International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 22 Issue 1 – MAY 2016.

DIMENSIONALITY REDUCTION USING F-SCORE ANALYSIS BASED ON SUPPORT VECTOR MACHINE

Ms. V. Sivapriya[1] Ms. R. K. Deebika

PG Student, Computer Science and Engineering, Vivekanandha College of Technology for Women, Tiruchengode
M.E, Assistant Professor, Information Technology, Vivekanandha College of Technology for Women, Tiruchengode

vsp.priyasiva53@gmail.com
deebikark@gmail.com

***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 a day, real world data such as electrocardiogram signals, speech signals, and 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. Dimensionality reduction is a process of extracting the essential information from the data.

A. Dimensionality Reduction

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 = (x1 . . . . xD) is transformed into d dimensional data Y = (y1..... yd). 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. The 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. This paper mainly focuses on to improve the accuracy of the classifier by reducing the dimension of the original data.

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.

The feature selection is done by F-score Analysis on images. 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 analyzing 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.

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. Xiaokang [1] uses gradient profile sharpness is an edge sharpness metric, which is extracted from two gradient description modelsare proposed for two profile description models, which can keep profile shape and profile gradient magnitude sum consistent during profile transformation[2] Real-time Artifact-free Image Upscaling method able to obtain artifact-free enlarged images preserving relevant image features and natural texture. Gao [3] and Mallet [4] both are represents neighbor-embedding-based (NE) algorithms for super-resolution (SR) have carried out independent processes to synthesize high-resolution (HR) image patches[5] used Context-Aware Sparse Decomposition for Image Denoising and Super-Resolution.

Yang [6] Sapiro [7] the local patches are reconstructed by geometric and provide general and computational efficient solutions for inverse problems for piecewise linear. Yang[8] to represent Landmark images by retrieving correlated web images.It retrieves more accurate high frequency details from the registered images.

Linear Discriminant Analysis is one of the techniques which reduce the data by finding the linear discriminants. Zizhu [10] uses Linear Discriminant Analysis (LDA) to reduce the dimensions on linear data. Principal Component Analysis (PCA) is an unsupervised technique projects the uncorrelated data. The major problem of PCA is sensitive to outliers. Robust Principal Component Analysis (RPCA) is used to overcome the problem of outliers.It finds the subset by analyzing 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. The limitations of F-score analysis is, it not suitable for redundant data.

Support Vector Machine (SVM) is an effective classifier, which is used to handle linear and non –linear data. To improve the accuracy of the classification, Recursive Support Vector Machine (RSVM) is used to extracting the support vectors.

In real world, most of the data are in the form of non-linear and high dimensionality. Handling all these data for the analysis makes complexity. To reduce the complexity the dimensions of the data should be reduced.

III. METHODOLOGY

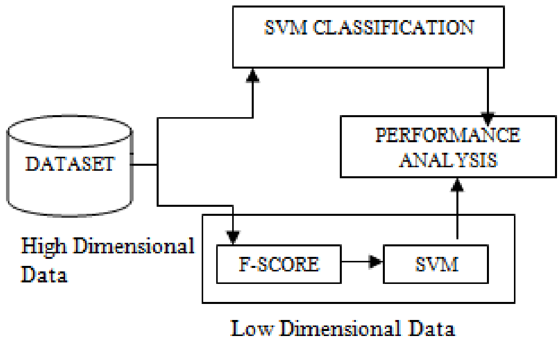

Fig. 1 Dimensionality Reduction using F-score Analysis

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=(x1…xD) be a high dimensional data and each xi 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,

$$T= \{(x1, y1), (x2, y2), (xn, yn)\}\ xi \epsilon Rn$$

Where, yi ϵ $\{-1, +1\}$ denotes the corresponding output. The optimal hyper plane is identified by Eq. (1),

$$Y = wT + b = 0 \qquad (1)$$

Where, w ϵ Rn and b ϵ R. The empirical risk is measured with the soft margin loss function by introducing the regularization terms and the slack variables Ψ = (Ψ1…Ψn). The soft margin function is expressed in Eq. (2).

$$\sum_{I=1}^{n} \max(0, 1 - \mathrm{yi}\,(\mathrm{w}^t \mathrm{xi} + \mathrm{b})) \qquad (2)$$

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.

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

$$\text{s.t } (wtx) + b \geq 1 - \Psi i,\ \Psi i \geq 0,\ i - 1, n \qquad (3)$$

$$\mathrm{wT} + b = 1 \ (4)$$

$$\mathrm{wT} + b = 1 \ (5)$$

Where, C>0 is the constant parameter. Minimization of the regularization term $\frac{1}{2}||\mathrm{w}||^2$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 used to measure the accuracy of the classification. 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)

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

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. Result on these data shows the effectiveness of the proposed feature selection technique in terms of accuracy.

Histogram equalization (HE) method is increase the contrast of input image having a low dynamic range as well as preserves the mean brightness of the image. The performance of the thresholding algorithm by preserving the edges as well as removing the noise. The Histogram Equalization enhance the hidden image detail and to increase the contrast of input image.

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.

B.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.

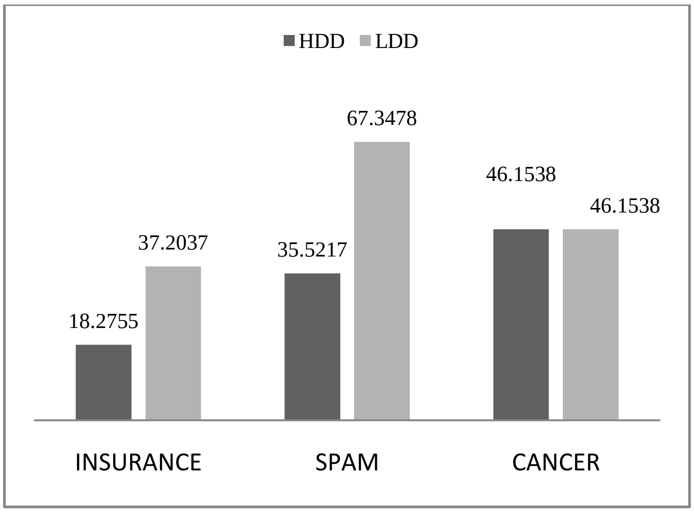

Fig 2. Accuracy with High Dimensional data and Low dimensional data

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). Then implement the analysis on classification; it can also be applied on regression problems.

VI. FUTURE WORK

In future, the images are tested with Added Color-Cast Enhanced with Spray-Based Methods. This work builds and expands on a previous article Random sprays are a two-dimensional collection of points with a given spatial distribution around the origin. So the noise is removed by traditional methods. Wavelet Shrinkage is a method of removing noise from images by shrinking the empirical wavelet coefficients in the wavelet domain and it is a nonlinear image denoising procedure to remove the noise.

REFERENCES

[1] Qing Yan, Yi Xu, Xiaokang Yang and Truong Q. Nguyen, (2015) “Single Image Superresolution Based on Gradient Profile Sharpness,” IEEE Trans. Image Process., Vol. 24, no. 10.

[2] Giachetti. A and Asuni. N, (2011) “Real-time artifact-free image upscaling,” IEEE Trans. Image Process., vol. 20, no. 10, pp. 2760–2768.

[3] Gao. X, Zhang. K, Tao. D, and Li. X, (2012) “Image super-resolution with sparse neighbor embedding,” IEEE Trans. Image Process., vol. 21, no. 7, pp. 3194–3205.

[4] Mallat. S and Yu. G,(2010) "Super-resolution with sparse mixing estimators," IEEE Trans. Image Process., vol. 19, no. 11, pp. 2889–2900.

[5] Ren. J, Liu. J, and Guo.Z , (2013) "Context-aware sparse decomposition for image denoising and super-resolution," IEEE Trans. Image Process., vol. 22, no. 4, pp. 1456–1469.

[6] Yang. S, Wang. M, and Y. Sun. Y, (2012) "Single-image super-resolution reconstruction via learned geometric dictionaries and clustered sparse coding," IEEE Trans. Image Process., vol. 21, no. 9, pp. 4016–4028.

[7] Yu. G, Sapiro. G, and Mallat. S, (2012) "Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity," IEEE Trans. Image Process., vol. 21, no. 5, pp. 2481–2499.

[8] Yue. H, Sun. X, Yang. J, and Wu. F, (2013) "Landmark image super-resolution by retrieving web images," IEEE Trans. Image Process., vol. 22, no. 12, pp. 4865–4878.

[9] Zhang. T, and Fan. Q, (2013) "An adaptive wavelet image denoising scheme based on local variance"IEEE Trans. Image Process.,vol. 21, no.3.

[10] Bin Zou; Luoqing Li; Zongben Xu; Tao Luo; Yuan Yan Tang , " Generalization Performance of Fisher Linear Discriminant Based on Markov Sampling " , IEEE Transactions on Neural Networks and Learning Systems, page(s): 288 - 300 Volume: 24, February : 2013