dimensionality Reduction using F-score Analysis

Fig 1 shows the architecture for dimensionality Reduction using F-Score Analysis on Support Vector Machine classification.

A. Dataset

The initial process of this paper is, collecting the dataset which has high dimensionality. In this paper, the data is downloaded from the UCI Repository. The high dimensional data is directly processed with the SVM classification.

B. SVM Classification

Support Vector Machine is an effective supervised classifier which classifies the data into two or more classes based on the hyper plane. Generally, SVM is used for two class classification and its class may be 0 or 1 otherwise -1 or 1. Let us consider $X=(x_1 \ldots x_D)$ be a high dimensional data and each x_i has its own class labels Y= [-1, 1].

In the case of linear data, SVM tries to find the hyper plane with minimum distance from the data points from the boundary. If the data is non-linearly distributed, the data is transformed by using non-linear transformation functions. The training set and the corresponding output is defined as,

$$T= \{(x_1,y_1), (x_2,y_2) ,\ldots,(x_n,y_n)\} \quad x_i \epsilon R^n$$

Where, $y_i \epsilon \{-1, +1\}$denotes the corresponding output.

The optimal hyper plane is identified by Eq. (1),

$$Y = w^T + b = 0 \qquad (1)$$

Where, $w \epsilon R^n$ and $b \epsilon R$. The emprical risk is measured with the soft margin loss function by introducing the regularization terms and the slack variables $\Psi = (\Psi_1 \ldots \Psi_n)$. The soft margin function is expressed in Eq. (2).

$$\sum_{i=1}^{n} \max\big(0, 1 - y_i(w^t x_i + b)\big) \qquad (2)$$

The Support Vector Machine Problem is defined using the regularization term is expressed in (3) and (4) and (5) represents the supporting hyper planes which are parallel to the decision plane.

$$\min \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \Psi i$$

$$\text{s.t } y_i(w^t x_i) + b \geq 1 - \Psi_i, \ \Psi_i \geq 0, \quad i - 1, .., n \qquad (3)$$

$$w^T + b = 1 \qquad (4)$$

$$w^T + b = 1 \qquad (5)$$

Where, C>0 is the constant parameter. Minimization of the regularization term $\frac{1}{2}||w||^2$ maximizes the margin between the parallel hyper planes.

Confusion Matrix is one of the methods to measure the accuracy of the classification.

TABLE 1
CONFUSION MATRIX

		Predicted	
		Negative	Positive
Actual	Negative	A	B
	Positive	C	D

Table 1 shows the confusion matrix which is used to measure the accuracy of the classification. From the above table, the Accuracy is calculated using (6),

$$Accuracy = \frac{A+D}{A+B+C+D} \qquad (6)$$

Where, A and D be the number of correct predictions that are negative and positive respectively and B and C denotes the number of false predictions.

The analyses with these high dimensional data is not produce good performance and increase the complexity of the system. To reduce complexity the relevant features are selected which leads to improve the overall performance of the system.

C .F-score Analysis

F-Score Analysis is a simple and effective technique for feature selection which makes selection by measuring the discrimination of two sets of real numbers. It gets the high dimensional data as input then finds the subset by spanning all the data point and maximizes the distance between classes and minimizes the distance within classes. F-score value for each i^{th} attribute is defined in (4),

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{j=1}^{n_+}(x_{j,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{j=1}^{n_-}(x_{j,i}^{(-)} - \bar{x}_i^{(-)})^2} \qquad (4)$$

Where,$\bar{x}_i$, $\bar{x}_i^{(+)}$,$\bar{x}_i^{(-)}$ are the average of the i^{th} feature, positive instances and negative instances respectively. $X_{j,i}^{(+)}$ is the i^{th} feature of the j^{th} positive instance and i^{th} feature of the negative instance is represented as $x_{j,i}^{(-)}$. The numerator indicates the discrimination between the positive and negative sets of instances and the denominator indicates the one within each of sets.

Algorithm

1) Import the high dimensional data as an input.
2) X= Input data;
3) Calculate the f-score value for each attributes in X.
4) Do the following steps for minimum 3 times.
5) Choose threshold value among the f-score value in X.
6) For each threshold,

a) Select features which are below the threshold.
b) Split the data into train data and valid data
c) X = train data; Go to step 5;

7) Choose the threshold with lowest average validation error.
8) Drop features whose f-score values are below the threshold.

The data with low dimensions are again processed with the Support Vector Machine. SVM works on the new data and the performance of the classification is evaluated by measuring the accuracy.

IV.EXPERIMENTAL RESULTS

In this section, the performance of the SVM with high dimensional data and the low dimensional data is evaluated. The result shows the better performance with the low dimensional data which are the more relevant for the analysis.

In this paper, we utilize three datasets, 'Insurance Bench Mark', 'Spam Base' and 'Lung-Cancer dataset' from the UCI repository. Result on these data shows the effectiveness of the proposed feature selection technique in terms of accuracy.

In Insurance Bench Mark dataset, there are 5822 instances and 86 attributes to analyse whether the person is eligible to get insurance. In each record of the dataset, 85 variables represent the personal details of each person. 86th attributes represents the class label. In Spam Base dataset, 4600 records with 58 attributes to analyse whether the mail is spam. 32 instances and 57 attributes are presented in the Lung – cancer dataset. This paper is done on Matlab environment.

A. Result of SVM

These high dimensional data is processed on the Support Vector Machine Classification. Accuracy is taken as a, metric to evaluate the performance of the SVM classification. SVM with original data produce the accuracy as 18.2755, 35.5217 and 46.1538 for Insurance Bench Mark, Spam Base and Lung-cancer datasets respectively. The results are shown in table 2.

TABLE 2
ACCURACY OF SVM CLASSIFICATION

DATASETS	No. of. attributes	ACCURACY OF SVM
Insurance Bench Mark	86	18.2755
Spam Base	58	35.5217
Lung-cancer	57	46.1538

B. Result of F-score Analysis

The high dimensional data is processed with the F-score Analysis. The f-score value for each attribute is measured and new dataset is selected based on this F-score values. The new subset represents the more relevant information which has more influence on the analysis.

TABLE 3
REDUCTION PERCENTAGE

Dataset	No. Of attributes	Reduced attributes	Reduction percentage
Insurance bench mark	86	31	36.47
Spam	58	30	52.63
Cancer	57	38	66.66

Table 3 shows the total number of attributes in the original data set and the number of features in reduced dataset. After performing F-score analysis, 86 is reduced into 31, 58 is reduced into 30 and 57 is reduced into 38 for Insurance, Spam and Cancer datasets respectively.

C. SVM with F-score Values

The low dimensional data which is selected from the F-score analysis is processed on the SVM classification and the accuracy is measured with the low dimensional data. The accuracy with the low dimensional data is obtained as 37.2037, 67.3478 and 46.1538 for Insurance Benchmark, Spam and Cancer datasets respectively. In table 3, HDD represents the number of variable in High dimensional space and LDD represents the number of variables which are selected from the F-score analysis.

TABLE 4
ACCURACY OF SVM WITH LOW DIMENSIONAL DATA

DATASETS	HDD	LDD	ACCURACY OF SVM
Insurance Bench Mark	86	31	37.2037
Spam Base	58	30	67.3478
Lung-cancer	57	38	46.1538

From the table 4, it comes to know the accuracy of the SVM is improved with the low dimensional data rather than the high dimensional original data.

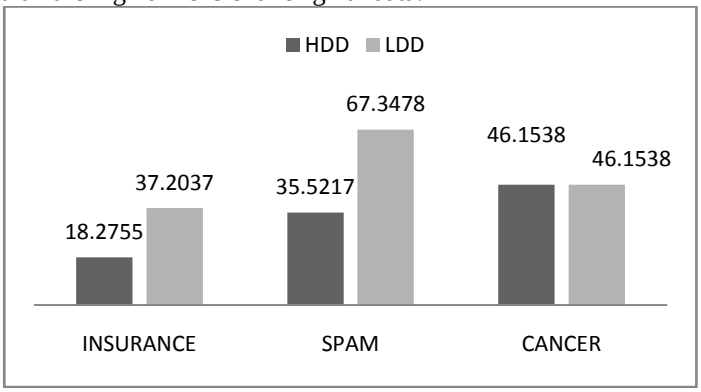

Fig 2. Accuracy with High Dimensional data and Low dimensional data

Fig 3 shows the comparison between the accuracy of SVM with high dimensional data and low dimensional data. Accuracy with the low dimensional data is more efficient than the high dimensional data.

V.CONCLUSION

In this paper, F-score Analysis is used as a feature selection technique to reduce the dimensions of the data which was validated on SVM classifier. The F-score feature selection works well by selecting the subset from the data based on the threshold value thereby eliminating the unwanted data. Though it improves the performance, there exists a problem that which is not suitable for redundant data.

This work can be continued by implementing hybrid techniques (F-score with machine learning techniques like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), etc). Here we implement the analysis on classification; it can also be applied on regression problems.

REFERENCES.

[1] Zizhu Fan., "Local Linear Discriminant Analysis Framework Using Sample Neighbors", IEEE Transactions on Neural Networks, , On page(s): 1119 - 1132 Volume: 22, July 2011

[2] Quanxue Gao., "Joint Global and Local Structure Discriminant Analysis", IEEE Transactions on Information Forensics and Security, April 2013

[3] Jieping Ye; Qi Li, "A two-stage linear discriminant analysis via QR-decomposition", IEEE Transactions on Pattern Analysis and Machine Intelligence Volume: 27 , Issue: 6 Publication Year: 2009.

[4] Jing Peng,; Riedel, N., "Discriminant Learning Analysis" , IEEE Transactions on Systems Man and Cybernetics, Volume.38, 2011.

[5] Yuxi Hou; Iickho Song; Hwang-Ki Min, " Complexity-Reduced Scheme for Feature Extraction With Linear Discriminant Analysis " , IEEE Transactions on Neural Networks and Learning Systems, June 2012

[6] Bin Zou; Luoqing Li, " Generalization Performance of Fisher Linear Discriminant Based on Markov Sampling " , IEEE Transactions on Neural Networks and Learning Systems, February : 2013

[7] Y. Zhang and D. Y. Yeung, "Semisupervised generalized discriminant analysis " , IEEE Transaction on Neural Networks, volume. 22, pages.1207 -1217 publication year : 2011

[8] Xu Chunming, "Robust two-dimensional principle component analysis ", IEEE transaction on signals and control, Year: 2010 , Page(s): 452 - 455

[9] Ran He; Bao-Gang Hu,"Robust Principal Component Analysis Based on Maximum Correntropy Criterion", IEEE Transactions on Image Processing,On page(s): 1485 - 1494 Volume: 20, Issue: 6, June 2011

[10] N. Kwak, "Principal component analysis based on L1-norm maximization", IEEE Transaction on Pattern Analysis and Machine Intelligence, volume. 30, no. 9, pages.1672 -1680 Year : 2008.

[11] Peng Toa, Huang Yi.et.al, "A method based on weighted F-score and SVM for feature selection", IEEE Transaction on Pattern Analysis and Machince Intelligence. Volume. 33, May 2013.

[12] Quanquan GuZhenhui Li,"Generalized Fisher Score for Feature Selection", IEEE transcation on Machine Learning 2008

[13] Ding,"Feature Selection Based F-Score and ACO Algorithm in Support Vector Machine", IEEE symposium on Knowledge Acquisition 2009.

[14] Kemal Polat, Salih Güneş " new feature selection method on classification of medical datasets: Kernel F-score feature selection" from Science Direct

[15] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[16] C. J. C. Burges, "A tutorial on support vectormachines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2,pp. 121–167, Jun. 1998.

[17] Set of tutorials on SVM's and kernel methods. [Online]. Available: http://www.kernel-machines.org/tutorial.html

[18] I. El-Naqa, Y. Yang, P. Galatsanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Trans. Med. Imag.*, vol. 21, no. 12, pp. 1552– 1563, Dec. 2002.

[19] J. Robinson and V. Kecman, "Combining support vector machine learning with the discrete cosine transform in image compression," *IEEE Trans. Neural Netw.*, vol. 14, no. 4, pp. 950–958, Jul. 2003.

[20] Thembisile Mazibuko, Daniel J. Mashao, " Feature Extraction and Dimensionality Reduction in SVM Speaker Recognition", IEEE Transcation on Machine Learning, 2008