Cork Stopper Classification Using Feature Selection Method and SVM Based Classifier

Aborisade, D.O and Adebayo I.G

Department of Electronic/Electrical Engineering, Ladoke Akintola University of Technology, Ogbomoso. Oyo-state
doaborisade@yahoo.com, Isaiahadebayo@yahoo.com

Abstract

Classifying cork stopper into group required large set of visual features. Selecting an optimal feature subset from large input feature set speeds up classification task and improve the classifier accuracy. Traditional feature selection methods, such as sequential forward selection, sequential backward selection, and sequential forward floating search are costly to implement. This paper we propose a feature selection method known as principal feature analysis that exploits the structure of the principal components of a feature set to find a subset of the original features information and support vector machines (SVMs) for classification. The experimental result show that the proposed method for SVM based classifier is lot faster than PCA and ICA based methods. It is also leads to better performance when the same number of principal/independent components is used and consistently picks the best subset of features in terms of sum-squared-error compared to competing methods.

Keywords: *Principal component analysis, Covariance matrix,* eigenvector, *eigenvalue, Independent component analysis, Support vector Machine, Dimensionality reduction.*

1. Introduction

Cork is a light, soft material produce in the Mediterranean countries. Cork stoppers are used for closing bottles, especially wine bottles because, unlike man-made material, they avoid leakage even in presence of irregularities in the bottle neck. They show an excellent chemical inertia and avoid any reaction with the wine that changes its taste. Their best characteristic is the ability to allow gaseous exchange with the outside atmosphere. This is an important step that makes it possible to improve the wine quality as time goes by [1].

Cork stoppers are routinely classified in different quality classes (see Fig. 1) and inspected for different defects like small holes due to insect attacks, stopper breaking or cracks (see Fig.2). Although it does not seem hard for human beings to visually analyze the cork stopper, it turns out difficult to precisely formulate the characteristics of the cork surface that are relevant to these tasks due to the porosity of the natural material. It is difficult even for the cork quality experts to exactly define all cork features necessary in the process of cork quality grading [2].

In an effort to tackle these drawbacks, automatic cork stopper quality classification is very much essential. The problem of cork analysis is based on considering a large set of visual features which may includes relevant and irrelevant features that can degrade the generalization performance of classifiers [3, 4]. However, for a real-time cork stopper classification problems high dimensional feature vectors is not suitable and thus some kind of dimensionality reduction must be performed. An approach to dimensionality reduction is

feature selection, which consists of determining an optimal feature subset from a large input feature set. The benefits of using feature selection are reduction in the time and cost of feature acquisition, as well as reduction in classifier training and testing time. It is also helpful in improving classifier accuracy, provided that noisy, irrelevant or redundant features are eliminated [5].

Most research work on feature selection for classifier design uses principal component analysis (PCA) and Independent Component Analysis (ICA) techniques for representing data in a reduced dimension [6, 7, 8]. This might be because feature selection methods, such as sequential forward selection (SFS), sequential backward selection (SBS), and sequential forward floating search (SFFS) [9, 10], require training and testing a classifier at each feature addition or elimination step, hence are costly to implement.

In this paper, to address the above mentioned problems, we propose another computationally efficient feature selection method that exploits the structure of the principal components of a feature set to find a subset of the original features that contains most of the essential information. The proposed method when used with support vector machine (SVM)-based classifiers on different kinds of features is much faster in term of speed with minimum error, compared to feature selection method based on PCA and ICA

The paper is organized as follows: After a brief discussion of PCA and ICA based feature selection in Section 2, Section 3 proposes the PFA based method. Experimental results are showed in Section 4 and Section 5 contained the conclusions.

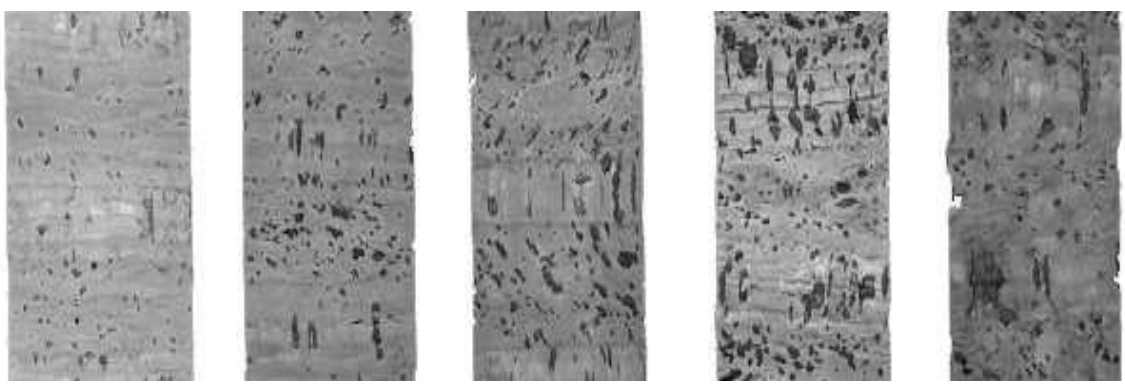

Figure 1. Surfaces of cork stoppers of 5 quality groups ordered from best to worst quality (from left to right).

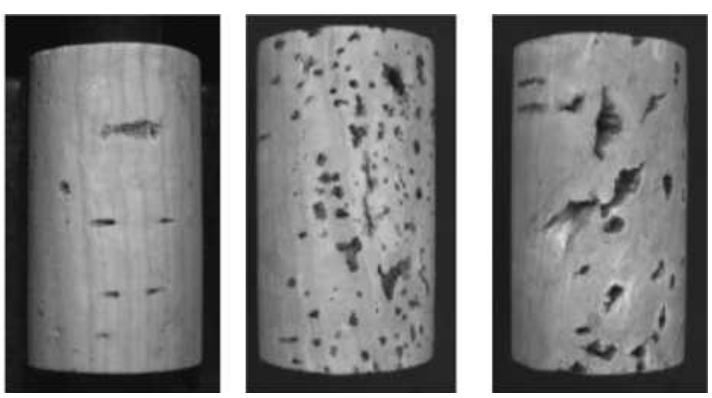

Figure 2. Cork stopper without defects, woody stopper and a stopper with a crack.

2. Overview of PCA and ICA

A. Principal Component Analysis

Principal component analysis is a classic tool for analyzing large scale multivariate data. PCA describes as an unsupervised feature selection technique provides mapping of $n-$ dimensional data space onto $q-$ dimensional data space where $q < n$ [11]. In the context of object classification, a set of N sample images $x_i \in \Re^n$ where $i = 1, \cdots, N$ is considered and there is a mapping from the original $n-$ dimensional image space to $q-$ dimensional feature space where $q < n$. The new feature vectors ξ are defined by the following linear transformation

$$\xi = \varphi_q{}^T x \tag{1}$$

with $\varphi_q{}^T \varphi_q = I_q$ where φ_q is a matrix with orthonormal columns $[\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_m]$, and I_q is the $m \times m$ identity matrix.

The mean image of all the samples is denoted by

$$m = E(x) = \frac{1}{N} \sum_1^N x_i \tag{2}$$

where $E(x)$ is the expectation of x. The difference, ∂ of each set of sample image from the mean image of all samples is computed:

$$\partial_i = x_i - m \tag{3}$$

Then the covariance matrix is estimated by

$$\begin{aligned} C_X &= E\{(x-m)(x-m)^T\} \\ &= \frac{1}{N} \sum_1^N (x_i - m)(x_i - m)^T \\ &= \frac{1}{N} \sum_1^N \partial_i \partial_i{}^T = AA^T \end{aligned} \tag{4}$$

where, $A = [\partial_1 \partial_2 \cdots \partial_N]$.

Eigenvectors ℓ_i and corresponding eigenvalues λ_i of the covariance matrix C_x are the solutions of the equation

$$C_x \ell_i = \lambda_i \ell_i, \quad i = 1, \cdots, n \tag{5}$$

Since C_x is very large $(N \times N)$, solving its eigenvector and eigenvalues is a non-trivial task. Instead, eigenvectors v_i and corresponding eigenvalues λ_i of matrix AA^T $(M \times M)$ can be computed. After that, ℓ_i can be computed from v_i as follows:

$$\ell_i = \sum_{j=1}^{R} v_{ij} \partial_j, \quad j = 1, \cdots, R \tag{6}$$

To reduce dimensionality, there is a need to keep only a smaller number of eigenvectors R_K $(R << M)$ corresponding to the largest eigenvalues. A new sample image x, after subtracting the mean $(\partial = x - m)$ can then be reconstructed in eigenspace by the formula:

$$\tilde{\partial} = \sum_{i=1}^{R_K} w_i \ell_i \tag{7}$$

where $w_i = \ell_i^T x$ are coefficients of the projection (i.e., eigenfeatures) which allow images to be represented as a linear combinations of the eigenvectors.

B. Independent Component Analysis

Independent component analysis (ICA) is a statistical technique for decomposing a complex dataset into independent sub-parts [7]. It develops from blind source separation of independent sources from their observed linear mixtures [12]. The model of ICA can be expressed as follows:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{8}$$

where; $\mathbf{x} = (x_1, x_2, \cdots, x)^T$ is the measured data matrix, $\mathbf{A}$ is $m \times n$ nonsingular mixing matrix, and $\mathbf{s} = (s_1, s_2, \cdots, s_n)$ is the source matrix containing statistically independent source vectors in its rows. The aim of ICA algorithm is to find the components s_i as independent as possible so that the set of observed signals can be expressed as linear combination of statistically independent components. The original source vector $\mathbf{s}$ is recovered with the help of the following linear transformation

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{9}$$

where; $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ are statistically independent (called also the estimated sources of the s_i's); and $\mathbf{W}$ is equal to the pseudo inverse matrix of *A*. The estimation of sources and basis functions could be done by maximization of negentropy $J(s_i)$ [13]:

$$\begin{aligned} \max J(s_i) &= J(\mathrm{w}_i^{\mathrm{T}} \mathrm{x}) \\ &= \left[\mathrm{E}\left\{\mathrm{G}(\mathrm{w}_i^{\mathrm{T}} \mathrm{x})\right\} - \mathrm{E}\{\mathrm{G}(\mathrm{v})\}\right]^2 \end{aligned} \tag{10}$$

where w_i is the ith row of mixing matrix $\mathbf{W}$, G is a contrast function and v is a standardized Gaussian variable.

3. The Proposed Method

Both the PCA Based Dimensionality reduction and the ICA Based Dimensionality reduction method choose a subset of feature by computing its PC projection to a smaller dimensional space. The drawback of these methods is the complexity of finding the subset. The goal of the proposed PFA method is to exploits the structure of the principal components of a feature set to find optimal subset of the original features in terms of sum-squared-error compared to competing methods using the same number of principal/independent components.

In order to simply formulate what PFA does, let $X = [f_1, f_2, \cdots, f_d]^T$ denote a set of d features in $\Re^n$ space with zero mean (i.e., $f_j \in \Re^n$) ; f_j are column vectors and X is of size $d \times n$. Let ψ be the covariance matrix (correlation matrix) of X . Let Φ be a matrix whose columns are the orthonormal eigenvectors of the matrix ψ [14, 15]:

$$\psi = \Phi D_\lambda \Phi^T \tag{11}$$

where D_λ *is a diagonal matrix whose diagonal elements are the eigenvalues of* ψ *:*

$$D_\lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix} \tag{12}$$

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Let Φ_m be the first m columns of Φ and let the row of the matrix Φ consists of m feature vectors

$$\phi'_1, \phi'_2, \cdots, \phi'_k, \cdots, \phi'_d \in \Re^m, \ m < n \tag{13}$$

Each vector ϕ'_i represents the projection of the ith feature (variable) of the vector X to the lower dimensional space, that is, the m elements of ϕ'_i correspond to the weights of the ith feature on each axis of the subspace. Thus, for any vector X , the reduced image m-vectors are given by

$$\rho_i = \Phi' X \tag{14}$$

The key observation of this method is that features that are highly correlated or have high mutual information will have similar absolute value weight vectors ϕ'_i (changing the sign has no statistical significance [16]). On the two extreme sides, two independent variables have maximally separated weight vectors; while two fully correlated variables have identical weight vectors (up to a change of sign). To find the best subset we use the structure of the rows ϕ'_i to find the subsets of features that are highly correlated and to choose one feature from each subset. The chosen features represent each group optimally in terms of high spread in the lower dimension, reconstruction and insensitivity to noise. The algorithm may be summarized as follows:

1. Compute the correlation matrix from the patterns of the training set.
2. Compute the eigenvalues and corresponding eigenvectors of the correlation matrix as defined in Equation (11).
3. Choose the subspace dimension m and construct the matrix Φ_m and Φ' from Φ.
4. Cluster the vectors $|\phi'_1|, |\phi'_2|, \cdots, |\phi'_k|, \cdots, |\phi'_d| \in \Re^m$ using K-Means algorithm. Choosing n greater than m is usually necessary if the same variability as the PCA is desired.
5. For each cluster, find the corresponding vector ϕ'_i which is closest to the mean of the cluster.
6. Compute the representation of the features in the lower dimensional space as defined in (14).

4. Support Vector Machine

The effectiveness of the feature space depends on how well different classes can be separated in the space. To compare the proposed method with both PCA and ICA based feature selection methods, we used Support vector machine (SVM) based classifier. SVMs are primarily two-class classifiers that have been shown to be an attractive and more systematic approach to learn linear or non-linear decision boundaries [17, 18]. The classifier constructs an optimal separating hyper-plane between the classes in the dataset by maximizing the distance of either class from the hyper-plane using the Gaussian radial basis kernel. This is equivalent to performing structural risk minimization to achieve good generalization [17, 18]. Finding the optimal hyper-plane implies solving a constrained optimization problem using quadratic programming [4].

The dimensionality of the feature space is determined by the number of support vectors extracted from the training data. The SVM can locate all the support vectors, which exclusively determine the decision boundaries. To estimate the misclassification rate (risk), the so called leave-one-out procedure is used. It removes one of N_i training samples, performs training using the remaining training samples, and tests the removed sample with the newly derived hyper-plane. It repeats this process for all of the samples, and the total number of errors becomes the estimation of the risk [19].

5. Experimental Results

We have performed a number of experiments and comparisons to demonstrate the performances of the proposed feature selection method based on SVM. For training sixty samples for each of the five classes are used and seventy samples for each class are used for testing.

In this experiment LibSVM [20] was used to train SVM classifiers with a Gaussian radial basis kernel on selected feature subsets. The results in Figure 3 shows the performances of the SVM classifiers trained on subset features selected by our method using top-50 eigenvector compare to that of PCA and ICA based feature selection. Our method performs better than PCA or ICA when the same number of principal/independent components is used. The speed of our method per second using SVM classifier is also compared to that of PCA and ICA feature selection in Table 1. Our algorithm is a lot faster than PCA and ICA feature selection when the number of features in each subset is large enough, e.g., 80.

6. Conclusion

We have presented a systematic feature subset selection method using PFA. To evaluate the proposed method, we considered cork stoppers classification problems. In this case, we used PFA for feature selection and SVMs classifier with a Gaussian radial basis kernel. Experimental results on different kinds of features used for cork stopper classification shows that the proposed method is capable of removing redundant and irrelevant features, outperforming traditional approaches. Further analysis illustrate that the proposed method improves the performance of classifier both in terms of accuracy and complexity (i.e., number of features).

Acknowledgement

The authors would like to acknowledge the anonymous reviewers for their helpful comments on improving this paper.

References

[1] M. A. Fortes, "Cork and corks", European Rev., vol. 1, Pp. 189-195, 1993.

[2] Jordi Vitria, Marco Bressan and Petia Radeva, "Bayesian Classification of Cork Stoppers using Class-Conditional Independent Component Analysis", Journal of the Royal Statistical. Society, vol. 39, Pp. 1- 38, 1977.

[3] Bins, J., Draper, B.A." Feature selection from huge feature sets", Proc. Intl. Conf. on Computer Vision (ICCV). Vol. 2. Pp. 159-165, 2001.

[4] Sun, Z., Bebis, G., Miller, R. "Object detection using feature subset selection", Pattern Recognition 37**,** Pp. 2165-2176, 2004.

[5] Zehra Cataltepe, Hakki Murat Genc, and Thomas Pearson, "A PCA/ICA Based Feature Selection Method and Its Application for Corn Fungi Detection", 15th European Signal Processing Conference, Poland. Pp.970-974, 2007.

[6] Arnaz, M., and Robert X. G.,"PCA-Based Feature Selection Scheme for Machine Defect Classification",IEEE Transactions on Instrumentation and Measurement, Vol. 53, No. 6, Pp. 1517-1525, December 2004.

[7] Xiao benlina *et a*l.,"Study on Independent Component Analysis-Application in Classification and Change Detection of Multispectral Images", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B7. Pp. 871-876, Beijing 2008.

[8] Imran S. Bajwa *et al*., "Feature Based Image Classification by using Principal Component Analysis", ICGST-GVIP Journal, Volume (9), Issue (II). Pp. 11-17, April 2009.

[9] Jain A. and Zongker D., "Feature selection: Evaluation, application, and small sample performance", IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(2), Pp. 153-158, 1997.

[10] Pudil P., Novovicova J., and Kittler J., "Floating search methods in feature selection", Pattern Recognition Letter **15**(11), Pp. 119-1125, 1994.

[11] Krzanowski, W.J., "Selection of Variables to Preserve Multivariate Data Structure, Using Principal Component Analysis", Applied Statistics- Journal of the Royal Statistical Society Series C, vol.36, Pp. 22-33, 1987.

[12] Hyvarinen, A., Oja, E., "Independent component analysis: Algorithms and applications", Neural Networks 13 (4–5), 411–430, 2000.

[13] A. Hyvarinen, J. Karhunen, E. Oja, "Independent Component Analysis", John Wiley & Sons, New York: 2001.

[14] Yijuan Lu ***et al.*** **"**Feature Selection Using Principal Feature Analysis", ACM Multimedia, Augsburg, Germany, September 23-29, 2007.

[15] Omid Khayat et al.," Image classification using principal feature analysis", 7th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED'08), University of Cambridge, UK, Pp.198-203, Feb 20-22, 2008.

[16] Jolliffe, I.T., "Principal Component Analysis", Springer-Verlag, New-York, 1986.

[17] V. Vapnik, The Nature of Statistical Learning Theory, Springer, Berlin, 1995.

[18] C. Burges, Tutorial on support vector machines for pattern recognition, Data Mining Knowledge Discovery 2 (2) Pp. 955-974, 1998.

[19] V. C. Chen," Evaluation of Bayes, ICA, PCA and SVM Methods for Classification", RTO SET Symposium on Target Identification and Recognition Using RF Systems, Oslo, Norway, Pp. 37.1-37.12, October 2004.

[20] Chang, C.C., Lin, C.J, " LIBSVM: a library for support vector machines", (2001) Software available at http://www.csie.ntu.edu.tw/" cjlin/libsvm.

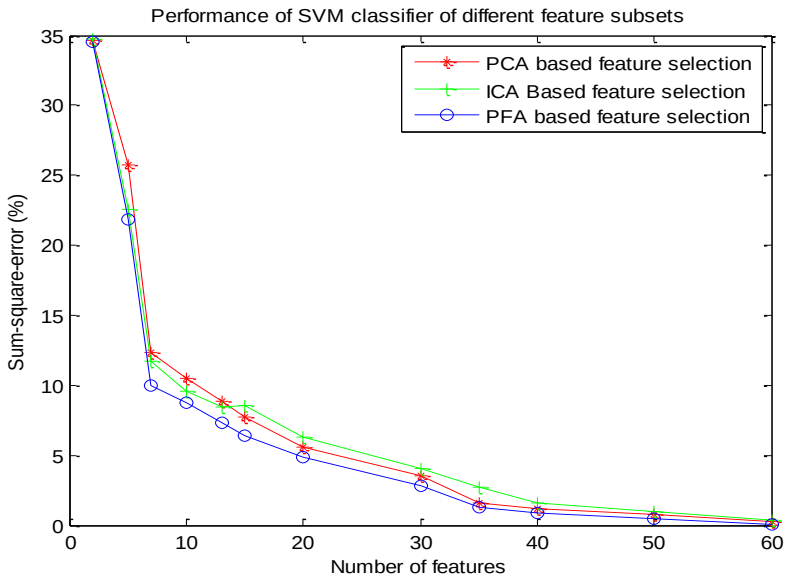

Figure 3. Sum-square-error curves for SVM classifiers trained on different feature subsets selected from different selection method.

Table 1. Time required for feature selection and training using SVM

No. of Features	PCA Based Feature Selection	ICA Based Feature Selection	PFA Based Feature Selection
2	492	1034	192
5	493	1034	193
7	493	1036	193
10	495	1038	195
13	496	1038	196
15	496	1039	196
20	498	1043	198
30	498	1048	198
35	501	1053	200
40	501	1054	200

Authors

Aborisade, David. O received the B.Eng. degree in Electronic and Electrical Engineering Technology from Federal University of Technology, Owerri, in 1989. He received M.Eng. and Ph.D. Degrees in Electrical Engineering from University of Ilorin, in 1995 and 2006, respectively. He is currently a Senior Lecturer with the Department Electronic and Electrical Engineering, Ladoke Akintola University of Technology, Ogbomoso. His research interests include computer vision, pattern recognition, image and signal processing, neural networks, and fuzzy logic.

Adebayo, I. G got his first degree in the year 2004 with Second class upper division in Electronic and Electrical Engineering from Ladoke Akintola University of Technology (LAUTECH), Ogbomoso. He is about to complete his M.Tech. Degree programme in Electronic and Electrical Engineering from LAUTECH., Ogbomoso. He is currently a Graduate Assistant in the Department Electronic and Electrical Engineering, LAUTECH., Ogbomoso. His major area of research is in Power System Engineering.