Universum linear discriminant analysis

X.H. Chen, S.C. Chen and H. Xue

Universum learning has been used for classification and clustering, and obtains favourable improvements with the help of Universum – the samples that do not belong to either class of interest. In this reported work, universum learning is extended to dimensionality reduction by incorporating it with linear discriminant analysis (LDA). However, for the C-class problem, LDA can get at most $C-1$ projection directions due to the rank limitation. The $C-1$ directions are not enough for sufficient discrimination, which has motivated the adaption of the one-against-one trick to decompose the original C-class LDA into $0.5C(C-1)$ binary LDA ones for getting more directions. Uniiversum learning is then introduced to each binary LDA and the method is termed as universum linear discriminant analysis (ULDA). ULDA aims to find discriminant directions by maximising the distance between two target classes and simultaneously minimising the distance between the Universum and the mean of the target classes. The experiments on UCI datasets demonstrate the advantages and effectiveness of the ULDA.

Introduction: Universum learning was first proposed for training a classifier with the help of Universum – the samples that belong to the same application domain as the training data, but do not belong to either target classes [1, 2]. For example, to recognition handwritten digits 5 and 8, one can introduce other handwritten digits as Universum. Weston *et al.* [1] proposed Universum support vector machines (USVM), which constructs a regularisation term based on the Universum in addition to the standard SVM. Sinz *et al.* [2] discussed the properties of USVM and presented a least squares version of the USVM, which has a closed-form solution. Zhang *et al.* [3] designed a graph-based semi-supervised classifier, which learns from the labelled, unlabelled and the Universum simultaneously, and further extended universum learning to document clustering [4]. Shen *et al.* [5] designed a new boosting algorithm, UBoost, which adds a regularisation term based on the Universum to the original optimisation objective. Although universum learning has shown better performance in classification and clustering, to our knowledge there is no work on applying it to dimensionality reduction (DR) till now. Inspired by its success, we try to extend universum learning to dimensionality reduction by incorporating it into linear discriminant analysis (LDA) to acquire a novel DR method.

LDA [6] is a well-known supervised DR method and has been used in extensive applications, such as biometrics and bioinformatics. It aims to seek a lower-dimensional subspace by maximising the ratio of the between-class scatter to the within-class scatter, finally boiling down to solving a generalised eigenvalue problem. Due to the rank limitation of the between-class scatter matrix, LDA can get at most $C-1$ discriminant directions, which are not enough for sufficient discrimination especially for the case where the dimension is greatly larger than the number of classes. For example, for the high-dimensional Yale Face containing 15 individuals, we can get only 14 discriminant directions with LDA despite the original high dimensionality. In order to extract more meaningful directions, we decompose original multi-class LDA into several binary LDA ones with the one-against-one trick which has been widely used in multi-class classifier design [7]. Accordingly, we obtain $0.5C(C-1)$ discriminant directions where each one is constructed with the corresponding two classes samples out of C classes. Obviously, the mathematical formulation of binary LDA indicates that each direction w_{ij}, obtained from the samples composed by the ith and the jth classes, just ensures maximal separability between such a pair of classes and neglects the other classes' information. The targeted classes $i\&j$ may be mixed with the rest of the classes on the discriminant direction. To avoid the possible deterioration of performance, we attempt to resort to the universum learning. First, we select the samples out of C classes except classes $i\&j$ as Universum. Secondly, we try to find discriminant direction w_{ij} by maximising the distance between classes $i\&j$ and minimising the distance between the Universum and the mean of classes $i\&j$ simultaneously. Finally, we obtain a projection matrix by combining the obtained $0.5C(C-1)$ discriminant directions together. The result is a new feature vector with dimensions of $0.5C(C-1)$, which is greater than $C-1$ of the original C-class LDA.

To the best of our knowledge, ULDA is the first DR method combining universum learning and the one-against-one decomposition trick together. Now it is worth-highlighting our contributions as follows:

1. Compared to original C-class LDA, each discriminant direction of ULDA can be explicitly obtained in closed-form which avoids the complex eigenvalue problem. Furthermore, the extracted features of ULDA are $0.5C(C-1)$, greater than that of LDA, therefore it is meaningful for the case of the dimensionality greatly larger than the number of classes.
2. Compared to multi-class LDA using the one-against-one strategy (OAO-LDA), ULDA exploits not only labelled data from any pair of classes but also Universum formed by the rest-class samples. The better performance of ULDA shown in the experiments exactly accords with the intuition that the Universum can serve for DR.
3. For ULDA, the universum learning and the one-against-one decomposition trick are interdependent of each other. With such a trick, we can naturally select the remained samples outside of the target classes as Universum. With the universum learning idea, we can likewise improve the performance of OAO-LDA.
4. More importantly, in the existing works [1, 2], Universum has to be constructed carefully and explicitly beyond the target classes before the training. It is time-consuming and task-dependent. By contrast, ULDA directly selects the remained samples except two target classes as Universum which is reasonable, natural and time-saving.

Universum linear discriminant analysis (ULDA): Given a dataset consisting of samples $\{(x_i, y_i)\}_{i=1}^n$ and $x_i \in R^D, y_i \in \{1, \cdots, C\}$ is the class label to x_i, and D the data dimensionality. The matrix $X=[x_1, \ldots, x_n]$ can be partitioned as $X=[X_1, \ldots, X_C]$, where $X_i \in R^{D^* ni}$, n_i is the size of the ith class, and $n_1+\ldots+n_C=n$. The objective of LDA is described as follows:

$$\max_W \frac{|W^T S_B W|}{|W^T S_w W|} \tag{1}$$

where $S_B = 1/n \sum_{i=1}^{C} n_i(u_i-u)(u_i-u)^T$, $S_w = \sum_{i=1}^{C} S_i$, $S_i = 1/n_i \sum_{x \in X_i} (x-u_i)(x-u_i)^T$, and u is the mean across all the samples, and u_i is the means of the ith class.

For the multi-class problem, the optimal solution $\{w_i | i=1, \ldots, C-1\}$ is a set of the eigenvectors of the following generalised eigenvalue problem:

$$S_B w_i = \lambda_i S_w w_i, \; i=1, \ldots, C-1 \tag{2}$$

Due to the fact that $rank(S_B) \le C-1$, at most $C-1$ eigenvectors can be obtained corresponding to the nonzero eigenvalues. In particular, for the two-class case, $S_B=(u_1-u_2)(u_1-u_2)^T$ and just one projection vector can explicitly be obtained in the following closed-form formulation

$$w=(S_1+S_2)^{-1}(u_1-u_2) \tag{3}$$

In general, such limited $C-1$ directions are not necessarily enough for sufficient discrimination.Though there are many strategies of increasing discriminant ability, one of the most convenient and naive strategies may be to increase the number of projection directions, for which we adopt the commonly-used one-against-one trick to decompose original C-class LDA into $0.5C(C-1)$ binary LDA ones, each *just* concerning a pair of classes, e.g. classes $i\&j$. Now we take the samples out of C classes except X_i, X_j as Universum and mandatorily make them close to $(u_i+u_j)/2$ on the discriminant direction. The purpose of doing so is to force these data (not belonging to the classes of interest) to lie as maximally nearby the class boundary in projected space as possible, and thus preventing them from falling in the discriminant region of the classes of interest. For modelling such an intuition, what we need to do is minimise the distance between $(u_i+u_j)/2$ and the Universum in the projected space, as a result, leading to the following formulation:

$$\min_{w_{ij}} \sum_{x \notin X_i \cup X_j} w_{ij}^T [x-(u_i+u_j)/2][x-(u_i+u_j)/2]^T w_{ij} \tag{4}$$

Let $A_{ij} = \sum_{x \notin X_i \cup X_j} [x-(u_i+u_j)/2][x-(u_i+u_j)/2]^T$ and combine the Fisher criterion of binary LDA and (4), we establish the ULDA objective:

$$\max_{w_{ij}} \frac{w_{ij}^T(u_i - u_j)(u_i - u_j)^T w_{ij}}{w_{ij}^T(S_i + S_j + \lambda A_{ij})w_{ij}} \quad (5)$$

where λ is a tuning parameter which controls the balance between the target classes $i\&j$ and the Universum. Through optimising (5), we seek a direction along which the distance between two targeted classes $i\&j$ is maximised and meanwhile the distance between $(u_i + u_j)/2$ and the Universum is minimised. Due to the same form as the objective of binary LDA, we directly obtain its closed-form projection direction w_{ij}

$$w_{ij} = (S_i + S_j + \lambda A_{ij})^{-1}(u_i - u_j) \quad (6)$$

With the discriminant directions obtained from the above binary LDAs, we construct a total project matrix $W_{opt} = [w_{12}, \cdots, w_{1c}, w_{23}, \cdots, w_{2c}, \cdots, w_{c-1,c}]$, this way we can finally get $0.5C(C-1)$ features far greater than $C-1$, which breaks the rank limitation of the original multi-class LDA.

Experiments and analyses: To evaluate our ULDA, we performed experiment comparisons on 13 multi-class UCI datasets. Besides LDA and ULDA, we also present results of OAO-LDA, which is designed specially for our task. We randomly chose half of each class for training and the remaining for testing. In the experiment, the parameter λ was selected from$\{2^{-5}, 2^{-4}, \ldots 2^{4}, 2^{5}\}$ and searched by cross-validation for optimising performance. We used the 1NN classifier to perform final classification and repeated the process 10 times. Table 1 reports the average results where the best performances are highlighted in bold.

Table 1: Comparison of average (%) and variance (10^{-4}) among LDA, OAO-LDA and ULDA on UCI datasets

Dataset (Class/Dim/Num)	LDA	OAO-LDA	ULDA
Banlance (3/4/625)	87.53 ± 4.83	87.95 ± 3.31	**89.33** ± 3.54
Cmc (3/9/1473)	42.46 ± 4.34	43.60 ± 3.61	**44.85** ± 3.12
Dermat (6/33/366)	**96.32** ± 0.48	95.33 ± 2.16	95.55 ± 2.11
Ecoli (6/6/332)	80.37 ± 4.77	80.61 ± 6.92	**82.20** ± 1.14
Glass (6/29/214)	54.76 ± 40.0	60.28 ± 16.2	**62.86** ± 12.0
Iris (3/4/150)	93.33 ± 3.56	96.26 ± 3.08	**98.53** ± 0.90
Waveform (3/21/5000)	81.22 ± 7.96	**81.76** ± 0.41	81.63 ± 0.62
Lense (3/4/24)	79.09 ± 56.0	68.12 ± 39.3	**85.45** ± 40.4
Soybean (4/35/47)	97.39 ± 5.04	98.69 ± 17.0	**1.00** ± 0.01
Teaching (3/5/151)	55.60 ± 79.0	54.40 ± 35.0	**57.60** ± 39.4
Thyroid (3/5/215)	93.27 ± 7.14	95.42 ± 5.52	**96.18** ± 3.0
Vehicle (4/18/846)	73.84 ± 3.05	63.72 ± 5.77	**77.92** ± 4.22
Wine (3/13/178)	**97.84** ± 1.56	96.70 ± 4.42	97.23 ± 1.8

From Table 1, we obtain several attractive observations as follows.

1. ULDA prominently outperforms LDA on 11 out of 13 datasets, especially achieving the maximum improvement of 8% on Glass, of 6% on Lense and of 5% on Iris. ULDA improves more than 2% on the seven datasets and obtains comparable results on the other two datasets.
2. From columns 2 and 3 in Table 1, we observe that the performance of OAO-LDA is better than or comparable to LDA on eight datasets partially due to the increase of projection directions. And it is consistent with the fact that some binary target classes may be mixed with some of the rest of the samples on the direction of OAO-LDA.
3. Compared with OAO-LDA, ULDA provides better results on 12 out of 13 datasets and exceeds OAO-LDA to different degrees from 1 to 13%. Especially on Lense and Vehicle, ULDA improves more than 13% in accuracy. The results state that Universum can indeed be used to improve the performance of DR as well.

Conclusion: A novel and efficient supervised DR method, ULDA, is proposed. ULDA incorporates the universum learning idea and the one-against-one decomposing strategy for multi-class LDA to a set of binary LDAs and inherits both the merits of universum learning and binary LDA in the form of a solution. Experimental results on 13 of UCI datasets demonstrate the effectiveness of our ULDA, which extends a new line of applying the Universum concept. Our future work will include universum-motivated feature selection and metric learning.

Acknowledgments: This work was partially supported by the NSFC (grants 61170151, 60905002) and NUAA Funding (grant NP2011030).

19 July 2012
doi: 10.1049/el.2012.2506
One or more of the Figures in this Letter are available in colour online.

X.H. Chen and S.C. Chen (*Department of Computer Science and Engineering, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, People's Republic of China*)

E-mail: s.chen@nuaa.edu.cn

H. Xue (*School of Computer Science and Engineering, Southeast University, Nanjing 210096, People's Republic of China*)

References

1 Weston, J., Collobert, J.R., Sinz, F., Botton, L., and Vapnik, V.: 'Inference with the universum'. Int. Conf. on Machine Learning, Pittsburgh, PA, USA, 2006, pp. 1009–1016
2 Sinz, F.H., Chapelle, O., Agarwal, A., and Schölkopf, B.: 'An analysis of inference with the universum'. Conf. Neural Information Processing Systems, Cambridge, MA, USA, 2008, pp. 1369–1376
3 Zhang, D., Wang, J.D., Wang, F., and Zhang, C.: 'Semi-supervised classification with universum'. Proc. SIAM Int. Conf. on Data Mining, Georgia, USA, 2008, pp. 323–333
4 Zhang, D., Wang, J.D., and Si, L.: 'Document clustering with universum'. SIGIR'11, Beijing, China, 2011
5 Shen, C., Wang, P., and Wang, H.: 'UBoost: boosting with the universum', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (4), pp. 825–831
6 Fisher, R.: 'The use of multiple measurements in taxonomic problems', *Ann. Eugenics*, 1936, **7**, (2), pp. 179–188
7 Debnath, R., Takahide, N., and Takahashi, H.: 'A decision based one-against-one method for multi-class support vector machine', *Pattern Anal. Appl.*, 2004, 7, pp. 164–175