To Increase Quality of Feature Reduction Approaches Based on Processing Input Datasets

Shervan Fekri Ershad
Department of computer science, engineering and IT
Shiraz University
Shiraz/Iran
shfekri@shirazu.ac.ir

Sattar Hashemi
Department of computer science, engineering and IT
Shiraz University
Shiraz/Iran
S_hashemi@shirazu.ac.ir

***Abstract* - Feature extraction is an important concept which is used for reducing features to decrease the complexity and time of classification. So far some methods have been presented for solving this problem but almost all of them just presented a fix output for each input dataset that some of them aren't satisfied cases for classification. In this paper first we present a new concept called Dispelling Classes Gradually (DCG) to increase separability of classes based on their labels. Next we will use this method to process input dataset of the feature reduction approaches to decrease the misclassification error rate of their outputs more than when output is achieved without any processing. In addition our method has a good quality to collate with noise based on adapting dataset with feature reduction approaches. The results compare two conditions (With process and without that) to support our idea.**

Keywords- Feature reduction, Dispelling classes gradually, Feature Extraction

I. INTRODUCTION

Since the middle of 20th century when artificial intelligence science was established, classification methods were too important. By the pass of time datasets that were presented for classification got more complex than past. One of the items that increase the complexity of dataset is the number of features. So the AI's scientists induct a new concept called feature reduction in the literature of AI (Fukunnaga 1991) [4]. If we describe input data as a matrix like X= $\{x_1 \ldots . x_n\} R^{m*N}$, where N is the sample number and m is the original feature dimensions, So the purpose of linear feature extraction is to search for a projection matrix W $\in R^{m'*m}$ that transforms $x_i \in R^m$ into a desired low-dimensional representation $y_i \in R^{m'}$, where $m' \ll m$ *and* y_i =Wx_i.

Some of the popular feature extraction methods are principal component analysis (PCA) (Joliffe, 1986) [8], linear discriminant analysis (LDA) (Fukunnaga, 1991) Spectral regression discriminant analysis (SRDA) (CAI Et Al, 2008) and Renyi's entropy discriminant analysis (REDA) (Xiao-Tong Bao-Gang, 2009) [1]. All of these methods are good to compute the quality of our idea.

Now the output dataset Y is in a low-dimensional so the complexity of classification Y and the time of that are less. It's easy to know that feature extraction wastes some pieces of the dataset's information. All of the methods for feature extraction are common in this object, so algorithms try to find a better formula for matrix W to decrease the amount of waste information to decrease the miss classification error rate of output. Almost none of the previous methods have a good generalization for each datasets. In this paper we present a method to adapt the input dataset with feature extraction methods. First we present a new concept called Dispelling Classes Gradually (DCG) to increase separability of classes based on their labels and then we process the original input dataset with DCG and create a new dataset as an input dataset of the feature extraction methods to decrease the amount of misclassification error rate on its output rather than original situation.

II. DCG CONCEPT

In the numerical datasets, misclassification error rate decreases with increasing the distance between classes. So it's necessary and beneficial to increase these distances. In the basic mathematics it's determined that if two different numbers like P_1 *and* P_2 be subtracted two another fix numbers like Q_1 *and* Q_2 , respectively, and this action get repeated for α times, where$Q_1 < Q_2$, it's authenticate that in extreme , the distance between $P_1 and P_2$will be more than first generation. This principal is shown in figure (1).

If we assume class labels to numeric labels and then assign Q_i for class labels, separability of classes increases at extreme situation by subtracting each class label from its sample value in available features. This subject is described in (1).

$$\left(X_i - \alpha L_{X_i}\right) - \left(X_j - \alpha L_{X_j}\right) > X_i - X_j \qquad (1)$$

Where, X_i is the i_{th} sample and X_j is the j_{th} sample and L_{X_i} is the class label of X_i and L_{X_j} is the class label of X_j . Figure (2) shows the results of applying (1) on Iris (UCI Dataset) on two features.

978-1-61284-486-2/11/$26.00 ©2011 IEEE

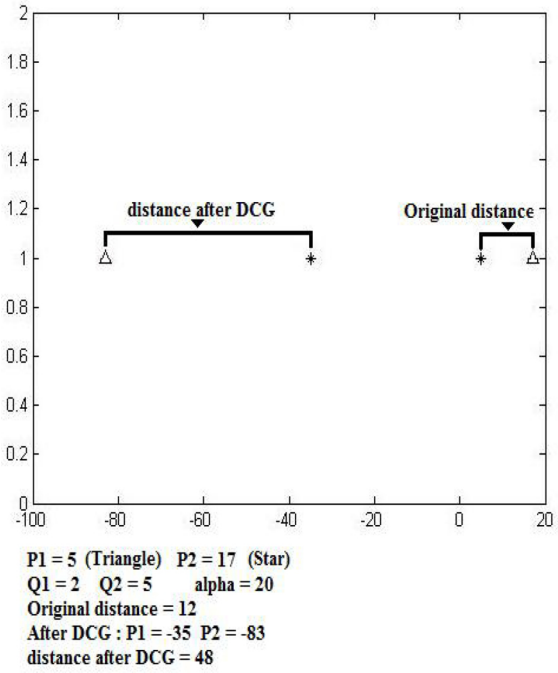

Fig. 1. Description of DCG

There is a same authentication for negative numbers of α. If the numbers of α be negative just changes the orientation of classes move, but result of applying DCG is same as positive α values. According to figure (2) and formula (1), the distances between all samples of each class doesn't change also variance of each class, but separability of classes increase by dispelling mean of the classes.

III. PROPOSED APPROACH

According to previous part, DCG is too beneficial to increase separability of classes. But DCG does its process based on class labels. So on test set, DCG is not useful because we don't know sample labels but for train set it's useful. According to introduction part, all of feature extraction methods try to calculate a formula to create projection matrix W based on train set. So if DCG increases separability of train set and then calculates output dataset based on projection matrix, then misclassification error rate of output decreases and output set Y classifiers easily get more than output of original input set. To implement this method, sample labels assume to be represented as a matrix C= $[c_1 \dots . c_n] \in R^{N*N_c}$ where N is the number of samples and N_c is the number of labels and the elements of the indicator vector C_i is set to be 1 to N_c . Then do X-C for α times. This method has 3 important advantages. (A) Adapting dataset with feature extraction method (B) the results never are worse than original manner. (C) Increasing the noise tolerability of dataset.

(A) According to introduction part, if the input dataset of feature reduction be original just there is a fix output. The misclassification error rate of this fix output probably won't be desired. But with using our method it's possible to adapt input set with feature reduction method more than primitive manner. By using optimal α it's possible to create a better output with lower misclassification error rate than original output.

(B) If the number of α be zero, original input and processed input are same so the feature extraction's outputs of them are same and there is not any difference between misclassification error rates.

(C) One of the motivations that is presented for feature extraction methods is noise tolerability. In almost the entire feature extraction methods if noise sits on input dataset, the misclassification error rate of output gets more than original manner. But our method can cover this motivation to a large extent. Our method does this action based on two features that are hidden in its formula.

First feature is α. If the input dataset is noisy we can compute the optimal α to adapt feature reduction method with noisy dataset. It's useful to calculate an output with error rate close the situation that input set is not noisy. Second feature is C. one of the methods that is presented to decrease the effect of noise is using fuzzy. In matrix C there is a concept like fuzzy which is hidden. When we subtract sample's values of their labels in reality we subtract sample's values of all of the labels just with different coefficient. Because all of labels are coefficients for the other ones.

$$X_i = X_i - \alpha L_{X_i} \quad , \quad X_j = X_j - \alpha L_{X_j}$$

$$\text{If} \quad L_{X_i} = \beta L_{X_j} \quad \rightarrow \quad X_i = X_i - \beta\alpha L_{X_j} \qquad (2)$$

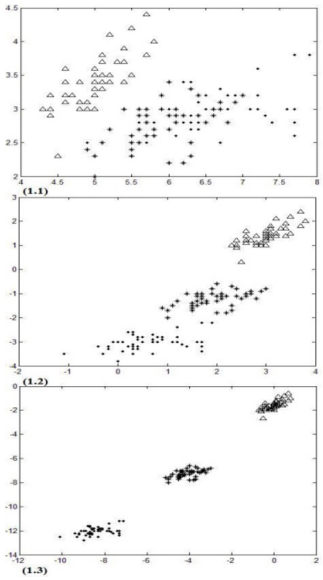

Fig. 2. Effect of DCG on first and second features of Iris dataset (1.1) Original (1.2) α=2 (1.3) α=5

IV. JUSTFICATION

The main question which is mentioned in this article is that why applying DCG concept on input dataset could increase the quality of feature reduction approaches. This question will be answered true this section. Almost all feature reduction approaches's purpose is to describe new dimensions which have two below mentioned signification based on finding the mean, variance, and eigenvectors of each class in input dataset. (A) The first signification is that new dimensions are less than original dimensions (B) the second signification is that separability of classes ability is saved as far as possible.

As described in section two after processing the input dataset by DCG, both slope of each eigenvector and variance of each class remain constant and this happened because the transfer amount of each sample is equal to transfer amount of the same class's samples. Consequently, the main section of feature reduction approaches which is finding slope of eigenvectors, and variance of classes will be remained constant and only mean of each class will be changed. This subject is shown in figure (3). As it is shown in figure (3) , the slope of eigenvectors of dataset is not changed after processing the dataset by DCG.

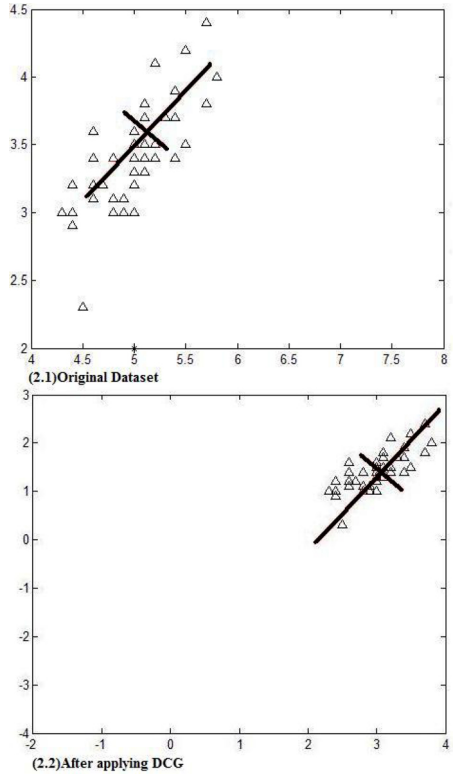

Fig3. Effect of DCG on Mean and eigenvectors of one of the classes of iris dataset

According to proposed approach, it is asserted by transferring coordination of different classes's mean, and distance between mean, it is possible to find the new coordination for each means and in this situation classification accuracy rate of feature reduction approach output is higher than original input dataset. Briefly, this article is mainly trying to show that by using DCG for processing input dataset, the feature reduction output error rate will be less than when input dataset is not processed.

V. PARAMETER SELECTION

To process input dataset of feature extraction methods there is just an input parameter. The number of DCG Loop called α is an important parameter. A should tune correctly to fulfill our aims. To select α there are two important points to pay attention. (I) according to DCG part, DCG avouch the separability of classes in infinite situation .So there are some values for α at each dataset that α doesn't dispel classes enough and it's possible that some of its values decrease the separability of classes. Because the move orientations of classes for dispelling are hidden, maybe the move orientation of a class is like another one and of course their speeds are never the same, so for some of α values, it's possible that classes come near each other and their distances are less than primitive situation. This range of α is called loop's problem maker range (LPMR), so we need the values that are out of this range. (II) But the second point to select α refers to feature extraction method. Using DCG method to process input datasets certainly changes the values of samples in each feature so it's important that these transforms don't cause problem for feature extraction algorithms. Because in some of the feature extraction algorithms there are some formula that don't work correctly with every values.

For example one of the new feature extraction methods that is presented in ICML 2009 by Xiao-Tong and Bao-Gang is REDA [1]. This method calculates the best projection matrix W in an iterative manner. In their algorithm there is a formula that computes an item by this form: exp (-$\|Wx - C_i\|^2/\sigma^2$). we know exponential of numbers of more than 30 is too huge and less than -13 is considered as zero for all cases. So it's important that new value of X after process don't be out of this bound. For solving this problem we can compute the good range of α by (3).where m_i is the minimum value in samples of i^{th} sample's label

$$\theta_{min} < (W(m_i - \alpha L_{X_i}) - C_i)^2/\sigma^2 < \theta_{max} \qquad (3)$$

According to these points and basic theory of the DCG concept, the misclassification error rates of feature extraction's outputs are different for different numbers of α. In the table (1) for example we computed the accuracy of the SRDA's outputs of the input dataset (Haber-man) after applying DCG based on numbers of α between numbers 1 to 30.

TABLE I. FEATURE EXTRACTION OUTPUT'S ACCURACY AFTER APPLYING DCG BASED ON DIFFERENT A

Haber-Man(after Applying DCG)			
SRDA			
Number of loops	KNN (Accuracy ± std-dev)	Number of loops	KNN (Accuracy ± std-dev)
α = 1	64.27	α = 16	63.89
α = 2	64.06	α = 17	64.29

α = 3	65.09	α = 18	64.80
α = 4	64.32	α = 19	65.09
α = 5	64.68	α = 20	65.87
α = 6	64.27	α = 21	66.37
α = 7	64.05	α = 22	66.10
α = 8	63.98	α = 23	66.13
α = 9	63.77	α = 24	66.30
α = 10	63.87	α = 25	65.43
α = 11	63.69	α = 26	65.80
α = 12	63.99	α = 27	65.43
α = 13	63.76	α = 28	64.80
α = 14	64.23	α = 29	65.20
α = 15	63.67	α = 30	65.06

According to the table (1), we plan the figure (4) based on numbers of α Between 1 to 60. According to the figure (4), the amount of the feature extraction's outputs accuracies has tolerance, so there are some local optimums and a global optimum. The authors suggest hill climbing algorithms or simple genetic algorithm (J.Holland 1970's) with a fitness function based on output set's misclassification error rate to compute best value of α (Global optimum).

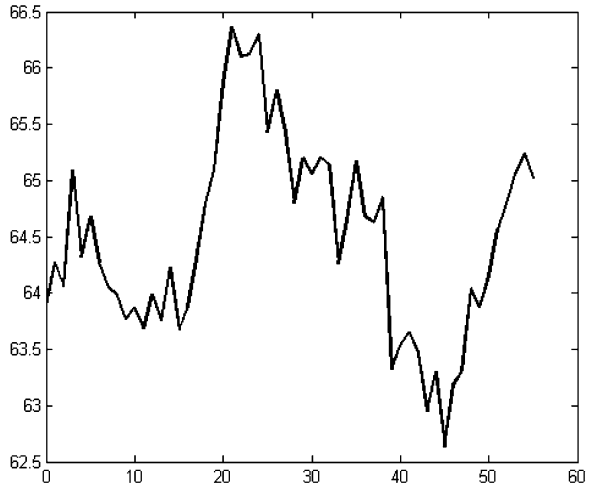

Fig.4. Tolerance of feature extraction output's accuracy based on different α

VI. RESULTS

In this part we compare the misclassification error rate of my idea and original methods. In each table of this part, first we computed the misclassification error rate of feature extraction's output of one of the original UCI datasets in second columns of each row. SRDA (Cai et al, 2008), PCA (Joliffe, 1986) and REDA-SRDA (Xiao-Tong Bao-Gang, 2009) are our feature extraction cases. Next there is misclassification error rate of feature extraction's output of that dataset but after applying DCG on its. Parameter α is computed with a SGA (simple genetic algorithm) for each datasets and highlighted in third columns of each rows. All of the misclassification error rates computed with k-nearest neighborhood classifier and we tried to use best number of k for each dataset to decrease the error rate.

TABLE II. **TABLE 1.** PERFORMANCE COMPARISON ON HUBER-MAN

Huber-Man (UCI Dataset)			
Original Input Dataset		**Input Dataset after Appling DCG**	
Classifier / Methods	**KNN : Accuracy ± std-dev**	**Classifier / Methods**	**KNN : Accuracy ± std-dev**
PCA	70.45±0.17	**PCA α=6**	**71.42±0.28**
SRDA	63.89±0.24	**SRDA α=22**	**66.20±0.30**
REDA-SRDA	64.75±0.29	**REDA-SRDA α=28**	**66.70±0.37**

TABLE III. PERFORMANCE COMPARISON ON BREAST-CANCER

Breast-Cancer (UCI Dataset)			
Original Input Dataset		**Input Dataset after Appling DCG**	
Classifier / Methods	**KNN : Accuracy ± std-dev**	**Classifier / Methods**	**KNN : Accuracy ± std-dev**
PCA	80.86±0.19	**PCA α=5**	**87.06±0.15**
SRDA	88.69±0.19	**SRDA α=0**	**88.69±0.19**
REDA-SRDA	91.34±0.50	**REDA-SRDA α=14**	**92.15±0.20**

TABLE IV. PERFORMANCE COMPARISON ON GLASS

Glass (UCI Dataset)			
Original Input Dataset		**Input Dataset after Appling DCG**	
Classifier / Methods	**KNN : Accuracy ± std-dev**	**Classifier / Methods**	**KNN : Accuracy ± std-dev**
PCA	64.46±0.32	**PCA α=2**	**65.52±0.24**
SRDA	56.69±0.25	**SRDA α=57**	**63.23±0.20**
REDA-SRDA	65.58±0.20	**REDA-SRDA α=0**	**65.58±0.20**

TABLE V. PERFORMANCE COMPARISON ON LUNG-CANCER

Lung-Cancer (UCI Dataset)			
Original Input Dataset		**Input Dataset after Appling DCG**	
Classifier / Methods	**KNN : Accuracy ± std-dev**	**Classifier / Methods**	**KNN : Accuracy ± std-dev**
PCA	35.29±0.30	**PCA α=2**	**42.22±1.40**
SRDA	52.23±1.30	**SRDA α=47**	**54.50±0.90**
REDA-SRDA	49.36±0.60	**REDA-SRDA α=76**	**53.79±0.70**

(1) Haber-Man (306 instances 3 attributes 2 classes)

(2) Breast-Cancer(699 instances 10 attribute 2 classes)
(3) Glass (214 instances 9 attribute 7 classes)
(4) Lung-Cancer(32 instances 56 attributes 3 classes)

VII. CONCLUSION

In this paper we presented a new concept call DCG to increase separability of classes. DCG is provided by subtracting features values of their labels. So we processed the input dataset of feature extraction methods based on DCG. The results showed that output of feature reduction methods have misclassification error rate when the input is processed less than original dataset. Next we presented some ways to compute optimum numbers of DCG's loops and we mentioned some motivations for that. The results showed the quality of our idea.

VIII. FUTURE WORK

One interesting future research direction is to study how to compute numbers of DCG's loop without any algorithms just with formula.

ACKNOWLEDGMENT

The authors like to say thanks to Mr. Vahid Motaghed (Master student of Shiraz international university) for his ideas.

REFERENCES

[1] Xiao-Tong Y., Bao-gang H., (2009) robust feature extraction via information theoretic learning, International Conference on machine learning.

[2] Cai, D., He, X., & Han, J. (2008).An efficient algorithm for large-scale discriminant analysis. IEEE Transactions on Knowledge and Data Engineering, 20(1), 1–12.

[3] Cayton, L., & Dasgupta, S. (2006). Robust Euclidean Embedding. International Conference on Machine Learning (pp. 169–176).

[4] Friedman, J. (1989). Regularized discriminative analysis. Journal of American Statistical Association, 84(405), 165–175.

[5] Fukunnaga, K. (1991). Introduction to statistical pattern recognition. Academic Press.

[6] He, X., & Niyogi, P. (2004). Locality preserving projections. Advances in Neural Information Processing Systems 16. Cambridge, MA: MIT Press.

[7] Hild-II, K., Erdogmus, D., Torkkola, K., & Principe, C. (2006). Feature extraction using information theoretic learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(9), 1385– 1392.

[8] Huber, P. (1981). Robust statistics. Wiley. Jenssen, R., Eltoft, T., Girolami, M., & Erdogmus, D. (2006). Kernel maximum entropy data transformation and an enhanced spectral clustering algorithm. Advances in Neural Information Processing Systems 19 (pp. 633–640). Cambridge, MA: MIT Press.

[9] Joliffe, I. (1986). Principal component analysis. Springer-Verlag. Kaski, S., & Peltonen, J. (2003). Informative discriminant analysis. International Conference on Machine Learning (pp. 329–336).

[10] Liu, W., Pokharel, P. P., & Principe, J. C. (2007).Correntropy: Properties and applications in nongaussian signal processing. IEEE Transactions on Signal Processing, 55(11), 5286–5298.

[11] Mizera, I., & Muller, C. (2002). Breakdown points of Cauchy regression-scale estimators. Statistics and Probability Letters, 57, 79–89.

[12] Principe, J., Xu, D., & Fisher, J. (2000). Information Theoretic learning, Unsupervised Adaptive Filtering. New York: Wiley.

[13] Principe, J., Xu, D., & Fisher, J. (2000). Information Theoretic learning, Unsupervised Adaptive Filtering. New York: Wiley.

[14] Renyi, A. (1961). On measures of information and entropy. Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability (pp. 547–561).

[15] Rockfellar,R. (1970). Convex analysis, Princeton Press. Torkkola, K. (2003). Feature extraction by nonparametric mutual information maximization. Journal of Machine Learning Research, 3, 1415–1438.