Exploiting Quadratic Mutual Information for Discriminant Analysis

Vasileios Gavriilidis and Anastasios Tefas

Aristotle University of Thessaloniki, Department of Informatics,
Thessaloniki, Greece
vgavril@csd.auth.gr, tefas@aiia.csd.auth.gr

Abstract. Novel criteria that reformulate the Quadratic Mutual Information according to Fisher's Discriminant Analysis are proposed for supervised dimensionality reduction. The proposed method uses a quadratic divergence measure and requires no prior assumptions about class densities. The criteria are optimized using gradient ascent with initialization using random or LDA based projections. Experiments on various datasets are conducted and highlight the superiority of the proposed approach compared to the standard QMI criterion.

Keywords: Renyi Entropy, Parzen estimator, Feature transform, Feature extraction, Mutual information.

1 Introduction

Dimensionality reduction is a commonly used step in machine learning, especially when dealing with a high dimensional space of features. The original feature space is mapped onto a new, reduced dimensionality space and the examples to be used by machine learning algorithms are represented in that new space. Dimensionality reduction saves memory usage for storing training patterns and reduces the computation required for distance calculation. This way we improve performance and alleviate the effect of the curse of dimensionality [2]. Apart from time, dimensionality reduction is also crucial in terms of separability, thus a good selection or extraction lies to the criterion to be evaluated and enhanced.

Feature extraction, uses a transform to lower dimensions such as a projection matrix, which maximizes or minimizes a given criterion. The data transformation may be linear, as in Principal Component Analysis (PCA) [6] or Independent Component Analysis (ICA) [1], but many nonlinear dimensionality reduction techniques also exist such as kernel PCA [10].

In pattern classification, we are interested in methods that best separate the classes. Such a technique is Linear Discriminant Analysis (LDA), where a transform is produced that enhances the discrimination between data in different classes [5]. LDA assumes that samples are normally distributed, although techniques have been proposed for bypassing that problem [7]. In addition, LDA is limited to the number of features it can produce which is, $N_c - 1$ where N_c is the number of classes, but extensions have been proposed to overcome this [8].

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 90–97, 2012.

Information theory provides us measures that can be used to optimize class separability. Mutual information between the class labels and the transformed data to fewer dimensions acts as a more general criterion that overcomes many limitations of the methods discussed above. An even more sophisticated approach is given in [12]. This approximation is inspired by the quadratic Renyi entropy, it is differentiable and it can both avoid the knowledge of density of the classes and be applied to large training datasets. It can provide the ability to perform linear mappings for clustering [13] and even feature extraction [9].

A combination of QMI and LDA is introduced in this paper to provide a novel dimensionality reduction method that enhances class separability. The proposed approach uses the definition of QMI in order to reformulate criteria inspired by LDA. The proposed criteria are given in the form of ratios that enforce the within class similarity and between class dissimilarity. The novel optimization criteria then can be efficiently optimized using gradient ascent and update rules are derived for the projection matrix.

The manuscript is organized as follows. The derivation of QMI starting from Shannon's entropy definition is described in Section 2. The novel criteria that are inspired by LDA and combine measures that appear in QMI criterion are presented in Section 3. Classification results using nearest neighbor classifier in several datasets from the UCI machine learning repository are given in Section 4. Finally, conclusions are drawn in Section 5.

2 Prior Work and Problem Statement

Assume a random variable Y that models, $\mathbf{y}_i \in R^d$, that represent the projected input data, $\mathbf{x}_i \in R^D$, where $D > d$, and a discrete-value random variable C representing class labels taking values from 1 to N_c. The projected data are calculated by the product of each sample $\mathbf{x}_i$ with the projection matrix $\mathbf{W} \in R^{D \times d}$ hence, $\mathbf{y}_i = \mathbf{W}^T \mathbf{x}_i$.

Using Shannon's definition [11], the entropy of the discrete distribution, which is a measure of the randomness or unpredictability of a sequence of symbols, is $H(C) = -\sum_c P(c) log\left(P(c)\right)$, where P denotes a probability while a lower case p denotes probability density. In general, the mutual information expresses the reduction in uncertainty about one variable due to the knowledge of the other variable, hence it can measure dependence between two variables, in our case the difference $H\left(C\right) - H\left(C|Y\right)$ is the uncertainty about the class C by observing the feature vector $\mathbf{y}$. It is defined as:

$$\begin{aligned} I\left(C,Y\right) &= H\left(C\right) - H\left(C|Y\right) \\ &= \sum_c \int_{\mathbf{y}} p\left(c,\mathbf{y}\right) log\left(\frac{p\left(c,\mathbf{y}\right)}{P\left(c\right)p\left(\mathbf{y}\right)}\right) d\mathbf{y} \end{aligned} \tag{1}$$

Torkkola [12] has proposed the quadratic mutual information between the data sample and the corresponding class labels to be calculated as:

$$I_T\left(C,Y\right) = V_{IN} + V_{ALL} - 2V_{BTW} \tag{2}$$

where:

- V_{IN} is expressing the interactions between pairs of samples inside each class, summed over all classes.
- V_{ALL} is expressing the interactions between all pairs of samples, regardless of class, weighted by the sum of squared class priors.
- V_{BTW} is expressing the interactions between samples of a particular class against all samples weighed by the class prior and summed over all classes.

Details can be found on [12]. The objective is to find a transform g such that $\mathbf{y} = g(\mathbf{x}_i)$ maximizes $I_T(C, Y)$.

3 Update Methods Inspired by Discriminant Analysis

The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible. In general, in order to use LDA for multiple classes, we first define the scatter matrices, $\mathbf{S}_B$ and $\mathbf{S}_W$ which are the between classes scatter matrix and the within class scatter matrix, respectively. Those matrices are computed as:

$$\mathbf{S}_B = \sum_{c=1}^{N_c} P(c)(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T \tag{3}$$

$$\mathbf{S}_W = \sum_{c=1}^{N_c} \sum_{i \in c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T \tag{4}$$

where $\boldsymbol{\mu}$ is the overall mean vector of the data, $\boldsymbol{\mu}_c$ is the mean vector of class c and $\mathbf{x}_i$ is the i-th vector that belongs to class c. Scatter matrices are quite similar to V_{IN} and V_{BTW}, however, whereas V_{IN} and V_{BTW} represent similarity, $\mathbf{S}_B$ and $\mathbf{S}_W$ represent dissimilarity. The corresponding criterion that needs maximization in the case of LDA is the following:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{5}$$

That is, based on (5), we propose a criterion inspired by QMI that has the form of (5). In order to increase the sample interactions inside each class while decreasing the sample interactions of different classes, we propose the transform of (2) to any of the following two criteria:

$$I_B(C, Y) = \frac{V_{IN}}{V_{BTW}} \tag{6}$$

$$I_A(C, Y) = \frac{V_{IN}}{V_{ALL}} \tag{7}$$

Let N be the number of samples, J_p the number of samples for each class, c_p, and $G(\mathbf{y}, \boldsymbol{\Sigma})$, be a n–dimensional Gaussian function where $\boldsymbol{\Sigma}$ is the covariance

matrix. The prior probability of each class is $P(c_p) = \frac{J_p}{N}$, thus, $\sum_{p=1}^{N_c} J_p = N$. The Parzen density estimation, that corresponds to the density of each class, the joint density, as well as the density of all classes is given by:

$$p(\mathbf{y}|c_p) = \frac{1}{J_p} \sum_{j=1}^{J_p} G\left(\mathbf{y} - \mathbf{y}_{pj}, \sigma^2 I\right) \tag{8}$$

$$p(c_p, \mathbf{y}) = \frac{1}{N} \sum_{j=1}^{J_p} G\left(\mathbf{y} - \mathbf{y}_{pj}, \sigma^2 I\right) \tag{9}$$

$$p(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} G\left(\mathbf{y} - \mathbf{y}_i, \sigma^2 I\right) \tag{10}$$

respectively, where I is the identity matrix. Now analysing, and computing components in (6) and (7) while using Parzen density estimation given in (8), (9) and (10) we obtain:

$$V_{IN} = \sum_c \int_{\mathbf{y}} p\left(c, \mathbf{y}\right)^2 \, dy = \frac{1}{N^2} \sum_{p=1}^{N_c} \sum_{k=1}^{J_p} \sum_{l=1}^{J_p} G\left(\mathbf{y}_{pk} - \mathbf{y}_{pl}, 2\sigma^2 I\right) \tag{11}$$

$$V_{ALL} = \sum_c \int_{\mathbf{y}} P\left(c\right)^2 p\left(\mathbf{y}\right)^2 \, dy = \frac{1}{N^2} \left(\sum_{p=1}^{N_c} \left(\frac{J_p}{N} \right)^2 \right) \sum_{k=1}^{N} \sum_{l=1}^{N} G\left(\mathbf{y}_k - \mathbf{y}_l, 2\sigma^2 I\right) \tag{12}$$

$$V_{BTW} = \sum_c \int_{\mathbf{y}} p\left(c, \mathbf{y}\right) P\left(c\right) p\left(\mathbf{y}\right) \, dy = \frac{1}{N^2} \sum_{p=1}^{N_c} \frac{J_p}{N} \sum_{j=1}^{J_p} \sum_{k=1}^{N} G\left(\mathbf{y}_{pj} - \mathbf{y}_k, 2\sigma^2 I\right) \tag{13}$$

It is straightforward to show that if all classes have the same number of samples then $V_{ALL} = V_{BTW}$ that is, if all classes have the same probability to occur then (12) becomes equal to (13).

All different measures given in (2), (6) and (7), need a maximization update rule for the given projection matrix $\mathbf{W}$. The projections of the input data, derive directly from the projection matrix $\mathbf{W}$, thus:

$$\mathbf{W} = \arg\max_{\mathbf{W}} \left(I(\{c_i, \mathbf{y}_i\}) \right) \tag{14}$$

Unfortunately, the optimization of the criterion in (14) can not be solved analytically, hence, a numerical optimization is needed. Using gradient ascent with learning rate, ρ for updating $\mathbf{W}$, the update rule of the projection matrix can be the following:

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \rho \frac{\partial I}{\partial \mathbf{W}} \tag{15}$$

Using the chain rule, one can obtain the following:

$$\frac{\partial I}{\partial \mathbf{W}} = \sum_{i=1}^{N} \frac{\partial I}{\partial \mathbf{y}_i} \frac{\partial \mathbf{y}_i}{\partial \mathbf{W}} = \sum_{i=1}^{N} \frac{\partial I}{\partial \mathbf{y}_i} \mathbf{x}_i^T \tag{16}$$

hence, we need to find the derivatives of (2), (6) and (7). We also impose the constraint, $\mathbf{W}^T\mathbf{W} = I$, in order to have an orthonormal subspace as solution and prevent convergence to trivial infinite solutions, hence after updating $\mathbf{W}$ Gram–Schmidt orthonormalization is used.

Derivatives represent the direction where each sample would likely move after the transformation is applied. Firstly, we know that the derivative of the potential between two samples is computed as:

$$\frac{\partial}{\partial \mathbf{y}_i} G\left(\mathbf{y}_i - \mathbf{y}_j, 2\sigma^2 I\right) = G\left(\mathbf{y}_i - \mathbf{y}_j, 2\sigma^2 I\right) \frac{(\mathbf{y}_j - \mathbf{y}_i)}{2\sigma^2} \tag{17}$$

We can now perform gradient ascent using the derivatives of (11), (12) and (13), which are given by:

$$\frac{\partial}{\partial \mathbf{y}_{ci}} V_{IN} = \frac{1}{N^2\sigma^2} \sum_{k=1}^{J_c} G\left(\mathbf{y}_{ck} - \mathbf{y}_{ci}, 2\sigma^2 I\right)(\mathbf{y}_{ck} - \mathbf{y}_{ci}) \tag{18}$$

$$\frac{\partial}{\partial \mathbf{y}_{ci}} V_{ALL} = \frac{1}{N^2\sigma^2} \left(\sum_{p=1}^{N_c} \left(\frac{J_p}{N}\right)^2\right) \sum_{k=1}^{N} G\left(\mathbf{y}_k - \mathbf{y}_i, 2\sigma^2 I\right)(\mathbf{y}_k - \mathbf{y}_i) \tag{19}$$

$$\frac{\partial}{\partial \mathbf{y}_{ci}} V_{BTW} = \frac{1}{N^2\sigma^2} \sum_{p=1}^{N_c} \frac{J_p + J_c}{2N} \sum_{j=1}^{J_p} G\left(\mathbf{y}_{pj} - \mathbf{y}_{ci}, 2\sigma^2 I\right)(\mathbf{y}_{pj} - \mathbf{y}_{ci}) \tag{20}$$

We can now calculate the gradient of (2), (6) and (7) as follows:

$$\frac{\partial I_T}{\partial \mathbf{y}_i} = \frac{\partial V_{IN}}{\partial \mathbf{y}_i} + \frac{\partial V_{ALL}}{\partial \mathbf{y}_i} - 2\frac{\partial V_{BTW}}{\partial \mathbf{y}_i} \tag{21}$$

$$\frac{\partial I_B}{\partial \mathbf{y}_i} = \frac{V_{BTW}}{V_{BTW}^2}\frac{\partial V_{IN}}{\partial \mathbf{y}_i} - \frac{V_{IN}}{V_{BTW}^2}\frac{\partial V_{BTW}}{\partial \mathbf{y}_i} \tag{22}$$

$$\frac{\partial I_A}{\partial \mathbf{y}_i} = \frac{V_{ALL}}{V_{ALL}^2}\frac{\partial V_{IN}}{\partial \mathbf{y}_i} - \frac{V_{IN}}{V_{ALL}^2}\frac{\partial V_{ALL}}{\partial \mathbf{y}_i} \tag{23}$$

These gradients can be calculated using (18), (19) and (20). Using (21), (22) and (23) in (16) and the result in (15) we derive the update rules for updating the projection matrix $\mathbf{W}$ until convergence.

Except for the calculation of gradients, we also need to somehow initialize $\mathbf{W}$. There are many options, among them initializations based on linear feature extraction, that one can find in the literature, like PCA, ICA and LDA can be used. In the proposed approach we use random values and LDA as proposed in [12].

4 Experimental Results

We compared the performance of dimensionality reduction using the standard QMI definition against the criteria proposed in (6) and (7) which are called for simplicity MI_B and MI_A, respectively. To do so, several databases from the UCI

machine learning repository [4] have been used in the experiments. The datasets that were used are presented in Table 1.

All these datasets were scaled to the interval $[0, 1]$. To evaluate the test error on the various experiments we used 5×2 fold cross validation. Moreover, the classifier that was used was a k–nearest neighbor and we provide results for $\mathbf{W}$ initializations using both random projections and LDA projections. We should also mention that the same random initializations have been used for all criteria.

In addition, the number of dimensions that we tested are from 1 to $N_c - 1$ due to LDA limitations, although we do not show all the results in CMU face images because of the large number of classes it possess. As explained earlier, measures MI_B in (6) and MI_A in (7) have no difference when all classes have the same number of samples, so test results of these measures are substituted with the test results of the measure MI_{BA} which can be either one. Tables 2 - 4 show classifications test error results, where in the first column the measure to be maximized and the initialization of $\mathbf{W}$ is given.

Table 1. UCI Machine Learning Repository Data Sets Characteristics

Database	Samples	Dimension	Classes
CMU faces	640	960	20
Balance	625	4	3
Ionosphere	351	34	2
Ecoli	336	7	8
Wine	178	13	3
Iris	150	4	3

CMU Face Images. This data consists of 640 greyscale face images of people taken with varying pose, expression, eyes (wearing sunglasses or not). There are 32 images for each person capturing every variation combination. Each sample image was resized to 30×32. Observing Table 2, we can notice that the proposed MI_{BA} criterion is much better when $\mathbf{W}$ is initialized by LDA. In addition we performed Dietterich f statistical test [3], and gained a value over 8, on LDA initialization, hence the error rates difference between the criteria is statistically significant.

Ecoli. This dataset contains protein localization sites. As can be seen in Table 3, proposed criteria are better than the standard QMI criterion in both initialization methods. In addition, in figure 1 a projection in two dimensions revealing

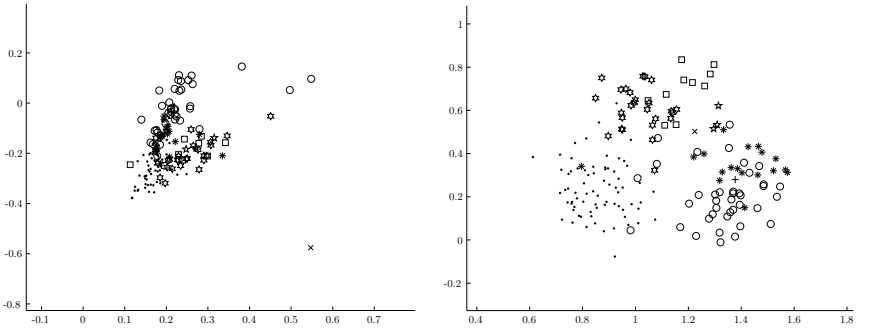
Table 2. Error rates on CMU faces Images

Dimension	1	2	3	4	8	12	15	16	17	18	19
QMI, Random	52.50	28.94	22.69	15.06	17.50	19.62	13.56	15.62	13.81	15.44	**10.94**
MI_{BA}, Random	74.19	59.75	48.06	42.00	19.19	19.94	13.56	15.62	13.81	15.44	**10.94**
QMI, LDA	49.50	24.00	12.00	7.19	5.94	4.50	4.87	6.12	3.44	4.94	4.56
MI_{BA}, LDA	27.25	4.38	3.44	3.25	2.81	2.25	2.69	**1.94**	2.94	3.81	5.25

information about class separability and compactness is given. MI_A has produced a projection that attempts to separate all the classes. Standard QMI is very compact but fails to provide any separability between classes, while MI_A is superior for classification but is not as compact as the standard QMI.

Table 3. Error rates on Ecoli

Dimension	1	2	3	4	5	6	7
QMI, Random	36.63	32.29	25.90	20.60	19.76	17.23	15.66
MI_B, Random	35.90	23.37	17.71	17.47	16.75	15.78	15.66
MI_A, Random	32.77	22.89	18.31	16.99	16.99	**15.42**	15.66
QMI, LDA	36.27	28.31	29.40	22.53	20.72	17.59	13.13
MI_B, LDA	36.51	20.72	18.31	16.87	17.35	15.90	**13.01**
MI_A, LDA	36.39	21.33	18.07	16.87	17.95	16.14	13.13



(a) QMI (b) MI_A

Fig. 1. Projected samples of the Ecoli dataset using the same random initialization. The QMI criterion is given in (a) and MI_A criterion in (b).

Datasets with Small Number of Classes. Balance, Ionosphere, Wine and Iris are some datasets that possess a small number of classes and all results are shown in Table 4. Overall, MI_B or MI_A is superior against standard QMI.

Table 4. Error rates on databases with small number of classes

Database	Balance		Ionosphere	Wine		Iris	
Dimension	1	2	1	1	2	1	2
QMI, Random	23.78	**20.19**	42.40	33.18	**30.68**	4.53	4.80
MI_B, Random	65.77	36.73	44.11	32.95	30.91	4.27	**3.73**
MI_A, Random	66.09	37.24	**42.29**	32.95	31.36	4.27	**3.73**
QMI, LDA	8.97	10.26	39.20	32.73	30.45	8.53	3.47
MI_B, LDA	9.29	10.26	33.83	34.77	**29.55**	3.47	**2.67**
MI_A, LDA	**8.91**	10.26	**32.57**	34.77	29.77	3.47	**2.67**

5 Conclusions

A novel method for dimensionality reduction and feature extraction inspired by mutual information between features and class labels and using Linear Discriminant Analysis criteria is proposed. As it has been illustrated, we can substitute the standard QMI with the MI_{BA} criterion which attains in most cases better classification and separability characteristics. Future work will be directed on overcoming the limitation of LDA initialization.

References

1. Comon, P.: Independent component analysis, a new concept? Signal Processing 36(3), 287–314 (1994); Higher Order Statistics
2. Corporation, B.: Dynamic programming. Princeton University Press, City (1957)
3. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10(7), 1895–1923 (1998)
4. Frank, A., Asuncion, A.: UCI machine learning repository (2010), `http://archive.ics.uci.edu/ml`
5. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press (1990)
6. Kittler, J., Devijver, P.A.: Statistical properties of error estimators in performance assessment of recognition systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI 4(2), 215–220 (1982)
7. Kumar, N., Andreou, A.G.: Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. Speech Communication 26(4), 283–297 (1998)
8. Okada, T., Tomita, S.: An optimal orthonormal system for discriminant analysis. Pattern Recognition, 139–144 (1985)
9. Ozertem, U., Erdogmus, D., Jenssen, R.: Spectral feature projections that maximize shannon mutual information with class labels. Pattern Recogn. 39, 1241–1252 (2006)
10. Schölkopf, B., Smola, A., Müller, K.R.: Kernel principal component analysis (1999)
11. Shannon, C.: A mathematical theory of communication. Bell Systems Techn. Journal 27, 623–656 (1948)
12. Torkkola, K.: Feature extraction by non-parametric mutual information maximization. Journal of Machine Learning Research 3, 1415–1438 (2003)
13. Vera, P.A., Estévez, P.A., Principe, J.C.: Linear Projection Method Based on Information Theoretic Learning. In: Diamantaras, K., Duch, W., Iliadis, L.S. (eds.) ICANN 2010, Part III. LNCS, vol. 6354, pp. 178–187. Springer, Heidelberg (2010)