# Weighted Discriminant Embedding: Discriminant Subspace Learning for Imbalanced Medical Data Classification

Tobey H. Ko[1], Zhonglei Gu[2], Yang Liu[2,3]
[1]Department of Industrial and Manufacturing Systems Engineering, University of Hong Kong, HKSAR, China
[2]Department of Computer Science, Hong Kong Baptist University, HKSAR, China
[3]Institute of Research and Continuing Education, Hong Kong Baptist University, Shenzhen, China
tobeyko@hku.hk,csygliu@comp.hkbu.edu.hk,cszlgu@comp.hkbu.edu.hk



## ABSTRACT

A model designed for automatic prediction of diseases based on multimedia data collected in hospitals is introduced in this working notes paper. In order to perform the automatic diseases prediction efficiently, while using as few data as possible for training, we develop a two-stage learning strategy, which first performs the weighted discriminant embedding (WDE) to project the original data to a low-dimensional feature subspace and then utilizes the cost-sensitive nearest neighbor (CS-NN) method in the learned subspace for disease prediction. The proposed approach is evaluated on the MediaEval 2018 Medico Multimedia Task.

## 1 INTRODUCTION

Aiming at improving the efficiency of detecting medical abnormalities in the machine intelligence assisted medical diagnosis, and using as little information as possible, the MediaEval 2018 Medico Multimedia Task [3] seeks to design an integrated approach to assist the medical experts' decision-making process using a combination of video and image information, as well as other sensory information. In this paper, a two-stage learning strategy is introduced to facilitate efficient detection of diseases using multimedia and sensory information. The first stage consists of a dimensionality reduction process which projects the original data to a low-dimensional feature representation using weighted discriminant embedding (WDE), which improves the efficiency of the learning process while also preserving the key discriminant information of the original data. Then, the cost-sensitive nearest neighbor (CS-NN) method is employed to make the prediction in the learned subspace.

## 2 WEIGHTED DISCRIMINANT EMBEDDING

Let $\mathcal{X}$ be the training set: $\mathcal{X} = \{(\mathbf{x}_1, l_1), \cdots, (\mathbf{x}_n, l_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^D$ $(i = 1, ..., n)$ denotes the feature representation of the $i$-th sample, $l_i \in \{1, \cdots, C\}$ denotes the label of $\mathbf{x}_i$, $n$ denotes the number of data samples in the set, $C$ denotes the number of classes, and $D$ denotes the original dimension of data. Given the training set, weighted discriminant embedding (WDE) aims to learn a transformation matrix $\mathbf{W} \in \mathbb{R}^{D\times d}$ $(d \leq D)$, which is capable of projecting the original high-dimensional data to a low-dimensional subspace $\mathcal{Z} = \mathbb{R}^d$, where the weighted discriminant information could be preserved.

In this year's Medico task, the sample numbers in different classes are highly imbalanced. To enhance the algorithm's power in making correct detection on rarer classes, we expect that data samples belonging to the same class, especially for the rarer class, should be close to each other as much as possible in the learned subspace, while nearby data samples from different classes, again, especially for rarer classes, should be separated from each other as much as possible in the learned subspace.

To minimize the weighted intra-class scatter, we present the following objective:

$$\mathbf{W} = \arg\min_{\mathbf{W}} tr\Big( \sum_{i,j=1}^{n} A_{ij} \mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W} \Big), \quad (1)$$

where $A_{ij} = (I_i + I_j)/2$ if $l_i = l_j$; and 0 otherwise. Here $I_i$ indicates the importance of class $l_i$ and is defined using the entropy-based formulation [2]:

$$I_i = -\frac{(1 - p_i)^2}{p_i} \log(p_i), \quad (2)$$

where $p_i$ denotes the proportion of class $l_i$ in the dataset. In Eq. (2), small proportion indicates high importance. Eq. (1) could be rewritten as:

$$\mathbf{W} = \arg\min_{\mathbf{W}} tr(\mathbf{W}^T \mathbf{L}_A \mathbf{W}), \quad (3)$$

where $\mathbf{L}_A$ is a Laplacian matrix [1] defined as $\mathbf{L}_A = \mathbf{D}_A - \mathbf{A}$, with $\mathbf{D}_A$ being a diagonal matrix defined as $(D_A)_{ii} = \sum_{j=1}^{n} (A)_{ij}$ $(i = 1, \cdots, n)$.

Similarly, we define the following objective function to maximize the weighted inter-class scatter:

$$\mathbf{W} = \arg\max_{\mathbf{W}} tr\Big( \sum_{i,j=1}^{n} B_{ij} \mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W} \Big), \quad (4)$$

where $B_{ij} = N_{ij}(I_i + I_j)/2$ if $l_i \neq l_j$; and 0 otherwise. Here $N_{ij} = exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2)$ is utilized to measure the closeness between two data samples. Eq. (4) could be rewritten as:

$$\mathbf{W} = \arg\max_{\mathbf{W}} tr(\mathbf{W}^T \mathbf{L}_B \mathbf{W}), \quad (5)$$

where $\mathbf{L}_B = \mathbf{D}_B - \mathbf{B}$, with $\mathbf{D}_B$ being a diagonal matrix defined as $(D_B)_{ii} = \sum_{j=1}^{n} (B)_{ij}$ $(i = 1, \cdots, n)$.


*MediaEval'18, 29-31 October 2018, Sophia Antipolis, France*

We integrate Eqs. (3) and (5) to form a unified objective function of WDE:

$$\mathbf{W} = \arg\max_{\mathbf{W}} tr\left(\frac{\mathbf{W}^T\mathbf{L}_B\mathbf{W}}{\mathbf{W}^T\mathbf{L}_A\mathbf{W}}\right). \quad (6)$$

Then the optimal $\mathbf{W}$ that maximizes the objective function in Eq. (6) is composed of the normalized eigenvectors corresponding to the $d$ largest eigenvalues of the following eigen-decomposition problem:

$$\mathbf{L}_B\mathbf{w} = \lambda\mathbf{L}_A\mathbf{w}. \quad (7)$$

For a high-dimensional data sample $\mathbf{x}_i$, it can be mapped to the subspace by $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$.

## 3 RESULTS AND ANALYSIS

To evaluate our approach, we test its performance on the MediaEval 2018 Medico Multimedia Task. The task contains both development set (with $5,293$ samples) and test set (with $8,740$ samples). For each sample, we use six types of features: the 168-D JCD feature; the 18-D Tamura feature; the 33-D ColorLayout feature; the 80-D EdgeHistogram feature; the 256-D AutoColorCorrelogram feature; and the 630-D PHOG feature. The totally dimension is $1,185$.

We participate in two subtasks: 1) Classification of diseases and findings; and 2) Fast and efficient classification. For both tasks, we submit 5 runs.

- For Run 1 (on both subtasks), we use all the data from the development set for training;
- For Run 2 (on both subtasks), we randomly select 50% data for each class from the development set for training;
- For Run 3 (on both subtasks), we randomly select 50% data for each class from the development set, together with the remaining data in the "out-of-patient" and "instruments" classes, for training;
- For Run 4 (on both subtasks), we randomly select 25% data for each class from the development set for training;
- For Run 5 (on both subtasks), we randomly select 25% data for each class from the development set, together with the remaining data in the "out-of-patient" and "instruments" classes, for training.

In the training stage, we use the training data to learn the transformation matrix $\mathbf{W}$ via WDE. We set $\sigma = 1$ and the subspace dimension $d = 50$. In the test stage, we use the obtained $\mathbf{W}$ to map both training and test data to the 50-D subspace, and then use the cost-sensitive nearest neighbor (CS-NN) method for the final classification in the learned subspace, where the cost of misclassifying the data of class $c$ ($c = 1, \cdots, C$) to other classes is defined as $cost_c = n/n_c$, with $n$ and $n_c$ being the total number of the training data and the number of data in class $c$, respectively.

Tables 1 and 2 report the results of our approach on subtask 1 and subtask 2, respectively. Although the accuracy looks good, the overall performance is far from satisfactory as the results on other four important criteria are relatively low.

**Table 1: Results of our approach on the first subtask of MediaEval 2018 Medico Multimedia Task.**

| | Recall | Precision | Accuracy | F1 Score | Rk |
|---|---|---|---|---|---|
| Run 1 | 0.5001 | 0.4917 | 0.9471 | 0.4830 | 0.5357 |
| Run 2 | 0.4415 | 0.4294 | 0.9384 | 0.4251 | 0.4612 |
| Run 3 | 0.3947 | 0.3670 | 0.9320 | 0.3728 | 0.4035 |
| Run 4 | 0.3553 | 0.3333 | 0.9256 | 0.3324 | 0.3511 |
| Run 5 | 0.3019 | 0.2814 | 0.9186 | 0.2812 | 0.2918 |

**Table 2: Results of our approach on the second subtask of MediaEval 2018 Medico Multimedia Task.**

| | Recall | Precision | Accuracy | F1 Score | Rk |
|---|---|---|---|---|---|
| Run 1 | 0.5005 | 0.4917 | 0.9471 | 0.4830 | 0.5357 |
| Run 2 | 0.4181 | 0.3857 | 0.9337 | 0.4251 | 0.4193 |
| Run 3 | 0.4259 | 0.4085 | 0.9350 | 0.4040 | 0.4348 |
| Run 4 | 0.3430 | 0.3107 | 0.9231 | 0.3135 | 0.3293 |
| Run 5 | 0.3257 | 0.3053 | 0.9227 | 0.3057 | 0.3246 |

The reason might be that the proposed WDE is a linear mapping method, which is not sufficient to capture the complex discriminant information embedded in the high-dimensional feature space. This motivates us to consider extending our method to the nonlinear case to improve the performance. Furthermore, by comparing the performance on Run 2 (Run 4) and that on Run 3 (Run 5), we observe that even we use all the data from the minority classes (i.e., the "out-of-patient" and "instruments" classes), the performance is not improved. The reason might be that the number of data in these two classes are too small to represent the "real" distribution of the classes. On possible solution is to employ the oversampling technology to reasonably and faithfully generate samples for minority classes.

## 4 CONCLUSION

In this paper, we propose a subspace learning method called weighted discriminant embedding (WDE), aiming at discovering the discriminant subspace for imbalanced dataset. After dimensionality reduction, the cost-sensitive nearest neighbor is utilized for classification. We plan to extend our work from two aspects. First, we will generalize our approach to nonlinear case to enhance its data representation ability. Second, we will incorporate some oversampling methods into our approach to make it stronger for imbalanced learning problem.

## ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61503317, in part by the General Research Fund (GRF) from the Research Grant Council (RGC) of Hong Kong SAR under Project HKBU12202417, and in part by the SZSTI Grant with the Projct Code JCYJ20170307161544087.

## REFERENCES

[1] M. Belkin and P. Niyogi. 2003. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* 15, 6 (2003), 1373–1396.

[2] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr. 2017. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2999–3007.

[3] K. Pogorelov, M. Riegler, P. Halvorsen, T. de Lange, K. R. Randel, D.-T. Dang-Nguyen, M. Lux, and O. Ostroukhova. Medico Multimedia Task at MediaEval 2018. In *Proceedings of the MediaEval 2018 Workshop*. CEUR-WS, Sophia Antipolis, France, 29–31 October, 2018.