A New Feature Extraction Method based on Clustering for Face Recognition

Sabra El ferchichi[1], Salah Zidi[2], Kaouther Laabidi[1], Moufida Ksouri[1], Salah Maouche[2]

[1] LACS, ENIT, BP 37, le Belvédère 1002 Tunis, Tunisia
[2]LAGIS, USTL, Villeneuve d'Ascq, 59650 Lille, France

{sabra.elferchichi, kaouther.laabidi, moufida.ksouri}@enit.rnu.tn,
{salah.zidi, salah.maouche}@univ-lille1.fr

Abstract. When solving a pattern classification problem, it is common to apply a feature extraction method as a pre-processing step, not only to reduce the computation complexity but also to obtain better classification performance by reducing the amount of irrelevant and redundant information in the data. In this study, we investigate a novel schema for linear feature extraction in classification problems. The method we have proposed is based on clustering technique to realize feature extraction. It focuses in identifying and transforming redundant information in the data. A new similarity measure-based trend analysis is devised to identify those features. The simulation results on face recognition show that the proposed method gives better or competitive results when compared to conventional unsupervised methods like PCA and ICA.

Keywords: Feature extraction, dimensionality reduction, similarity measure, clustering, face recognition.

1 Introduction

The recent development in information technology has induced a rapid accumulation of high dimensional data such as face images and gene expression microarrays. Typically, in many pattern recognition problems, the huge number of features that represent data makes discrimination between patterns much harder and less accurate [1]. Thus, feature extraction is an important preprocessing step to machine learning and data mining. It has been effective in reducing dimensionality and increasing learning accuracy [2]. It tries to re-describe data in a lower dimensional feature space with respect to its underlying structure and generalization capabilities [2]. Abundant techniques for feature extraction problem were developed in the literature [3-8]. Principal Component Analysis (PCA) [3-5], remains the standard approach for feature extraction. It performs a linear transformation, derived from the eigenvectors corresponding to the largest eigenvalues of the covariance matrix. PCA implicitly assume that pattern elements are random variables with Gaussian distribution. Thus, in the case of non-Gaussian distribution, largest variances would not correspond to

PCA basis vectors [4]. Independent Component Analysis (ICA) [5] has been proposed to minimize both second-order and higher-order dependencies in the input data and attempts to find the basis along which the projected data are statistically independent. Unlike unsupervised method like PCA and ICA, Linear Discriminant Analysis (LDA) method [7-8], exploit label classes to make feature extraction [7]. It extracts features that minimize the within-class scatter matrix and maximize the between-class scatter matrix.

In this work, our interest is to perform a feature extraction procedure without assuming any form of knowledge, neither about data distribution nor about its class distribution, unlike already discussed methods. Actually, in high dimensional data, many features have the same tendencies along the data set: they describe the same variations of monotonicity (increasing or decreasing). Thus we can consider that these features give very similar information for the learning process. Such features are then considered redundant and useless information. Once groups of similar features have been settled, feature extraction can be realized through a linear transformation of each group of similar features. Hence, a clustering technique based on a new measure of similarity between features was used to identify and gather similar features. A linear transformation is finally made on each identified groups of original features to obtain a set of new features. Performance of our method is assessed through face recognition problem. The rest of the paper is organized as follows: in section 2, we devise the new similarity measure based trend analysis to evaluate similarity between features and we detail the proposed Feature Extraction Method based Clustering (FEMC). In section 3, experiments are carried out on face recognition problem through Yale and ORL data sets. We compare FEMC with conventional unsupervised and supervised feature extractor such as PCA, ICA and LDA. Finally in section 4, a brief conclusion is drawn with some future work.

2 Clustering based Feature Extraction algorithm

Actually, redundant information is an intrinsic characteristic of high dimensional data which complicate learning task and degrade classifier performance. For this reason, eliminating redundant features is one clue to reduce the dimension without loss of some important information. One existing solution is to use a filter method to select relevant features. Although redundant information is not relevant for discrimination task but it has implicit interaction with the rest of features. Eliminating them from feature space may lead to eliminate some predictive information of the inherent structure of data and thereby don't lead necessarily to a more accurate classifier. FECM seeks to transform redundant features such that the amount of predictive information lost, is minimized. Intuitively, redundancy in a dataset can be expressed by similar features in term of behavior along the data set. They describe very similar or mostly the same variations of monotonicity along the data set. Thereby we consider that they incorporate the same discriminating information. Our method seeks to identify this form of redundancy and incorporate it by a linear transformation into the new set of features. We based our method on clustering technique using a new similarity measure to identify linear or complex relations that would exist

between features. Once groups of similar features are formed, a linear transformation is realized to extract a new set of features. Actually, there exists another work that exploit clustering algorithm as a feature extraction technique to find new features [9]. It focuses on the prediction of HIV protease resistance to drugs. K-means based a biological similarity function was used. However, it can be used only for a specific purpose and can't be extended to other classification problems, unlike the general concept developed in this work.

2.1 Formulation

Merely, the aim of feature extraction is finding a transform T such that $y_i = T(x_i); 1 \le i \le L.$ Where, $x_1, x_2, ...x_L \in \mathbb{R}^D$ is the D-dimensional data set and $y_i \in \mathbb{R}^{d \ll D}$ is the transformed sample composed of d new features $v_1, v_2, ...v_d \in \mathbb{R}^L$ where $d \ll D$. Each feature vector v_i is constituted by the different values corresponding to each instance or sample $x_{1 \le j \le L}$. FECM uses clustering process to partition the feature space into $d \ll D$ clusters. Actually, clustering is supposed to discover inherent structure of data [1], [8]. Its goal isn't to find the best partition of a given samples but to approximate the true partition of the underlying space. Each obtained cluster C_k is composed of n_k similar features according to the new similarity measure defined in the next section. It is represented each by its centroid g_k, obtained by applying the transform f defined by:

$$g_k = f(S_k) = \frac{1}{n_k} \sum_{s=1}^{n_k} v_s \tag{1}$$

Where, S_k is the set of n_k feature vectors $v_{s \in S_k}$ belonging to the cluster C_k. The obtained set of centroids $\{g_k\}_{1 \le k \le M}$ is then considered as the set of the d new features.

2.2 Similarity measure

Distance or similarity relationships between pairs of patterns are the most important information for clustering process to approximate true partition in a dataset. FECM focuses on defining a similarity measure that characterizes similarity in the behavior of each features pair. We propose to analyze their tendencies through studying their variations of monotonicity along the data set rather than difference between their real values. Thus, conventional distance like Euclidean distance used normally in clustering algorithm is not suitable for our objective. Using Euclidean distance may lead to erroneous results since it computes the mean of difference between each value of two vectors without use of tendency information about them. In fact, two features may have the same mean (or closer means) but they differ completely in their trend. Actually, a trend is a semi-quantitative information, describing the evolution of the qualitative state of a variable, in a time interval, using a set of symbols such as

{Increasing, Decreasing, Steady}[10]. To determine the trend of a feature vector in each point, we compute firstly, the corresponding first order derivative of a feature vector v at each sample x: $\frac{dv(i)}{dx} = \frac{v_i - v_{i-1}}{x_i - x_{i-1}}$. Then, we determine the sign of the derivative in each point by:

$$if \begin{cases} \frac{dv}{dx} \prec 0 \\ \frac{dv}{dx} \succ 0 \\ \frac{dv}{dx} = 0 \end{cases} \quad then \quad \alpha = \operatorname{sgn}\left(\frac{dv}{dx}\right) = \begin{cases} -1 & (decrease) \\ 1 & (increase) \\ 0 & (steady) \end{cases} \tag{2}$$

A feature vector $v \in \mathbb{R}^L$ is then being represented as an L-dimensional vector composed of L variables $\alpha \in \{1, -1, 0\}$. Distance function devised to compare two feature vectors relies on verifying difference in the sign of tendency between two feature vectors. It is the squared sum of the absolute difference between occurrences of a specified value of α for two given feature vectors. It was inspired from the Value Difference Metric (VDM) [11]. Thus, the location of a feature vector within the feature space is not defined directly by the values of its components, but by the conditional distributions of the extracted trend in each component. Hence, this makes the proposed metric independent from the order of data and has a generalization capability. It is given by the following expression:

$$\begin{aligned} d(v_i, v_j) &= \sqrt{\delta_1(v_i, v_j) + \delta_{-1}(v_i, v_j) + \delta_0(v_i, v_j)} \\ \delta_\alpha(v_i, v_j) &= \left| p(v_i / \alpha) - p(v_j / \alpha) \right| \\ p(v_i / \alpha) &= \frac{\text{Occurrence of } \alpha \text{ in } v_i}{L} \end{aligned} \tag{3}$$

$p(v_i / \alpha)$ is determined by counting haw many times the value α occurs in the feature vector v_i for the learning data set. In fact, in this work we have computed the occurrence of the pair of variables $(\alpha, \beta) \in \{10, 11, 1-1, -10, -11, -1-1, 00, 01, 0-1\}$ in each vector v_i instead of computing only the probability of the single variable $\alpha \in \{1, -1, 0\}$. A similarity matrix M between all features vectors is then generated such that $M(i, j) = d(v_i, v_j);\ i, j = 1...n$.

2.3 Feature extraction schema

Feature extraction process is given by the pseudocode below. The similarity matrix M is computed based on the metric defined previously by (3). Then a clustering strategy based on C-means clustering, is performed. Clusters are initialized randomly.

Then they are sequentially enlarged by selecting the ε first ranked features in the similarity matrix M : set of the most similar features to the corresponding centroid g_j. Hence, this process allows an overlap between clusters. To remedy to this, we perform an intersection between each pair of the obtained clusters to determine common features and decide at which cluster they finally belong. Each common feature is then assigned to the closest cluster according to the Euclidean distance

$$\forall v_i \in \{C_{k1} \cap C_{k2}\},\ v_i \in C_h = \arg\min \left\| g_{j\in\{k1,k2\}} - v_i \right\|^2 \tag{4}$$

Where h is the either the index k_1 or k_2. Clusters centers are then re-computed using the current cluster memberships and the process is stopped when all d clusters are constructed.

```
{Input: Raw Data}
{Output: New features= cluster 'centers}
   Compute matrix of distance M(i,j) = d(v_i, v_j); i,j = 1...n
Clustering
  {d = N div ε : number of clusters
  ε : number of preselected features
  C: initial number of features
  Idx: index of initial centroid
   While C > 1
     Cluster C_k = select the ε first features from M
     C= C- ε
   End
   Intersection between d final clusters
   Update clusters centers}
```

3 Experimental Results

Face recognition is a technically difficult task due to the varying conditions in the data capturing process like variations in pose, orientation, expressions and illumination conditions. We assess the feasibility and performance of our proposed method of feature extraction on the face recognition task using the Yale and ORL datasets, presented in the Table 1. Each image from was down-sampled into a manageable size for computational efficiency. The classification performances of FECM were compared with those of PCA, ICA and LDA. A leave-one-out schema

was used to obtain the performances and K-Nearest Neighbor classifier (KNN), known for its standard performance, was used as classifier system for both datasets.

Table 1. Data sets information

Data sets	No. of features	No. of instances	No. of classes
Yale data set	783	165	15
ORL data set	952	400	40

Table 2. Classification accuracy on Yale Dataset

Methods	Error rate %	Number of features
KNN	21.82	783
PCA	24.85	30
ICA	23.03	30
LDA	8.48	14
FEMC	21.00	14

Performances of FECM for ORL dataset are detailed in the Table 3. Because there must be at least 40 PCs to get 39 features, the first 40 PCs are retained as input for LDA and ICA. Note that the number of extracted features by LDA is 39 because there are 40 classes. Our approach is close to PCA and ICA in terms of classification accuracy with smaller number of features which is 23, but LDA stay having the best accuracy classification as in Yale dataset due to its supervised nature.

Table 3. Classification accuracy on AT&T Dataset

Methods	Error rate %	Number of features
KNN	3.00	952
PCA	4.00	40
ICA	4.25	40
LDA	2.00	39
FEMC	5.00	23

4 Conclusion

This paper deals with the important problem of extracting discriminant features for pattern classification. Feature extraction techniques often trust in some interestingness criterion to search for a lower dimensional representation. However, because the true structure of the data is unknown, it is inherently ambiguous what constitutes a good low dimensional representation. This makes it difficult to define a proper interestingness criterion. In this work, we propose a new feature extraction approach based on feature clustering. The main motivation behind it was to identify redundancy in feature space and reduce its effect without losing some important information for

classification task. Similar feature are recognized through analyzing their monotonicity along the data set and a new similarity measure is then devised. The proposed approach applies clustering technique into feature space to determine its underlying groups of features. Each obtained cluster is represented by one feature, computed as the mean of all features grouped in the cluster. The main difficult in the proposed method is its dependence on the partition resulting from the clustering process. In validation strategy, clustering algorithm analyses only training set, which changes every time, so clustering rules changes every time. That can produce some instability in the results of the method. Hence, further work to verify robustness of the method towards noise in data has to be engaged.

For the face recognition task, our approach produces a lower number of features than PCA and ICA, and achieves better or competitive classification accuracy. Unlike LDA, we didn't make use of class information in our procedure. It would be interesting to introduce this specific information in our procedure to approach semi supervised learning task.

References

1. Fern, X.Z., Brodley, E.C.: Cluster Ensembles for High Dimensional Clustering: an empirical study. Technical report CS06-30-02 (2004)
2. Torkkola K.: Feature Extraction by Non-parametric Mutual Information Maximization. In Journal of Machine Learning Research, vol.3, pp.1415-1438 (2003)
3. Saul, L.K., Weinberger, K.Q., Sha, F., Ham, J., Lee, D.D.: Spectral Methods for Dimensionality Reduction. In O. Chapelle, B. Schoelkopf, and A. Zien (eds.), Semi supervised Learning, MIT Press. Cambridge, MA (2006)
4. Delac, K., Grgic, M., Grgic, S.: Independent Comparative Study of PCA, ICA and LDA on FERET data set. In Wiley Periodicals, Inc. Int J Imaging Syst Technol, vol.15, pp. 252-260 (2006)
5. Kwak, N.: Feature Extraction for Classification Problems and its Application to Face Recognition. In Journal of Pattern Recognition, vol.41, pp.1701-1717 (2008)
6. Wang, J., Lin, Y., Yang, W., Yang, J.: Kernel Maximum Scatter Difference based Feature Extraction and its application to Face Recognition. In Pattern Recognition Letters, vol.29, pp.1832-1835 Elsevier, Amsterdam, PAYS-BAS (2008)
7. Xiang, C., Huang, D.: Feature Extraction using Recursive Cluster-based Linear Discriminant with Application to Face Recognition. In IEEE Transcations on Image Processing, vol.15, pp.3824-3832 (2006)
8. Von Luxburg, U., Budeck, S., Jegelka, S., Kaufmann, M.: Consistent Minimization of Clustering Objective Functions. In Neural Information Processing Systems NIPS (2007)
9. Bonet, I., Saeys, Y., Grau Abalo, R., M.Garcia, M., Sanchez, R., Van de Peer, Y.: Feature Extraction Using Clustering of Protein. In the 11th Iberoamerican congress in pattern recognition, CIARP (2006)
10. Charbonnier, S., Gentil, S.: A trend-based Alarm System to Improve Patient Monitoring in Intensive Care Units. In Control Engineering Practice, vol.15, pp.1039-1050 Elsevier, Kidlington, ROYAUME-UNI (2007)
11. Payne, R.T., Edwards, P.: Implicit Feature Selection with the Value Difference Metric. In the 13th European Conference on Artificial Intelligence, pp.450-454 (1998)