Guorong Xuan, Peiqi Chai, Minhui Wu,"Bhattacharyya Distance Feature Selection ",
Proceedings of the 13th International Conference on Pattern Recognition, vol. 2, pp. 195-199, 1996

Bhattacharyya Distance Feature Selection

Xuan Guorong Chai Peiqi Wu Minhui
Department of Computer Science, Tongji University
Shanghai 200092, P. R. of China

Abstract

A recursive algorithm named Bhattacharyya distance feature selection for selecting a real-optimum feature under normal multidistribution is presented. The key of this method is to change the problem of minimizing the criterion of sum of the upper bound of error probability of every two class pair in subspace to a problem of solving nonlinear matrix equation in multiclass problem under orthonormal coordinate system. The recursive algorithm is considered as finding the optimal solution of transformation matrix from the nonlinear matrix equation. The theoretical analysis and experimental results show that under normal multidistribution the performance of proposed algorithm is superior to the performance of any previous one .

Keywords: Bhattacharyya distance, Upper bound of error probability, Nonlinear matrix equation, Recursive algorithm

1. Introduction

The problem of feature selection is very important in pattern recognition. The probability of error is the optimum measure of feature effectiveness. When the criterion is not related to the probability of error directly, upper bound of error probability can be used instead. Foley and Sammon [1] proposed a feature selection method based on the generalized Fisher criterion under the orthonormal condition. Okada and Tomita [2] proposed a feature selection method based on the orthonormal discriminant vector (ODV) In 1984, which generalized the method of Foley and Sammon. However, these criteria do not have a direct relationship to either probability of error or upper bound of error probability. Bhattacharyya distance can be used as the criterion , which is related to the upper bound of error probability . Bhattacharyya distance can be divided into two terms separately. Fukunaga [3] or Henderson [4] proposed analytical procedure to select a near-optimum feature under Bhattacharyya distance. This is an approximate method or named suboptimum feature selection under special condition. The method is only available when either term of the Bhattacharyya distance criterion formula is dominant. A recursive algorithm named Bhattacharyya feature selection to select a real-optimum feature under normal multidistribution is presented in this paper. This is an accurate method under common condition. The method is available for any value of terms of the Bhattacharyya distance criterion formula. The method of Bhattacharyya feature selection is to change the problem of minimizing the criterion of sum of the upper bound of error probability of every two class pair in subspace to a problem of solving nonlinear matrix equation under orthonormal coordinate system. The recursive algorithm is considered as finding the optimal solution of transformation matrix from the nonlinear matrix equation. The effectiveness of the new Bhattacharyya feature selection is shown by some numerical examples.

2. Bhattacharyya Distance and Error Probability

For two-class problem under normal distribution, the upper bound of error probability in the n-dimensional original space can be shown as Chernoff Bound [8]

$$e_{chernoff} = P(w_i)^{1-s} P(w_j)^{s} \exp\left[-m_{ij}(s)\right]$$

where $P(w_i)$, $P(w_j)$ are the prior probability for class i and j. If s=1/2, Chernoff Bound is equal to e_{pij} approximately.

$$e_{pij} = \left\{ P(w_i) \cdot P(w_j) \right\}^{\frac{1}{2}} \exp\left[-m_{ij}(1/2)\right]$$

$$\because \left\{ P(w_i) \cdot P(w_j) \right\}^{\frac{1}{2}} \le \frac{1}{2}$$

the upper bound of error probability can be changed to e_{ij}

$$e_{pij} \le e_{ij} = \frac{1}{2} \exp\left[-m_{ij}(1/2) \right]$$

The Bhattacharyya distance between class i and j

$$m_{ij}(1/2) = \frac{1}{8} tr\left[S_{wij}^{-1} S_{bij} \right] + \frac{1}{2} \ln \frac{|S_{wij}|}{|S_{wi}|^{\frac{1}{2}} \cdot |S_{wj}|^{\frac{1}{2}}}$$

where:

Within-class scatter matrix for class i.

$$S_{wi} = E\left[(X - M_i)(X - M_i)^t \right]$$

Within-class scatter matrix for class j

$$S_{wj} = E\left[(X - M_j)(X - M_j)^t \right]$$

Within-class scatter matrix for class i and j

$$S_{wij} = \frac{1}{2}(S_{wi} + S_{wj})$$

Between-class scatter matrix.

$$S_{bij} = (M_i - M_j)(M_i - M_j)^t$$

It is obvious, that the Bhattacharyya distance formula has two terms. The first term is considered as Mahalanobis distance, which related to both mean and variance of the sample distribution of two classes. The second term is only related to variance of the sample distribution of these two classes.

For L-class normal multidistribution oriented problem in the n-dimensional original space ,we have the sum of the upper bound of error probability of every two class pair as

$$e = \sum_{i>j}^{L} \sum_{j=1}^{L} e_{ij} = \frac{1}{2} \sum_{i>j}^{L} \sum_{j=1}^{L} \exp\left[-m_{ij}(1/2) \right]$$

where $e_{ij}(i, j = 1,2,\cdots, L)$ is the upper bound of error probability of every two class pair.

In selected m-dimensional feature space (m<n), the Bhattacharyya distance between class i and j

$$m_{mij}(1/2) = \frac{1}{8} tr\left[S_{wmij}^{-1} S_{bmij} \right] + \frac{1}{2} \ln \frac{|S_{wmij}|}{|S_{wmi}|^{\frac{1}{2}} \cdot |S_{wmj}|^{\frac{1}{2}}}$$

where

Within-class scatter matrix for class i

$$S_{wmi} = A^t S_{wi} A$$

Within-class scatter matrix for class j

$$S_{wmj} = A^t S_{wj} A$$

Within-class scatter matrix for class i and j

$$S_{wmij} = A^t S_{wij} A$$

where A is the transformation matrix of $n \times m$.

For L-class normal multidistribution oriented problem in m-dimensional selected space ,we have the sum of the upper bound of error probability of every two class pair as [5]

$$e_m = \sum_{i>j}^{L} \sum_{j=1}^{L} e_{mij} = \frac{1}{2} \sum_{i>j}^{L} \sum_{j=1}^{L} \exp\left[-m_{mij}(1/2) \right]$$

where $e_{mij}(i, j = 1,2,\cdots, L)$ is the upper bound of error probability of every two class pair in the m-dimensional selected space.

3. Bhattacharyya Distance Feature Selection for 2-Distribution

The aim of Bhattacharyya distance feature selection is to find a linear transformation under the orthonormal condition of coordinate axes that maps a n-dimensional feature space to a m-dimensional feature space in terms of minimizing the upper bound of error probability.

The derivative of $m_{mij}(1/2)$ with respect to A can be obtained as(see formula A16 and formula A27 at pp. 566-568 in [8])

$$\frac{dm_{mij}(1/2)}{dA}$$

$$= \frac{1}{8} \cdot \frac{d}{dA} \left\{ tr\left[\left(A^t S_{wij} A \right)^{-1} \left(A^t S_{bij} A \right) \right] \right\}$$

$$+ \frac{1}{2} \frac{d}{dA} \left[\ln \frac{|A^t S_{wij} A|}{|A^t S_{wi} A|^{\frac{1}{2}} \cdot |A^t S_{wj} A|^{\frac{1}{2}}} \right]$$

$$= \frac{1}{4} q_{mij} - \frac{1}{2} \left(S_{wi} A S_{wmi}^{-1} + S_{wj} A S_{wmj}^{-1} \right)$$

where

$$q_{mij} = S_{wij} \left\{ (S_{cij} + 4I) \cdot A - A \cdot S_{cmij} \right\} S_{wmij}^{-1} \quad (1)$$

$$S_{cij} = S_{wij}^{-1} \cdot S_{bij}$$

$$S_{cmij} = S_{wmij}^{-1} \cdot S_{bmij} = \left(A^t S_{wij} A \right)^{-1} \left(A^t S_{bij} A \right)$$

For minimizing the upper bound of error probability, we have

$$\frac{de_{mij}}{dA} = 0$$

where

$$e_{mij} = \frac{1}{2}\exp\left[-m_{mij}(1/2)\right]$$

$$\frac{de_{mij}}{dA} = -\frac{1}{2}\frac{d\left(m_{mij}(1/2)\right)}{dA}\exp\left[-m_{mij}(1/2)\right]$$

$$= -\frac{1}{8}\left(q_{mij} - 2S_{wi}AS_{wmi}^{-1} - 2S_{wj}AS_{wmj}^{-1}\right)\cdot \exp\left[-m_{mij}(1/2)\right]$$

and q_{mij} from formula (1)

The matrix equation for Bhattacharyya distance feature selection under 2-distribution can be obtained as

$$U = 0$$

where

$$U = \left(q_{m} - 2S_{wi}AS_{wmi}^{-1} - 2S_{wj}AS_{wmj}^{-1}\right)\cdot \exp\left[-m_{mij}(1/2)\right]$$

4. Bhattacharyya Distance Feature Selection for L-Distribution

For L-class normal multidistribution oriented problem in m-dimensional selected space , the derivative of sum of the upper bound of error probability e_m with respect to A can be obtained as

$$\frac{de_m}{dA} = \sum_{i>j}^{L}\sum_{j=1}^{L}\frac{de_{mij}}{dA}$$

$$= \frac{1}{2}\sum_{i>j}^{L}\sum_{j=1}^{L}\frac{d\left(m_{mij}(1/2)\right)}{dA}\cdot\exp\left[-m_{mij}(1/2)\right] = 0$$

where

$$e_m = \sum_{i>j}^{L}\sum_{j=1}^{L}e_{mij} = \frac{1}{2}\sum_{i>j}^{L}\sum_{j=1}^{L}\exp\left[-m_{mij}(1/2)\right]$$

The matrix equation for Bhattacharyya distance feature selection under multidistribution is

$$U = 0$$

where

$$U = \sum_{i>j}^{L}\sum_{j=1}^{L}\left(q_{mij} - 2S_{wi}AS_{wmi}^{-1} - 2S_{wj}AS_{wmj}^{-1}\right)\cdot \exp\left[-m_{mij}(1/2)\right]$$

and q_{mij} from formula (1)

5. Recursive Algorithm

5.1 Recursive algorithm formula

Let us construct an identical equation as

$$AV = AV$$

According to the matrix equation for Bhattacharyya distance feature selection $U = 0$, we have

$$AV = AV + a\cdot W\cdot U$$

$$A = \left(AV + a\cdot W\cdot U\right)\cdot V^{-1}$$

where a is the length of recursive step

$$V = \sum_{i>j}^{L}\sum_{j=1}^{L}S_{wmij}$$

$$W = \sum_{i>j}^{L}\sum_{j=1}^{L}S_{wij}^{-1}$$

Because the transformation matrix A should be under the orthonormal condition of coordinate axes, we must orthonormalize A as

$$A = C\cdot Q$$

where

$$C = \left(A\cdot V + a\cdot W\cdot U\right)\cdot V^{-1}$$

Q is the Gram-Schmidt orthonormalization matrix of $m\times m$, which orthonormalizes C to an orthonormal matrix A

5.2 Recursive procedure

- **To Initialize:**

Set initial transformation matrix

$$A_0 = \begin{bmatrix} 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ \vdots & \vdots & \cdots \\ 0 & 0 & \cdots \end{bmatrix}$$

of $n\times m$ .In fact, we can set A_0 as an arbitrary matrix ,whose rank is identical to m ; select an appropriate length of step a and a threshold d (i.e. 0.000001). We set initial step k=0, where k is the sequence number of recursion.

- **To Generate Successor**

When step is k, we have

$$C_k = \left(A_k\cdot V_k + a\cdot W\cdot U_k\right)\cdot V_k^{-1}$$

$$A_{k+1} = C_k\cdot Q_k$$

where

$$V_k = \sum_{i>j}^{L}\sum_{j=1}^{L}A_k^{t}S_{wij}A_k$$

$$U_k = \sum_{i>j}^{L}\sum_{j=1}^{L}\left(q_{kij} - 2S_{wi}AS_{wmki}^{-1} - 2S_{wj}AS_{wmkj}^{-1}\right)\cdot \exp\left[-m_{mkij}(1/2)\right]$$

where

$$q_{kij} = S_{wij}\left\{\left(S_{cij} + 4I\right)\cdot A_k - A_k\cdot S_{cmkij}\right\}S_{wmkij}^{-1}$$

In the m-dimensional feature space, the Bhattacharyya distance between class i and j

$$m_{mkij}(1/2) = \frac{1}{8}tr\left[S_{wmkij}^{-1}S_{bmkij}\right] + \frac{1}{2}\ln\frac{|S_{wmkij}|}{|S_{wmki}|^{\frac{1}{2}}\cdot|S_{wmkj}|^{\frac{1}{2}}}$$

where:

Within-class scatter matrix for class I

$$S_{wmki} = A_k^{\,t}S_{wi}A_k$$

Within-class scatter matrix for class j

$$S_{wmkj} = A_k^{\,t}S_{wj}A_k$$

Within-class scatter matrix for class i and j

$$S_{wmkij} = A_k^{\,t}S_{wij}A_k$$

Between-class scatter matrix for class i and j.

$$S_{bmkij} = A_k^{\,t}S_{bij}A_k$$

- **To Obtain Optimal Solution**

The recursion is continued until

$$d_k < d$$

where

d_k is the norm of difference between new transformation matrix and before one

$$d_k = norm\left(A_{k+1} - A_k\right).$$

The optimal solution of transformation matrix A is obtained by recursion, that is $A = A_k$

6. Example

In this section, we show some numerical examples in order to illustrate the effectiveness of our algorithm.

Example 1: an eight-dimensional four-class case, n=8, C=4 ,the data set of within-class covariance matrices and mean vectors of the normal distribution is provided by T.Marill[1].The dimension of subspace is 2,m=2.

Example 2: an eight-dimensional two-class case, n=8, C=2 , the data set of within-class covariance matrices and mean vectors is same as the data set of class 1 and 2 in Example 1. The dimension of subspace is also 2, m=2.

- Table 1 shows the experimental results of the new algorithm and the traditional ones. From this table, we see that, for all cases, the upper bound of error probability of Bhattacharyya feature selection is smaller than those from traditional ones.
- Table 2 shows effectiveness and convergence under different length of recursive step. we see that if the length of step is too large, the recursion will become divergent.

Table 1. values of the upper bound of error probability

	upper bound of error probability	
Feature Selection	Example 1 : (n=8, m=2, C=4)	Example 2 : (n=8, m=2, C=2)
Original space	0.3353	0.0479
Bhattacharyya	0.8343	0.1021
Mahalanobis	0.8473	0.1080
ODV	0.8858	0.1254

Table 2. length of step, total number of recursive steps (Example 1 , Bhattacharyya feature selection)

length step a	recursive steps	threshold d	time cost (486/33, MATLAB)
0.01	319	0.000001	66 second
0.02	170	0.000001	35 second
0.04	89	0.000001	20 second
0.06	60	0.000001	14 second
≥ 0.08		divergent	

The accuracy of the result of transformation matrix equation is satisfied, because the U_k of matrix equation is very close to zero.

7. Conclusion

A new recursive algorithm under normal multidistribution for Bhattacharyya distance feature selection is proposed. The recursive algorithm is considered as finding the optimal solution of transformation matrix of the recursive algorithm the nonlinear matrix equation for multiclass problem with orthonormal coordinate axes. The experimental results are shown that for all cases, the upper bound of error probability in Bhattacharyya distance feature selection is smaller than those from traditional ones. The experimental results also show that convergence can be obtained by length of recursive step chosen and the accuracy of the result of transformation matrix equation is satisfied. Under normal multidistribution, the theoretical analysis and experimental results show that the performance of proposed algorithm is superior to that of any previous one.

Reference

[1] H.Foley and J.W.Sammon, Jr., " An Optimal set of Discriminant Vectors", IEEE Trans. on Computer, Vol.24, 1975, pp.281-189

[2] T.Okada and S.Tomita, " An Optimal Orthonormal System for Discriminant Analysis", Pattern Recognition, Vol. 18, No2, 1985, pp.139-144

[3] K.Fukunaga, " Introduction to Statistical Pattern Recognition", Academic Press, Inc., New York, 1972

[4] T.L.Henderson and D.G.Lainiotis, " Comments on Linear Feature Selection", IEEE Trans. Information Theory, IT-15, 1969, pp.729-730

[5] D.G.Lainiotis, " A Class of Upper Bound on Probability of Error for Multihypotheses Pattern Recognition", IEEE Trans. Information Theory, IT-15, 1969, pp.730-731

[6] T.Marill and D.M.Green, " On the Effectiveness of Receptors in Recognition Systems", IEEE Trans., IT-9,1963, pp.11-27

[7] Yi-Tzun Chien, " Interactive Pattern Recognition",1978

[8] K.Fukunaga, " Introduction to Statistical Pattern Recognition", 2nd Edition, Academic Press, Inc., Boston, 1990

[9] Xuan Guorong, " A New Feature Selection Method Based on Mahalanobis Distance", Proc. of 7th International Conference on Pattern Recognition, Montreal, Canada, July 30, 1984, pp.131-133

[10] Xuan Guorong, " The Optimal Characteristics of Mahalanobis Distance Feature Selection", Proc. of 2nd International Conference on Computer Society of IEEE, June 23, 1987, pp.914-919