IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Recognise the Degraded Devnagari Script by Dimensionality Reduction Linear and quadratic Classifiers using Fisher Linear Discriminant

Sushilkumar N. Holambe[1*], Ulhas B Shinde[2]

[1]Computer Science & Engineering Department, College Of Engineering, Osmanabad-413512(M.S.)India.
[2]CSMSS,CSCOE, Aurangabad-431001(M.S.) India

Available online at: www.ijcseonline.org

Accepted: 18/Nov/2018, Published: 30/Nov/2018

Abstract— In this paper we are implementing parametric classifier Linear and quadratics using fisher linear discriminant for find the misclassification rate using cross validation, useful in recognizing the degraded devnagari script scan document.Dimensionality reduction is the process of transforming input data into a lower dimensional space where a more efficient classifier can be built are divided in two groups: Feature extraction, which map input data using linear transformation i.e. a transformation matrix and feature selection, which performs the mapping by selecting a subset of the original features.Feature extraction methods are supported by fisher's linear discriminant function.Feature selection is use to choose an optimal subset according to some criterion of cardinality m among the d input features. In feature ranking each Feature is evaluated individually according to the chosen criterion, and the values are then sorted the m features with the best value of the criterion are retained for classification. Also we focus on learning machine stages which consists of two stages: dimensionality reduction and classification.

Keywords— Linear, Quadratic, Fisher Linear Discriminant, Cross validation, Feature Extraction, Dimensionality.

I. INTRODUCTION

Pattern recognition deals with classification problems that we would like to delegate to a machine, for example, scanning for abnormalities in smear test samples, identifying a person by voice and a face image for security purpose, detecting fraudulent credit card transaction, and so on. Each object (test sample, person, transaction is described by a set of p features and can be thought of as a point in some p dimensional feature. The term 'pattern' to denote the p-dimensional data vector

x= (x1,............xp)^T

of measurements (T denotes vector transpose), whose component xi are measurements of the features of an object , thus the features are the variables specified by the investigator and thought to be important for classification. In discrimination, we assume that there exist C groups or classes, denote and associated with each pattern x is a categorical variable z that denotes the class or group membership; that is if z=1 then the pattern belongs to $\omega_i \ i \in \{1,\ldots\ldots\ldots\ldots C\}$,

the p features submitted to its input. For designing a classifier, also called discriminant analysis, we use a labeled data set, Z, of n objects, where each object is described by its, feature values and true class label.

The fundamental idea used in statistical pattern recognition is Bayes decision thory [1]. The C classes are treated as random entities that occur with prior probabilities

$p(\omega_i), i, \ldots c.$ The posterior probability of being in class ω_i for an observed data point x is calculated using Bayes rule

$$p(\omega_i|x) = \frac{p(x|\omega_i)p(\omega_i)}{\sum_{j=I}^{c} p(x|\omega_i)p(\omega_j)}, \qquad (2)$$

Where $p(x|\omega_i)$ is the class conditional probability density function (pdf) of x, given class ω_i. According to the Bayes rule, the class with the largest posterior probability is selected as the label of x, Ties are broken randomly. The bayes rule guarantees the minimum misclassification rate. Sometimes the misclassifications cost differently for different classes. Then we can use a loss matrix $\Lambda = [\lambda_{ij}]$, where λ_{ij} is a measure of the loss incurred if we assign class label ω_i when the true label is ω_j The minimum risk classifier assigns x to class with the minimum expected risk

$$R_x(\omega_i)=\sum_{j=1}^{c}\lambda_{ij}\;p(\omega_j|x). \qquad (3)$$

In most real life problems we do not have a ready made classification algorithm. We can only provide a rough guidance in a linguistic format and pick out features that we believe are relevant for the task. The classifier has to be trained by using a set of labeled examples. The training depends on the classifier model. [1,5]

II. FISHER LINEAR DISCRIMINANT

Fisher Linear Discriminant (FLD) is an example of a class specific subspace method that finds the optimal linear projection for classification. Rather than finding a projection that maximizes the projected variance as in principal component analysis, FLD determines a projection, $y = W_F^T\ X$, that maximizes the ratio between the between class scatter and the within –class scatter. Consequently, classification is simplified in the project space.

Consider a C-class problem, with the between-class scatter matrix given by

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

and the within- class scatter matrix by

$$S_w=\sum_{i=1}^{c}\ \sum_{x_k\in X_1}(x_k-\mu_i)\,(x_k-\mu_i)^T \quad (5)$$

Where μ is the mean of all samples, μ_i is the mean of classes i, and N_i is the number of samples in class i. The optimal projection W_f is the projection matrix which maximizes the ratio of the determinant of the between –class scatter to the determinant of the within class scatter of the projections

$$Wf=\max_{w}\frac{W^r S_B W}{W^t S_w W}=[\omega_1\ \omega_2\ ....\omega_m\ ] \quad (6)$$

Where $\{\omega_i\ |i\ =1,2,..,m\}$ is the set of generalized eigenvectors of S_B and S_W, corresponding to the m largest generalized eigenvalues $\{\lambda_i\ |i=1,2,...,m\}$. However, the rank of S_B is c- 2 or less since it is the sum of c is c-1. To avoid the singularity, one can apply PCA First to reduce the dimension of the feature space to N-c, and then use FLD to reduce the dimension to c-1. The class membership of a sample was then determined using the maximum a posteriori probability, or equivalently by a likelihood ratio test. [3]

III. LINEAR AND QUADRATIC CLASSIFIERS

A quadratic form in x defines the decision boundary of a quadratic classifier, derived through Bayesian error minimization. Assuming that the distribution of each class is gaussian, the classifer output is given by

$$f(x)=\frac{1}{2}(x-\mu_1)T\Sigma_1^{-1}(x-\mu_1)-$$
$$\frac{1}{2}(x-\mu_2)T\Sigma_2^{-1}(x-\mu_2)+\frac{1}{2}\ In\left|\frac{\Sigma_1}{\Sigma_2}\right| \qquad (7)$$

Where μ_i and $\sum\ _i\ (i=1,2)$ are the mean and covariance matrix of the respective gaussian distributions

A linear classifier is a special case of the quadratic form, based on the assumption that $\sum_1 = \sum_2 = \sum,$. Which simplifies the discriminant to

$$f(x)=(\mu_2-\mu_1)\sum^{-1}x+\frac{1}{2}(\mu_2^T\sum^{-1}\mu_1-\mu_2^T\sum^{-1}\mu_2) \qquad (8)$$

For both classifiers, the sign of F (x) determines class membership and is also equivalent to a likelihood ratio test. [1,5,2]

IV. DIMENSIONALITY REDUCTION

Dimensionality reduction is the process of transforming input data into a lower dimensional space where a more efficient classifier can be built are divided in two groups: Feature extraction, which map input data using linear transformation, (i.e. a transformation matrix,), and feature selection, which performs the mapping by selecting a subset of the original features.

Feature extraction methods supported by fisher’s linear discriminant function The dimensionality reduction methods can be used in a stand alone mode through the dedicated menus), or as a part of the cross validation analyses. The stand alone mode may be used to explore the relationships among input features. However, it is not particularly well suited for classifier design since it may produce heavily biased error estimates. For classifier design, cross validation should be applied to the combination of dimensionality reduction and classifier. [3,4]

V. FEATURE EXTRACTION

Given input vectors X of dimension d, feature extraction seeks to find an optimal m x d transformation matrix W which maps X into the m-dimensional space of vectors y, where m< d:

$$Y = W_{opt} X, \qquad W_{opt} = \arg\max_{w} J(W) \qquad (9)$$

Where J(W) is the criterion used to evaluate the discriminatory potential of the selected subset of m features. Feature extraction methods differ in the choice of criterion J(W). One popular criterion is the ratio of between class and within-class scatter matrices S_B and S_W of the transformed vectors. Multiple discriminant Analysis, the measure is defined as

$$J(W) = \left| S_w^{-1} S_b \right| \qquad (10)$$

This quantity, in effect, measures the relative average distance among class centres in the mapped space. In order for the mapping to ensure that an accurate classifier can be built in the space of vectors Y, the average distance should be maximised. It can be shown [1,5] that the matrix W which maximises J(W) of eof equation (10) can be computed by eigen analysis of the scatter matrices in the input space, and the resulting transformation is known as fisher's linear discriminant.

VI. FEATURE SELECTION

The goal of feature selection is to choose an optimal subset (according to some criterion) of cardinality m among the d input features The benefit of feature selection is that, once a reliable subset has been identified, only m measurements need to be collected to predict class labels for new samples. In contrast, feature extraction always requires the full complement of d features. This difference may be a significant factor if obtaining the measurements is costly.

The definition of a feature selection method requires specification of the following two components.

- Search algorithm determines which subsets are evaluated
- The definition of a criterion for evaluating the fitness of each subset

The optimal subset maximizes the value of the chosen criterion.

The search problem is clearly NP complete, since evaluation of all possible subset of cardinality m requires $\binom{d}{m}$ evaluations of the criterion, if m is known. If m is unknown, which is normally the case, the number of evaluations to determine the optimal m and he optimal subset of size m jumps to 2^d . Both numbers are astronomically large for most realistic data sets.

VII. FEATURE RANKING

- Feature ranking. Each Feature is evaluated individually according to the chosen criterion, and the values are then sorted; the m features with the best value of the criterion are retained for classification. The method is conceptually simple and computationally attractive, but is only optimal in the unlikely case of independent, features.
- Forward Selection. Start with an empty subset and add one feature at a time. The added feature is one, which optimizes the value of the criterion for the newly formed subset. The process stops when m features have been added.
- Backward elimination. Start with all d features and remove one feature at a time. The subtracted feature minimally reduces the value of the feature selection criterion for the new subset. The process stops when m features remain.

In addition to solving the search problem, feature selection requires the choice of a criterion for evaluating the fitness of a particular feature subset. A variety of criteria have been proposed and used with these and other feature selection algorithms. When considering the criteria, a distinction must be made between feature ranking and feature subset selection because individual feature ranking can use specialized criteria not available for more than one feature.

VIII. DIMENSIONALITY REDUCTION IN CROSS VALIDATION

In many applications, the learning machine consists of two stages: dimensionality reduction and classification. The Structure is shown in figure.1.

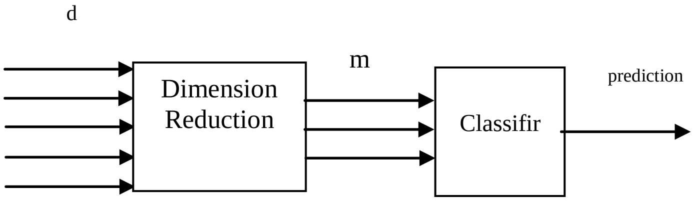

Figure.1 Dimensionality Reduction in Cross Validation

The combined learning machine consisting of dimensionality reduction and classification steps. D is dimensionality of input vectors, m the dimension of transformed vectors. Multiple arrows are meant to convey the effect of reducing the dimensionality of the input data

Dimensionality reduction as a stand alone process independent of the design of the classifier stage leads to over fitting. The reason is as follows. Independent feature subset selection / mapping Dimensionality reduction in cross validation

It is performed using the training dataset. If the resulting features are then used to cross validate a classifier (Which is again done on the training dataset), the same data would have been used for training (the dimensionality reduction stage), and testing (the cross validation of the classifier). This is over fitting, and may produce heavily biased error estimates. In this scenario, each cross validation learning dataset will first be used to compute the dimensionality reduction parameters, and then the computed transformation will be applied to the learning and validation dataset. This is followed by building a classifier using the mapped learning dataset, and testing it using the mapped validation dataset. Thus, validation subset is never used for either dimensionality reduction nor classifier learning, but exclusively for testing this is a statistically rigorous approach to machine learning which avoid over fitting.

IX. RESULTS

Numbers of cross validation subsets:3

Normalized data: Yes

Table 1 :Dimensionality reduction method :FLD

	Quadratic Classifier Error rate	Liner Classifier Error rate
Cumulative	2.67	2.67
Class1	0.00	0.00
Class2	6.00	0.00
Class3	2.00	4.00

Normalized data :- NO

Table 2 :Dimensionality reduction method :- None

	Quadratic Classifier Error rate	Liner Classifier Error rate
Cumulative	2.67	2.67
Class1	0.00	0.00
Class2	6.02	4.10
Class3	2.01	4.02

X. CONCLUSIONS

When trained, some classifiers can provide us with an interpretable decision strategy ,where as other classifiers behave as black boxes.Even when we can verify the logic of the decision making, the ultimate judge of the classifier performance is the classification error.

Estimating the misclassification rate of our classifier is done through the training protocol. Part of the data set, is used for training and the remaining part is left for testing. The most popular training/testing protocol is cross-validation. The error of the classifier is the averaged testing error.

REFERENCES

[1] Duda, R.O., Hart, P.E. & Stork, D.G. (2001). Pattern Classification, 2nd Edition, John Wiley & Sons, New York.

[2] Jain, A.K., Duin, R.P.W. & Mao, I. (2000). Statistical pattern recognition: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 22(1), 4–37.

[3] , "Any discrimination rule can have an arbitrarily bad probability of error for finite sample size," IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-4, pp. 154–157, 1982.

[4] "Automatic pattern recognition: A study of the probability of error," IEEE Trans. Pattern Anal. Machine Intell., vol. 10, pp. 530–543, 1988.

[5] L. Devroye, L. Gy¨orfi, and G. Lugosi, A Probabilistic Theory of Pattern Recognition. New York: Springer-Verlag, 1996.

[6] C.R. Rao, The utilization of multiple measurements in problems of biological classification, Journal of the Royal Statistical Society, Series B 10 (1948) 159–203.

[7] O.S. Hamsici, A.M. Martinez, Bayes optimality in linear discriminant analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 647–657.

[8] A.M. Martinez, M. Zhu, Where are linear feature extraction methods applicable, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 1934–1944.

[9] S. Petridis, S.J. Perantonis, On the relation between discriminant analysis and mutual information for supervised linear feature extraction, Pattern Recognition 37 (2004) 857–874.

[10] K. Fukunaga, J.M. Mantock, Nonparametric discriminant analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 5 (1983) 671–678.

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

[12] R. Lotlikar, R. Kothari, Fractional-step dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 623–627.

[13] N. Kumar, A.G. Andreou, Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition, Speech Communication 26 (1998) 283–297.

[14] R.P.W. Duin, M. Loog, Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2004) 732–739.

[15] T.M. Cover, J.A. Thomas, Elements of Information Theory, second ed., Wiley,New York, 2006.

[16] M. Ben-Bassat, User of distance measures, in: P. Krishnaiah, L. Kanal (Eds.), Handbook of Statistics, North-Holland, Amsterdam, 1982, pp. 773–791.

[17] P.A. Estevez, M. Tesmer, C.A. Perez, J.M. Zurada, Normalized mutual information feature selection, IEEE Transactions on Neural Networks 20 (2009) 189–201.

[18] N. Kwak, C.-H. Choi, Input feature selection by mutual information based on Parzen window, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 1667–1671.

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

[20] K.E. Hild, D. Erdogmus, J.C. Principe, An analysis of entropy estimators for blind source separation, Signal Processing 86 (2006) 182–194.

[21] D. Erdogmus, J.C. Principe, Generalized information potential criterion for adaptive system training, IEEE Transactions on Neural Networks 13 (2002) 1035–1044.

[22] P. Viola, W.M. Wells III, Alignment by maximization of mutual information, International Journal of Computer Vision 24 (1997) 137–154.

[23] D.W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, Wiley, 1992.

[24] A.W. Bowman, A. Azzalini, Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations, Oxford University Press, New York, 1997.

[25] M. Loog, On the equivalence of linear dimensionality-reducing transformations, Journal of Machine Learning Research 9 (2009) 2489–2490.

[26] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, 2001. /http://www.csie.ntu.edu.tw/+cjlin/libsvmS.

[27] J. Li, X. Li, D. Tao, KPCA for semantic object extraction in images, Pattern Recognition 41 (2008) 3244–3250.

[28] Y. Xu, D. Zhang, J.-Y. Yang, A feature extraction method for use with bimodal biometrics, Pattern Recognition 43 (2010) 1106–1115.

[29]Senhadji, L., Carrault, G., Bellanger, J. J., & Passariello, G., "Comparing wavelet transforms for recognizing cardiac patterns", IEEE Engineering in Medicine and Biology Magazine, vol. 14, p. 167-173, 1995.

[30]Gautam, Mayank Kumar, "Performance Analysis of ECG Signal Using by Wavelet Transform, Independent Component Analysis and Fast Fourier Transform", IJSRCSEIT, vol. 1, pp. 95-98, Sept-Oct.2016.

[31] Amirtharajan R and Rayappan JBB ,‖ An intelligent chaotic embedding approach to enhance stego image quality‖,Inform Science, vol. 193, pp. 115-124, June 2012. [32] Bassam Jamil Mohd, Saed Abed, Thaier Al- Hayajneh and Sahel Alouneh, ―FPGA Hardware of the LSB Steganography Method,‖ IEEE Transaction on consumer Electronics, vol. 978, no. 1, pp. 4673–1550,2012.

AUTHOR'S PROFILE

Sushilkumar N. Holambe Department of CSE College of Engineering Osmanabad-413512 (M.S.) India
snholambe@yahoo.com

Dr. Ulhas B. Shinde Principal, CSMSS,CSCOE, Aurangabad-431001(M.S.) India
dhindeulhas@gmail.com