2011 First International Conference on Informatics and Computational Intelligence

Hirarchical Harmony Linear Discriminant Analysis

Masoud Niazi [€], Naemeh Ganoodi[Φ], Mona Yaghoubi[Ψ]
[€] Department of Computer Engineering, Islamic Azad University-Mashad branch, Mashad, Iran
[Φ] Faculty of Science, University of Birjand, Birjand, Iran
[Ψ] Department of Computer Engineering, Islamic Azad University-Neyshabour branch, Neyshabour, Iran
[€]masood.niazi@gmail.com, [Φ]n_ganoodi@hotmail.com, [Ψ]mona.yaghoobi@gmail.com

***Abstract*—Linear Discriminate Analysis is commonly used in feature reduction. Based on the analysis on the several limitations of traditional LDA, this paper makes an effort to propose a new way named Hierarchical Harmony Linear Discriminant Analysis (HH-LDA), which computes between class scatter matrixes optimally. It is reached by combining hierarchical scheme and Harmony Search (HS) algorithm. In this paper, a pre-processing step is proposed in order to increase accuracy of classification. The aim of this approach is finding a transformation matrix causes classes to be more discriminable by transforming data into the new space and consequently, increases the classification accuracy. This transformation matrix is computed through two methods based on linear discrimination. In the first method, we use class dependent LDA to increase classification accuracy by finding a transformation that maximizes the between-class scatter and minimizes within-class scatter using a transformation matrix. Because LDA cannot obtain optimal transformation, in the second method, Harmony Search is used to increase performance of LDA. Obtained results show that utilization of these pre-processing causes increasing the accuracy of different classifiers.**

Keywords-Harmony Search; Linear Discriminant Analysis; classification; pre-processing

I. INTRODUCTION

Linear Discriminate Analysis (LDA) is a well-known method for dimensionality reduction and classification that projects high-dimensional data onto a low-dimensional space where the data achieves maximum class separability [6,7,8]. The derived features in LDA are linear combinations of the original features, where the coefficients are from the transformation matrix [21]. The optimal projection or transformation in classical LDA is obtained by minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination. The optimal transformation is readily computed by solving a generalized eigen value problem. The original LDA formulation, known as the Fisher Linear Discriminate Analysis (FLDA) [1], deals with binary-class classifications. The key idea in FLDA is to look for a direction that separates the class means well (when projected onto that direction) while achieving a small variance around these means. Based on the fact that linear discriminate analysis does not lead to an optimal transforming necessarily, LDA causes three major problems. The first problem is observed when the number of features is more than the number of samples which is called Small Sample Size (SSS) Problem. A solution to the SSS problem is the one that use PCA as a pre-processing step to discard the null space of the within-class scatter matrix for dimension reduction [3]. The next problem is about the distribution of each class. LDA is optimal just in the case that each class has Gaussian distribution and also all classes have the same within-class covariance while having different means. Evolutionary algorithms are suggested to be used to solve this problem[20,21] The optimality criterion of LDA is the last problem. The formula is essentially based on the distance of samples, is not directly related to the classification accuracy [2]. Especially, since the between-class scatter is defined as the sum of all scatters between the means of any two classes, its maximization does not necessarily guarantee expected separation between any two classes in the output space.[21]

In this paper we propose a method to solve the problem of between class scatter by using hierarchical LDA. In proposed approach we build hierarchical scheme to calculate transformation matrix for each obtain class. Also pre-processing step is suggested before classification to compute a transformation matrix using training data with the aim of increasing discrimination of classes by transforming them into a new space. The transformation is independent of classifier and classifier type has no effect on computation of the transformation matrix, i.e., this transformation could be used for all classifiers.

The remaining of this paper is organized as follows: In the second section, proposed approach is investigated. In the third section, evaluation methods of proposed approach are discussed. The forth section represents empirical results and finally the last section contains the conclusion

II. LINEAR DISCRIMINATE ANALYSIS

Using LDA, we obtain a transformation matrix in order to map features from an h-dimensional space to a k-dimensional space ($k<h$) such that the samples or patterns of classes are well-separated in the new space. This transformation can be performed in a Class-Dependent (CD) or Class-Independent (CI) manner. Assume that W_{CI} and $W_{i,CD}$ are Class-Independent LDA (CI-LDA) transformation matrix and Class-dependent LDA (CD-LDA) transformation matrix in class i respectively. These matrices are computed by maximizing Fisher's criteria as shown in Eq. 1 [6,7,8]. where $S_{W,CI}$ is within-class scatter matrix for all classes

978-0-7695-4618-6/11 $26.00 © 2011 IEEE
DOI 10.1109/ICI.2011.22

(class-independent) and $S^{i}_{W,CD}$ is within-class scatter matrix for class *i* (class-dependent), as in Eq. 2 and 3.

$$J_{CI}(W) = \frac{\left|W_{CI}^{T} S_B W_{CI}\right|}{\left|W_{CI}^{T} S_{W,CI} W_{CI}\right|}$$
$$J_{CD}(W_i) = \frac{\left|W_{i,CD}^{T} S_B W_{i,CD}\right|}{\left|W_{i,CD}^{T} S^{i}_{W,CD} W_{i,CD}\right|} \quad (1)$$

$$S_{W,CI} = \sum_{i=1}^{c}\sum_{j=1}^{N_i}(x_i(j)-\mu_i)(x_i(j)-\mu_i)^T \quad (2)$$

$$S^{i}_{W,CD} = \sum_{j=1}^{N_i}(x_i(j)-\mu_i)(x_i(j)-\mu_i)^T \quad (3)$$

Where $x_i(j)$ shows j^{th} sample of class *i*, *c* is number of all classes, N_i is number of samples in class *i* and μ_i is the mean of class *i*. Between-class scatter matrix S_B is computed as Eq. 4.

$$S_B = \sum_{i=1}^{c}(\mu_i-\mu)(\mu_i-\mu)^T \quad (4)$$

Where μ represents the mean of all classes. The transformation matrix W_{CI} and $W_{i,CD}$ are found by eigen-decomposition of matrix $S^{-1}_{W,CI} \times S_B$ and matrix $S^{i^{-1}}_{W,CD} \times S_B$, respectively. New discriminative features *y* are derived from the original features *x* by $y = W \times x$.

III. Proposed Approach

LDA does not guarantee to find optimal subspace because of using sum of between class scatter as the scatters between the means of any two classes. In this paper, we suggest computing between class scatter in different steps. For this goal, hierarchical scheme and Harmony search are suggested to be used to divide all classes into two classes: A and B in each step. Class dependent transformation matrix is computed for obtained A and B classes (see Figure 1). Finally, for each class transformation matrix is computed by multiplying all transformations which are in parent nodes. By instance, in Figure 1, $W_{c3} = W_{135} * W_3$

As this method is used to improve performance of class dependent LDA, we call it Hierarchical Harmony search-Class-Dependent LDA (HHCD-LDA).

By computing proposed class dependent transformation matrix, class separation is increased. We also suggest using transformation matrix as a preprocessing in classification. The pre-processing step is used before classification in order to increase classification accuracy in which a transformation matrix is computed using training data and then data are prepared for classification by transforming it into the new space. In this step, the transformation is computed which causes maximum discriminated classes.

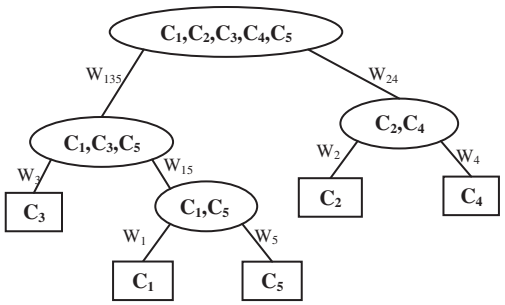

Figure 1. Example of five classes are divided by proposed approach

Two linear discriminate based methods are used to compute the transformation matrix. In the first method, LDA is used to compute the transformation matrix. LDA has been used for classification and dimension reduction. In this paper, LDA is used as pre-processing before classification to increase class discrimination.

Algorithm: Transformation of test set

Input: Test Set
Output: Transformed test set
Step0: for each class c_i of training data, compute its mean value $M=\{m_1,...,m_c\}$
Step1: transform each m_i to the new space $m_i' = m_i \times T_i$
Step2: for each $x \in X_{Test}$, do [step 3-5]
Step3: Compute set $X'=\{x'_1,...,x'_c\}$ where each x'_i is the transformation of x using the transformation matrix T_i
Step4: Construct the distance vector $D=\{d_1,...,d_c\}$ in which d_i represents the distance between x'_i and the mean value m'_i (computed in step 0).
Step5: Select x'_j as the transformed value of x where d_j is the minimum value in the distance vector D (i.e. $min(d_1,...,d_c)$)
Step6: End

Algorithm 1: Pseudo code for test set transformation

A. *Improving LDA Performance by HS*

Harmony search (HS) algorithm was recently developed in an analogy with music improvisation process where music players improvise the pitches of their instruments to obtain better harmony [19]. Harmony search algorithm had been very successful in a wide variety of optimization problems [2,18], presenting several advantages with respect to traditional optimization techniques such as the following [18]: (a) HS algorithm imposes fewer mathematical requirements and does not require initial value settings of the decision variables. (b) As the HS algorithm uses stochastic random searches, derivative information is also unnecessary. (c) The HS algorithm generates a new vector, after considering all of the existing vectors, whereas the genetic algorithm (GA) only considers the two parent vectors. These

features increase the flexibility of the HS algorithm and produce better solutions. The steps in the procedure of harmony search are shown in Fig. 2 [18].

1) Construction of the hypothesis space

The first step in HS is to define the encoding to describe any potential solution as a numerical vector. We use a binary vector to express an individual code which represents the status of each class. The length of individuals in the hypothesis space is the number of classes we decide to divide into two classes.

2) Fitness Function

The role of the Fitness function is to measure the quality of solutions. Each individual is a transformation matrix which is evaluated by the fitness function J_{CD} as below:

$$J_{CD}(W_i) = \frac{\left|W_{i,CD}^T S_B W_{i,CD}\right|}{\left|W_{i,CD}^T S_{W,CD}^i W_{i,CD}\right|} \quad (5)$$

Harmony search finds transformation matrix W that maximizes Eq. 5.

IV. EVALUATION OF PROPOSED APPROACH

To evaluate suggested methods, we use four datasets: Magic-Gamma-Telescope (M.G.T), Glass, Vowel and Waveform. The datasets used in this paper are obtained from UCI database [9], the characteristics of which are shown in Table 1.

TABLE I. THE CHARACTERISTICS OF UTILIZED DATASET

Dataset	Attribute	Samples	Classes
M.G.T.	10	19020	2
Glass	9	214	6
Vowel	12	9906	11
Waveform	40	5000	3

Also, different classifiers are used to evaluate proposed approach. Classifiers can be divided into 7 categories [12]: function-based classifier, rule-based classifier, tree-based classifier, lazy classifier, combine classifier, Baysian Classifier and Interval Classifier. One arbitrary method is selected from each category to be utilized in this paper which are RBFNetwork [10], JRIP [13], Naïve Bayes Tree (NBTree) [12], *k*-Nearest Neighbor (*k*-NN) [12], Classification Via Regression (CVR) [14], Naïve Bayes (NB) [11] and Hyper pipes (Hyper-p) [12] respectively. In k-NN, the parameter k is assumed to be 3. Evaluation criterion of SD [16] and Dunn [15] are also used to depict dispersion of data before and after transformation.

A. Dunn Index

Dunn is a validity index which attempts to identify compact and well-separated classes defined by Eq. 6 for a specific number of classes.

$$D_{nc} = \min_{i=1,\ldots,nc}\left\{\min_{j=i+1,\ldots,nc}\left(\frac{d(c_i,c_j)}{\max_{k=1,\ldots,nc} diam(c_k)}\right)\right\} \quad (6)$$

where *nc* is number of classes, *d(ci,cj)* is the dissimilarity function between two classes C_i and C_j defined by Eq. 7.

$$d(c_i,c_j) = \min_{x\in c_i, y\in c_j} d(x,y) \quad (7)$$

$$diam(c) = \max_{x,y\in c} d(x,y) \quad (8)$$

And *diam(c)* is the diameter of a class, which may be considered as a measure of dispersion of the classes. The diameter of a class *c* can be defined as Eq. 8. It is clear that if the dataset contains compact and well-separated classes, the distance between the classes is expected to be large and the diameter of the classes is expected to be small. Thus, based on the Dunn's index definition, we may conclude that large values of the index indicate the presence of compact and well-separated classes.

B. SD Index

The SD validity index is defined based on the concepts of the average scattering for classes and total separation between classes. Average scattering for classes is defined as Eq. 9 where V_i is the center of i^{th} class and X is overall data.

$$Scat(nc) = \frac{1}{nc}\sum_{i=1}^{nc}\left\|\sigma(v_i)\right\| / \left\|\sigma(X)\right\| \quad (9)$$

$$Dis(nc) = \frac{D_{\max}}{D_{\min}}\sum_{k=1}^{nc}\left(\sum_{z=1}^{nc}\left\|v_k - v_z\right\|\right)^{-1} \quad (10)$$

Total separation between classes is defined as Eq. 10. Where D_{max} and D_{min} are the maximum and minimum distance between class centers respectively.

$$D_{\max} = \max(\left\|v_i - v_j\right\|)\forall i, j \in \{1,2,\ldots,nc\} \quad (11)$$

$$D_{\min} = \min(\left\|v_i - v_j\right\|)\forall i, j \in \{1,2,\ldots,nc\} \quad (12)$$

Now, validity index SD is defined based on Eq. 9 and 10 as follows:

$$SD(nc) = \lambda.Scat(nc) + Dis(nc) \quad (13)$$

The first term *Scat(nc)* defined as Eq. 9 shows the average compactness of classes and intra-class distance. A small value of this term indicates compact classes. As the within-classes scatter increases and so classes become less compact, the value of *Scat(nc)* increases. The second term *Dis(nc)* shows the total separation between *nc* classes and inter-class distance. *Dis(nc)* is influenced by the geometry of the class centers. It increases with the number of classes. Based on Eq. 9 and 10, we can say that small values of the index show the presence of compact and well-separated classes. As the two terms of *SD* have the different ranges, the weighting factor λ is used in order to incorporate both terms in a balanced way. The number of classes, *nc*, minimizes the above index. So, it can be considered as an optimal value for the number of classes present in the dataset.

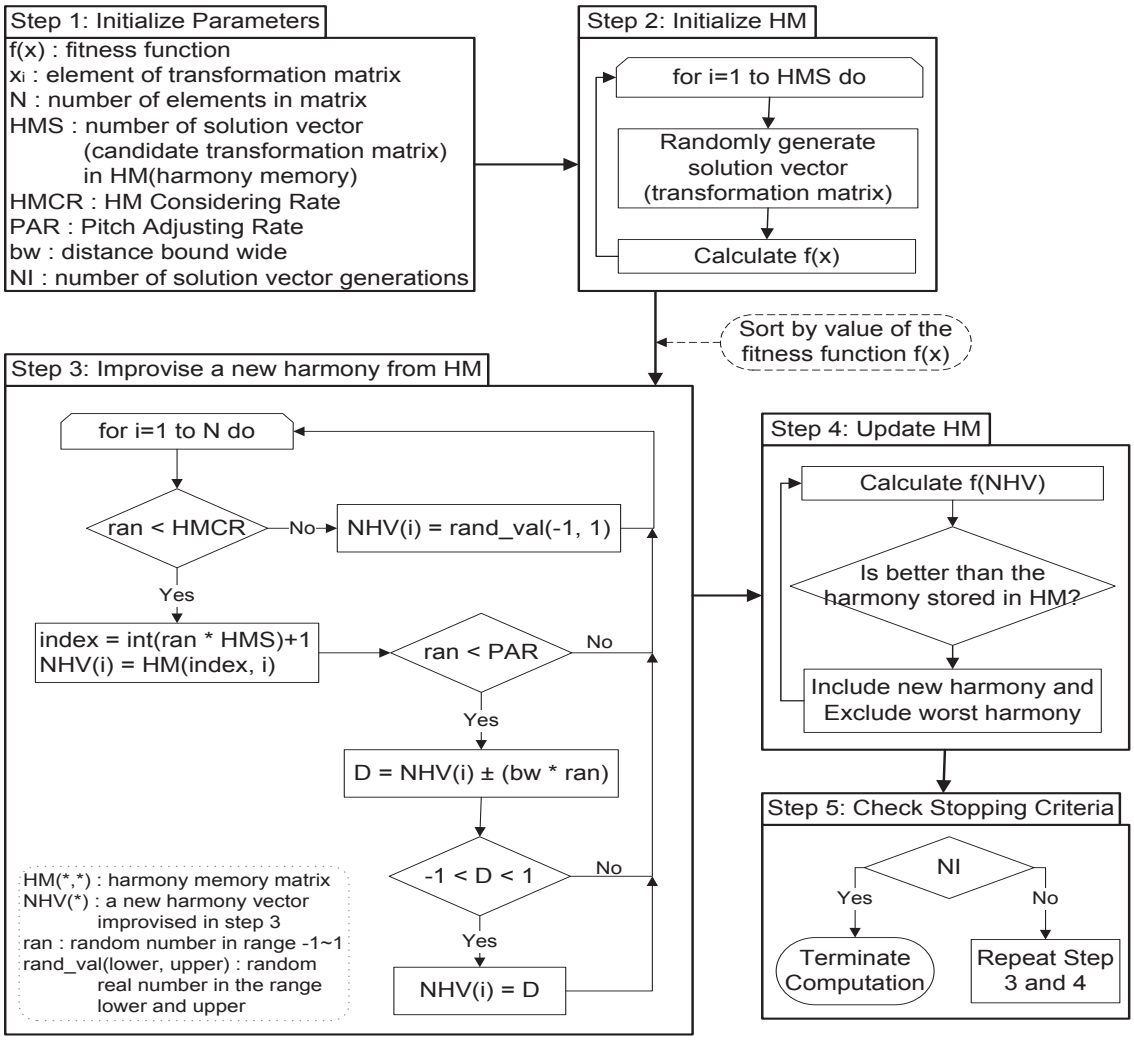

Figure 2. Harmony Search Approach

V. EXPERIMENTAL RESULT

We use 66% of the data as training set and the remaining are used in evaluation progress. An implementation of pre-processing is developed using Microsoft VC++6 and WEKA[17] is used for data classification. Yet, as shown in Table 4, classification accuracy increases in most of classifiers using the proposed pre-processing step. The results also show that HS obtain better transformation matrix and is more effective than standard CD-LDA.

TABLE II. COMPARISON OF INDEX DUNN BEFORE AND AFTER UTILIZING PRE-PROCESSING

	Normal	CD-LDA	HHCD-LDA
M.G.T.	0.482	0.493	0.502
Vowel	0.450	0.481	0.492
Glass	0.341	0.317	0.332
Waveform	0.986	0.988	0.996

TABLE III. COMPARISON OF INDEX SD BEFORE AND AFTER UTILIZING PRE-PROCESSING

	Normal	CD-LDA	HHCD-LDA
M.G.T.	0.557	0.489	0.408
Glass	0.31	0.241	0.251
Vowel	0.069	0.039	0.041
Waveform	0.371	0.36	0.322

For evaluating our pre-processing approaches, we computed Dunn and SD indexes for our dataset before and after applying transformation matrix. As shown, in Table 2 and Table 3, classes' dispersion after transformation are decreased.

TABLE IV. CLASSIFICATION ACCURACY WITH PRE-PROCESSING METHODS. THE BESTS ARE BOLD, THE SECONDS ARE UNDERLINE.

Classifier		Normal	CD-LDA	HHCD-LDA
NB	M.G.T.	72.491	86.856	**88.897**
	Vowel	59.643	**80.415**	78.261
	Glass	49.315	83.561	**85.742**
	Waveform	80.588	93.588	**94.113**
RBFN	M.G.T.	85.155	89.748	**91.758**
	Vowel	86.350	78.388	**88.427**
	Glass	57.534	82.191	**84.11**
	Waveform	83.352	92	**93.22**
3-NN	M.G.T.	82.542	89.516	**90.196**
	Vowel	89.614	81.602	**93.768**
	Glass	63.013	**91.780**	90.32
	Waveform	77.705	93.176	**93.765**
CVR	M.G.T.	85.588	**90.985**	90.907
	Vowel	75.964	83.382	**89.020**
	Glass	58.904	80.821	**84.201**
	Waveform	81.941	93.882	**94.721**
HayperP	M.G.T.	66.599	70.249	**72.625**
	Vowel	34.421	49.258	**84.866**
	Glass	52.054	47.945	**54.201**
	Waveform	45	48.882	**51.012**
NBTree	M.G.T.	85.170	**90.211**	89.052
	Vowel	83.382	85.223	**88.130**
	Glass	61.643	**78.082**	77.201
	Waveform	80	93.588	**95.201**
JRIP	M.G.T.	84.231	90.258	**91.170**
	Vowel	62.908	81.602	**87.240**
	Glass	61.643	80.821	**82.221**
	Waveform	79.764	93.588	**95.876**

VI. CONCLUSION

In this paper, two pre-processing methods are suggested before classification to increase classification accuracy. In this pre-processing step, transformation of data into a new space causes increasing discrimination of classes which is performed by using two methods based on linear discrimination. In the first method, class dependent linear discriminate analysis is used to compute transformation matrix, and in the second method, Harmony search is combined with hierarchical scheme to increase performance of LDA. To evaluate suggested methods, we use Magic gamma telescop, Glass, Vowel and Waveform from UCI machine learning dataset. The results show that utilizing these pre-processing methods increases classification accuracy. The results also show that combination of HS and LDA computes more effective transformation than LDA solely.

REFERENCES

[1] R. Fisher, "The use of multiple measurements in taxonomic problems.", Annals of Eugenics, 7, pp. 179–188, 1936.

[2] J.H. Kim, Z.W. Geem, E.S. Kim, "Parameter estimation of the nonlinear Muskingum model using harmony search", J. Am. Water Resour. Assoc. 37 (5) , pp. 1131–1138, 2001.

[3] P. Belhumeur, J. Hespanha, , D. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection." IEEE Trans. Pattern Anal. Machine Intell 19 , 7, pp. 711–720, 1997.

[4] A. P. Engelbrecht, "Particle swarm optimization: Where does it belong?". In Proceedings of the IEEE Swarm Intell. Symp., pp 48–54, 2006.

[5] A. Baraldi, and P. Blonda, "A Survey of Fuzzy Clustering Algorithms for Pattern Recognition—Part I and II". IEEE Tran. on Pattern Analysis and Machine Intelligence, 2005, vol 27.

[6] R. O. Duda, P. E. Hart and D. Stork , Pattern Classification. Wiley, 2000.

[7] K. Fukunaga, "Introduction to Statistical Pattern Recognition.", Academic Press, San Diego, California, USA, 1990.

[8] T. Hastie, and R. Tibshirani, "Discriminant analysis by Gaussian mixture*s",* Journal of the Royal Statistical Society series B, 1996, 58:158–176.

[9] http://archive.ics.uci.edu/ml/datasets.html

[10] J.C., Luo, C.H., Zhou, and Y., Leung , "A Knowledge Integrated RBF Network for Remote Sensing Classification*"*. In Proceedings of the 22nd Asian Conference on Remote Sensing, Singapore, November 5-9, 2001.

[11] G.H. John, and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers". In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. San Mateo, pp. 338-345, 1995.

[12] H. Witten, and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques. second edition, Elsevier Science & Technology, 2005.

[13] W.W. Cohen," Fast Effective Rule Induction". In Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann Publishers, Tahoe City, California,1995, pp. 115-123.

[14] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I.H. Witten, "Using model trees for classification." . In Proceedings of the Machine Learning ,1998, vol.32, no.1,pp. 63-76.

[15] J.C. Dunn, "Well Separated Clusters and Optimal Fuzzy Partitions", Journal of Cybernetics, 1974, pp. 95-104.

[16] M. Halkidi, , M. Vazirgiannis and I. Batistaks "Quality Scheme Assessment in the Cluster Process". In Proceedings of the PKDD, Lyon, France, 2000.

[17] http://www.cs.waikato.ac.nz/ml/weka/

[18] K.S. Lee, Z.W. Geem, "A new meta-heuristic algorithm for continues engineering optimization: harmony search theory and practice", Comput. Meth. Appl. Mech. Eng. 194 , 2004, pp. 3902–3933.

[19] Z.W. Geem, J.H. Kim, G.V. Loganathan, "A new heuristic optimization algorithm: harmony search", Simulation 76 (2), 2001, pp. 60–68.

[20] H.Moeinzadeh, M-M.Mohammadi, A.Akbari, B.Nasersharif , "Evolutionary-Class Independent LDA As a Pre-Process for Improving Classification ", 11th Annual conference on Genetic and Evolutionary Computation (GECCO 2009), Montreal, Canada

[21] H.Moeinzadeh, E.Asgarian, M.Zanjani, A.Rezaee, M.Seidi , "Combination of Harmony Search and Linear Discriminate Analysis to Improve Classification" , 3th IEEE International Conferance on Modelling & Simulation (AMS 2009), Baundung, Indonesia