Reduced Data Based Improved MEB/L2-SVM Equivalence

Lachachi Nour-Eddine and Adla Abdelkader

Computer Science Department, Oran University, Algeria
{Lach_Nour,AekAdla}@Yahoo.fr

Abstract. As a powerful tool in machine learning, support vector machine (SVM) suffers from expensive computational cost in the training phase due to the large number of original training samples. In addition, Minimal Enclosing Ball (MEB) has a limitation with a large dataset, and the training computational increases as data size becomes large. This paper presents an improved two approaches based SVMs reduced to Minimal Enclosing Ball (MEB) problems. These approaches find the concentric balls with minimum volume of data description to reduce the chance of accepting abnormal data that contain most of the training samples. Our study is experimented on speech information to eliminate all noise data and reducing time training. Numerical experiments on some real-world datasets verify the usefulness of our approaches for data mining.

Keywords: Support Vector Machines (SVMs), Minimal Enclosing Ball (MEB), Core-set.

1 Introduction

The theory of Support Vector Machines was introduced by Vapnik and was developed from the theory of Structural Risk Minimization [1]. SVMs learn the boundary regions between samples belonging to two classes by mapping the input samples into a high dimensional space, and seeking a separating hyperplane in this space. The separating hyperplane is chosen in such a way as to maximize its distance from the closest training samples (a quantity referred to as the margin).

Training a SVM involves solving a constrained quadratic programming problem, which requires large memory and enormous amounts of training time for large-scale problems. Goal is to find a separation hyperplane which implicates a $N \times N$ matrix density, where N is the number of points in the dataset. This needs more computational time and memory for large datasets, so the training complexity of SVM is highly dependent on the size of a dataset. Then, we partition the data in several data sources and we train them by Support Vector Machines (SVMs) using Fuzzy C-Mean Clustering Algorithm.

Here, we improve a technique cited in [2]. The improvement has based from an entropy algorithm that consider both Lagrangian duality and the Jaynes' maximum

J.F. Martínez-Trinidad et al. (Eds.): MCPR 2014, LNCS 8495, pp. 1–10, 2014.

entropy principle. The idea is to use information entropy and maximum entropy formalism in the solution of nonlinear programming problems [3].

Computation of such SVMs lead to find a Core-Set for the image of the data in a feature space. Thus, an alternative method is presented based on an equivalence between SVMs and Minimal Enclosing Ball (MEB) problems from which important improvements on training efficiency has been reported [4,5] for large-scale datasets. The study focus on multi-class problems where two methods explored to extend binary SVMs to the multi-category setting, which preserve the equivalence between the model and MEBs.

Algorithms to compute SVMs based on the MEB equivalence are based on the greedy computation of a core-set, a typically small subset of the data which provides the same MEB as the full dataset. Then, we formulate new multiclass SVM problem using core-sets for reduce large datasets which can be considered optimally matched to the input demands of different background architectures of speaker verification or Language Identification systems. The core idea of these two approaches cited above is to adopt multiclass SVMs formulation and Minimal Enclosing Ball to reduce dataset without influence data noise.

Along the whole paper, we define the variables as follows:

- f: Index of feature; F: number of features.
- n: Index of feature dimension; N: dimensionality of feature.
- l: Index of Classe; L: number of Classes.

2 L2-Support Vector Machines (L2-SVMs)

In [2] it is shown that for a binary classification, the L2-SVM build a separating hyperplane $f(z)$ by solving the following quadratic program:

$$\begin{gathered} \min_{w,b,\rho,\xi} \frac{1}{2}\left(\|w\|^2 + b^2 + C\sum_{f=1}^{F}\xi_t^2\right) - \rho \\ st:\ y_f f(z_f) \geq \rho - \xi_f \quad f = 1,\dots,F \end{gathered} \tag{1}$$

And for a given training task having L classes, these label vectors are chosen out of the definite set of vectors $\{y_1, y_2, \dots, y_F\}$. Hence, we can define the primal for the learning problem for the L2-SVM Multi-class classification as

$$\begin{gathered} \min_{w,b,\rho,\xi} \frac{1}{2}\left(\|W\|^2 + \|b\|^2 + C\sum_{f=1}^{F}\xi_f^2\right) - \rho \\ st\colon y_f^T\left(W^T z_f + b\right) \geq \rho - \xi_f^2 \geq 0 \quad f = 1,\dots,F \end{gathered} \tag{2}$$

where $z_f = \phi(x_f)$

After introducing Lagrange multipliers, for the both problem, we conduct to solve

$$\begin{gathered} \min_{\alpha} \sum_{f=1}^{F}\sum_{f'}^{F} \alpha_f \alpha_{f'} K_{ff'} \\ st\colon \alpha_f \geq 0,\ \sum_{f=1}^{F}\alpha_f = 1 \end{gathered} \tag{3}$$

where $K_{ff'} = y_f y_{f'} k(x_f, x_{f'}) + y_f y_{f'} + \frac{\delta_{ff'}}{C}$ for binary classification and $K_{ff'} = y_f^T y_{f'} k(x_f, x_{f'}) + y_f^T y_{f'} + \frac{\delta_{ff'}}{C}$ for multi-class classification. $\delta_{ff'}$ is the Kronecker delta function $(\delta_{ff'} = 1 \; if \; f = f' \; otherwise \; 0)$ and $k(x_f, x_{f'})$ implements the dot-product $z_f^T z_{f'}$.

We note that there are two types of extensions to build Multi-Class SVMs [6,7]. The first is One-Versus-One approach (OVO) that use several binary classifiers, separately trained and joined into a multi-category decision function. The second is One-Versus-All approach (OVA), where a different binary SVM is used to separate each class from the All.

3 Minimal Enclosing Balls (MEB)

MEB has originally introduced to estimate the support of a high dimensional distribution [8]. Suppose we have a set of F independent and identically distributed observations $\{x_f\}_{f=1}^{F}$ from an unknown distribution function P. The MEB algorithm seeks to find a minimal region R, which surrounds almost all the data points. This approximated region lead to enclose with high probability the test examples presuming that they are drawn from the same probability distribution P.

Denoting training data set as $S = \{\tilde{z}_f = \phi(x_f)\}_{f=1}^{F}$. Let $\tilde{Z}$ a space equipped with a dot product $\tilde{z}_f^T \tilde{z}_{f'}$ that corresponding to norm $\|\tilde{z}\|^2 = \tilde{z}^T \tilde{z}$. We define the ball $\mathcal{B}(c, R)$ of center $c \in \tilde{Z}$ and radius R in $\mathbb{R}$ as the subset of points $\tilde{z} \in \tilde{Z}$ for which $\|\tilde{z} - c\|^2 \leq R^2$. The minimal-enclosing ball of a set of points $S = \{\tilde{z}_f\}_{f=1}^{F}$ in $\tilde{Z}$ is in turn the ball $\mathcal{B}^*(S, c^*, R^*)$ of smallest radius that contains S, that is, the solution to the following optimization problem.

$$\begin{gathered} \min_{R,c} R^2 \\ st\colon \|\tilde{z} - c\|^2 \leq R^2 \quad \forall\, \tilde{z} \in S \end{gathered} \tag{4}$$

After introducing Lagrange multipliers, we obtain from the optimality conditions the following dual problem

$$\begin{gathered} \min_{\alpha} \sum_{f=1}^{T} \sum_{f'=1}^{F} \alpha_f \alpha_{f'} \tilde{z}_f^T \tilde{z}_{f'} - \sum_{f=1}^{F} \alpha_f \tilde{z}_f^T \tilde{z}_f \\ st\colon \alpha_f \geq 0,\ \sum_{f=1}^{F} \alpha_f = 1 \end{gathered} \tag{5}$$

if we consider that $\sum_{f=1}^{F} \alpha_f \tilde{z}_f^T \tilde{z}_f = \kappa$ a, we can drop it from the dual objective in Eq. (1), we obtain a simpler QP problem

$$\begin{gathered} \min_{\alpha} \sum_{f=1}^{F} \sum_{f'=1}^{F} \alpha_f \alpha_{f'} \tilde{z}_f^T \tilde{z}_{f'} = \min_{\alpha} \sum_{f=1}^{F} \sum_{f'=1}^{F} \alpha_f \alpha_{f'} \widetilde{K}_{ff'} \\ st\colon \alpha_f \geq 0,\ \sum_{f=1}^{F} \alpha_f = 1 \end{gathered} \tag{6}$$

In [9] it is shown that the primal variables c and R can be recovered from the optimal α as $c = \sum_{f=1}^{F} \alpha_f \tilde{z}_f$, $R = \sqrt{\sum_{f=1}^{F} \sum_{f'=1}^{F} \alpha_f \alpha_{f'} \tilde{z}_f^T \tilde{z}_{f'}}$ and the main appeal of the L2-SVM implementation Eq. (3) is that it supports a convenient equivalence to a Minimal Enclosing Ball (MEB) problem Eq. (6) when the kernel used in the SVM is normalized, that is $k(x, x) = \kappa \ \forall\, x \in X$, where κ is a constant. The advantage of this equivalence is that the Bãdoiu and Clarkson algorithm [10] can efficiently approximate the solution of a MEB problem with any degree of accuracy.

Core-Set

Bãdoiu and Clarkson [10] define the Core-Set of S as a set $C_S \subset S$ if the Minimal Enclosing Ball computed over C_S is equivalent to the Minimal Enclosing Ball considering all the points in S. A ball $\mathcal{B}(c, R)$ is said an ϵ-*approximation* to the Minimal Enclosing Ball $\mathcal{B}^*(S, c^*, R^*)$ of S if $R \leq R^*$ and it contains S up to precision ϵ, that is $S \subset \mathcal{B}(c, (1+\epsilon)R)$. Consequently, a set $C_{S,\epsilon}$ is called a ϵ-*core-set* if the Minimal Enclosing Ball of $C_{S,\epsilon}$ is a ϵ-approximation to $\mathcal{B}^*(S, c^*, R^*)$.

In [10] the most usual version of the algorithm is presented.

4 Improved MEB/L2-SVM Equivalence Algorithm

Calculating the Lagrange multipliers leads to a simpler QP problem Eq. (3) or Eq. (6) with non-negative constraints and one normality condition, which is one of the difficulties in the original MEB algorithm. The improved MEB algorithm present a simple and efficient algorithm, which takes advantage of the features of problem Eq. (6). We derive an entropy-based algorithm for the considered problem by means of Lagrangian duality and the Jaynes' maximum entropy principle. The idea is to use the information entropy and maximum entropy formalism in the solution of nonlinear programming problems [3].

Consider that the MEB QP problem Eq. (6) written as the following form:

$$\begin{aligned} &\min_\alpha L(\alpha) = \sum_{f=1}^{F} \sum_{f'=1}^{F} \alpha_f \alpha_{f'} \widetilde{K}_{ff'} \\ &st\colon \alpha_{f'} \geq 0,\ \sum_{f'=1}^{F} \alpha_{f'} = 1 \text{ and } \tilde{z}_f^T \tilde{z}_{f'} = \widetilde{K}_{ff'} \end{aligned} \tag{7}$$

From the constraints of optimization problem Eq. (7), we know that the dual variables go into the range [0, 1] and sum to one, so they meet the definition of probability. Our approach to the solution of Eq. (7) based on a probabilistic interpretation show that the center of the ball represents the mean vector of the images of all data points and the Lagrange multiplier α_f represents the probability that x_f is a support vector SV. Hence, we may consider searching for the MEB as a procedure of probability assignments, which should follow the Jaynes' maximum entropy principle [3]. Thus instead of QP problem Eq. (7), we construct a composite minimization problem:

$$\begin{aligned} \min\ & L_{\mathcal{P}}(\alpha) = L(\alpha) + H(\alpha)/\mathcal{P} \\ st\colon & \alpha_{f'} \geq 0,\ \sum_{f'=1}^{F} \alpha_{f'} = 1 \end{aligned} \tag{8}$$

Where $\mathcal{P}$ is a non-negative parameter, and

$$H(\alpha) = \sum_{f'=1}^{F} \alpha_{f'} \ln \alpha_{f'} \tag{9}$$

From information theory perspectives, $H(\alpha)$ represents an information entropy of the multipliers $\alpha_{f'}$. The additional term $H(\alpha)/\mathcal{P}$ is commensurate with the application of an extra criterion of minimizing the multipliers entropy to the original MEB QP problem Eq. (7). It is intuitively obvious that the entropy term on the solution of Eq. (8) will diminish as p approaches infinity.
To solve this problem we introduce the Lagrangian

$$L_{\mathcal{P}}(\alpha, \beta) = L(\alpha) + \frac{H(\alpha)}{\mathcal{P}} + \beta\left(\sum_{f'=1}^{F} \alpha_{f'} - 1\right) \tag{10}$$

where β is a Lagrange multiplier. Setting to zero the derivative of $L_{\mathcal{P}}(\alpha, \beta)$ with respect to α and β, respectively, leads to

$$\frac{\partial L}{\partial \alpha_{f'}} - \frac{1}{\mathcal{P}}\left(1 + \ln \alpha_{f'}\right) + \beta = 0 \tag{11}$$

and

$$\sum_{f'=1}^{F} \alpha_{f'} = 1 \tag{12}$$

Solving Eq. (11) for $\alpha_{f'}$, $f' = 1,2, \dots, F$

$$\alpha_{f'} = e^{\left(\mathcal{P}\left(\frac{\partial L}{\partial \alpha_{f'}} + \beta\right) - 1\right)} \tag{13}$$

Substituting α from Eq. (13) into Eq. (12), we obtain

$$e^{(\mathcal{P}\beta - 1)} \sum_{f'=1}^{F} e^{\left(\mathcal{P}\frac{\partial L}{\partial \alpha_{f'}}\right)} = 1 \tag{14}$$

Between Eq. (13) and Eq. (12), we eliminate the term $e^{(\mathcal{P}\beta - 1)}$ to give

$$\alpha_{f'} = \frac{e^{\left(\mathcal{P}\frac{\partial L}{\partial \alpha_{f'}}\right)}}{\sum_{f'=1}^{F} e^{\left(\mathcal{P}\frac{\partial L}{\partial \alpha_{f'}}\right)}} \tag{15}$$

By optimization problem Eq. (7), we have

$$L_{\alpha_{f'}}(\alpha) \equiv \frac{\partial L}{\partial \alpha_{f'}} = 2 \sum_{f=1}^{F} \alpha_f K_{ff'} \tag{16}$$

Thus, we obtain the iterative formula

$$\alpha_{f'}^{(k+1)} = \frac{e^{\left(\mathcal{P}^{(k)} L_{\alpha_{f'}}\left(\alpha^{(k)}\right)\right)}}{\sum_{f'=1}^{F} e^{\left(\mathcal{P}^{(k)} L_{\alpha_{f'}}\left(\alpha^{(k)}\right)\right)}} \tag{17}$$

Based on formulas Eq(s) (8) – (11), we obtain the entropy-based iterative algorithm for the solution of optimization problem Eq. (7) as follows:

Algorithm 1 Entropy-based iterative algorithm

1: Let $\mathcal{P}^{(0)} = 0$; from Eq. (15) we get $\alpha_{f'}^{(0)} = 1/F$;
$f' = 1,2,\ldots,F$;
let $\Delta\mathcal{P} \in (0,+\infty)$ and set $k = 0$

2: Based on formulas Eq. (16) and Eq. (17),
compute $\alpha_{f'}^{(k+1)}$, $f' = 1,2,\ldots,F$; let $\mathcal{P}^{(k+1)} = \mathcal{P}^{(k)} + \Delta\mathcal{P}$

3: if Stop criteria satisfied, the stop; otherwise,
we set $k = k + 1$, then return to step **2**

In short, we start with rough estimates of Lagrange multipliers, calculate improved estimates by iterative formula Eq. (17), and repeat until some convergence criterion has met.

Here, we note an important deduction that through the improved estimation of Lagrange multipliers, the Bãdoiu and Clarkson algorithm [10] is improved.

5 Reduced Data Approaches

The key idea of our method is to cast a L2-SVM as a MEB problem reduced in a Core-Set by using a feature space $\tilde{Z} = \phi(X)$ where the training examples are embedded via a mapping ϕ. Hence, we first formulate an algorithm to compute the MEB of the images $\tilde{S}$ of S in $\tilde{Z}$ when S is decomposed in a collection of subsets S_p. Then we will instantiate the solution for classifiers supporting the reduction to MEB problems. The algorithm is based on the idea of computing Core-Sets $\mathcal{C}_p$ for each set $\tilde{S}_p = \phi(S_p)$ and taking its union $\mathcal{C} = \cup_p \mathcal{C}_p$ as an approximation to a Core-Set for $\tilde{S} = \cup_p S_p$. In a first step the algorithm extracts a Core-Set for each subset S_p. In the second step, the MEB of the union of the Core-Sets is computed.

The decomposition of S in a collection of subsets S_p by Fuzzy C-Means (FCM) method clustering which allows one piece of data to belong to two or more clusters [11,12].

From the section 2 the kernel $\tilde{k}(x_f, x_{f'}) = y_f y_{f'} k(x_f, x_{f'}) + y_f y_{f'} + \frac{\delta_{ff'}}{C}$ for the binary case (OVO approach) and the kernel $\tilde{k}(x_f, x_{f'}) = y_f^T y_{f'} k(x_f, x_{f'}) + y_f^T y_{f'} + \frac{\delta_{ff'}}{C}$ in the multi-category case (OVA approach). In addition, for the both binary (OVO) and multi-category (OVA) Multi-Class case, we depict algorithm 2 and algorithm 3 respectively.

Algorithm 2 Computation of the MEB using OVO approach

1: **for** Each subset S_p , $p = 1, \dots, P$ **do**
2: **for** Each Class $l = 1, \dots, L-1$ **do**
3: **for** Each Class $l' = l+1, \dots, L$ **do**
4: Let $S_p^{ll'}$ the subset of S_p corresponding to class l and l'.
5: Label $S_p^{ll'}$ using the standard binary codes $+1$ and -1 for class l and l' respectively
6: Compute a core-set $C_p^{ll'}$ of $S_p^{ll'}$ [10] using the Kernel $\tilde{k}(x_f, x_{f'}) = y_f y_{f'} k(x_f, x_{f'}) + y_f y_{f'} + \frac{\delta_{ff'}}{C}$
7: **end for**
8: **end for**
9: Take the union of the core-set inferred for each pair of classes $C_p = C_p^{ll'} \cup \dots \cup C_p^{ll'}$
10: **end for**
11: Join core-set $C_S = C_1 \cup \dots \cup C_P$.
12: Compute the minimal enclosing ball of C_S using the same kernel $\tilde{k}$

Algorithm 3 Computation of the MEB using OVA approach

1: **for** Each subset S_p , $p = 1, \dots, P$ **do**
2: Label each example $x_f \in S_p$ with the code y_{fl} assigned to the class of x_f and let y_f such label
3: Compute a core-set C_p of S_p [10] using the kernel

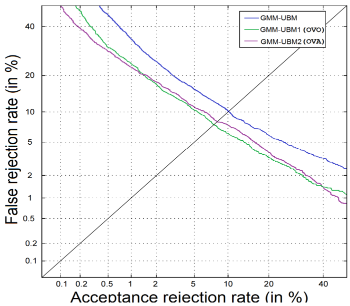
$$\tilde{k}(x_f, x_{f'}) = y_f^T y_{f'} k(x_f, x_{f'}) + y_f^T y_{f'} + \frac{\delta_{ff'}}{C}$$

4: **end for**
5: Join the core-sets $C_S = C_1 \cup \dots \cup C_P$.
6: Compute the minimal enclosing ball of C_S using the same kernel $\tilde{k}$

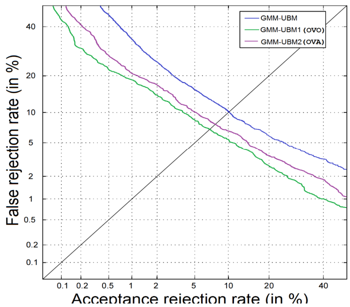
6 Experiments

This section presents the performance of text-independent speaker verification task based on the Gaussian Mixture Model – Universal Background Model (GMM-UBM) system described in [13]. As in [2], we compare the performance of speaker verification system with three UBMs, the first one was created directly from the Speaker Recognition corpus [14], consists of telephone speech. The two last later is the reduced first one from the application of our two algorithms developed above. The kernel used for the two algorithms is the Gaussian Radial Basis Function with a fixed value of σ with 0.50. We have trained a 512-mixture gender-independent from each UBM with diagonal covariance matrices. Speaker GMMs has trained by adapting only the mean vectors from the UBM using a relevance factor r of 16. The experiment based on the improved estimation of Lagrange multipliers cited previously is compared with the result in [2] where we used the same corpus.

The two figures below shows the detection error tradeoff (DET) curves for the three systems. The (A) represent the result issue from our previous study for the same corpus in [2] and (B) represent our experiences for the improved system enounced in this paper. In (A) we see that the system based reduced GMM-UBM2 from One-Versus-All multiclass L2-SVM outperforms the GMM-UBM with an equal-error-rate (EER) of 8.55 %, compared to 10,13 % of the GMM-UBM. The system based reduced GMM-UBM1 from One-Versus-One multiclass L2-SVMs exhibits the best performance with an EER of 7.60 %. On the other hand, in (B), the same system give an improved rate with an equal-error-rate (EER) of 6.15 %, compared to 10.13 % of the GMM-UBM. However, the system based reduced GMM-UBM2 from One-Versus-All L2-SVMs had given an EER of 7.80 %, that is also give a performance from the result issued in [2] for the same approach.



(A)



(B)

Fig. 1. Detection error tradeoff (DET) curves for the speaker verification system using three UBMs

In comparison from the two curves, we note that the best result given is for the system based reduced GMM-UBM1 from One-Versus-One in the order of 1.45 %.

7 Conclusion

In this paper, we proposed two algorithms that compute an approximation to the minimum enclosing ball of a given finite set of vectors based Core-Set for reducing huge dataset. Both algorithms is especially well-suited for large-scale instances of the Minimal Enclosing Ball (MEB) problem and can compute a small core set whose size depends only on the approximation parameter.

We have explored two improved methods based on the computation of Core-Sets to train multi-category SVM models when the set of examples is fragmented. The main contribution has been shown through our experiments, that the improved methods proposed give the best performance with a reproduction of high solution accuracy where the noisy sample in huge data set are eliminated, without complex and costly computation. SVMs based on Core-Sets have shown however important advantages in large-scale applications, which can hence be extended to distributed data-mining problems. A real contribution of this work has been an improved direct implementation of multi-category SVMs based Core-Sets supporting a reduction to a Minimal-Enclosing Ball (MEB) problem. Although the Core-Sets method exhibits always better prediction accuracy used with the OVO scheme, the direct implementation shows a lower complexity and it is better than the previous direct implementation proposed for MEB based SVMs.

References

1. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
2. Nour-Eddine, L., Abdelkader, A.: Reduced universal background model for speech recognition and identification system. In: Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Olvera López, J.A., Boyer, K.L. (eds.) MCPR 2012. LNCS, vol. 7329, pp. 303–312. Springer, Heidelberg (2012)
3. Templeman, A.B., Li, X.S.: A maximum entropy approach to constrained nonlinear programming. Engineering Optimization 12, 191–205 (1987)
4. Kocsor, A., Kwork, J., Tsang, I.: Simpler core vector machines with enclosing balls. In: ICML 2007, pp. 911–918. ACM (2007)
5. Cheung, P.M., Kwok, J., Tsang, I.: Core vector machines: Fast SVM training on very large datasets. Journal of Machine Learning Research (6), 363–392 (2005)
6. Nour-Eddine, L., Abdelkader, A.: Multi-class Support Vector Machines Methodology. In: 1er Congrès International sur les Modèles, Optimisation et Sécurité des Systèmes, ICMOSS 2010, Tiaret, Algérie, pp. 325–329 (2010)
7. Hsu, C., Lin, C.: A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks 13(2), 415–425 (2002)
8. Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation 13, 1443–1471 (2001)

9. Tsang, I., Kwok, J., Cheung, P.M.: Core vector machines: Fast SVM training on very large data sets. Journal of Machine Learning Research 6, 363–392 (2005)
10. Bãdoiu, M., Clarkson, K.L.: Optimal core-sets for balls. Computer Geometry Theory Application 40(1), 14–22 (2008)
11. Dunn, J.C.: A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics 3(3), 32–57 (1973)
12. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
13. Alkanhal, M., Alghamdi, M., Muzaffar, Z.: Speaker Verification based on Saudi Acceted Arabic Database. In: 9th International Symposium on Signal Processing and its Applications, ISSPA 2007, Sharjah, United Arab Emirate, pp. 1–4 (February 2007)
14. Speaker corpus in, http://www.ll.mit.edu/mission/communication/ist/corpora/SpeechCorpora.html