Proceedings of the 2017 International Conference on Machine Learning and Cybernetics, Ningbo, China, 9-12 July, 2017

HEART DISEASE CLASSIFICATION BASED ON FEATURE FUSION

TING-TING ZHAO[1], YU-BO YUAN[1], YING-JIE WANG[2], JU GAO[2], PING HE[3]

[1]East China University of Science and Technology, Shanghai, 200237, China
[2]ShuGuang Hospital of Shanghai University of Traditional Chinese Medicine, Shanghai, 201203, China
[3]Shanghai Hospital Development Center, Shanghai, 20051, China
E-MAIL: ybyuan@ecust.edu.cn, strawc277@163.com,
wangyingjie7421@163.com, gaoju@mail.sh.cn, 2265137200@qq.com

Abstract:

Heart disease classification is one of the most important topics in clinical decision support systems (CDSS). However, the performance of classification is greatly affected by feature selection. Canonical correlation analysis (CCA) is a popular method to extract effective features from two relevant data sets. In this paper, we employ discriminant minimum class locality preserving canonical correlation analysis (DMPCCA) to get useful features and finish disease classification by support vector machine (SVM). Normalized mutual information based on entropies and information gains are used to divide the data sets into two views (X1 and X2). Features extraction and fusion are implemented by different methods, including CCA, DMPCCA and PCA. We select two data sets to test the performance. One (1329 patients) is from Shanghai Shuguang Hospital, the another one is UCI heart disease data set (270 patients). The experimental results show that the performance of DMPCCA is the best.

Keywords:

Machine Learning; Heart Disease Classification; Feature Fusion; CCA; DMPCCA

1. Introduction

Cardiovascular diseases (CVD) are the leading cause of death globally [1]. Diagnosis of CVD is a complicated and important task that needs to be executed accurately and efficiently [18]. In order to improve the quality of health care and relieve the pressure of medical service, many data mining techniques are applied for clinical decision making supported by clinical decision support systems (CDSS) [18]. Classification is one of the most important algorithms in CDSS. The performance of classification is greatly affected by feature selection. It is a challenging problem to get good features. It is the motivation of this paper.

Feature fusion is an effective method to get excellent features. PCA is a powerful technique to find a subspace whose basis vectors correspond to the maximum-variance directions in the original space [16][17]. CCA is a classical method that has been widely used in information fusion and dimensionality reduction [3] [5]. DMPCCA, a novel method we proposed in 2016, is improved on the basis of CCA [7]. The proposed method introduces local structure information and global discriminant information into the classical CCA and considers a optimal combination of intra-class locality preserving, global discriminant ability and the maximal correlation between two sets. Essentially, CCA and DMPCCA are also tools for dimensionality reduction by finding a pair of project vectors. Dimensionality reduction will inevitably lead to information loss, PCA will be compared with CCA and DMPCCA in this paper.

We focus on feature selection of heart disease for heart disease classification. The contributions are listed as follows:(1)Normalized mutual information based on entropies and information gains are employed to get effective views of features; (2)Three important operators including PCA, CCA and DMPCCA are used to fuse the features; (3)A new heart disease data set from Shanghai Shuguang Hospital is released.

2. Proposed technical framework for heart disease classification

Our basic framework is based on Shanghai Shuguang Hospital. Figure 1 is the framework of Shuguang heart failure classification. We collected medical data of heart failure patients from Shanghai Shuguang hospital. After data cleaning and data supplementing, we got a heart failure sample database which contains demographic data, diagnosis information, medications data and assay data. In this paper, we use DMPCCA method to

978-1-5386-0408-3/17/$31.00 ©2017 IEEE

do disease classification with the idea of feature fusion on this sample database.

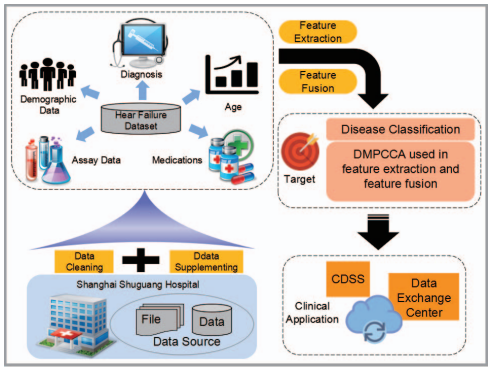


FIGURE 1. The framework of disease classification(Shuguang HF data set)

2.1 Data sets

Our heart disease data set is described as follows

$$\{(x_1, y_1), \cdots, (x_m, y_m)\}, x_i \in \mathcal{R}^n, y_i \in \{-1, 1\}. \quad (1)$$

In which, m is the number of patients. n is the number of features. -1 indicates that one patient is without heart disease. 1 indicates that one patient is with heart disease. The data set is denoted as

$$\mathbf{S} = (\mathbf{X}, \mathbf{Y}), \mathbf{X} \in \mathcal{R}^{m \times n}, \mathbf{Y} \in \mathcal{R}^m. \quad (2)$$

We used the Heart Disease Data Set from the University of California, Irvine (UCI) Machine Learning Repository and a real world data set about HF patients from ShuGuang Hospital. The former is called as UCI dataset, the latter as HF dataset.

UCI dataset contains 14 attributes, 270 records, of which 120 were healthy and the other 150 diagnosed as heart disease. The attributes(A) are: A1-Age; A2-Sex; A3-Chest Pain Type(cp); A4-Resting Blood Pressure (restbp); A5-Serum Cholesterol in mg/dl (chol); A6-Fasting Blood Sugar (fbs); A7-Resting Electrocardiographic Results (restecg); A8-Maximum Heart Rate Achieved (mhr); A9-Exercise Induced Angina(exe); A10-Oldpeak; A11-Slope; A12-Number of Major Vessels Colored by Fluoroscopy (numv); A13-Thalium Heart Scan (thal); A14-Diagnosis of Heart Disease.

HF dataset contains 23 attributes, 1329 records of HF patients in Shuguang Hospital.There are 873 patients with HF and 456 patients without HF. The overall dataset contains patient's age, sex, diseases prevalence and some selected assay results. The parameters(P) are: P1-Age; P2-Sex; P3-Leukocyte (WBC); P4-Hemoglobin (HGB); P5-Red Blood Cells (RBC); P6-Blood Platelet (PLT); P7-Neutrophils (NEUT); P8-Basophils(BA); P9-Lymphocytes (LY); P10-Monocytes (MONO); P11-Globulin (GLB); P12-Alanine Aminotransferase (ALT); P13-B-type Natriuretic Peptide (BNP); P14-Troponin (CTN); P15-New York Cardiac Function Classification (NYHA); P16-Hypertension; P17-Diabetes; P18-Arrhythmia; P19-Coronary Heart Disease (CHD); P20-Angina; P21-Bronchitis; P22-Lung Infection; P23-Diagnosis of Heart Failure. P1-P2 are demographics, P3-P14 are assay index, P15-P22 are HF patients' diagnosis.

2.2 Analysis of risk factors of heart disease

Two important methods are employed to analysis the risk factors. One is normalized mutual information based on entropies and the other is information gains. For one patient, it denoted as $(x_1, x_2, \cdots, x_n, y)$.

2.2.1 Normalized mutual information based on entropies

For one feature x_i, the entropy is defined as follows

$$H(x_i) = -\sum_{k=1}^{K} p_k log(p_k). \quad (3)$$

In which, p_k is the frequency of k-th information in feature x_i.

Normalized mutual information[10] is defined as follows

$$c_{ij} = \frac{H(x_i) + H(x_j) - H(x_i, x_j)}{H(x_j)}. \quad (4)$$

2.2.2 Information gains

$$H(Y) = -\sum_{l=1}^{L} P_l \log(P_l). \quad (5)$$

The conditional entropy of attribute X is:

$$H(Y|X) = -\sum_{i=1}^{n} P_i H(Y|X = x_i). \quad (6)$$

To sum up, we can get the calculation method of information gain ratio[8] of various attributes X:

$$\begin{aligned} G(X) &= H(Y) - H(Y|X) \\ &= -\sum_{l=1}^{L} P_l \log(P_l) + \sum_{i=1}^{n} P_i H(Y|X = x_i). \end{aligned} \quad (7)$$

2.3 Two-views of heart disease and fusion

With normalized mutual information and information gains, we segmented features $\mathbf{X}$ into two-views $\mathbf{X}_1, \mathbf{X}_2$.

$$\mathbf{X}_1 \cup \mathbf{X}_2 \subseteq \mathbf{X}. \tag{8}$$

2.3.1 DMPCCA

Based on the above analysis, we give the proposed DMPCCA. Given a set of pairwise samples $\boldsymbol{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n \in \mathcal{R}^p \times \mathcal{R}^q$, the goal of DMPCCA is to find two projecting vectors $\boldsymbol{\omega}_x \in \mathcal{R}^p$ and $\boldsymbol{\omega}_y \in \mathcal{R}^q$ for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, so as to maximize the correlation, meanwhile, the model can also preserve local structure information and maximize discriminant ability. The proposed objective function is as follows:

$$\rho = \max_{\boldsymbol{\omega}_x, \boldsymbol{\omega}_y} \frac{\boldsymbol{\omega}_x^T \boldsymbol{X}\boldsymbol{B}\boldsymbol{Y}^T \boldsymbol{\omega}_y}{\sqrt{\boldsymbol{\omega}_x^T \boldsymbol{X}\boldsymbol{S}_{xx}\boldsymbol{X}^T \boldsymbol{\omega}_x \cdot \boldsymbol{\omega}_y^T \boldsymbol{Y}\boldsymbol{S}_{yy}\boldsymbol{Y}^T \boldsymbol{\omega}_y}}. \tag{9}$$

Converting the objective function to the following optimization model:

$$\begin{aligned} \max_{\boldsymbol{\omega}_x, \boldsymbol{\omega}_y} \quad & \boldsymbol{\omega}_x^T \boldsymbol{X}\boldsymbol{B}\boldsymbol{Y}^T \boldsymbol{\omega}_y \\ s.t. \quad & \boldsymbol{\omega}_x^T \boldsymbol{X}\boldsymbol{S}_{xx}\boldsymbol{X}^T \boldsymbol{\omega}_x = 1, \ \ \boldsymbol{\omega}_y^T \boldsymbol{Y}\boldsymbol{S}_{yy}\boldsymbol{Y}^T \boldsymbol{\omega}_y = 1, \end{aligned} \tag{10}$$

obviously, it is a constrained optimization problem, the solution can be obtained through solving a generalized eigenvalue problem:

$$\begin{aligned} & \begin{bmatrix} 0 & \boldsymbol{X}\boldsymbol{B}\boldsymbol{Y}^T \\ \boldsymbol{Y}\boldsymbol{B}\boldsymbol{X}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega}_x \\ \boldsymbol{\omega}_y \end{bmatrix} \\ = \lambda & \begin{bmatrix} \boldsymbol{X}\boldsymbol{S}_{xx}\boldsymbol{X}^T & 0 \\ 0 & \boldsymbol{Y}\boldsymbol{S}_{yy}\boldsymbol{Y}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega}_x \\ \boldsymbol{\omega}_y \end{bmatrix}, \end{aligned} \tag{11}$$

where λ is the eigenvalue corresponding to the eigenvector.

Suppose the eigenvectors corresponding to d ($d < \min(p, q)$) generalized eigenvalues $\lambda_i, i = 1, 2, \ldots, d$ are $\boldsymbol{\Omega}_x = [\boldsymbol{\omega}_x^1, \ldots, \boldsymbol{\omega}_x^d] \in \mathcal{R}^{p \times d}$ and $\boldsymbol{\Omega}_y = [\boldsymbol{\omega}_y^1, \ldots, \boldsymbol{\omega}_y^d] \in \mathcal{R}^{q \times d}$, then for any sample $(\boldsymbol{x}, \boldsymbol{y})$, we can extract features as $\boldsymbol{\Omega}_x^T \boldsymbol{x} + \boldsymbol{\Omega}_y^T \boldsymbol{y}$ (serial combination) or $\begin{bmatrix} \boldsymbol{\Omega}_x^T \boldsymbol{x} \\ \boldsymbol{\Omega}_y^T \boldsymbol{y} \end{bmatrix}$ (parallel combination).

2.4 Support vector machine(SVM)

Given a set of pair-wise samples $\boldsymbol{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathcal{R}^m$ is a sample point in m-dimensional space, and $y_i \in \{+1, -1\}$ is the corresponding label. The direct way to separate these samples into two classes is to find a separating hyperplane.

For the linearly separable case, the SVM model is as follows:

$$\begin{cases} \min_{\boldsymbol{w}} & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ s.t & y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1, \qquad \forall i = 1, \ldots, n. \end{cases} \tag{12}$$

By transforming this optimization problem into its corresponding dual problem, the optimal discriminant vectors can be found through

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i, \tag{13}$$

where α_i and $\boldsymbol{x}_i$ are the dual variable and data sample(called support vector), respectively.

3 Algorithm and Experiment

3.1 Algorithm Flow

Figure 2 is the algorithm flow chart of our experiment. Shuguang HF data set and UCI data set are divided into two parts respectively. And then get protection vectors using CCA, DMPCCA and PCA. After the projecting original data, SVM finishs disease classification. We define UCI dataset (or HF dataset) as $D = (X, Y, label) \in R^{n \times (p+q+1)}$, where $X = [x_1, ..., x_p] \in R^{n \times p}$ and $Y = [y_1, ..., y_q] \in R^{n \times q}$. And n represents the number of samples, $(p + q)$ is the number of attributes. We elect the appropriate attributes to form X and Y.

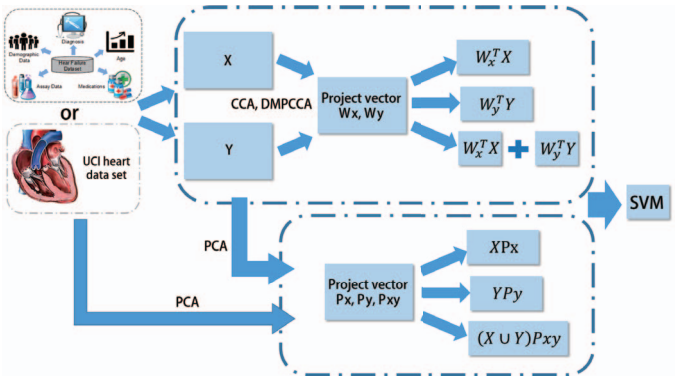


FIGURE 2. The algorithm flow chart

In David Blokh's research [10], the triple combined parameters:(A1, A12, A13), (A1, A3, A12), (A1, A2, A12) are the top-3 combinations that are strongly related to heart disease. So the choose X=(A1, A2, A3, A12), Y=(A4, A5, A6, A7, A8, A9,

A10, A11), we call this kind of combination as UCIdata-1. In order to make a contrast experiment, we choose the 4 attributes which are the least relevant attributes to heart disease to consist X. The other attributes consists of Y. That is to say, UCIdata-2: X=(A4, A5, A6, A7), Y=(A1, A2, A3, A8, A9, A10, A11, A12).

In Wang's paper [8], the author calculates information gain (IG) for each attribute. There are three types of attribute: demographic information, assay results, diagnosis of disease. The top-10 IG attributes are: A19, A15, A1, A13, A14, A16, A18, A7, A7, A4. So we have two combinations for the contrast experiment. The first one is to select attributes according to their attribute types: X=(P1, P2, P15, P16, P18, P19, P22), Y=(P4, P7, P8, P11, P12, P13, P14). We have removed some attributes whose IG are too small in this kind of combination called as HFdata-1. The second one is to choose the top-7 IG attributes to form X. And the last 7 IG attributes become members of Y. HFdata-2: X=(P1, P13, P14, P15, P16, P18, P19), Y=(P2, P3, P5, P6, P10, P11, P20).

4. Results

First, we will show some classification results of UCI heart dataset. Figure 3 is the F-measures of three methods on UCIdata-1 and UCIdata-2.

$$\mathbf{F-measure} = \frac{2Precision * Recall}{Precision + Recall} \tag{14}$$

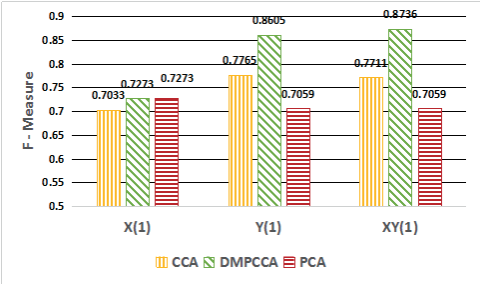
where $Precision$ and $Recall$ are calculated based on SVM heart disease classification. We compare one-dimensional and two-dimensional information of X, Y, XY using the three methods. It can be seen in Figure 3(a) and Figure 3(b) that the performance of DMPCCA is better than that of CCA and PCA when XY(1) and XY(1,2) as inputs of SVM. This is the performance of three methods using one dimension information and feature fusion plays an important role in disease classification problem. Thanks to DMPCCA, F-measure reach 0.8736 while PCA is 0.7952. When we do experiments on UCIdata-2, DMPCCA is also better than CCA and PCA (see Figure 3(c) and Figure 3(d)). XY is always better than X and Y both in one-dimension and two-dimension of information. So we can say that feature fusion is better than single aspect of features.

In order to compare the two feature combinations of UCI dataset (UCIdata-1 and UCIdata-2), we draw Figure 4. It is shown that whether it is the CCA method or DMPCCA method, UCIdata-1 is better than UCIdata-2. That is to say, it is effective to select features according to the information entropy and the combination in David Blokh's paper [10] is justified. In other words, UCIdata-1 is more able to stand for UCI data set. UCIdata-1 is produced according to normalized mutual information based on entropies.

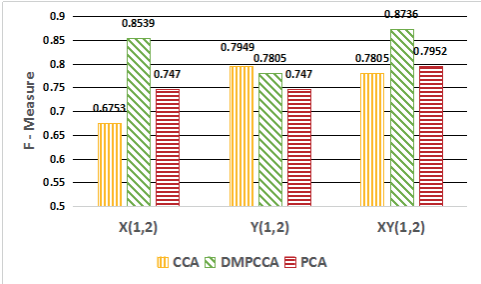
For HF dataset, we still find the DMPCCA is better than CCA and PCA in HF disease classification both on One-dimension and Two-dimension. In Figure 5(a) and Figure 5(b), XY(1,2) is better than XY(1) and XY is better than X and Y. It is obvious that the performance of DMPCCA is consistent with the UCI dataset and HF dataset. Features in HFdata-1 are selected according to the type of feature: features are divided into diagnosis information and assay data. Features in HFdata-2 are selected based on IG proposed in Wang's paper [8]. When we compared the two feature combinations of HF dataset in Figure 6, we found that HFdata-2 is better than HFdata-1 using CCA. In Figure 6(a) and Figure 6(b), CCA is a good choice for HFdata-2, features with high IG is helpful to CCA. As is to DMPCCA, HFdata-1 is better than HFdata-2 in Figure 6(c) and Figure 6(d). Because DMPCCA considers a optimal combination of intra-class locality preserving, global discriminant ability and the maximal correlation between two sets. If we use DMPCCA to solve feature fusion problem, we should select appropriate attributes for the purpose of maximizing patient's characteristics.

5. Conclusion

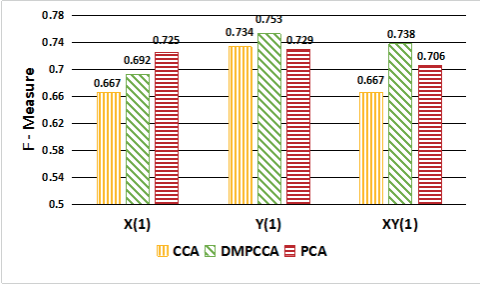
This article uses an improved CCA method for HF disease classification based on the idea of feature fusion. Our experiments use the UCI database and a real-world data extracted from Shanghai Shuguang Hospital. First, we describe the data set from two perspectives by selecting some special features according to normalized mutual information based on entropies and information gians. And then we use CCA, DMPCCA and PCA to do data projection. Finally, SVM completed the work of disease classification. our Experimental results show that DMPCCA is better than CCA in feature fusion. DMPCCA has the characteristics of feature fusion, so DMPCCA is better than PCA in dimensionality reduction. In addition, if we want to use DMPCCA, we should first describe the data in two perspectives, and selecting appropriate features to express data set can improve the performance of DMPCCA. The idea of feature fusion using DMPCCA can be applied to the future research about disease classification or prediction, our current research is the basis for the future realization of CDSS.



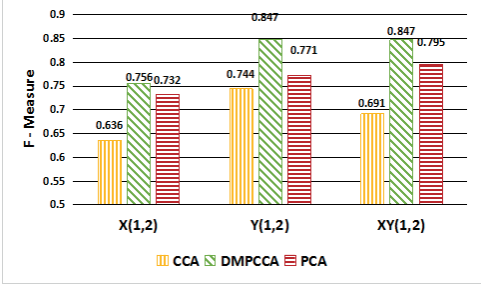
(a) One-dimension of UCIdata-1



(b) Two-dimension of UCIdata-1

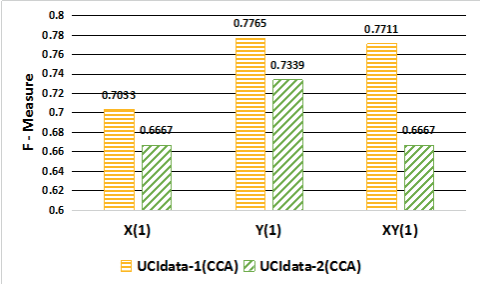


(c) One-dimension of UCIdata-2

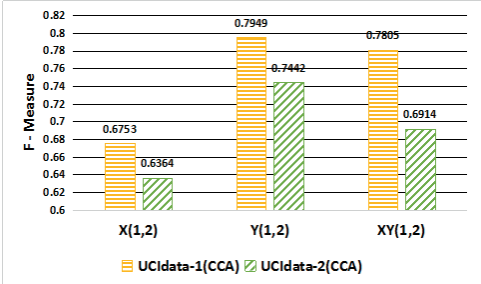


(d) Two-dimension of UCIdata-2

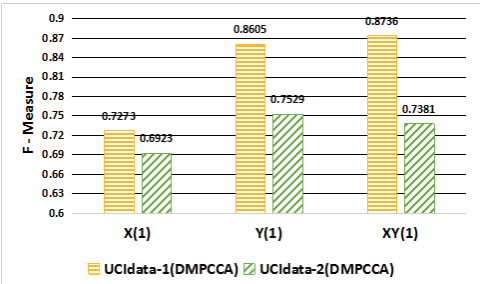
FIGURE 3. CCA, DMPCCA, PCA on UCI dataset



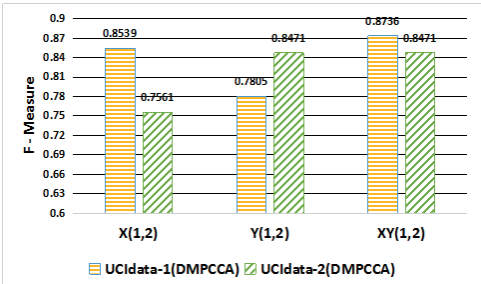
(a) UCIdata-1 VS UCIdata-2 using CCA



(b) UCIdata-1 VS UCIdata-2 using CCA

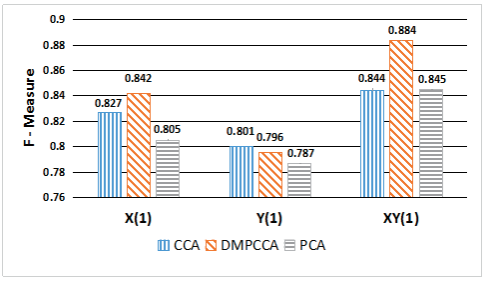


(c) UCIdata-1 VS UCIdata-2 using DMPCCA

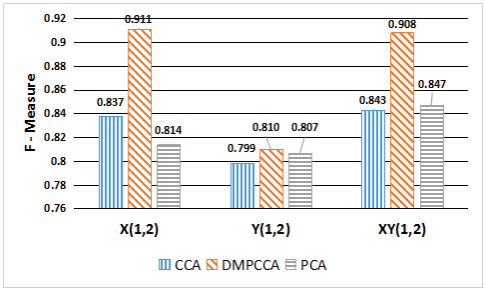


(d) UCIdata-1 VS UCIdata-2 using DMPCCA

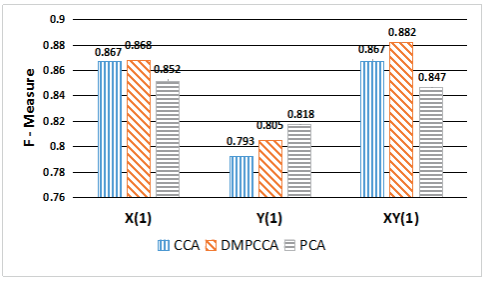
FIGURE 4. UCIdata-1 VS UCIdata-2 using CCA and DMPCCA



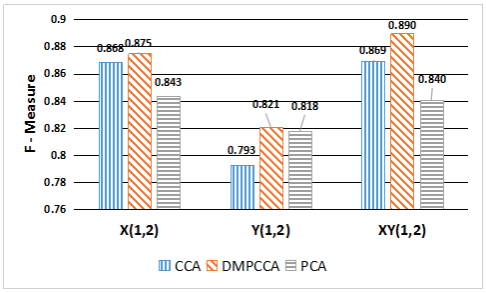
(a) One-dimension of HFdata-1



(b) Two-dimension of HFdata-1

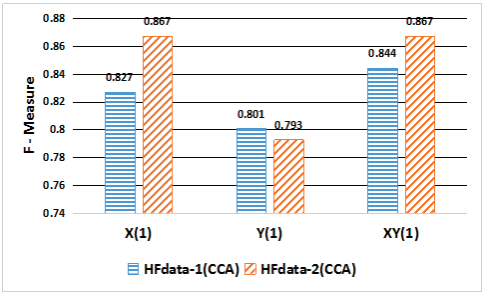


(c) One-dimension of HFdata-2

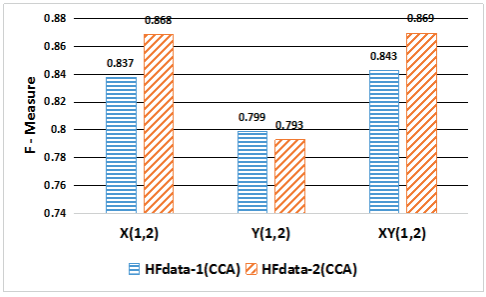


(d) Two-dimension of HFdata-2

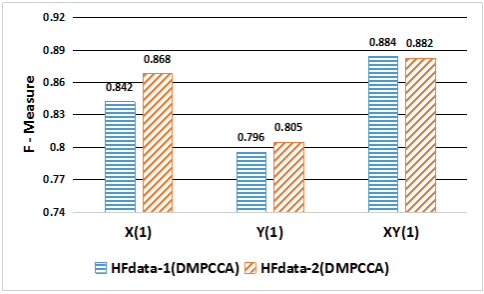
FIGURE 5. CCA, DMPCCA, PCA on HFdata-1 and HFdata-2



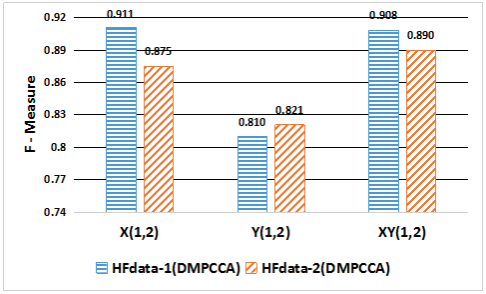
(a) HFdata-1 VS HFdata-2 using CCA



(b) HFdata-1 VS HFdata-2 using CCA



(c) HFdata-1 VS HFdata-2 using DMPCCA



(d) HFdata-1 VS HFdata-2 using DMPCCA

FIGURE 6. HFdata-1 VS HFdata-2 using CCA and DMPCCA

Acknowledgements

This research has been supported by the National High Technology Research and Development Program of China (863 Program) under Grant (No.2015AA020107).

References

[1] Shanthi Mendis, Pekka Puska, Bo Norrving, "Global Atlas on cardiovascular disease prevention and control", World Health Organization, pp. 3-18, 2011.

[2] Chen Weiwei, Fan Xiaohan, et al., "Report On Cardiovascular Disease In China(2015)", Chinese Circulation Journal, Vol 31, No. 6, June.2016.

[3] Peng Y., Zhang D., and Zhang J., "A New Canonical Correlation Analysis Algorithm with Local Discrimination", Neural Processing Letters, Vol 31, No. 1, pp. 1-15, Feb. 2010.

[4] Sheng Wang, Jianfeng Lu, et al., "Canonical principal angles correlation analysis for two-view data", Journal of Visual Communication and Image Representation, Vol 35, pp. 209-219, Feb. 2016.

[5] Baiying Lei, Siping Chen, et al., "Discriminative Learning for Alzheimer's Disease Diagnosis via Canonical Correlation Analysis and Multimodal Fusion", Front. Aging Neurosci., 17 May 2016.

[6] David R. Hardoon, Sandor Szedmak, et al., "Canonical Correlation Analysis: An Overview with Application to Learning Methods", Neural Computation, Vol 16, No.12, pp. 2639-2664, Dec. 2004.

[7] Yuan Yubo, Ma Chenglong, Pu Dongmei, "A Novel Discriminant Minimum Class Locality Preserving Canonical Correlation Analysis And Its Applications", Journal of Industrial and Management Optimization, Vol 12, No.1, pp. 251-268, JAN. 2016.

[8] Zhang Huanhuan, Wang Pengcheng, et al., "Risk Factors Of Heart Failure For Patients Classification With Extreme Learning Machine", Proceeding of ICMLC2016 Conference, Jeju, South Korea, pp. 814-819, July 2016.

[9] Anne Nakano, Kenneth Egstrup, et al., "Age and sex related differences in use of guideline-recommended care and mortality among patients with incident heart failure in Denmark", Age and Ageing, Vol 45, pp. 635-642, July 2016.

[10] David Blokh, Ilia Stambler, "Information Theoretical Analysis of Aging as a Risk Factor for Heart Disease", Aging and Disease, Vol 6, No.3, pp. 196-207, June 2015.

[11] Costas Sideris, Mohammad Pourhomayoun, et al., "A flexible data-driven comorbidity feature extraction framework", Computers in Biology and Medicine, Vol 73, pp. 165-172, 2016.

[12] Ronald Gijsen, Nancy Hoeymans, et al., "Causes and consequences of comorbidity: A review", Journal of Clinical Epidemiology, Vol 54, pp. 661-674, 2001.

[13] Sangeeta C. Ahluwalia, Cary P. Gross, et al., "Change in Comorbidity Prevalence with Advancing Age Among Persons with Heart Failure", Comorbidity Prevalence in Persons with Heart Failure, Vol 26, No.10, pp. 1145-51, May. 2011.

[14] Davide L. Vetrano, Andrea D. Foebel, et al., "Chronic diseases and geriatric syndromes: The different weight of comorbidity", European Journal of Internal Medicine, Vol 27, pp. 62-67, May. 2016.

[15] The Chinese medical association cardiovascular epidemiology branch, The cardiovascular disease magazine editorial board, "China heart failure treatment guidelines 2014", Chinese journal of practical rural doctors, Vol 42, pp. 675-690, 2015.

[16] Jian Yang, David Zhang, et al. "Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 26, NO.1, pp. 675-690, Jan. 2004.

[17] Aleix M., Avinash C. Kak, "PCA versus LDA", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 23, No.2, pp. 228-233, Feb. 2001.

[18] Aditya Methaila, Prince Kansal, et al., "Early Heart Disease Prediction using Data Mining Techniques", Computer Science and Information Technology, Vol 4, No.8, pp. 53-59, 2014.

[19] J. Jabez Christopher, H. Khanna Nehemiah , et al., "A Swarm Optimization approach for clinical knowledge mining", Comput Methods Programs Biomed, Vol 121, No.3, pp. 137-138, 2015.

[20] Xiekai Zhang, Shifei Ding , et al., "An improved multiple birth support vector machine for pattern classification", Neurocomputing, Vol 225, pp. 119-128, 2017.