MFSPFA: An Enhanced Filter based Feature Selection Algorithm

V. Arul Kumar
Research Scholar
Dept. of Computer Science
St. Joseph's College (Autonomous)
Trichy, TN, India

L. Arockiam
Associate Professor
Dept. of Computer Science
St. Joseph's College (Autonomous)
Trichy, TN, India

ABSTRACT

Feature Selection is the process of selecting the momentous feature subset from the original ones. This technique is frequently used as a preprocessing technique in data mining. In this study, a new feature selection algorithm is proposed and is called Modified Fisher Score Principal Feature Analysis (MFSPFA). The new algorithm is developed by combining the proposed Modified Fisher Score (MFS) and Principal Feature Analysis (PFA). The proposed algorithm is tested on publicly available datasets. The experimental results show that, the proposed algorithm is able to reduce the futile features and improves the classification accuracy.

General Terms

Data Mining, Classification, Filter Approach, Feature Selection Algorithms.

Keywords

Feature Selection, Modified Fisher Score, Principal Component Analysis, Principal Feature Analysis

1. INTRODUCTION

Feature Selection is the process of selecting the subset of relevant features by removing redundant, irrelevant and noisy data from the original dataset. Feature selection methods fall into two categories: Filter approach and Wrapper approach. In filter approach, the features were selected based on criteria which are independent of the particular learning algorithm to be applied to the data. In wrapper approach, the feature selection is based on a wrapper, which is a subset of attributes and are evaluated with a learning algorithm [1].

Feature selection algorithms are categorized into Supervised Algorithms [2, 3], Unsupervised Algorithms [4, 5] and Semi-supervised Algorithms [6, 7]. In Supervised Learning, all instances are associated with the class labels. In Unsupervised Learning, no class labels are available for the instances. In Semi-supervised Learning, few instances have class labels and the remaining instances do not have the class labels [8]. The selection criterion is a key component in feature selection to select the best features. In earlier period, various selection criteria have been proposed for the filter based feature selection. They are Mutual Information [9], ReliefF [10], Laplacian Score [11], Fisher Score [12], SPEC [13], Hilbert Schmidt Independence Criterion (HSIC) [14], and Trace Ratio [15].

In feature selection technique, the most relevant features are selected. The noisy and irrelevant features are removed in the supervised method. In unsupervised method, the redundant features are removed by finding the similarity or correlation measure between the features. In the recent study, the supervised and unsupervised methods are combined to find the best feature set from the original set [1]. The problem still persists in obtaining the best feature subset and the classification accuracy.

In this paper, a new feature selection algorithm is proposed by combining the MFS and PFA. MFS is a supervised method that removes the noisy and irrelevant features which have the less discriminant information. The PFA is an unsupervised method which selects the relevant features using the correlation or similarity measure and also removes the redundant features. The proposed algorithm is validated with the publicly available datasets. The result shows that the proposed algorithm can largely reduce the feature dimensions and also it improves the classification accuracy.

The rest of the paper is organized into different sections: Section 2 shows the existing feature selection criteria, Section 3 describes the Principal Component Analysis (PCA), Section 4 describes the Principle Feature Analysis (PFA), Section 5 deals with the proposed algorithm, Section 6 presents our experimental results, and Section 7 gives the concluding remarks.

2. EXISTING FEATURE SELECTION CRITERIA USED FOR FINDING FEATURE SUBSET

In this section, existing feature selection criteria are discussed. This includes various algorithm like ReliefF [10], Laplacian Score [11], Fisher Score [12], SPEC [13], Hilbert Schmidt Independence Criterion (HSIC) [14] and Trace Ratio [15].

2.1 Feature selection using Fisher Score

Fisher score is one of the simplest filter algorithms for feature selection [12]. In this criteria, the features are selected which have the similar values in the same class and the dissimilar values in different classes. The Fisher score is calculated using the formula

$$FS = \frac{\sum_{k=1}^{m} s_k \left(\mu_{i,k} - \mu_i\right)^2}{\sum_{k=1}^{m} s_k \sigma_{i,k}^2} \qquad (1)$$

where

μ_i is the mean of the features

s_k is the number of samples in the k^{th} class

$\mu_{i,k}$ is the mean of the features in the k^{th} class

$\sigma_{i,k}^2$ is the variance of the features in the *k*th class

Fisher trace criterion

The Fisher trace criterion [16] is calculated by the formula

$$FS = tr\left\{(\tilde{S}_b)(\tilde{S}_c + \gamma X)^{-1}\right\} \qquad (2)$$

where

$\tilde{S}_b$ is the scatter matrix between-class

$\tilde{S}_c$ is the total scatter matrix

γX is the positive regularized parameter

tr is the trace value of the given scatter matrix

$$\tilde{S}_b = \sum_{n=1}^{m} k_n\,(\tilde{\mu}_n - \tilde{\mu})(\tilde{\mu}_n - \tilde{\mu})^T \qquad (3)$$

$$\tilde{S}_c = \sum_{j=1}^{s}(x_j - \tilde{\mu})(x_j - \tilde{\mu})^T \qquad (4)$$

where

$\tilde{\mu}$ is the total mean of the reduced data

k_n is the size of the *n*th class reduced data

$\tilde{\mu}_n$ is the mean of the *n*th feature

x_j is the number of samples of the *j*th class in the reduced data

2.2 Feature selection using Laplacian Score

Laplacian Score [11, 17] is proposed to select features that have the strong locality preserving ability. The Laplacian for the nth feature is calculates as:

$$L = \frac{\tilde{f}_n^{\,T} L \tilde{f}_n}{\tilde{f}_n^{\,T} D \tilde{f}} \qquad (5)$$

where

$\tilde{f}_n = \tilde{f}_n - \frac{\tilde{f}^T D1}{1^T D1} 1$

L is the Laplacian matrix

D is the degree diagonal matrix

2.3 Feature selection using Trace Ratio Criterion

The trace ratio criterion for feature selection is proposed by Feiping Nie et al. in [15]. Let us consider the data matrix $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$, based on the undirected graph, two weighted matrices are defined Z_p and Z_q. Z_p reflects the within-class or local affinity relationship of instances and Z_q reflects the between-class or global affinity relationship of instances,

$$L_p = D_p - Z_p \qquad (6)$$

$$L_q = D_q - Z_q \qquad (7)$$

where

L_p and L_q are the Laplacian matrices

D_p and D_q are the Diagonal matrices

Let $W \in \mathbb{R}^{d\times n}$, be a selection matrix. It denotes the column vector by $w_i \in \mathbb{R}^d$.

$W = [w_{s(1)}, w_{s(2)}, \ldots\ldots, w_{sn}]$, where s is a permutation of {1, 2… d}, the trace ratio criterion is calculated by

$$\mathcal{J}(W_s) = \frac{tr(W_s^T X Z_q X^T W_s)}{tr(W_s^T X Z_p X^T W_s)} \qquad (8)$$

The above equation is reduced to $U = XZqXT \in \mathbb{R}^{d\times d}$ and $V = XZpXT \in \mathbb{R}^{d\times d}$. Equation (8) can be rewritten as,

$$\mathcal{J}(W_s) = \frac{tr(W_s^T U W_s)}{tr(W_s^T V W_s)} \qquad (9)$$

The score of the features is calculated by

$$score(F_s) = \frac{W_s^T U W_s}{W_s^T V W_s} \qquad (10)$$

2.4 Feature selection using Hilbert-Schmidt Independence Criterion (HSIC)

HSIC is proposed by Gretton in [18] to measure the dependency between two kernals. This criterion is used to select the subset of features in which, the HSIC criterion get maximized from the obtained kernals. The criterion is calculated by

$$S(F) = \frac{1}{n(n-3)}\left[Trace(H_F H) + \frac{1^T H_F 1 1^T H 1}{(n-1)(n-2)} - \frac{2}{n-2} 1^T H_F H 1\right] \qquad (11)$$

where

F is a subset of original features

H_F is the kernel obtained from subset of original features

The Forward Selection or Backward Elimination is used to select the best feature set using the HSIC criterion.

2.5 Feature selection using ReliefF

ReliefF is proposed in [10] for the supervised filter approach algorithm using the feature weights. Let us consider the n instances that are randomly selected from the original features. The evaluation criterion ReliefF is calculated by

$$RF = \frac{1}{2}\sum_{m=1}^{n} d\left(f_{m,t} - fPQ(x_m), t\right) - d\left(f_{m,t} - fPR(x_m), t\right) \qquad (12)$$

where

f_m, t represents the value of instances x_m

$fPQ(x_m), t$ and $fPR(x_m), t$ represents value on the tth features of the nearest points to x_m

d represents the distance measurement

ReliefF criterion handles the multiclass problem by extending the above criterion and it is rewritten as,

$$RF = \frac{1}{n} \cdot \sum_{m=1}^{n} \left\{ -\frac{1}{y_{x_m}} \sum_{x_j \in PQ(x_m)} d(f_{m,t} - f_{j,t}) + \sum_{z \neq z_{x_m}} \frac{1}{y_{x_m,z}} \frac{P(z)}{1 - P(z_{x_m})} \sum_{x_j \in PR(x_m,z)} d(f_{m,t} - f_{j,t}) \right\} \quad (13)$$

where

z_{x_m} is the class label of the instances x_m

P(z) is the probability of an instance in class z

PQ(x) or PR(x,z) represents a set of nearest points to x

y_{x_m} and $y_{x_m,z}$are the sizes of the set PQ(x) and PR(x,z)

3. PRINCIPAL COMPONENT ANALYSIS (PCA)

The Principal Component Analysis [19] is a well established technique which is used to reduce the large dimension of dataset into small dimension of dataset that represents the most of the information in the original data matrix. To reduce the dimension, it forms n × m covariance matrix and it computes the Eigenvalues and Eigenvectors. Finally, it retains the *q* Eigenvector with the largest Eigenvalues. An *n* dimension vector *y* is projected to *q* dimensional subspace and the vector *p* is computed by

$$p = X^T(y - \mu) \quad (14)$$

The main disadvantage of PCA is that it does not explicitly specify which features are important.

4. PRINCIPAL FEATURE ANALYSIS (PFA)

The PFA [20] is dimensionality reduction technique which is derived from the most popular statistical technique called Principal Component Analysis (PCA). The PFA technique is used to select the features by dividing the features into different groups and it consequently selects a feature from each group.

Let *X* be an *n*-dimensional random feature vector and *D* is the covariance matrix of *X*. Consider *A* is an orthnormal matrix composed of the Eigenvectors of *D*.

$$\mathrm{D} = A\Lambda A^T \quad (15)$$

Let A_m be the first m column of matrix *A*. Let *Z1, Z2, Z3…Zn Є* R_m be the rows of A_m. To identify the finest subset, row vector Z_i is used to cluster the highly correlated features and select one feature from each cluster. The PFA algorithm can be summarized with the following five steps:

Step 1: Calculate the sample Covariance Matrix.

Step 2: Calculate the Eigenvalues and Principal Components of the Covariance Matrix.

Step 3: Matrix A_m is constructed from *A* by choosing the subspace dimension *m*. This can be chosen by deciding how much of the variability of the data is desired to be retained.

Step 4: Using the K-means algorithm the vectors $|Z_1|, |Z_2|, \ldots, |Z_3| \in \boldsymbol{R}^{mx}$ are divided into *p* Clusters. Euclidean distance is used as a distance measure in K-means algorithm.

Step 5: From each cluster, find the corresponding vector Z_i which is closest to the mean of the cluster. Choose the corresponding features, S_i, as a principal feature. This step will yield the choice of *p* features.

5. PROPOSED WORK

5.1 Modified Fisher Score (MFS)

In the existing feature selection algorithms, Fisher score is used as a criterion for selecting the features. In the proposed work the Fisher score is modified by calculating the Euclidean Norm for the given input matrix and this value is used as Fisher criterion to select the best features. The MFS is calculated by the given formula

$$M_{FS} = \|(Q_m)(Q_n + \delta Pr)^{-1}\| \quad (16)$$

$$Q_m = \sum_{i=1}^{n} s_i\,(\mu_i - \mu)(\mu_i - \mu)^T \quad (17)$$

$$Q_n = \sum_{j=1}^{m} (z_j - \mu)(z_j - \mu)^T \quad (18)$$

$$\mu = \sum_{i=1}^{n} s_i\,\mu_i \quad (19)$$

where

μ_i is the mean of *i*th class

s_i is the size of the *i*th class

μ is the total mean

Q_m is between-class scatter matrix

Q_n is total scatter matrix

δPr is a positive parameter

z_j is the number of samples of the *j*th class

5.2 MFSPFA Algorithm

The main goal of the feature selection algorithm is to find most relevant features by eliminating irrelevant, redundant and noisy features. To overcome these issues a new algorithm MFSPFA is proposed by merging supervised method (MFS) and unsupervised method (PFA). The proposed MFS is a supervised feature selection criterion. It is used to remove the noisy and irrelevant features but it is unable to remove the redundant features present in the data set. PFA removes the redundant features by exploring the correlation analysis between the features. In the proposed algorithm, MFS is used to choose the *m* features that have the most discriminant information. In MFS, the features are retained by gathering the MFS value in descending order until the percentage value does not exceed the users' threshold value of 0.95. Then PFA method is used to divide the selected *m* features into *n* clusters. Finally, the best feature subset is obtained by choosing one feature in each cluster. The MFSPFA algorithm is summarized in the following steps;

Step 1: Compute the MFS measure for each feature.

Step 2: Arrange the features in the descending order based on the MFS measure.

Step 3: Select the best *m* features till the MFS measure surpasses the users' threshold value.

Step 4: The selected *m* features are divided into *n* clusters using PFA method.

Step 5: Finally, select a feature from each cluster and form a best feature subset.

6. EXPERIMENTS

In this section, the experiment is carried out to prove the efficiency of the proposed feature selection algorithm. The experiment is carried out on a well known publicly available dataset for UCI Machine Learning Repository. The efficiency of the proposed algorithm is analyzed in terms of Classification accuracy. The K–Nearest Neighbor classifier is used to evaluate the classification accuracy of proposed algorithm. The proposed algorithm is compared with two existing feature selection algorithms; they are FS+PFA and LDDR.

In the FS+PFA and LDDR algorithms, trace values are calculated for the given input matrix and the values are used as a Fisher criterion for selecting the relevant features. In the proposed algorithm, instead of trace value, the Euclidean norm is calculated for the given matrix and this value is used as a Fisher criterion for selecting the most pertinent features.

Table 1. Classification accuracy of feature selection algorithms

Dataset	Features	Accuracy Percentage (%)		
		FS+PFA	LDDR	MFSPFA
Ionosphere	32	92.00	93.11	**95.28**
USPS	256	94.32	95.15	**96.11**
Isolet	617	88.13	90.22	**92.86**
Multi-features	649	83.58	85.13	**90.18**
Leukemia	12558	89.29	91.86	**93.47**

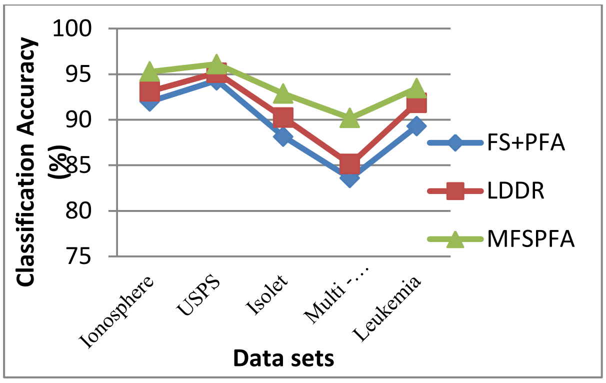

Fig 1: Comparative Analysis of feature selection algorithms

The experimental results of the proposed feature selection algorithm are depicted in table 1. Fig 1 shows the comparative analysis of the feature selection algorithms. From the results, the proposed algorithm shows an improved classification accuracy compared to the existing feature selection algorithms.

7. CONCLUSION

In this paper, a new feature selection algorithm is proposed by combining the MFS and PFA. The aim of the proposed algorithm is to obtain the best feature subset by eliminating the irrelevant, redundant and noisy features. The MFA criterion eliminates the irrelevant and noisy features by analyzing its discriminant information. Whereas PFA eliminates the redundant features by analyzing the correlation measure between the features. The proposed algorithm is evaluated using various datasets. The results show that the proposed algorithm performs better than the existing FS+PFA and LDDR algorithms. This work can be extended by combining the remaining supervised and unsupervised learning algorithms to improve the classification accuracy

8. REFERENCES

[1] Boyang Li,Qiangwei Wang,Jinglu Hua, Feature Subset Selection: A Correlation -Based SVM Filter Approach, IEEJ Transactions On Electrical And Electronic Engineering, Volume 6, 2011,pp. 173-179.

[2] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization, Journal of Machine Learning Research, Volume 13, 2012, pp. 1393-1434.

[3] J. Weston, A. Elisseff, B. Schoelkopf, and M. Tipping, Use of the zero norm with linear models and kernel methods, Journal of Machine Learning Research, Volume 3, 2003, pp. 1439–1461.

[4] Julia Handl, Joshua Knowles, Feature Subset Selection in Unsupervised Learning via Multiobjective Optimization, International Journal of Computational Intelligence Research, Volume 2, Number3, 2006, pp. 217–238, ISSN:0973-1873.

[5] Duy-Dinh Le, Shin'ichi Satoh1, An Efficient Feature Selection Method for Object Detection, Springer LNCS, Volume 3686, 2005, pp. 461–468.

[6] Gauthier Doquire, Michel Verleysen, Graph Laplacian for Semi-supervised Feature Selection in Regression

Problems, Springer LNCS, Volume 6691, 2011, pp. 248–255.

[7] Jidong Zhaoa,Ke Lua, Xiaofei He, Locality sensitive semi-supervised feature selection, Journal of Neurocomputing, Volume 71, 2008, pp. 1842-1849.

[8] Novi Quadrianto, Alex J. Smola, Tib´erio S. Caetano, Quoc V. Le, Estimating Labels from Label Proportions, Journal of Machine Learning Research, Volume 10, 2009, pp. 2349-2374.

[9] Li juan Wang, An improved multiple fuzzy NNC system based on mutual information and fuzzy integral, International Journal of Machine Learning and Cybernetics, Volume 2, Number 1, 2011, pp.25-36.

[10] Yuxuan SUN, Xiaojun LOU, Bisai BAO, A Novel Relief Feature Selection Algorithm Based on Mean-Variance Model, Journal of Information & Computational Science, Volume 8, Number 16, 2011, pp. 3921–3929.

[11] Xiaofei He, Ming Ji, Chiyuan Zhang, Hujun Bao, A Variance Minimization Criterion to Feature Selection Using Laplacian Regularization, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 33, Number 10, October 2011, pp. 2013-2025.

[12] Jian-Bo Yang, Kai-Quan Shen,Chong-Jin Ong,Xiao-Ping Li, Feature Selection for MLP Neural Network: The Use of Random Permutation of Probabilistic Outputs, IEEE Transactions on Neural Networks, Volume 20, Issue 12, December 2009,pp. 1911-1922.

[13] Zheng Zhao, Huan Liu, Spectral feature selection for supervised and unsupervised learning, Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 1151-1157, ISBN: 978-1-59593-793-3.

[14] P. Daniusis,P. Vaitkus, Supervised Feature Extraction Using Hilbert-Schmidt Norms, Springer LNCS, Volume 5788, 2009, pp.25-33.

[15] Feiping Nie, Shiming Xiang,Yangqing Jia, Changshui Zhang, Shuicheng Yan, Trace ratio criterion for feature selection, 23rd National Conference on Artificial Intelligence, Volume 2, 2008, pp. 671-676, ISBN: 978-1-57735-368-3

[16] P. E. H. R. O. Duda and D. G. Stork, Pattern Classification. Wiley-Interscience Publication, 2001.

[17] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, Advances in Neural Information Processing Systems, Volume 18, 2005, pp. 507-514, ISBN : 978-0-262-23253-1

[18] A. Gretton, O. Bousquet, A. Smola, B. Schoelkopf, Measuring Statistical Dependence with Hilbert-Schmidt Norms, Proceding of the 16th International Conference Algorithmic Learning Theory, 2005, pp. 63-78. DOI:270036

[19] Fengxi Song, Zhongwei Guo,Dayong Mei, Feature Selection Using Principal Component Analysis, IEEE International Conference onSystem Science, Engineering Design and Manufacturing Informatization, Volume 1, 2010, pp.27-30.

[20] Yijuan Lu, Ira Cohen, Xiang Sean Zhou, Qi Tian, Feature Selection Using Principal Feature Analysis, Proceedings of the 15th international conference on Multimedia, Germany, 2007, pp. 301-304, ISBN: 978-1-59593-702-5.