Optimization Methods for Identity Verification System Using Biometric Features

Mohamed A. El-Sayed1,2 and Mohammed A. Abdel-Latif3

1Computer Science Department, College of Computers and IT, Taif University, Hawia, Taif, KSA.
2Mathematics Department, Faculty of Science, Fayyoum University, Fayyoum, Egypt.
3Department of Mathematics, Faculty of Science, South Valley University, Qena, Egypt.

Abstract
The paper will present an approach to solve the real-world biometric identification problems, with an objective to circumvent the curse of dimensionality in biometric identification system through very large database size. The main objective of the proposed approach is obtaining of effective in reducing the complexity and hence improving the performance of user identification. This will be achieved through the following steps: Firstly, reduction of features extraction, we will use feature selection algorithm based on several objective optimization techniques. Secondly, acceleration of classification, we will use recent method to determine clusters of users having significant similarities and then apply Support Vector Machine (SVM) classifiers on each cluster of users independently. Hence reducing the dimensionality of the dataset for training SVMs and improving the performance of user identification. The experimental results of the proposed technique are presented and compared.

Keywords: identity verification, biometric authentication, optimization methods, features optimization.

INTRODUCTION
In biometric authentication systems, an important issue is the curse of dimensionality problem because large dimensionality data can both dramatically increase the computational cost of the authentication task and decrease the authentication accuracy by limiting the generalization capabilities of the classifier [1]. The curse of dimensionality can affect the performance of the biometric authentication system when the dimensionality of the biometric template is very large and/or when the number of classes (users) increases. Feature Selection (FS), also called dimensionality reduction, is usually performed before classification and clustering processes. It consists of determining pertinent features in a dataset in order to increase the classes/clusters identification performance. FS also allows getting less complex representations that can have higher quality when compared to the representation with the original features in high dimensional datasets with large precision. It is the most popular method for reducing the dimensionality while representing different feature vectors into well-separated classes.

PCA, LDA or LPP can subsequently be performed in the feature space resulting in Kernel Principal Component Analysis (KPCA), Generalized Discriminant Analysis (GDA), and Kernel Locality Preserving Projection (KLPP). Several studies showed that KPCA, GDA and KLPP are able to extract nonlinear features and thus provide better recognition rates in applications such as face recognition [2]. The main disadvantage of the PCA, LDA, ICA, and LPP is their linearity. Recently, kernel methods have been successfully applied to solve pattern recognition problems because of their capacity to handle nonlinear data. By mapping sample data to a higher dimensional feature space, effectively a nonlinear problem defined in the original space is turned into a linear problem in the feature space [3].

Since the popular back-propagation neural network may be trained to recognize feature vectors, the direct application of training to recognize feature vectors is often unbearable due to the size of the input features as it would lead to a complex network which would be hard to train. Thus, the feature selection technique should be applied before execution of classification. The training complexity of SVMs is highly dependent on the size of the dataset, for additional specifics about SVM can be obtain in [4, 5].

These methods treat the problem as a single objective to identify the relevant features. However, the classification performance is not always outstanding because FS has two crucial conflicting objectives of maximizing the classification performance and minimizing the number of features.

El-Sayed et al [6] offered a technique of authentication based on identification of retinal features. This technique includes blood-vessel segmentation, creation of feature pattern and then matching of these features. The commonly obtainable databases DRIVE, VARIA and STARE are used to calculate the performance analysis.

Reducing the number of features while improving the classification rate involves optimizing simultaneously two conflicting objectives or more. Different multi-objective feature selection techniques have been addressed these recent
years, the most recent are described below. In [7], the authors proposed a new feature selection method to handle the problem of Maximum Entropy based Named Entity Recognition. They applied NSGA-II to optimize recall and precision classification quality measures. The proposed algorithm was evaluated using two distinct languages (Bengali and English). The obtained results were encouraging. The authors extended this work in [8] by using three different classifiers (maximum entropy, conditional random field, and support vector machine) and three resource constrained languages (Bengali, Hindi and Telugu). One advantage of this proposed technique is that it can be applied to any language without prior knowledge because it only needs language-independent features. The results demonstrated improvement in recall and precision measures.

In [9], a multi-objective feature selection method is presented to address the network anomaly detection problem. The aim was to accurately detect normal and abnormal behaviors in order to act in real time. To achieve this goal, it is crucial to treat the most important features subset instead of using the whole dataset. The authors used NSGA-II to reduce the complexity of Growing Hierarchical Self-Organizing Map classifier and improve the classification performance. To test the proposed algorithm, they used KDD-NCL dataset. The obtained results were successfully compared with those of the state-of-the-art feature selection methods, which proved a high classification performance.

In [10], the problem of feature selection and classification is treated by suggesting a new feature selection method based on the modified micro Genetic Algorithm. This algorithm employs voting-based elite-selection for selection. In addition, three objective functions are optimized: maximizing the specificity rate, maximizing the sensitivity rate and minimizing the number of selected features. The proposed algorithm was initially tested with two UCI datasets and compared with PCA, Bagging, Boosting and Rotation forest methods. The results of experiments indicated that the proposed method is able to provide accurate classification results with smaller numbers of features.

In [11], the authors proposed a new feature selection method based on weighting concept. They applied multi-objective evolutionary algorithm (MOEA/D) to optimize inter-class and intra-class distances. The obtained data with the reduced features are classified using k-nearest neighbor (k-NN) classifier. The proposed algorithm was tested with several UCI datasets. The results were successfully compared with different existing single- and multi-objective feature selection methods.

The rest of this paper is organized as follows: in Section 2, a presentation of optimization techniques based clustering for SVM classifiers. Section 3 deals with proposed approach of feature selection using SVM and projection pursuit indices. In Section 4, GA related proposed approach is presented. Section 5 describes the feature selection and classification using SVM, it discusses the experimental results and the effectiveness. Finally, conclusions are drawn in Section 6.

**OPTIMIZATION BASED CLUSTERING FOR SVM CLASSIFIERS**

Several attempts have been made to solve the training problem of SVMs but mainly in the case of a two-class SVM. Multi-class SVM is more complex for large datasets than binary classification, even in the case that both have the same amount of training data [12]. Platt [13] proposed Sequential Minimization Optimization (SMO) technique to solve SVM training problem. SMO consists of breaking the large QP problem into series of smallest possible QP problems and then solve analytically these problems using Lagrange Multipliers while satisfying Karush-Kuhn-Tucker (KKT) conditions. This method does not only require a memory space to save all the sub problems but also the KKT conditions are required to be checked for each sub problem. Inspired from the core set-based approximate Minimum Enclosing Ball algorithms in computational geometry, Tsang et al. have developed an approximation algorithm for SVM training called Core Vector Machines [14]. Generally speaking, in an optimization problem, a core set is a subset of input points such that a good approximation to the original input can be found by solving the optimization problem directly on the core set.

Another solution consists in scaling down the training data before inputting to the SVM. The simplest method would be just to sample from the original dataset and use the sample to train SVM. Pavlov et al. [15] used boosting to combine a large number of SVMs, each is trained on only a small data subsample. While this method is not optimal in general, it allows for very fast training of SVMs, has substantially lower memory cost and yields performance close to that of the full SMO. Alternatively, Collobert et al. [16] proposed a parallel SVM training and classification algorithm that each subset of a dataset is trained with SVM and then these small SVM classifiers are combined into a final single classifier using a neural-network-based method. Lee and Mangasarian [17] proposed the reduced SVM (RSVM), which uses a random rectangular subset of the kernel matrix to greatly reduce the size of the quadratic program to be solved and simplify the characterization of the nonlinear separating surface.

Instead of random sampling, one can intelligently sample a small number of training data for SVM training. Authors in [18] and [19] have proposed to use active learning to select training samples that are expected to minimize the estimated future error. This heuristic has shown that SVM trained on a well-chosen subset of the available offers better performance at less time and lower cost than the naive approach of training on all available data. However these methods are not immune to selecting outliers. In [20], authors used squashing that allows to scale the data down while preserving its statistical properties. Training SVMs on squashed datasets has been fast,
decreased memory requirement and produced classification accuracy that is closed to that of the full data. In [21], authors used editing by selectively removing samples from the training set using probabilistic estimates. The removal procedure aimed at creating a separable distribution of training examples without modifying the location of the decision boundary. Boley and Cao [22] used clustering to partition the training data into several pair-wise disjoint clusters by exploiting the distributional properties of the training data. Then, the representatives of these clusters are used to train an initial support vector machine, based on which are approximately identified the support vectors and non-support vectors. After replacing the cluster containing only non-support vectors with its representative, the number of training data can be significantly reduced, thereby speeding up the training process. Similarly, Yu et al. [23] applied a hierarchical micro-clustering algorithm that scans the entire dataset only once to provide an SVM with high quality samples that carry the statistical summaries of the data such that the summaries maximize the benefit of learning the SVM. Their method has shown to be scalable in terms of the training efficiency while maximizing the performance of SVMs.

Projection Pursuit (PP) has effectively been developed in several domains such as clustering [24], regression analysis [25], classification and others. In classification PP has been employed to find low-dimensional projections that expose dissimilarities between categorized groups. In [26], the authors combined PP with Neural Network method (called Projection Pursuit Networks) that reduces the dimensionality and uses clustering. They suggested a new (one, two)-dimensional PP index based on the quality of projected clusters. The index is dedicated to identify interesting linear projections by localizing clusters that contain different classes. The authors suggested further studies to improve PP network.

In [27], PP was applied to circumvent the curse of dimensionality where a new PP index based on Penalized LDA (called PDA) was proposed for classification. The results showed that the performance of PDA index depends on the value of the used parameter. The PDA index is mainly recommended to be used for the SSS problem. In [28], a new supervised classification method (called PPTree) was addressed. It consists of combining PP with Tree classification method using visualization representation. A recursive binary partition method along with PP were applied. The performance of PPTree highly depends on the characteristics of the datasets and the choice of PP index in each node of the tree. In [29], a framework to circumvent the curse of dimensionality in classification was developed. It includes two phases: the compaction phase to reduce the dimension of the data and PP phase to find the optimal projections utilizing the compacted illustration as input. The obtained experimental results were encouraging. However, the choice of the compaction method and the PP index depends on the dataset. In [30], the paper depicted visual statistical inference technique to solve the curse of dimensionality problem. PP is applied to reduce the dimension of the data using LDA [31] and PDA [27] indices. Then, PP optimization procedure in "tourr" package was performed.

To summarize, most of the proposed approaches using PP are concerned with solving the curse of dimensionality problem related to the small sample size data and none of the approaches uses SVM for the classification. Henceforth, we are interested in this paper to investigate the use of PP to solve the training complexity of SVMs in the case of increased number of classes.

PROPOSED APPROACH OF FEATURE SELECTION USING PP-SVM

SVM has been successfully applied to various classification problems, but several large-scale problems such as biometrics based user identification are overly difficult to solve using traditional SVMs [32, 33].

Like divide and conquer rule, in our general scheme, the complicated large-scale database can be decomposed into several relatively small clusters (clustering step), and for each cluster an SVM is used independently for classification (sub-classification step). More specifically, in this research project we did investigate the effectiveness of the PP to find clusters of users among large-scale datasets. The idea of PP is to locate the projection(s) from high-dimensional space to low-dimensional space that reveal the most details about the structure of the data set. Once an interesting set of projections has been found, existing clusters can be extracted and analyzed separately. PP may be applied several times to get different groups levels until the classification rate is satisfactory.

The indices is used in the proposed feature selection technique employed in this research. We focused on one-dimensional indices and we used the Friedman index [34, 35] and the kurtosis index [36, 24]. Note that other projection indices can be used such as the discriminant index [37] and within-between index [27] among others.

The projection pursuit index constitutes the core of a projection pursuit technique as it gives in some sense a definition of the purpose of the procedure. Many projection pursuit indices I(a) estimators of a measure of distance between the distribution of the projection of the data in the direction a = (a1, ..., ap) and some "uninteresting distribution". Other projection indices are based on moments of the projected data. We only consider vector a such that aT a = 1.

The Friedman-Tukey index

The search for interesting structures is done via the following projection index 1FT(a) = s(a)d(a): It is composed of two
parts, s which depends only on the covariance structure and d which captures the “local clusters” of the data. The term s can be avoided if the P -dimensional data X are sphericical. In Jones and Sibson [38], with Y = Xa; yj = aTXj and h is a bandwidth, d is expanded as a kernel density estimate:

$$I_n(y) = \frac{1}{N} \sum_{j=1}^{N} K\left(\frac{y - y_j}{h}\right)$$ (1)

The Friedman-Tukey index can be written as:

$$I_{FT}(a) = \frac{1}{K^2 N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} K\left(\frac{y_i - y_j}{h}\right)$$ (2)

It turns out that this index is an estimate of:

$$I_{FT}(a) = \int f^2(y) dy = E_Y(f_Y(y))$$ (3)

It is minimized by a parabolic density which is close to a standard normal density. Thus a departure from a parabolic density is also a departure from the normal density. So, in our work, we focus on the univariate Friedman-Tukey index:

$$I_{FT}(a) = \frac{1}{K^2 N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} K\left(a^2(X_i - X_j)\right)$$ (4)

with K defined by $K(u) = \frac{35}{32}(1-u^2)^31,|u| \leq 1$ and $I_n$ denotes the dummy variable associated with the set π, and h = 3.12N$^{-\frac{1}{3}}$ and we look for a maximizing this index [39]. The main drawback of this index is that the projections associated with maxima generally lead to the detection of outliers but may miss other interesting structures such as clusters. Another drawback of this index is that it is time-consuming.

The Friedman or Legendre index

The Friedman index is based on the Legendre polynomials [34]. It measures the departure between the density of the projected data and the normal density which is assumed to correspond to a non-interesting projection. The formula is given as follows:

$$I_{h}^F(a) = \sum_{j=1}^{h} \left( \frac{1}{2j+1} \sum_{i=1}^{N} L_j(2\varphi(X_i)-1) \right)^2$$ (5)

Where φ is the univariate standard normal distribution. The recursive definition of the Legendre polynomials is given by:

- $L_0(r) = 1$, $L_1(r) = r$, $L_2(r) = \frac{1}{2}(3r^2 - 1)$, ...
- $L_j(r) = \frac{1}{j} (2j-1)rL_{j-1}(r) - (j-1)L_{j-2}(r)$, $j \geq 3$

The choice of the value of h depends on the data dimension P and the sample size N. In the present research h is fixed to 3 according to the recommendations given in [34] and [40].

GENETIC ALGORITHM

In our work, chromosome structure: each individual of the population is encoded in real numbers. The chromosome represents the projection vector which is defined as a normalized vector of P dimension. For example, the chromosome 0.45, 0.98, -1.50, 1.19, -0.65, 0.42 is a normalized projection vector of 6 dimension. The population initialization of size T is set randomly by the user. Evaluation function: This function takes a single solution as a parameter and returns a number indicating how good the solution is. The best solution will be selected by comparing the fitness value returned by all the possible solutions. The projection index is considered as the fitness function.

Selection: It is a genetic operator. An individual from the current population is selected to produce a new generation. Several selection methods exist. Some of them rank the fitness of each solution and select the best ones. Some others rank only a random sample of the population. However, this process may be very time-consuming. The selection allows the diversity of the population to be kept and the premature convergence to poor solutions to be prevented. The roulette wheel selection and the tournament selection are popular and well-studied selection methods. The investigated method is summarized in Algorithm 1 and it is clearly described in [41].

**Genetic Algorithm**

Input: Initial population of individual: randomly initialized and encoded in float point numbers.

Fitness function = a projection index.

Evaluate the fitness of each individual in that population.

Determine the best individual.

While a fixed number of iterations is not reached do

Apply a tournament selection of 3 participants.

Breed new individuals through genetic operators with pc = 0.65 and pm = 0.05.

Evaluate the fitness of new individuals

End

Replace least-fitness population with new individuals

Rank selection: when the fitnesses differ very much, the previous selection will come across problems. For example, if 90% of all the roulette wheel is occupied by the best chromosome, there will be little chance for the other chromosomes to be selected. For the rank selection method, the population is ranked. Then from this ranking, a fitness is
assigned to each chromosome. The worst will have fitness 1, second worst 2, … etc., and the best will have fitness T (the number of chromosomes in the population). With the second method, the selection will be more accurate. But this method can lead to slower convergence, because the best chromosomes are almost similar to the other ones.

The two-point crossover with pc = 0.65 are applied to all the population. In fact, we choose randomly a part of two parents of equal size. The two children will have the same genes of the two parents respectively except those of the selected part.

In order to ensure that the individuals are different, a mutation operator is applied. It is only and randomly implemented on a single individual and it represents occasional modification of one allele or some parts of the chromosome. The mutation rate is applied with pm less than 10% probability. The mutation method is essential to ensure genetic diversity within the population. The mutation operator is applied in our case to all the individuals with pm = 0.05 by choosing randomly one gene and replace it by a random real value.

EXPERIMENTAL RESULTS

Datasets

We used 4 various data sets related to user biometric identification and reduction dimension, like database about face, eyes, nose, lips, iris, ear and retinal. YALE [42], face database contains 165 grayscale images of 15 individuals. There are 11 images per subject, one per different facial expression or configuration, center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. The size of the cropped image is 200x160 pixels. ORL [43], there is a well-known face database which can be downloadable from the AT&T Laboratories, Cambridge University site. The database contains 400 face images from 40 individuals in different states. Each image was digitized and presented by a 92x112 pixel. VARIA [44], the database is a set of retinal images used for authentication purposes. The database currently includes 233 images, from 139 different individuals. The images have been acquired with a TopCon non-mydriatic camera NW-100 model and are optic disc centered with a resolution of 768x584. IIT Delhi Ear [45], the database consists of the ear image database collected from the students and staff at IIT Delhi, New Delhi, India. It is acquired from the 121 different subjects and each subject has at least 3 ear images. The database of 471 images has been sequentially numbered for every user with an integer identification/number. The resolution of these images is 272x204 pixels and cropped ear images of size 50x180 pixels. To ascertain the validity of the proposed approach, we consider in this paper biometric identification system based on face recognition with the following three experiments.

Experimental 1:

Figure 1 show the required time of each algorithm for dimension reduction of vector length of image. We test the algorithms to generate 5,10,…or 100 features for each data sets. We present the various techniques to reduce the dimensions of the original datasets. From the survey, it comes to know that, PCA, LLE, LDA, GDA and MPPCA are the powerful techniques to handle the linear types of data. KPCA, Landmark Isomap , TSNE and LTSA are effectively worked on non-linear data while computing low and high dimensional feature dataset. The drawback of these nonlinear embedding techniques are consumes large time while computing high dimensional feature dataset such as Isomap, SNE, SPE, SymSNE and TSNE, see Figure 18. But the nonlinear techniques are efficient compared with the linear techniques of extraction of good features in the non-linear real world data.
Experimental 2:
In order to show the effectiveness of which described in previous section compared, comparisons are conducted against six well-known state-of-the-art multi-objective feature selection techniques including K-Nearest Neighbor (KNN) [46, 47, 48], Decision Tree Method (DTM) [49, 50, 51], Naive Bayes Algorithm (NBA) [52, 53], Linear Discriminant Analysis (LDA) [54, 55, 56] and SVM, using UCI/Statlog.
Each experiment run with start parameters values depend on the dataset size. The control effect parameters $C_1$ and $C_2$ give the random values on $[1.5, 2.5]$. The inertia weight $\omega$ of the particle take the random value on $[0.1, 0.5]$ . The numbers $r_1$ and $r_2$ are uniformly distributed random on $[0,1]$. The search space lies in bound $[1,10]$. The number of particles is equal to 100. The number of generation is equal to 100. The mutation probability calculated by inverse product on number of features. The program run 100 times in order to make sure that the non-dominated solutions obtained are the best ones and cannot be improved. If the dataset is small, the generation number can be reduced.

The classification accuracy of the selected UCI/ Statlog datasets selected is performed using SVM classifier before the feature selection algorithm to validate the effectiveness of the proposed multi-objective feature selection technique. We used the library LIBSVM [4] implemented in Matlab for SVM classification. The types of kernel are polynomial, linear and radial basis functions are used in this experimentation. The applied SVM parameters depend on each of the UCI / Statlog datasets selected. Table 3 and Figure 19 show the performance of the previous techniques and proposed approach through the average classification error rate and classical techniques.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DTM</th>
<th>KNN</th>
<th>LDA</th>
<th>NBA</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balance</td>
<td>22.70</td>
<td>15.62</td>
<td>16.32</td>
<td>9.51</td>
<td>17.67</td>
</tr>
<tr>
<td>Diabetes</td>
<td>26.35</td>
<td>29.70</td>
<td>22.50</td>
<td>23.54</td>
<td>29.48</td>
</tr>
<tr>
<td>Glass</td>
<td>31.80</td>
<td>28.32</td>
<td>40.76</td>
<td>37.18</td>
<td>27.67</td>
</tr>
<tr>
<td>Heart</td>
<td>78.19</td>
<td>47.84</td>
<td>39.36</td>
<td>37.48</td>
<td>38.14</td>
</tr>
<tr>
<td>Hill-Valley</td>
<td>4.50</td>
<td>2.10</td>
<td>2.00</td>
<td>3.40</td>
<td>2.15</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>11.30</td>
<td>14.30</td>
<td>13.10</td>
<td>19.16</td>
<td>14.30</td>
</tr>
<tr>
<td>Iris</td>
<td>8.33</td>
<td>5.33</td>
<td>2.00</td>
<td>6.75</td>
<td>4.67</td>
</tr>
<tr>
<td>Liver</td>
<td>36.33</td>
<td>37.10</td>
<td>38.60</td>
<td>42.77</td>
<td>38.86</td>
</tr>
<tr>
<td>Musk1</td>
<td>8.60</td>
<td>7.20</td>
<td>8.80</td>
<td>9.13</td>
<td>6.52</td>
</tr>
<tr>
<td>Sonar</td>
<td>23.14</td>
<td>12.50</td>
<td>25.00</td>
<td>17.10</td>
<td>14.12</td>
</tr>
<tr>
<td>Vehicle</td>
<td>29.10</td>
<td>31.60</td>
<td>30.00</td>
<td>55.72</td>
<td>22.98</td>
</tr>
<tr>
<td>Error Rate Average</td>
<td>24.71</td>
<td>20.70</td>
<td>21.03</td>
<td>23.57</td>
<td>19.50</td>
</tr>
</tbody>
</table>

![Figure 2: The average of classification error rate.](image)
These results show clearly the efficiency of our proposed approach to provide good performance related to classification rate and execution time by reducing the feature cardinality. SVM cannot find the optimal hyper plane with small numbers of samples (only 150 samples in Iris dataset), also the result is not superior when the number of features is very small such as Balance and Iris datasets (where only 4 features), so the results are trained without feature selection method. But the rest of the datasets selected, the proposed approach improves classification accuracy rates compared to those without feature selection. In some cases of the classification result is not well improved such as Vehicle dataset, but the experiment time of the same dataset is reduced.

Figures 2-3 show the classification accuracy rate and some techniques on tested datasets. It show clearly the efficiency of our proposed approach in classification phase.

CONCLUSIONS

In this paper SVM approach used to solve the reduction of features extraction and acceleration of classification in biometric identification problems. The results indicate that separating the database into subsets will decrease the complexity time and good classification. The paper provides improvements over SVM both in terms of accuracy and time.

REFERENCES


[26] Grochowski, M. and Duch, W. Projection Pursuit Constructive Neural Networks Based on Quality of Projected Clusters. Proceeding of the International Conference on Artificial Neural Networks-ICANN,


