Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 396780, 7 pages
http://dx.doi.org/10.1155/2013/396780

Research Article

Unsupervised Optimal Discriminant Vector Based Feature Selection Method

Su-Qun Cao[1,2] and Jonathan H. Manton[2]

[1] *Faculty of Mechanical Engineering, Huaiyin Institute of Technology, Huai'an 223003, China*
[2] *Department of Electrical and Electronic Engineering, University of Melbourne, Victoria, VIC 3010, Australia*

Correspondence should be addressed to Su-Qun Cao; caosuqun@hyit.edu.cn

Received 22 July 2013; Accepted 21 September 2013

Academic Editor: Baochang Zhang

An efficient unsupervised feature selection method based on unsupervised optimal discriminant vector is developed to find the important features without using class labels. Features are ranked according to the feature importance measurement based on unsupervised optimal discriminant vector in the following steps. First, fuzzy Fisher criterion is adopted as objective function to derive the optimal discriminant vector in unsupervised pattern. Second, the feature importance measurement based on elements of unsupervised optimal discriminant vector is defined to determine the importance of each feature. The features with little importance measurement are removed from the feature subset. Experiments on UCI dataset and fault diagnosis are carried out to show that the proposed method is very efficient and able to deliver reliable results.

1. Introduction

Feature selection (FS) has become an active research topic in the area of pattern recognition, machine learning, data mining, intelligent fault diagnosis, and so forth. It is performed to choose a subset of the original features by removing redundant and noisy features from high-dimensional datasets in order to reduce computational cost, increase the classification accuracy, and improve result comprehensibility.

In the supervised FS algorithms, since class labels are available in supervised learning, various feature subsets are evaluated using some function of prediction accuracy to select only those features which are related to or lead to the decision classes of the data under consideration. There are numerous supervised feature selection methods [1–7] such as Fisher criterion [1, 2], Relief [3], and Relief-F [4].

However, for many existing datasets, class labels are often unknown or incomplete because large amounts of data make it difficult for humans to manually label the categories of each instance. Moreover, human labeling is expensive and subjective. Thus, it indicates the significance of unsupervised dimensionality reduction. Principal component analysis (PCA) [8] is often used in unsupervised pattern. However, PCA creates new features or principal components which are functions of original features. It is difficult to obtain intuitive understanding of the data using the new features only. Some unsupervised feature selection methods [8–14] have been proposed such as SUD [9]. SUD, which is a sequential backward selection algorithm to determine the relative importance of variables for Unsupervised Data, uses entropy similarity measurement to determine the importance of features with respect to the underlying clusters.

It is known to us that the famous Fisher criterion which can derive optimal discriminant vector is commonly used to realize feature dimension reduction in supervised pattern. In the unsupervised pattern, how to overcome the lack of the class information to realize feature selection is a worthy topic.

2. An Overview of Optimal Discriminant Vector

Fisher criterion is a discriminant criterion function that was first proposed by Fisher. It is based on the between-class scatter and the within-class scatter. By maximizing this criterion, one can obtain an optimal discriminant vector.

After the sample is projected to this vector, the within-class scatter is minimized and the between-class scatter is maximized [15].

Given c pattern classes $X^{(i)} = [x_i^1, x_i^2, ..., x_i^{N_i}]$ in the pattern set which contains N d-dimensional patterns, where $i = 1, 2, \ldots, c$, N_i is the number of all the patterns in the ith class; thus, $N = N_1 + N_2 + \cdots + N_c$. Fisher criterion is defined as follows:

$$J_{\mathrm{FC}}(\omega) = \frac{\omega^T S_b \omega}{\omega^T S_w \omega}, \tag{1}$$

where S_b is the between-class scatter matrix denoted by

$$S_b = \sum_{i=1}^{c} \frac{N_i}{N} (m_i - \overline{x})(m_i - \overline{x})^T, \tag{2}$$

and S_w is the within-class scatter matrix denoted by

$$S_w = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N_i} \left(x_j^i - m_i\right)\left(x_j^i - m_i\right)^T, \tag{3}$$

where m_i denotes the mean of the ith class, and $\overline{x}$ denotes the mean of all the patterns in the pattern set.

In order to seek an optimal discriminant vector ω by maximizing the Fisher criterion, the optimal discriminant vector ω^* can be obtained by solving the following eigensystem equation:

$$S_w^{-1} S_b \omega^* = \lambda \omega^*, \tag{4}$$

where λ is diagonal and consists of the corresponding eigenvalues. When the inverse of S_w exists, ω^* can be obtained by the maximum eigenvalue of $S_w^{-1} S_b{}^*$.

3. Unsupervised Optimal Discriminant Vector Based Feature Selection Method

Fisher criterion mentioned above can only be used in supervised pattern. This means that traditional optimal discriminant vector cannot be calculated directly by the unlabeled samples. Cao et al. [16] introduce fuzzy theory into Fisher criterion and define fuzzy Fisher criterion. Maximizing this criterion cannot only realize clustering but also obtain optimal discriminant vector.

Suppose that the membership function $u_{ij} \in [0, 1]$ with $\sum_{i=1}^{c} u_{ij} = 1$ for all j and the fuzzy index $m > 1$ is a given real value, where u_{ij} denotes the degree of the jth d-dimensional pattern belonging to the ith class; we can define the following fuzzy within-class scatter matrix S_{fw}:

$$S_{fw} = \sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij}^m \left(x_j - m_i\right)\left(x_j - m_i\right)^T \tag{5}$$

and the following fuzzy between-class scatter matrix S_{fb}:

$$S_{fb} = \sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij}^m (m_i - \overline{x})(m_i - \overline{x})^T. \tag{6}$$

Thus, we can derive fuzzy Fisher criterion as follows:

$$J_{\mathrm{FFC}} = \frac{\omega^T S_{fb} \omega}{\omega^T S_{fw} \omega}. \tag{7}$$

It is obvious that maximizing J_{FFC} directly in (7) is not a trivial task due to the existence of its denominator. However, we can reasonably relax this problem by applying the following Lagrange multipliers; λ and β_j $(j = 1, 2, \ldots n)$ together with the constraint $\sum_{i=1}^{c} u_{ij} = 1$ to (7):

$$F = \omega^T S_{fb} \omega - \lambda \omega^T S_{fw} \omega + \sum_{j=1}^{N} \beta_j \left(\sum_{i=1}^{c} u_{ij} - 1 \right). \tag{8}$$

Setting $\partial F / \partial \omega$ to be zero, we have

$$S_{fw}^{-1} S_{fb} \omega = \lambda \omega, \tag{9}$$

where ω is the eigenvector belonging to the largest eigenvalue λ of $S_{fw}^{-1} S_{fb}$.

Setting $\partial F / \partial m_i$ to be zero, we have

$$m_i = \frac{\sum_{j=1}^{N} u_{ij}^m \left(x_j - (1/\lambda)\,\overline{x}\right)}{\sum_{j=1}^{N} u_{ij}^m \left(1 - (1/\lambda)\right)}. \tag{10}$$

Here, m_i is a local maximum of F [17] proved in Appendix.

Setting $\partial F / \partial u_{ij}$ to be zero, we have

$$u_{ij} = \frac{\left(\omega^T \left(x_j - m_i\right)\left(x_j - m_i\right)^T \omega - (1/\lambda)\,\omega^T (m_i - \overline{x})(m_i - \overline{x})^T \omega\right)^{-1/(m-1)}}{\sum_{k=1}^{c} \left(\omega^T \left(x_j - m_k\right)\left(x_j - m_k\right)^T \omega - (1/\lambda)\,\omega^T (m_k - \overline{x})(m_k - \overline{x})^T \omega\right)^{-1/(m-1)}}. \tag{11}$$

When (11) is used, as stated previously, u_{ij} should satisfy $u_{ij} \in [0, 1]$; hence, in order to satisfy this constraint, we let $u_{ij} = 1$ and $u_{i'j} = 0$ for all $i' \neq i$, if

$$\omega^T \left(x_j - m_i\right)\left(x_j - m_i\right)^T \omega \leq \frac{1}{\lambda} \omega^T (m_i - \overline{x})(m_i - \overline{x})^T \omega. \tag{12}$$

With the above discussion, we can obtain the optimal discriminant vector ω in unsupervised pattern and then do feature selection based on ω. Now, let us illustrate this by the following experiment on 2-dimensional artificial dataset.

Figure 1 contains 168 2-dimensional samples. Through maximizing fuzzy Fisher criterion, we can obtain 2-class clustering result shown as red points and blue points, respectively, and can also get the vector $\omega = (\omega_1, \omega_2)' = (0.4562, -0.8899)'$ shown as a line in Figure 2. We project all samples to x-axis and y-axis. It is obvious that projective points in x-axis from different class are overlaping while those in y-axis are separated well. It means that Y feature is more important than X feature for leading to the decision classes. This is consistent with $|\omega_2| > |\omega_1|$ which gives us a revelation that we can apply the vector ω for feature selection.

Suppose $\omega = (\omega_1, \omega_2, \ldots \omega_d)'$; we define f_k as the k single feature importance measurement for comparison:

$$f_k = \frac{|\omega_k|}{\sum_{k=1}^{d} |\omega_k|}. \tag{13}$$

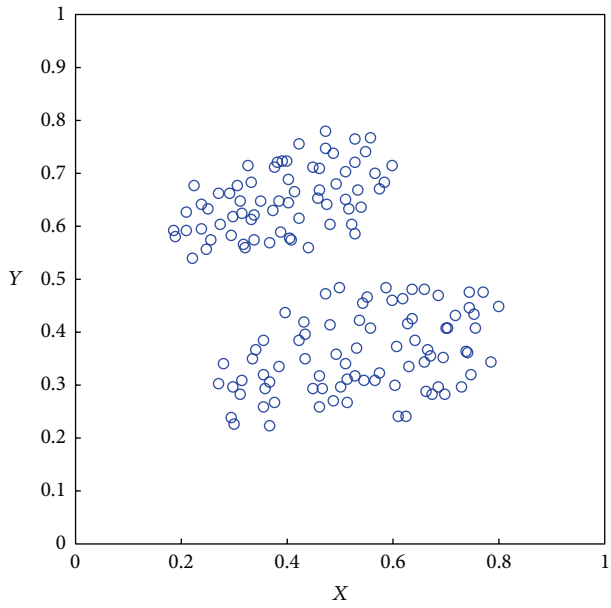


Figure 1: 2-dimensional artificial dataset.

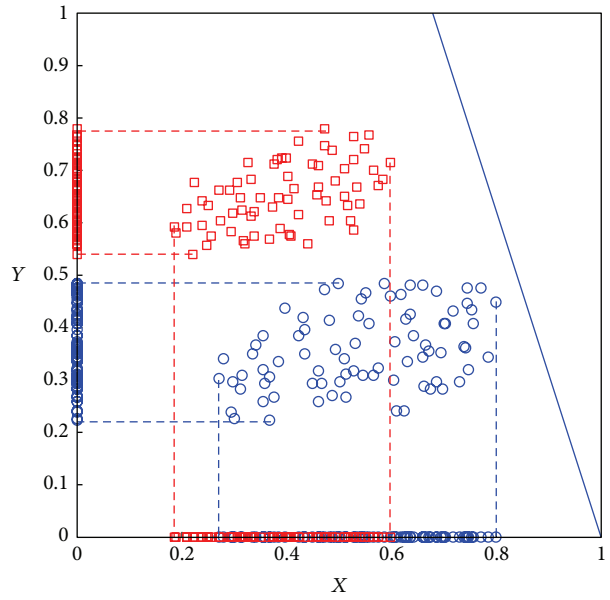


Figure 2: The clustering result based on fuzzy Fisher criterion (the solid line is the optimal discriminant vector ω).

To the above artificial dataset, $f_1 = 0.3389$ is the importance measurement of X feature and $f_2 = 0.6611$ is the importance measurement of Y feature.

Proposed Method

Step 1. Set the given threshold ε or the number of iterations α; initialize $U = [\mu_{ij}]_{c\times N}$ and $m = (m_1, m_2, \ldots, m_c)$ using K-means.

Step 2. Compute S_{fw}, S_{fb} using (5), (6), respectively.

Step 3. Compute the largest eigenvalue λ and the corresponding ω using (9).

Step 4. Update m_i and μ_{ij} using (10), (11), and (12), respectively.

Step 5. Compute J_{FFC}.

Step 6. If $J_{\mathrm{FFC}} < \varepsilon$ or the number of iterations $\geq \alpha$, go to Step 7; otherwise go to Step 2.

Step 7. Compute the feature importance measurements which are normalized as f_k. Then sort f_k by the descending order.

Step 8. Set the feature importance threshold θ.

Step 9. Find a feature subset size d_θ which is a minimize number making $\sum_{k=1}^{d_\theta} f_k$ no less than the threshold θ.

Step 10. Choose d_θ features corresponding to the sorted f_k in the descending order, that is, f_k $(k = 1, 2, \ldots, d_\theta)$, as the selected features and then terminate.

Different feature importance threshold θ leads to different feature subset size. In Step 7 of proposed method, features have already been sorted by the descending order. If the feature subset size d_θ is given from the start, we can simply select the first d_θ features. But if d_θ is not given, we can use θ to determine the feature subset size. The bigger θ is, the larger d_θ is. The recommended range of θ is from 0.8 to 0.95.

4. Experimental Results

4.1. Feature Selection on UCI Dataset Wine. In this experiment, the benchmarking UCI dataset *Wine* [18] was chosen to test the feature selection effectiveness of SUD, Relief-F and our method. We use the following Rand index [19] to evaluate the clustering performance of the dimension reduction data:

$$\mathrm{Rand}\left(P_1, P_2\right) = \frac{a+b}{n\times(n-1)/2}, \tag{14}$$

where P_1, P_2 denote the clustering results for the original dataset without noise and the corresponding noisy dataset, a denotes the number of any two patterns in the original dataset belonging to the same cluster in P_1, P_2, b denotes the number of any two patterns in the original dataset belonging to two different clusters in P_1, P_2, and n is the number of all patterns in the original dataset. Obviously, $\mathrm{Rand}(P_1, P_2) \in [0, 1]$. And $\mathrm{Rand}(P_1, P_2) = 1$ when P_1 is the same as P_2. The smaller $\mathrm{Rand}(P_1, P_2)$, the bigger the difference between P_1 and P_2. In other words, the corresponding algorithm has less robust capability in this case.

Table 1 illustrates the basic information of the dataset. We choose 130 samples which belong to class 1 and class 2 as testing dataset. The parameters for the proposed method are set as follows:

$$\varepsilon = 0.001, \qquad \alpha = 20, \qquad \theta = 0.90. \tag{15}$$

Table 1: Class distribution and features of *Wine* dataset.

Class	Number of samples	Features	
		Number	Name
Class 1	59	1	Alcohol
Class 2	71	2	Malic acid
Class 3	48	3	Ash
		4	Alcalinity of ash
		5	Magnesium
		6	Total phenols
		7	Flavanoids
		8	Nonflavanoid phenols
		9	Proanthocyanins
		10	Color intensity
		11	Hue
		12	OD280/OD315 of diluted wines
		13	Proline

Table 2: The feature importance measurement of *Wine* dataset.

Number	Number of features	The feature importance measurement f_k	$\sum_{k=1}^{d_\theta} f_k$
1	3	0.3180	0.3180
2	1	0.2709	0.5889
3	12	0.1793	0.7682
4	7	0.0520	0.8202
5	8	0.0445	0.8647
6	11	0.0364	0.9011
7	10	0.0301	0.9312
8	9	0.0243	0.9555
9	4	0.0241	0.9796
10	6	0.0088	0.9884
11	2	0.0057	0.9941
12	5	0.0049	0.9990
13	13	0.0009	0.9999

Table 2 lists the importance measurement of every feature computed by the proposed method. Due to the threshold θ, 6 features will be selected from original features. Figure 3 shows the Rand index values corresponding to the number of features using SUD, Relief-F and our method.

From Figure 3, we can easily find that data selected features by the proposed method have the best clustering result among these three algorithms.

4.2. Feature Selection for Fault Diagnosis. The steel plates faults dataset used in this experiment was donated by Semeion, Research Center of Sciences of Communication, Via Sersale 117, Rome, Italy [20, 21]. It classifies steel plates' faults into 7 different types: Pastry, Z_Scratch, K_Scratch, Stains, Dirtiness, Bumps, and Other_Faults. The dataset includes 1941 samples and every sample owns 27 independent features.

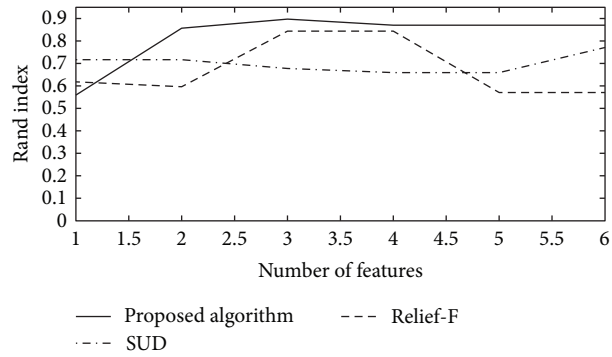


Figure 3: Rand index values corresponding to the number of features.

Table 3: Class distribution and features of steel plates dataset.

Class	Number of samples	Features	
		Number	Name
Pastry	158	1	X_Minimum
Z_Scratch	190	2	X_Maximum
K_Scratch	391	3	Y_Minimum
Stains	72	4	Y_Maximum
Dirtiness	55	5	Pixels_Areas
Bumps	402	6	X_Perimeter
Other_Faults	673	7	Y_Perimeter
		8	Sum_of_Luminosity
		9	Minimum_of_Luminosity
		10	Maximum_of_Luminosity
		11	Length_of_Conveyer
		12	TypeOfSteel_A300
		13	TypeOfSteel_A400
		14	Steel_Plate_Thickness
		15	Edges_Index
		16	Empty_Index
		17	Square_Index
		18	Outside_X_Index
		19	Edges_X_Index
		20	Edges_Y_Index
		21	Outside_Global_Index
		22	LogOfAreas
		23	Log_X_Index
		24	Log_Y_Index
		25	Orientation_Index
		26	Luminosity_Index
		27	SigmoidOfAreas

Table 3 shows class distribution and list of features. We choose 348 samples which belong to Pastry and Z_Scratch

Table 4: The feature importance measurement of steel plates dataset.

Number	Number of features	The feature importance measurement f_k	$\sum_{k=1}^{d_\theta} f_k$
1	19	0.2191	0.2191
2	15	0.1339	0.353
3	16	0.1283	0.4831
4	17	0.0777	0.559
5	13	0.0676	0.6266
6	12	0.0676	0.6942
7	20	0.0577	0.7519
8	27	0.0548	0.8067
9	21	0.051	0.8577
10	24	0.0343	0.892
11	22	0.0341	0.9261
12	23	0.0247	0.9508
13	25	0.0099	0.9607
14	2	0.0075	0.9682
15	1	0.0074	0.9756
16	26	0.007	0.9826
17	7	0.0044	0.987
18	6	0.0033	0.9903
19	4	0.0029	0.9932
20	3	0.0029	0.9961
21	14	0.0021	0.9982
22	9	0.001	0.9992
23	10	0.0005	0.9997
24	18	0.0001	0.9998
25	5	0.0001	0.9999
26	11	0	0.9999
27	8	0	0.9999

faults as testing dataset. The parameters for the proposed method are set as the previous experiment.

Table 4 lists the importance measurement of every feature computed by the proposed method. Due to the threshold θ, 11 features will be selected from original features. Figure 4 shows the Rand index values corresponding to the number of features using SUD, Relief-F, and our method.

Figure 4 shows that the proposed method is able to find the important features. It also shows that the performance of the proposed method without using class labels is very close to and sometimes better than that of SUD or Relief-F which ranks the original features using the class labels.

5. Conclusions

An efficient unsupervised feature selection method based on unsupervised optimal discriminant vector is developed to find the important features without using class labels. It adopts fuzzy Fisher criterion to derive the optimal discriminant vector in unsupervised pattern. It defines the single feature importance measurement based on unsupervised optimal discriminant vector to determine the importance of every feature. Two experiments on *Wine* dataset and fault diagnosis were carried out to show that the proposed method is able to find important features and is a reliable and efficient feature selection methodology compared to SUD and Relief-F. In the future, we will research how to introduce kernel techniques to the proposed method to enhance its applicability.

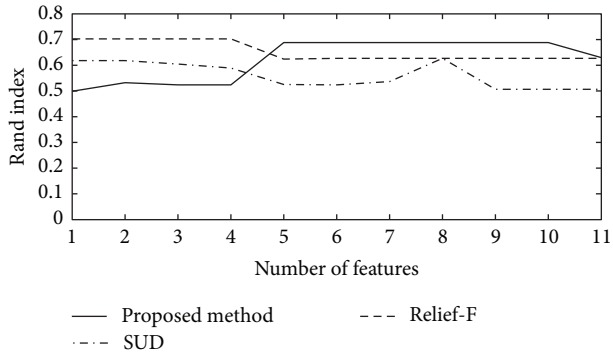


Figure 4: Rand Index values corresponding to the number of features.

Appendix

Proof of (10)

According to [22], we have

$$\frac{\partial\left(\omega^T S_{fb}\omega\right)}{\partial m_i} = 2\omega\omega^T\sum_{j=1}^{N}u_{ij}^m\left(m_i-\overline{x}\right),$$

$$\frac{\partial\left(\omega^T S_{fw}\omega\right)}{\partial m_i} = -2\omega\omega^T\sum_{j=1}^{N}u_{ij}^m\left(x_j-m_i\right),$$

$$\frac{\partial F}{\partial m_i} = 2\omega\omega^T\sum_{j=1}^{N}u_{ij}^m\left(m_i-\overline{x}\right)+2\lambda\omega\omega^T\sum_{j=1}^{N}u_{ij}^m\left(x_j-m_i\right)=0. \tag{A.1}$$

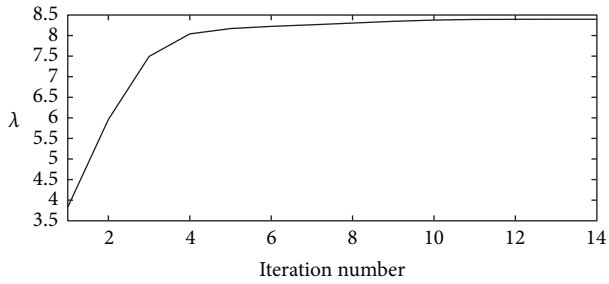
Premultiply by ω^T on both sides,

$$\omega^T\omega\omega^T\sum_{j=1}^{N}u_{ij}^m\left(m_i-\overline{x}\right)+\lambda\omega^T\omega\omega^T\sum_{j=1}^{N}u_{ij}^m\left(x_j-m_i\right)=0, \tag{A.2}$$

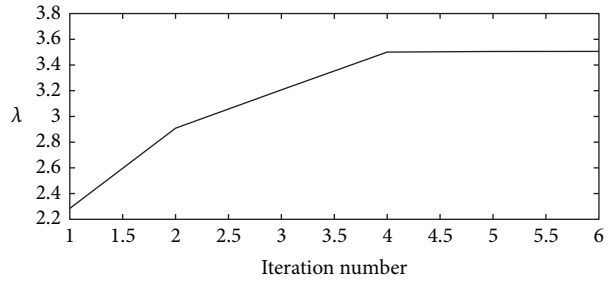
$$\omega^T\left(\sum_{j=1}^{N}u_{ij}^m\left(m_i-\overline{x}\right)+\lambda\sum_{j=1}^{N}u_{ij}^m\left(x_j-m_i\right)\right)=0. \tag{A.3}$$

We cannot solve m_i from the above equation. But it is obvious that the following equation is the particular solution of (A.3):

$$\sum_{j=1}^{N}u_{ij}^m\left(m_i-\overline{x}\right)+\lambda\sum_{j=1}^{N}u_{ij}^m\left(x_j-m_i\right)=0, \tag{A.4}$$



(a) *Wine* Dataset



(b) Steel plates dataset

Figure 5: The λ curve.

that is,

$$m_i = \frac{\sum_{j=1}^{N} u_{ij}^m \left(x_j - (1/\lambda)\,\overline{x}\right)}{\sum_{j=1}^{N} u_{ij}^m \left(1 - (1/\lambda)\right)}. \tag{A.5}$$

Now we proof that (A.5) is a local maximum of F.

We find

$$\begin{aligned} \frac{\partial^2 F}{\partial {m_i}^2} &= 2\omega\omega^T \sum_{j=1}^{N} u_{ij}^m - 2\lambda\omega\omega^T \sum_{j=1}^{N} u_{ij}^m \\ &= 2\left(1-\lambda\right)\omega\omega^T \sum_{j=1}^{N} u_{ij}^m. \end{aligned} \tag{A.6}$$

As $\omega\omega^T$ is positive semidefinite matrix, we have

$$\omega\omega^T > 0. \tag{A.7}$$

And it is obvious that

$$\sum_{j=1}^{N} u_{ij}^m > 0. \tag{A.8}$$

We track λ value in the experiments shown in Figure 5 and give the empirical evidence to proof $\lambda > 1$.

Thus,

$$\frac{\partial^2 F}{\partial {m_i}^2} = 2\left(1-\lambda\right)\omega\omega^T \sum_{j=1}^{N} u_{ij}^m < 0. \tag{A.9}$$

According to [23], (A.5) is the local maximum of F.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgments

This work is supported by the Major Program of the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (11KJA460001, 13KJA460001), China Spark Program (2013GA690404) Technology Innovation Project of Science and Technology Enterprises at Jiangsu Province in China (BC2012429), the Open Project from the Key Laboratory of Digital Manufacture Technology at Jiangsu Province in China (HGDML-1005), Huaian Science and Technique Program in China (HAG2010062), Huaian 533 Talents Project, Huaian International Science and Technology Cooperation Project (HG201309), Qing Lan Project of Jiangsu Province, and Jiangsu Overseas Research & Training Program for University Prominent Young & Middle-Aged Teachers and Presidents, China.

References

[1] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 178–188, 1936.

[2] J. Kittler, "On the discriminant Vector method of feature selection," *IEEE Transactions on Computers*, vol. 26, no. 6, pp. 604–606, 1977.

[3] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm," in *Proceedings of the 10th National Conference on Artificial Intelligence*, pp. 129–134, July 1992.

[4] I. Kononenko, "Estimating attributes: analysis and extension of RELIEF," *Proceedings of the European Conference on Machine Learning*, pp. 171–182, 1994.

[5] L. Yu and H. Liu, "Efficiently handling feature redundancy in high-dimensional data," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 685–690, August 2003.

[6] G. V. Lashkia and L. Anthony, "Relevant, irredundant feature selection and noisy example elimination," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 34, no. 2, pp. 888–897, 2004.

[7] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," in *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI '11)*, pp. 266–273, Barcelona, Spain, July 2011.

[8] A. R. Webb, *Statistical Pattern Recognition*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2002.

[9] M. Dash, H. Liu, and J. Yao, "Dimensionality reduction of unsupervised data," in *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*, pp. 532–539, November 1997.

[10] J. Basak, R. K. De, and S. K. Pal, "Unsupervised feature selection using a neuro-fuzzy approach," *Pattern Recognition Letters*, vol. 19, no. 11, pp. 997–1006, 1998.

[11] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301–312, 2002.

[12] M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature selection for clustering—a filter solution," in *Proceedings of the 2nd IEEE International Conference on Data Mining*, pp. 115–122, December 2002.

[13] S. Y. M. Shi and P. N. Suganthan, "Unsupervised similarity-based feature selection using heuristic Hopfield neural networks," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 3, pp. 1838–1843, July 2003.

[14] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *Journal of Machine Learning Research*, vol. 5, pp. 845–889, 2004.

[15] S. Z. Li and A. K. Jain, *Encyclopedia of Biometrics*, Springer, 2009.

[16] S.-Q. Cao, S.-T. Wang, X.-F. Chen, Z.-P. Xie, and Z.-H. Deng, "Fuzzy fisher criterion based semi-fuzzy clustering algorithm," *Journal of Electronics & Information Technology*, vol. 30, no. 9, pp. 2162–2165, 2008.

[17] X.-B. Zhi and J.-L. Fan, "Fuzzy fisher criterion based adaptive dimension reduction fuzzy clustering algorithm," *Journal of Electronics & Information Technology*, vol. 31, no. 11, pp. 2653–2658, 2009.

[18] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases," Department of Information and Computer Science, University of California, Irvine, Calif, USA, 1998, http://archive.ics.uci.edu/ml/.

[19] W. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[20] Center for Machine Learning and Intelligent Systems, the University of California, Irvine, Calif, USA, 2011, http://archive.ics.uci.edu/ml/datasets/Steel+Plates+Faults.

[21] M. Buscema, S. Terzi, and W. Tastle, "A new meta-classifier," in *Proceedings of the Annual North American Fuzzy Information Processing Society Conference (NAFIPS' 10)*, pp. 1–7, IEEE Press, Toronto, Canada, July 2010.

[22] J. H. Manton, "Differential calculus, tensor products and the importance of notation," 2013, http://arxiv.org/abs/1208.0197.

[23] J. Wilde, Unconstrained Optimization, 2011, http://www.econ.brown.edu/students/Takeshi_Suzuki/Math_Camp_2011/Unconstrained_Optimization-2011.pdf.