AN ADAPTIVE CLASSIFIER DESIGN FOR ACCURATE SPEECH DATA CLASSIFICATION

Omid Dehzangi, Ehsan Younessian
Nanyang Technological University, Singapore

Fariborz Hosseini Fard
SoundBuzz PTE LTD, Subsidiary of Motorola Inc., Singapore

Keywords: Nearest neighbor, Linear discriminant analysis, Adaptive distance measure, Weight learning algorithm.

Abstract: In this paper, an adaptive approach to designing accurate classifiers using Nearest Neighbor (NN) and Linear Discriminant Analysis (LDA) is proposed. A novel NN rule with an adaptive distance measure is proposed to classify input patterns. An iterative learning algorithm is employed to incorporate a local weight to the Euclidean distance measure that attempts to minimize the number of misclassified patterns in the training set. In case of data sets with highly overlapped classes, this may cause the classifier to increase its complexity and overfit. As a solution, LDA is considered as a popular feature extraction technique that aims at creating a feature space that best discriminates the data distributions and reduces overlaps between different classes of data. In this paper, an improved variation of LDA (im-LDA) is investigated which aims to moderate the effect of outlier classes. The proposed classifier design is evaluated by 6 standard data sets from UCI ML repository and eventually by TIMIT data set for framewise classification of speech data. The results show the effectiveness of the designed classifier using im-LDA with the proposed ad-NN method.

1 INTRODUCTION

The NN classifier is one of the oldest and the most successful methods of non-parametric pattern classification (Cover and Hart, 1998). However, it has some weaknesses in cases that patterns of different classes have overlap in some regions in the feature space. It also considers all the stored instances the same for classification, but the instances are different in being representative of their typical classes.

The performance of the NN classifier depends crucially on how to choose a suitable distance metric. Many methods have been developed to locally adapt the distance metrics such as the flexible metric method proposed in (Friedman, 1994), the discriminant adaptive method in (Hasti and Tibshirani, 1996) and the adaptive metric method in (Domeniconi et al., 2002). The common idea underlying these methods is that they estimate feature relevance locally at each query pattern. This leads to a weighted metric for computing the similarity between the query patterns and training data. In (Wang et. al., 2007), a simple locally adaptive distance measure is proposed that uses a heuristic measure to specify the weight of each training instance. The method we propose in this paper uses a locally adaptive metric to improve the performance of the basic NN classifier. An iterative learning algorithm is employed to incorporate a local weight to the Euclidean distance measure that attempts to minimize the number of misclassified patterns in the training set. In case of data sets with highly overlapped classes, examples in the overlapping area are considered to be noisy as for learning these examples. the learning algorithm would be in contradiction with other training examples or would need to increase its complexity in order to accommodate them. Learning these difficult examples may lead the algorithm to be unable to generalize well.

As a solution to this problem, linear discriminant analysis (LDA) is considered as one of the most traditional methods to find a linear feature transformation method, which maximizes the ratio of between-class scatter and the within-class scatter.

Dehzangi O., Younessian E. and Hosseini Fard F. (2009).
AN ADAPTIVE CLASSIFIER DESIGN FOR ACCURATE SPEECH DATA CLASSIFICATION.
In *Proceedings of the 6th International Conference on Informatics in Control, Automation and Robotics - Robotics and Automation*, pages 67-71
DOI: 10.5220/0002206200670071

The earliest of such methods, Fisher's Linear Discriminant Analysis (LDA) (Fisher, 1936), tries to find a linear combination of input variables that best discriminates between different class distributions and is still a powerful technique for feature extraction to reduce overlaps between different classes of data (Duda and Hart, 2001). However, LDA does not take into account the conjunctions between different pairs of classes in a multi-class problem (Loog et al., 2001). In such cases, if one or more classes are far away from others (i.e. outlier classes), there is no need to maximize their between-class scatter covariances in the transformed space. Thus, they do not contribute in the estimation of the uniform between-class covariance (Jarchi and Boostani, 2006). In this paper, an improved version of LDA is investigated that redefines the between-class scatter matrix by integrating a simple weight into it. In the transformed feature space, different classes of data have lower degrees of overlap with one another. Then, our proposed ad-NN classifier can be applied to the input patterns in the new space with lesser risk of overfiting. In order to assess our method, combination of the im-LDA and the ad-NN are applied on eight UCI ML driven data sets. The proposed classifier design is also applied on TIMIT speech data set, attempting to classify huge amount of speech frames with 60 different phoneme classes.

2 ADAPTIVE DISTANCE USING WEIGHTED INSTANCES

The nearest neighbour classifier assigns label of a test pattern according to the class label of its nearest training instance. To introduce the weighted version of NN rule, here, the notation of basic NN rule is briefly described. Assume that classification of patterns in an m-dimensional space is under investigation. Having a set of training instances $\{(X_i,C_k)\}$, where X_i, i=1,…,n are training feature vectors and C_k ,k=1,…,M are the labels. NN rule finds the nearest neighbor of a new test pattern X using a distance function and assigns X to C_w (the class label of the winner class). The Euclidean distance have been conventionally used to measure the distance between X and Y:

$$\mathrm{d}(X,Y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \qquad (1)$$

In the first step, the distance as a dissimilarity measure between query pattern X and the j^{th} instance X_j is changed to a similarity measure. This is done by a linear conversion as follows:

$$\mu(X, X_j) = 1 - \mathrm{d}(X, X_j) / \sqrt{m} \qquad (2)$$

where, $\sqrt{m}$ is the maximum distance which can ever occur Between two training instances in the whole training set, since the data is normalized in the range of [0, 1] in the first place. As a result, instead of finding minimum distance pattern to X, we search for instance X_j such that $\mu(X,X_j)$ is maximized. This can be interpreted as normalizing the distance from [$\sqrt{m}$, 0] to a real number in the interval [0, 1]. Now, we can introduce the weighted nearest neighbor rule. The pattern is classified according to consequent class of the winner instance X_{win}. The winner instance is specified using:

$$win = \underset{j=1,\ldots,n}{\operatorname{argmax}} \left\{ \mu(X, X_j) \,.\, w_j \right\} \qquad (3)$$

where w_j is the weight assigned to the j^{th} instance by the learning algorithm which is introduced in section 4.

2.1 Instance Weighting Algorithm

For an M-class problem, assume that n labeled patterns $\{X_i, i=1,2,\ldots,n\}$ are available from various classes. In this section, we propose an algorithm to learn the weight of each pattern using other labeled training instances. The calculated weight is optimal in the sense that it maximizes the classification rate of the classifier on the training data. At the beginning, a weight of one is assigned to each pattern in the whole training set (i.e. w_k=1, k=1,..., n).

In the following, the weight of each instance is specified assuming that the weights of all other training instances are given and fixed. The weight w_k of instance k with class label of Q can be found as follows. First, the problem is considered a 2-class problem of class Q as positive "p" and class $\bar{Q}$ as negative "n". The w_k is set to zero (i.e., instance X_k does not contribute in classification decision). Given current weights of all other training instances and a training set of P positive and N negative labeled patterns, the decision results of the weighted instance NN classifier described in section 3 can be grouped into four categories: true positives (TPs) denoting samples correctly labelled as positives; false positives (FPs) denoting negative samples incorrectly labeled as positives; true negatives (TNs) denoting samples correctly labeled as negatives; and

false negatives (FNs) denoting positive samples incorrectly labeled as negatives.

Training patterns of class Q which are classified correctly with the current values of instance weights are removed from the training set. These patterns will be classified correctly regardless of the value of w_k. Similarly, training patterns of class $\bar{Q}$ which are misclassified with the current assigned weights are also removed from the training set. These patterns will be misclassified regardless of the value of w_k. For each training pattern left in the training set (i.e. instances in TN and FN), a score is calculated using the following measure:

$$S(x_t) = \frac{\max\{\mu(X_t, X_j).w_j \mid j = 1,2,...,\mathrm{n},\ j \neq t\}}{\mu(X_t, X_k)} \quad (4)$$

where $\mu(X_i,X_j)$ represents the similarity of patterns X_i and X_j calculated using (2).

Those instances that have X_k as their nearest neighbor are called associates of X_k. It can be shown that X_t is an associate of X_k, if $S(X_t)$ is less than w_k. We have,

$$w_k > S(X_t) \Rightarrow w_k \cdot \mu(X_t, X_k) > S(X_t) \cdot \mu(X_t, X_k) \quad (5)$$

From (4) we have,

$$w_k.\mu(X_t, X_k) > \max\{\mu(X_t, X_j).w_j \mid j = 1,...,n \quad j \neq t,k\} \quad (6)$$

From (3) and (6), it is concluded that X_k is the winner instance to classify X_t. Given the weighted instance NN classifier introduced in previous section, associate set of instance X_k can be defined fomally as,

$$Associate\_set(X_k, w_k) = \{X_t \mid t = 1,...,n \quad t \neq k \quad w_k > S(X_t)\} \quad (7)$$

By altering w_k, the associate set of X_k is changed that cause modification in classification error rate of the classifier. Our aim is to determine w_k such that the error-rate of the classifier on training set is minimized given that the weights of all other training instances are given and fixed. We define an accuracy measure as,

$$Accuracy = \mathrm{TP} - \mathrm{FP} \quad (8)$$

We try to find a proper w_k such that associate set of X_k includes more FN instances which need to be classified as "*p*" and exclude more TN instances which are the instances of class "*n*". The optimal weight of the instance X_k is calculated by maximizing *Accuracy* measure assuming that the weights of all other instances are given and fixed. To do this, the set of patterns X_t are ranked in ascending order of their scores. We define a threshold initialized with zero. Then, assuming that X_t and X_{t+1} are two successive patterns in the ranked list, a threshold is computed as,

$$th = (\mathrm{Score}(X_t) + \mathrm{Score}(X_t{+}1))/2 \quad (9)$$

The threshold is then altered from the least score to the greatest and associated accuracy of the classifier with respect to the each threshold is measured. The value of the best threshold (i.e. leading to the best accuracy) is simply used as the weight of the instance X_k. The proposed instance weighting mechanism assigns a weight to each instance attempting to better discriminate between the patterns of the same class and patterns of all other classes. The search for the best combination of instance weights is conducted by optimizing each instance in turn assuming that the order of the instances to be optimized is fixed.

3 OVERVIEW OF LDA

The proposed algorithm in section 2.1 to learn the weight of each instance attempts to minimize the classification error in the training data. In case of data sets with highly overlapped classes, examples in the overlapping area are considered to be noisy as for learning these examples. Learning these difficult examples may lead the algorithm to be unable to generalize well. The goal of LDA is to find an optimal linear transformation of input feature vectors such that the class separability in the new space is maximized. In order to find an optimal linear discrimination transform, Fisher (Fisher, 1936) proposed a criterion that maximizes the ratio of between-class to within-class scatter matrices. The aim is to look for a linear discriminant transform w_{LDA},

$$w_{LDA} = \arg\max_{w}\{\varphi_{ij}\}, \quad (10)$$

where, φ_{ij} is the Fisher criterion that is determined as follows:

$$\varphi_{ij} = \frac{|\ W^T S_B W\ |}{|\ W^T S_W W\ |} \quad (11)$$

where S_B is between-class scatter matrix and S_W is within-class scatter matrix. Columns of W_{LDA} are eigenvectors corresponding to the (C-1) largest eigenvalues of $S_w^{-1}S_B$. Now, each sample X in data

set can be transformed to a new space by multiplying to the matrix w_{LDA} :

$$Y = w_{LDA}{}^{T}.X \quad (12)$$

Y is a new transformed sample corresponds to *X*.

3.1 The Improved LDA Method

By Projection of data in a lower dimensional space, LDA can also reduce the computation complexity caused by redundant information in the data which is useful for solving classification problems. However, LDA has some weaknesses. LDA considers all the classes the same to calculate between-class scatter matrix. If one or more classes are outliers, there is no need to maximize their between-class scatter covariance in the transformed space. Therefore, different weights should be integrated in the covariance estimation procedure. The aim in the weighted version of LDA is to alleviate the role of an outlier class. This is done by redefining between-class scatter matrix.

Assume that we use the between-class scatter matrix definition that is based on (Loog et al., 2001):

$$S_B = \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} P_i P_j (\mu_i - \mu_j)(\mu_i - \mu_j)^T \quad (13)$$

where, C is the number of classes, p_i is a prior probability for class i and p_j is a prior probability for class j, μ_i is the mean of training samples of class i.

Incorporating a weight $1/\varphi_{ij}$ into between-class scatter matrix leads to a new formula:

$$S_B = \frac{1}{N^2} \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} 1/\varphi_{ij} N_i N_j (\mu_i - \mu_j)(\mu_i - \mu_j)^T \quad (14)$$

where, φ_{ij} is Fisher discrimination value of the resulting w_{LDA} that is determined in (11), N is the number of training samples, N_i and N_j are the number of training samples of class i and j, respectively. By incorporating this weight, the farther the classes are from each other, the less their contribution is in the between class scatter matrix calculation.

4 EXPERIMENTAL RESULTS

In order to evaluate the performance of the designed classifier, two sets of experiments have been conducted.

4.1 Experiments using Standard Data

First, a number of standard data sets are used which derived from UCI ML repository (Merz and Murphy, 1996). Some statistics of the data sets are shown in Table 1.

Table 1: Statistics of the data sets used in our experiments.

Data set	# of attributes	# of patterns	# of Classes
Pima	8	768	2
Wine	13	178	3
Hepatitis	19	155	2
Image Seg.	15	210	7
Balance	4	625	3
Heart Clev.	13	303	5

To evaluate the classification accuracy on the data sets, the average result of ten trials of ten-fold cross validation is reported. Results, which are shown in the Table 2, illustrate that our proposed method is led to the best performance on all the data sets compared to basic combinations.

Table 2: Error rates of combination of LDA and NN along with our proposed method on 6 UCI ML data sets.

Data set	Basic NN	LDA+ Basic NN	Im-LDA+ Basic NN	Im-LDA+ ad-NN
Pima	29.06	25.96	25.54	22.47
Wine	5.09	3.23	2.88	2.32
Hepatitis	19.87	14.47	14.33	12.89
Image Seg.	8.63	7.48	6.21	5.32
Balance	18.69	17.51	15.67	14.19
Heart Clev	20.75	19.26	16.92	15.29
Ave. error	17.02	14.65	13.59	**12.08**

As it can be seen in Table 2, im-LDA outperforms LDA in multi-class data sets. This shows the effectiveness of the weight incorporated in the definition of between-class scatter matrix. Table 2 also shows that in the transformed feature space using im-LDA where different classes of data have lower degrees of overlap, ad-NN with adaptive local distance measure clearly improves the generalization ability of the basic NN.

4.2 Experiments using Real Data

In this section, we validate the proposed method on the TIMIT corpus (Garofolo, 1988) because of its high-quality phone labels. All results reported are framewise classification error rates for TIMIT complete test set (the 1344 si and sx sentences). The

speech waveforms are parameterized by a standard Mel-Frequency Cepstral Coefficient (MFCC) front end. The cepstral analysis uses a 25 msec Hamming window with a frame shift of 10 msec. Each input pattern X_i consists of the current frame of 12 MFCCs and energy plus delta and acceleration coefficients, and two context frames on each side, making a total of (13 + 13 + 13) * 5 = 195 components. This formulation was arrived at by experimentation with varying numbers of context frames left and right of the frame being classified. The training set has about 1.1 million frames and the test set has about 400 thousand frames. Each frame has an associated 1-of-60 phonetic label derived from the TIMIT label files.

Due to the large number of training data and large number of classes, TIMIT data set seems to be a suitable task to evaluate our proposed classifier. In Table 3, framewise classification error rates on the TIMIT test set using our classifier is compared to the existing methods.

Table 3: Framewise phoneme classification error rate on TIMIT test set.

Classifier	Error Rate
Recurrent Neural Nets (Schuster, 1997)	34.7%
Bidirectional LSTM (Graves, 2005)	29.8%
im-LDA + ad-NN	28.7%

The results show that the proposed classifier design outperforms previous works in classification of speech frames on TIMIT task.

5 CONCLUSIONS

In this paper, a novel classifier design based on combination of an improved version of LDA and an adaptive distance NN was presented. LDA, as a preprocessing step, was used to transform input data to a new feature space in which different classes of data has lower degrees of overlap. In the classification step, a novel learning algorithm was used to assign a weight to each stored instance, which was then contributed in the distance measure, with the goal of improvement in generalization ability of the basic NN. In this way, different weighs were given to the transformed samples based on a learning scheme which optimized the weights according to the classification error rate. Our proposed method was evaluated by various UCI ML data sets. Results showed that the proposed method improves the generalization ability of basic NN. By using TIMIT speech data set, the effectiveness of our approach in real problems like speech data classification was also proved.

REFERENCES

Cover, T.M., Hart, P.E., 1967. Nearest Neighbor Pattern Classification. *IEEE Transaction on Information Theory* 13, 21-27.

Friedman, J., 1994. Flexible metric nearest neighbor classification. Technical Report 113, Stanford University Statistics Department.

Hastie, T., Tibshirani, R., 1996. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18: 607-615.

Domeniconi, C., Peng, J., Gunopulos, D., 2002. Locally adaptive metric nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24: 1281-1285.

Wang, J., Neskovic, P., Cooper, L.N., 2007. Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recogition Letters*, 28: 207-213.

Fisher, R.A., 1936. The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, 7:179-188.

Duda, R.O., Hart, P.E., Stork, D., 2001. *Pattern Classification 2nd Edition.* Wiley, New York.

Loog, M., Duin, R.P.W., Haeb-Umbach, R., 2001. Multiclass linear dimension reduction by weighted pairwise fisher criteria, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23: 762-766.

Jarchi, D., Boostani, R., 2006. A New Weighted LDA Method in Comparison to Some Versions of LDA, *Transaction on Engineering and Computational Technology*, 18: 18-45.

Garofolo, J.S., 1988. *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database*, National Institute of Standards and Technology (NIST), Gaithersburgh, MD.

Merz, C.J., Murphy, P.M., 1996. *UCIRepository of Machine Learning Databases*. Irvine, CA: University of California Irvine, Department of information and Computer Science. Internet: http://www.ics.uci.edu/~mlearn/MLRepository.html

Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45: 2673-2681.

Graves, A., Schmidhuber, J., 2005. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *International Joint Conference on Neural Networks*.