Expert Systems with Applications: X 4 (2019) 100014

Contents lists available at ScienceDirect

Expert Systems with Applications: X

journal homepage: www.elsevier.com/locate/eswax

Mutual information inspired feature selection using kernel canonical correlation analysis

Yan Wang[a], Shuang Cang[b,*], Hongnian Yu[c,d,**]

[a] School of Electric and Information Engineering, Zhongyuan University of Technology, Zhengzhou, 450007, China
[b] School of Economics and Management, Yanshan University, Qinhuangdao 066004, China
[c] School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China
[d] School of Engineering and the Built Environment, Edinburgh Napier University, Edinburgh EH10 5DT, UK

ARTICLE INFO

Article history:
Received 22 September 2018
Revised 10 July 2019
Accepted 2 August 2019
Available online 3 August 2019

Keywords:
Feature selection
Joint redundancy
Kernel canonical correlation analysis
Mutual information
Incomplete Cholesky Decomposition

ABSTRACT

This paper proposes a filter-based feature selection method by combining the measurement of kernel canonical correlation analysis (KCCA) with the mutual information (MI)-based feature selection method, named mRMJR-KCCA. The mRMJR-KCCA maximizes the relevance between the feature candidate and the target class labels and simultaneously minimizes the joint redundancy between the feature candidate and the already selected features in the view of KCCA. To improve the computation efficiency, we adopt the Incomplete Cholesky Decomposition to approximate the kernel matrix in implementing the KCCA in mRMJR-KCCA for larger-size datasets. The proposed method is experimentally evaluated on 13 classification-associated datasets. Compared with certain popular feature selection methods, the experimental results demonstrate the better performance of the proposed mRMJR-KCCA.

1. Introduction

Hand-crafted features, as the inputs for most machine learning methods, are the quantitative and informative variables generated from the original data. Features can be time-domain (Machado, Gomes, Gamboa, Paixão, & Costa, 2015), frequency-domain (Suto, Oniga, & Sitar 2016), and hybrid (Montalto, Guerra, Bianchi, De Munari, & Ciampolini, 2015). The initial features usually include redundancy or may be too large to be efficiently dealt with, which results in several issues, such as higher computational cost involved in learning, low learning efficiency, over-fitting on unseen data, etc. (Chu, Liao, Ng, & Zhang, 2013; Gheid & Challal, 2016; Guyon & Elisseeff, 2003). Feature selection (FS), commonly used as a dimensionality reduction strategy, selects a smaller-size subset of the original feature set by removing the redundant and irrelevant features. The selected features are part of the original features without any feature transformation and maintain the physical meanings of the original features. In this way, FS helps users acquire a better understanding of their data by figuring out the most informative features, and hence to facilitate learning, enhance the generation performance and improve model interpretability (Tang, Alelyani, & Liu, 2014).

Supervised FS methods, designed for the classification or regression tasks, are generally seen as the following types: filter (Gheid & Challal, 2016), wrapper (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2013), and embedded approaches (Li, Cheng et al., 2017; Li, Zhu et al., 2017). Filter methods filter out irrelevant features by evaluating the relevance of a feature to the class label using a specific selection criterion (Urbanowicz, Meeker, LaCava, Olson, & Moore, 2017). A filter algorithm first ranks the original features based on the criterion, then selects the features with higher rankings. The above selection process is independent of any classifier, computationally efficient and usually obtains a trade-off between performance and efficiency.

Selection criteria play a critical role in filter-based FS methods. A range of criteria has been explored in the past decades, such as distance measure, similarity, dependency, mutual information (MI), correlation measure, canonical correlation analysis (CCA) (Dessì et al., 2015; Gheid et al., 2016, Li, Cheng et al., 2017; Li, Zhu et al., 2017). As the largest family in filter-based FS methods, an MI-based FS algorithm measures the importance of a feature by its selection criterion with the class label, assuming that the feature with a stronger correlation with the label will improve

[*] Corresponding author.
[**] Corresponding author at: School of Economics and Management, Yanshan University, Qinhuangdao 066004, China and School of Engineering and the Built Environment, Edinburgh Napier University, Edinburgh EH10 5DT, UK.
E-mail addresses: cangshuang@ysu.edu.cn (S. Cang), yu61150@IEEE.Org (H. Yu).

https://doi.org/10.1016/j.eswax.2019.100014

classification performance. The popular algorithms in MI family are minimum Relevance Maximum Relevance (mRMR) (Peng, Long, & Ding, 2005), Joint Mutual Information (JMI) (Bennasar, Hicks, & Setchi, 2015), Conditional Mutual Information Maximum (CMIM) (Gao, Ver Steeg, & Galstyan, 2016), etc. MI considers the correlation of variables in pairs and then uses a simple approximation strategy, i.e., the sum or the average, to approximate the relation between one variable (a feature or a label) and multidimensional variables (e.g., a set of features) (Brown, Pocock, Zhao, & Luján, 2012). As a result, the MI-based FS shares a common problem, i.e., it doesn't fully consider the complementarity within a set of variables. Different from the MI, the CCA measures the linear correlation between two sets of multidimensional variables by maximizing the correlation coefficients between them. The CCA may not extract a useful description of the data due to its linearity. The KCCA is a nonlinear correlation measurement by mapping the data into a higher-dimensional feature space with kernel tricks (Hardoon, Szedmak, & Shawe-Taylor, 2004). The CCA or the KCCA are easily employed as a feature selector (Mehrkanoon & Suykens, 2017; Yoshida, Yoshimoto, & Doya, 2017).

Inspired by MI-based FS methods and CCA-based measurements, this paper proposes and implements a new FS method, named mRMJR-KCCA. The mRMJR-KCCA maximizes the relevance between the feature candidate and the class labels and simultaneously minimizes the joint redundancy between the feature candidate and the already selected features by using KCCA. The proposed mRMJR-KCCA is experimentally evaluated over the 10 classification–related benchmark datasets from UCI[1] and our three ground-truth datasets involving 17 daily activities from 21 volunteers. We also compare mRMJR-KCCA with other available popular FS methods, including MCR-CCA and mRMR-CCA (Kaya, Eyben, Salah, & Schuller, 2014), Autoencoder (Wang, 2016), Sparse Filtering (Ngiam, Chen, Bhaskar, Koh, & Ng, 2011), four MI-based methods (Brown et al., 2012). The contributions of this paper are summarized as (1): mRMR uses the approximation of sum operation Σ when measuring the redundancy between the feature candidate and the already selected features in pairs, which somehow does not fully consider the complementarity within the already selected features. Our proposed mRMJR-KCCA introduces the measurement of KCCA into mRMR, which replaces the approximation of sum in mRMR with the KCCA analysis to measure the joint redundancy between the feature candidate and the already selected features. (2): We apply Incomplete Cholesky Decomposition (ICD) (Li, Bi, Kwok, & Lu, 2015) to reduce the dimensionality of the kernel matrix in the implementation of mRMJR-KCCA on the large-size ground truth datasets. (3): We also investigate the impact of the kernel parameter and the number of components decomposed from the kernel matrix by ICD on the classification accuracies.

The rest of the paper is organized as follows. Section 2 describes the fundamentals of MI and CCA and related studies. Section 3 presents the proposed method, mRMJR-KCCA, and its implementation. Section 4 gives the experimental results and the discussions. The conclusion is provided in Section 5.

2. Related works and fundamentals

2.1. Entropy and MI-based FS

This paper considers two groups of FS methods, and the first one is the MI-based FS. The MI is one of the most effective criteria to measure the correlation between variables. Let x and y be two discrete random variables, both x and y have N observations, the MI between x and y is defined as

$$I(x;y) = H(y) - H(y|x) = \sum_{x,y} p(x,y)\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where $H(y)$ represents the entropy of y which quantifies the degree of uncertainty in a discrete or discretized random variable y and $H(x|y)$ represents the conditional entropy of x given y; $p(.)$ is the probability mass function (Bennasar et al., 2015). The MI signifies how much information x and y share, which is nonnegative and equals zero if x and y are independent. The minimum Redundancy Maximum Relevance (mRMR) algorithm (Peng et al., 2005), which directly uses MI to value the redundancy and relevance of involved variables, is one of the most popular FS methods. The ranking criterion of the mRMR is

$$\mathrm{J}_{mRMR}(f_k) = \max_{f_l \in S, f_k \in F-S}\left[I(f_k;C) - \frac{1}{|S|}\sum I(f_k;f_l)\right] \tag{2}$$

where $I(;)$ is given in Eq. (1), f_k is a feature candidate; F is the whole feature set; S is the already selected feature set; f_l can be any feature in S; and C is the class labels. The second term in Eq. (2) considers the redundancy between the feature candidate and any already selected features in terms of paired variables, which doesn't fully consider the joint relevance and the conditional redundancy given the third or more variables. The improved mutual information measures can deal with the MI between three variables, one of which is Conditional Mutual Information Maximization (CMIM) (Brown et al., 2012). The corresponding criterion of the CMIM is

$$\mathrm{J}_{cmim}(f_k) = I(f_k;C) - \max[I(f_k;f_l) - I(f_k;f_l|C)] \tag{3}$$

where the additional term $I(f_k;\ f_l|C)$ includes the redundancy given the class labels C compared with the mRMR criterion. The other two typical MI-based methods are Joint Mutual Information (JMI) that includes the complementary information that is shared between the feature candidate and the already selected features given the class labels. The criterion of JMI is given in Eq. (4) below (Brown et al., 2012). Double Input Symmetrical Relevance (DISR) is the modification of JMI by estimating the normalization H(f_k, f_l; C).

$$\mathrm{J}_{JMI}(f_k) = max\sum_{f_l \in S} I(f_k, f_l; C) \tag{4}$$

where $I(f_k f_l;\ C)$ is the joint mutual information of variables f_k, f_l and C.

2.2. CCA and KCCA

CCA statistically finds the correlation between two sets of random variables X and Y(Hotelling, 1936). Denote $X = (x_1, \ldots x_p) \in R^{N\times p}$, $Y = (y_1, \ldots y_q) \in R^{N\times q}$. X and Y can be two feature spaces, or a feature space and a label space. To obtain the correlation between the two sets of variables, CCA finds a linear projection u in the space of X, and a linear projection v in the space of Y to maximize the following sample correlation in Eq. (5). Such that the projected data $u'X$ and $v'Y$ have a maximum correlation.

$$\rho_{CCA} = \underset{u\in R^p,\ v\in R^q}{\mathrm{argmax}} \frac{u'X'Yv}{\sqrt{(u'X'Xu)(v'Y'Yv)}} \tag{5}$$

CCA-based filter FS methods intend to use the correlation (measured by Eq. (5)) between the two projections of the variable sets to figure out the most important original features. Kaya et al. (2014) propose two CCA-based FS methods. The first method is called mRMR-CCA, which replaces the MI indicator with the CCA coefficient, as presented in Eq. (6). The second term in Eq. (6) is changed from a sum of paired redundancies in Eq. (2) to

[1] http://archive.ics.uci.edu/ml/.

redundancy which is handled once from multidimensional variables.

$$J_{mRMR-CCA}(f_k) = \max[\rho_{CCA}(f_k; C) - \rho_{CCA}(f_k; S)] \tag{6}$$

where ρ_{CCA} is given in Eq. (5). The second method in Kaya et al. (2014) is the Maximum Collective Relevance (MCR-CCA), similar to the JMI, which maximizes the collective correlation of the feature candidate and the already selected features against the class labels. The criterion of the MCR-CCA is

$$J_{MCR-CCA}(f_k) = \max[\rho_{CCA}(f_k \cup S; C)] \tag{7}$$

The CCA describes the linear correlation between two sets of variables, which are often insufficient to reveal the highly nonlinear correlation with many real-world data (Wang et al., 2015). The KCCA provides a nonlinear extension of CCA, which catches the nonlinear correlation by mapping the data into a higher-dimensional feature space before performing CCA (Sakar, Kursun, & Gurgen, 2012). The KCCA-applied correlation between two sets of random variables X and Y is thus to identify the weights α, β that maximize

$$\rho_{KCCA} = \underset{\alpha,\beta}{\operatorname{argmax}} \frac{\alpha' K_X K_Y \beta}{\sqrt{(\alpha' K_X K_X \alpha)(\beta' K_Y K_Y \beta)}} \tag{8}$$

where $K_X = XX'$ and $K_Y = YY'$ are the kernel matrices corresponding to the variable sets X and Y. However, the kernelized CCA problem in Eq. (8) causes an ill-posed inverse problem, and thus a regularization approach is needed to construct a meaningful estimator of the canonical correlation (Ashad Alam & Fukumizu, 2015; Bach & Jordan, 2002). The objective function for regularized kernel CCA becomes

$$\rho_{KCCA} = \underset{\alpha,\beta}{\operatorname{argmax}} \frac{\alpha' K_X K_Y \beta}{\sqrt{(\alpha' K_X K_X \alpha + \epsilon \alpha' K_X \alpha) \cdot (\beta' K_Y K_Y \beta + \epsilon \beta' K_Y \beta)}} \tag{9}$$

where ϵ is a regularization parameter that should be a small and positive value and approaches zero with an increasing sample size N (Lisanti, Masi, & Del Bimbo, 2014).

In KCCA, the inputs $X = \{x_p\}_1^N$ and $Y = \{y_q\}_1^N$ caused kernel matrix K_X and K_Y are both with the size of $N \times N$. Thus, solving Eq. (9) involves an eigenvalue problem of size $N \times N$, which is expensive both in memory (storing the kernel matrices) and in time with naively costs $\mathcal{O}(N^3)$ (Wang & Livescu, 2015). To overcome this issue, a range of kernel approximation techniques have been proposed to scale up KCCA, including singular value decomposition (SVD) (Chakraborty, Chatterjee, Dey, Ashour, & Hassanien, 2017), Nyström method (Patel, Goldstein, Dyer, Mirhoseini, & Baraniuk, 2016), Incomplete Cholesky Decomposition (ICD) (Li, Bi, Kwok, & Lu, 2015), and so on. After applying the above approximation methods, the efficiency of calculating KCCA can be much improved (Wang & Livescu, 2015).

3. The proposed KCCA based feature selection method

Over the last two decades, the KCCA has been using for various purposes in statistic and machine learning, such as feature learning (Sakar et al., 2012), computational vision (Bilenko and Gallant, 2016), statistical independence measurement (Lopez-Paz, Hennig, & Schölkopf, 2013) and so on. Lisanti et al. (2014) investigate matching people across cameras views by applying a learning method based on KCCA to find a common substance between their proposed descriptors, and their experimental results demonstrate the superiority of the proposed method. Sakar et al. (2012) propose a filter method for feature selection with the aim to find the unique information, which exploits correlated functions explored by KCCA as the inputs to mRMR. They demonstrate the effectiveness of their method on some benchmark datasets. Considering Eqs. (6)–(9), we propose a new kernel version FS method, i.e., mRMJR-KCCA, by applying KCCA in Eq. (9) to Eq. (6). The criterion of mRMJR-KCCA is

$$J_{mRMJR-KCCA}(f_k) = \max_{f_k \in F-S}[\rho_{KCCA}(f_k; C) - \rho_{KCCA}(S; f_k)] \tag{10}$$

where ρ_{KCCA} is the correlation coefficient calculated by KCCA between two sets of variables, given in Eq. (9). It is noted that we in fact useρ_{corr} (the Pearson's correlation) in calculating the first item (i.e., the relevance of the feature candidate and the target labels) in Eq. (10), since the CCA or KCCA essentially perform the calculation of the Pearson's correlation (Zou, Zeng, Cao, & Ji, 2016) when both X and Y are two vectors (such as f_k and C) in Eq. (5) or Eq. (8), The mRMJR-KCCA combines the idea of the mRMR and KCCA to maximize the relevance between the feature candidate and the target class labels, and simultaneously minimize the joint redundancy between the already selected features and the feature candidate.

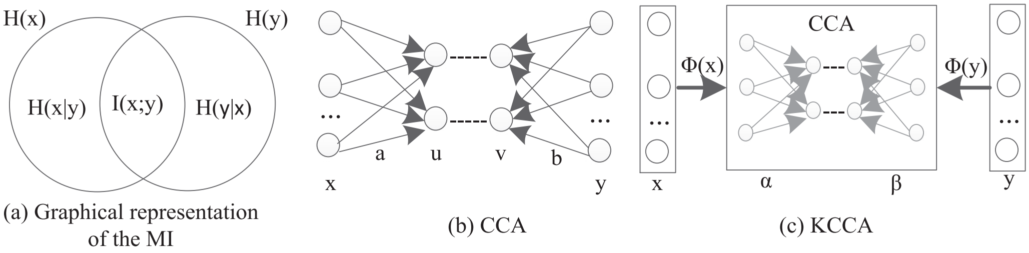
The MI between two variables in Eqs. (1) and (2) is the sum of MI between the discrete variates x and y if there are no higher order statistic dependencies than correlation (Fig. 1(a)). The CCA in Eq. (5) finds a pair of linear transformations from X and Y such that the correlation coefficient between extracted features is maximized (Fig. 1(b)). The KCCA in Eq. (8) finds pairs of nonlinear projections of the two views, and the optimal projections can maximize the correlation between X and Y by mapping the data-cases to feature vectors $\Phi(x)$ and $\Phi(y)$, as shown in Fig. 1(c).

The second term in Eq. (2) (Brown et al., 2012) is replaced from an approximation of sum of the paired redundancies with a new redundancy measurement in Eq. (6) (Kaya et al., 2014) which is handled once for multidimensional variables by CCA. Our proposed mRMJR-KCCA further changes the measurement of CCA in Eq. (6) to the KCCA, as presented in Eq. (10).

To implement the mRMJR-KCCA especially for the large-size datasets, we apply Incomplete Cholesky Decomposition (ICD) for kernel matrix approximation to improve the computation efficiency due to its accurate matrix approximation with far fewer samples (Patel et al., 2016). ICD generates a low-rank matrix $N \times M$ ($M \ll N$) by performing a standard Cholesky Decomposition but terminating the decomposition considering a small number of columns (M). So that the complexity to the eigenvalue problem of size $N \times N$ in Eq. (9) turns to $\mathcal{O}(M^2N)$ (Hardoon et al., 2004). Table 1 details the procedure to implement mRMJR-KCCA in this paper.

The mRMJR-KCCA algorithm ranks the features by the maximal relevance between the feature candidate and the target class labels and the minimal joint redundancy between the feature candidate and the already selected features, as presented in Eq. (10). It is noted that the nonlinear correlation coefficient is used to rank the feature candidates following Eq. (10), which is acquired by the transformation in KCCA. However, the coefficient is only for ranking the features, the selected features with higher ranking are still the original features instead of the transformed data. The steps of the mRMJR-KCCA algorithm in Table 1 are explained in detail below:

Step 1: Normalize features value to [0 1] range. This step ensures that all features have the same importance.
Step 2: Calculate the relevance score of each feature candidate with the class labels based on the first item in Eq. (10).
Step 3: Select the first feature f_s which has maximal relevance score in Step 2.
Step 4: Update $S = S \cup \{f_s\}$, $F = F|\{f_s\}$.
Step 5: Calculate the mRMJR-KCCA using Eq. (10). Also, the ICD is adopted to improve the implementation of KCCA in Eq. (10).



(a) Graphical representation of the MI

(b) CCA

(c) KCCA

Fig. 1. The representation of MI, CCA, and KCCA.

Table 1
Pseudocode of the mRMJR-KCCA.

```
Algorithm mRMJR-KCCA: Maximum Relevance and Minimum Joint Redundancy Kernel CCA

Input: an original feature set F, the number of features to be selected U
Output: a selected feature set S
  Initialize $F = \{f_1, f_2, \dots f_l, \dots f_n\}, S = \{\}, U$
  Normalize features to [0 1]
  Calculate $\rho_{KCCA}(f_n, C)$ using Eq. (9) for each $f_n$ with the class labels $C$
  Select the first feature $f_s$ with maximum $\rho_{KCCA}(f_n, C)$
  Update $S = S \cup \{f_s\}, \ F = F|\{f_s\}$
  If U<desired numbers
    Calculate mRMJR-KCCA: $\rho_{KCCA}(f_k; C) - \rho_{KCCA}(S; C)$ following Eq. (10)
    Select the next feature that maximizing mRMJR-KCCA
    Update $S$, $F$
  End
Write S to an excel file
```

Table 2
Descriptions of UCI datasets and ground-truth datasets used in the experiments.

Dataset	Data type	# Feature	# Class	# Instance	Year
1 Blood	Real	4	2	748	2008
2 Diabetes	Integer, Real	8	2	768	1990
3 Heart	Categorical, Real	13	2	270	N/A
4 Iris	Real	4	3	150	1988
5 Parkinsons	Real	22	2	195	2008
6 Seeds	Real	7	3	210	2012
7 Wdbc	Real	30	2	569	1995
8 Wine	Integer, Real	13	3	178	1991
9 Wine_red	Real	11	6	1599	2009
10 Wpbc	Real	33	2	198	1995
11 X_HAR	Real	75	17	32,844	2015
12 Y_HAR	Real	296	17	32,844	2015
13 Z_HAR	Real	371	17	32,844	2015

Step 6: Select the next feature which maximizes the mRMJR-KCCA.
Step 7: Go to Step 4 if the number of the already selected features is lower than the number of features to be selected.

It is noted that the main difference compared with the CCA approach in Kaya et al. (2014) is in Step 5. Due to applying Eq. (9) in the proposed Eq. (10), we utilize the ICD to approximate a kernel matrix and map the features into the nonlinear space especially for the larger-size dataset, such as the datasets of X_HAR, Y_HAR and Z_HAR in Table 2.

4. Experimentations and results

4.1. Benchmark datasets and learning algorithms

We employ 10 UCI benchmark datasets and three ground-truth datasets to evaluate the performance of mRMJR-KCCA. The datasets are all related to classification problems, covering both binary-class and multi-class; the data type includes real, integer and categorical; the number of original features ranges from 4 to 371; the sample number of each dataset varies from 150 to 32,844. The ground truth datasets 10–12 contain the daily activities performed in a home environment using five wearable sensors. The data sets 11, 12 and 13 record 17 activities from 21 subjects with 20 Hz sampling rate. X_HAR represents the feature set extracted from the wearable's attitude (roll, pitch, and yaw) and Y_HAR is the feature set generated from the sensor readings of an accelerometer, a gyroscope and a magnetometer, a barometer and a temperature individually. Z_HAR is the combination of X_HAR and Y_HAR. The details of all the datasets used in this work are shown in Table 2.

We experimentally evaluate mRMJR-KCCA using two learning algorithms on the selected subset of features, i.e., Support Vector Machines (SVM) and Random forest (RF) due to their excellent performance in classification applications (Alickovic, Kevric, & Subasi, 2018; Chernbumroong, Cang, & Yu, 2014; Sani, Massie, Wiratunga, & Cooper, 2017). The pair of parameters *gamma* and *c* in SVM, and the number of trees in RF are determined in 10-fold cross validation process individually. The results report the average accuracy from 10 times test. At the same time, we compare our proposed method with other available popular FS methods presented in Section 1.

4.2. Experimental results on the used datasets

The classification accuracies with SVM and RF are shown in Table 3 and Table 4, respectively, in which the best method for

Table 3
Classification accuracy (%) with SVM classification.

Dataset (# of the selected best features)	mRMJR-KCCA (proposed)	mRMR-CCA[a]	MCR-CCA[a]	Sparse Filtering[b]	Autoencoder[c]	mRMR[d]	JMI[d]	CMIM[d]	DISR[d]
Blood (4)	**77.94**	**77.94**	**77.94**	**77.94**	**77.94**	**77.94**	**77.94**	**77.94**	**77.94**
Diabetes (7)	77.98	77.98	**78.12**	72.26	70.18	77.98	77.79	77.99	77.79
Heart (5)	84.07	**84.93**	84.81	71.48	80.37	83.33	83.85	83.33	83.70
Iris (4)	**96.67**	**96.67**	**96.67**	**96.67**	**96.67**	**96.67**	**96.67**	**96.67**	**96.67**
Parkinsons (5)	92.21	91.74	91.24	91.26	**92.76**	92.21	90.74	89.58	90.21
Seeds (3)	93.81	91.43	93.81	94.76	**96.19**	94.29	92.86	93.81	93.81
Wdbc (12)	**97.71**	97.01	97.07	95.25	95.78	96.31	96.31	96.32	96.52
Wine (10)	**99.44**	97.78	**99.44**	97.78	96.22	96.11	**99.44**	**99.44**	**99.44**
Wine_red (4)	68.35	68.98	68.29	**70.1**	66.48	68.17	68.04	68.04	68.05
Wpbc (5)	80.82	79.26	80.37	76.82	78.82	**81.37**	78.26	78.82	78.79
X_HAR (20)	96.51	94.90	96.10	95.75	94.61	93.46	**96.82**	**96.82**	96.78
Y_HAR (20)	**97.29**	96.14	96.01	95.92	94.3	89.81	86.83	88.26	86.98
Z_HAR (30)	**98.50**	97.75	97.75	98.04	97.51	91.19	90.61	91.74	90.63
Rank*	**6**	3	4	3	4	3	4	4	3

Rank* denotes each FS method's ranking measured by the times of the FS method bests the others on the 13 datasets, i.e., the bigger number means higher ranking.
[a] Kaya et al. (2014).
[b] Ngiam et al. (2011).
[c] Wang (2016).
[d] Brown et al. (2012).

Table 4
Classification accuracy (%) with RF classification.

Dataset (# of the selected best features)	mRMJR-KCCA (proposed)	mRMR-CCA[a]	MCR-CCA[a]	Sparse Filtering[b]	Autoencoder[c]	mRMR[d]	JMI[d]	CMIM[d]	DISR[d]
Blood (3)	**75.94**	**75.94**	**75.94**	**75.94**	**75.94**	**75.94**	**75.94**	**75.94**	**75.94**
Diabetes (6)	76.29	**77.47**	77.07	71.62	68.22	77.46	76.68	77.46	76.51
Heart (3)	**84.44**	82.22	83.33	71.11	80.74	82.22	82.22	82.22	81.48
Iris (2)	**96.67**	**96.67**	**96.67**	**96.67**	**96.67**	**96.67**	**96.67**	**96.67**	**96.67**
Parkinsons (10)	**94.34**	92.26	92.79	90.26	89.18	90.13	90.66	92.26	91.68
Seeds (4)	92.86	90.48	94.29	93.81	**95.24**	94.76	94.29	94.29	94.29
Wdbc (5)	**96.84**	96.08	96.08	94.02	96.14	96.39	96.19	96.05	95.93
Wine (7)	97.75	95.57	**97.78**	96.6	96.86	96.29	**97.78**	**97.78**	**97.78**
Wine_red (8)	64.29	64.29	63.66	60.91	**70.98**	64.60	62.23	63.29	62.23
Wpbc (3)	76.87	76.76	76.79	76.76	**81.79**	76.76	76.32	77.29	76.29
X_HAR (30)	96.62	95.63	**96.65**	93.55	92.74	94.28	96.55	96.63	96.57
Y_HAR (30)	**97.80**	95.79	95.79	94.17	93.39	96.25	96.52	96.69	96.80
Z_HAR (30)	**98.80**	97.88	97.87	95.81	95.67	96.71	95.88	96.86	95.92
Rank*	**7**	3	4	2	5	2	3	3	3

Rank* denotes each FS method's ranking measured by the times of the FS method bests the others on the 13 datasets, i.e., the bigger number means higher ranking.
[a] Kaya et al. (2014).
[b] Ngiam et al. (2011).
[c] Wang (2016).
[d] Brown et al. (2012).

each dataset is highlighted in bold. Based on the SVM-based classification results in Table 3, the mRMJR-KCCA produces the best performance with the largest number (6) of higher ranking on the total 13 datasets. The mRMJR-KCCA bests the other FS methods on the datasets Blood, Iris, Wdbc, Wine, Y_HAR, and Z_HAR. The CCA-based methods show better performance than the MI-based methods regarding the Rank* in Table 3. The accuracies of MI-based methods on datasets Y-HAR and Z-HAR are much lower, which lowers down the Rank* of the MI-based methods. However, the mRMR presents the highest accuracy of 81.37% on the dataset Wpbc. On the datasets Blood and Iris, all the nine FS methods present the same performances since the original size of Blood and Iris is small (=4) and all the four features are used for classification respectively, the performances are therefore independent of the feature selection methods. The Autoencoder presents the highest accuracies of 92.76% and 96.19 on the datasets Parkinsons and Seeds respectively. The Sparse Filtering performs best on the dataset Wine_red (70.1%). Regarding the Rank* of each FS method in Table 3, the Autoencoder, MCR-CCA, JMI, and CMIM can still provide the performance four times better than the other methods.

Considering the RF classification results in Table 4, the mRMJR-KCCA and Autoencoder rank the first two on the 13 datasets regarding the Rank*, followed by the MCR-CCA. Meanwhile, the mRMJR-KCCA bests the other methods seven times with RF. The Autoencoder outperforms others four times with SVM in Table 3 and five times with RF in Table 4. The JMI, CMIM, DISR, and MCR-CCA perform best on the dataset Wine with RF classification. The Autoencoder and the Sparse Filtering obtains much lower results on datasets of Heart and Diabetes with both SVM and RF; this brings down the performance of the Autoencoder on the used datasets. The Autoencoder and Sparse Filtering fail to show their superiority in this paper, which could be attributed to the fact that we only use one-layer Sparse Filtering and Autoencoder. The superiority may be revealed when increasing the layers of Autoencoder and Sparse Filtering. The mRMR produces the highest accuracy of 70.98% on the dataset Wine_red with RF in Table 4, while it performs best (81.37%) on Wpbc in Table 3 with SVM. This implies that different classification methods can produce different results even on the same feature sets due to the parameters optimization or the intrinsic quality of a classification method. From the results in Tables 3 and 4, the performance of the mRMJR-KCCA remains consistent, which rank the first with both SVM and RF classification; the Autoencoder performs well in both Tables 3 and 4.

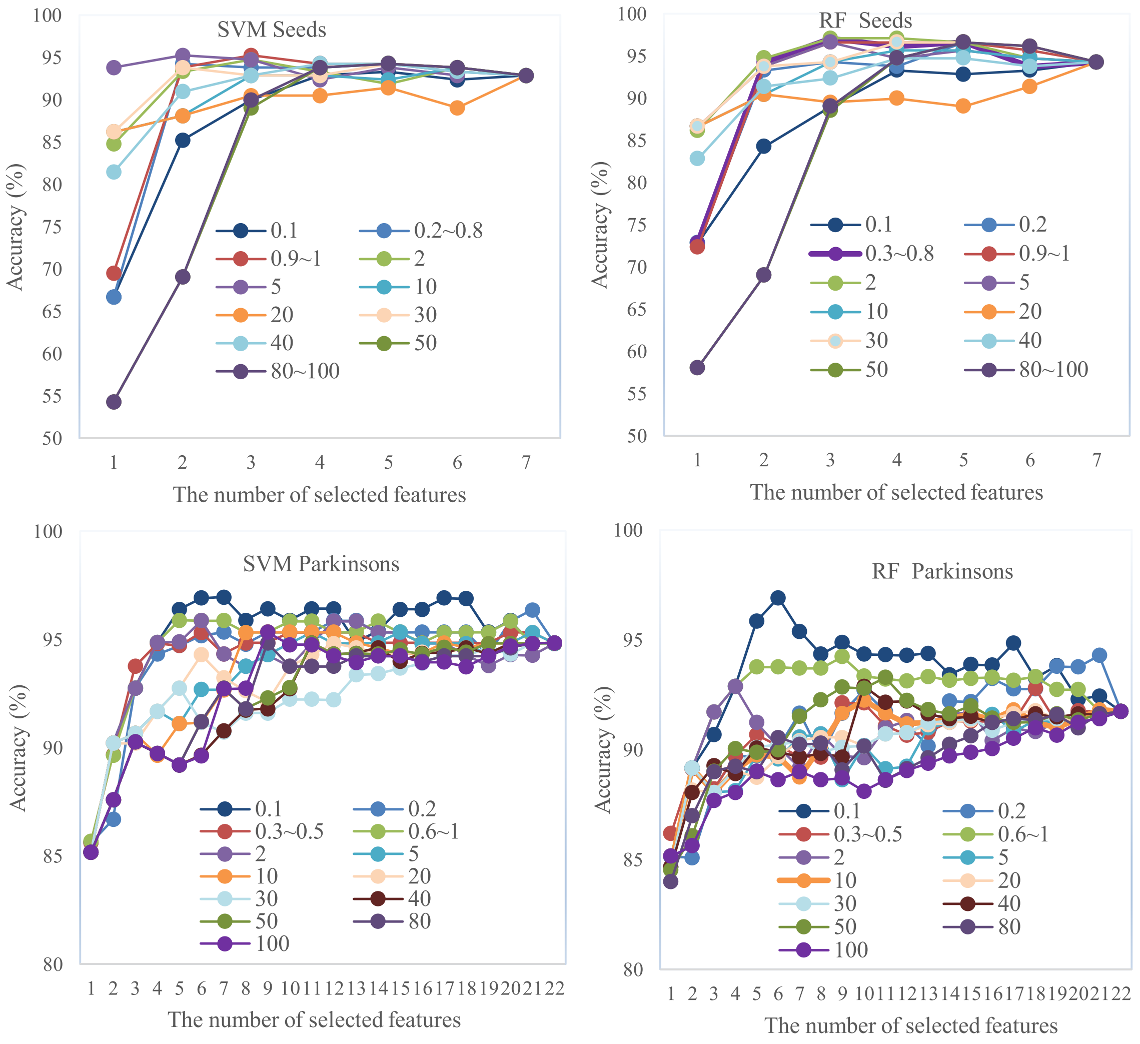


Fig. 2. Classification accuracy variations with the values of $\gamma(0.1 \sim 100)$ on datasets of Seeds and Parkinsons.

4.3. Impact of kernel parameter of on the obtained performance of KCCA

To produce kernel matrices in KCCA in this paper, we use a Gaussian RBF kernel, given in Eq. (11). Here, x and x' represent two feature vectors. The parameter γ in Eq. (11) differs from the choice of kernel bandwidth, which affects the shape of the distribution of canonical features.

$$k(x, x') = e^{-\gamma \|x - x'\|^2} \tag{11}$$

We therefore choose three datasets in Table 2 to explore the impact of the kernel parameter γ on different datasets in this section. Fig. 2 shows the variations of classification accuracy along with the different kernel parameter γ in mRMJR-KCCA on datasets of Seeds and Parkisons. Here, we set γ from 0.1 to 100 with different steps. Fig. 2 only presents part of the results based on the set γ values since some γ values yield similar results, e.g., $\gamma = 80$–100. The values of γ have different impacts on different datasets. For instance, the values of $\gamma = 0.9$,1 and 2 produce better performance on dataset Seeds with both SVM and RF classification, while the values of $\gamma = 0.1$ and 1 perform better on dataset Parkinsons. $\gamma = 1$ exhibits robust and steady performance on both datasets. It is noted that we set γ as 1 for most datasets in Tables 3 and 4. Fig. 3 presents the impact of γ on the accuracies of dataset X_HAR when we fix the number of selected features as 30, from which we can see that when $\gamma = 0.3$, 0.5, 0.9, 1 and 2, better and similar results with both SVM and RF are achieved. This further demonstrates $\gamma = 1$ exhibits better results for most of the datasets used in this paper. The choice of the γ values has different effects on the performance of the mRMJR-KCCA in Figs. 2 and 3. For other datasets in Table 2, the optimization of

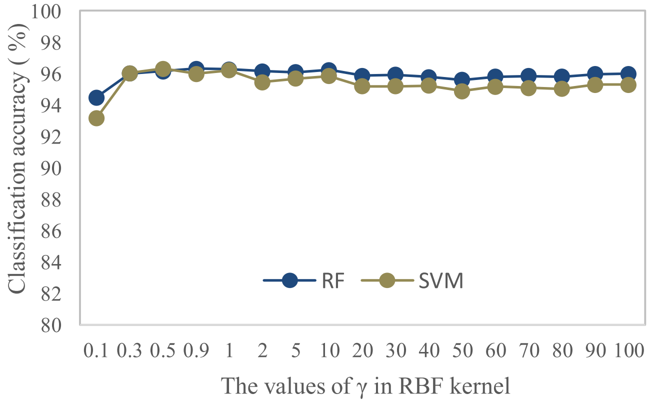


Fig. 3. Classification accuracies versus varied γ values in the RBF kerne l on X_HAR.

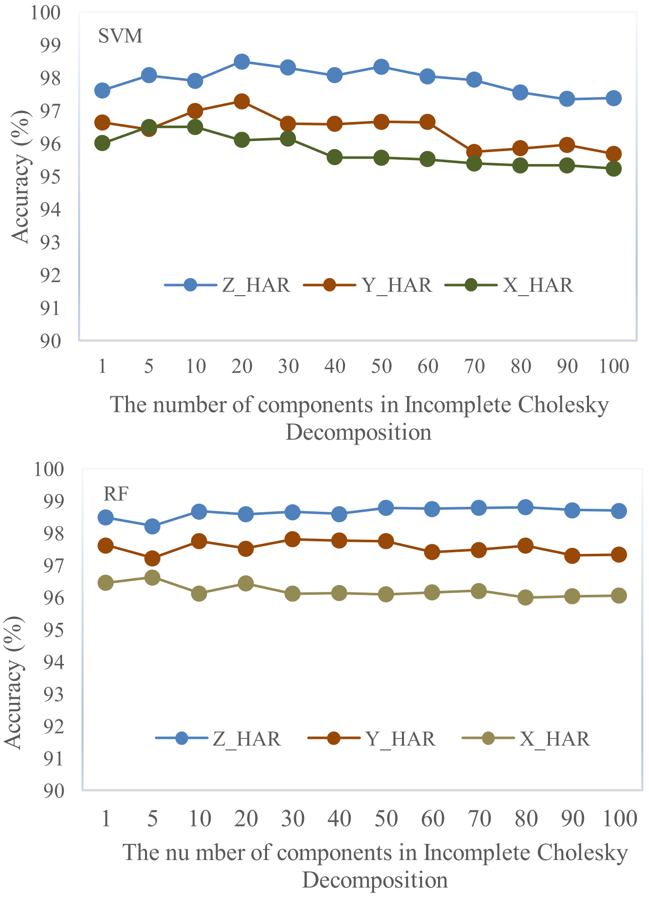


Fig. 4. Classification accuracies versus the number of components in ICD on datasets of X_HAR, Y-HAR, and Z_HAR.

the parameter γ in RBF kernel when using the mRMJR-KCCA can be studied by trials on each dataset or deploying some algorithms (such as genetic algorithm) to attain the optimized γ.

4.4. Impact of the number of the components decomposed in ICD from kernel matrices on the obtained performance

In Table 2, the sample sizes of the first 10 datasets can be easily dealt with to complete the full kernel matrix in KCCA. However, the sample sizes of the datasets of X_HAR, Y_HAR and Z_HAR are much larger (e.g., $N = 32{,}844$), which is memory intensive and computation expensive to realize a $\mathcal{O}(N^3)$ kernel matrix solution. A positive semi-definite matrix K can be decomposed as LL^*, where L is an $N \times N$ matrix, the decomposition in Incomplete Cholesky Decomposition (ICD) is to find a matrix $\tilde{L}$ of size $N \times M$, for small M, such that the difference $K\text{-}\tilde{L}\tilde{L}^T$ has norm less than a given value (Bach et al., 2002). This paper applies ICD on the KCCA for kernel matrix approximation, which reduces the computational complexity of KCCA to $\mathcal{O}(M^2N)$, here, M is the maximal rank of the solution. We set a range of M from 1 to 100 to investigate the impact of the number of the components in ICD on X_HAR, Y_HAR and Z_HAR using the top 30 selected features. Fig. 4 presents the effect of increasing the number of components decomposed in ICD on the performance of mRMJR-KCCA evaluated by SVM and RF. It can be seen in Fig. 4 that the number of components in ICD has a slight impact on datasets of X_HAR, Y_HAR and Z_HAR with RF classification, whilst, it has a bigger impact when using SVM classification. This may be attributed to that the optimal parameters in RF models are easier to obtain than the counterpart in SVM models.

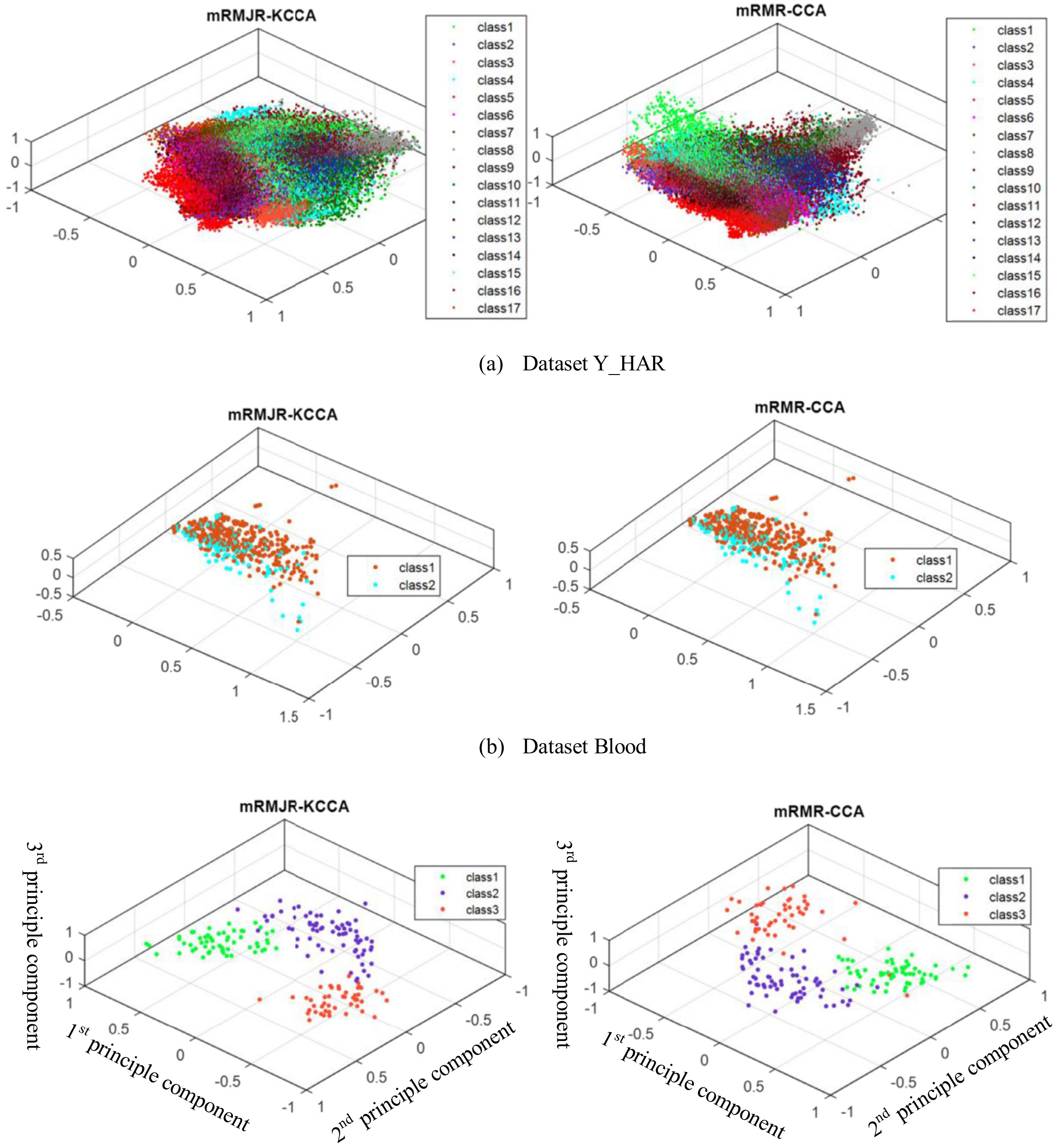
From Fig. 4, we also observe that increasing the number of components decomposed in ICD from kernel matrices does not necessarily increase the performance. When $M = 1$, 20 and 50, the better performances are achieved on mRMJR-KCCA and RF; when $M = 20$, the best performance is achieved with mRMJR-KCCA and SVM. Consequently, the impact of the number of the components in KCCA may depend on the dataset itself from the experimental results.

4.5. Impact of the features extracted by linear CCA and nonlinear KCCA on the performance

CCA finds pairs of basis vectors that maximise the correlation of a set of paired of variables, and these pairs can be considered as two views of the same object. The KCCA is a technique that generalises the linear CCA to nonlinear setting. This allows us to extract the nonlinear relation of two sets of variables. This paper uses the linear correlation coefficients in Eq. (6) for mRMR-CCA feature selection and nonlinear correlation coefficient in Eq. (10) for mRMJR-KCCA feature selection. Whilst, it is difficult to tell which real datasets imply linear or nonlinear correlation among the features. Tables 3 and 4 show that the mRMJR- KCCA produces the highest average performance and rank on the used benchmark datasets. However, the mRMJR-KCCA does not perform best on all the datasets. For example, the mRMR-CCA and mRMJR-KCCA perform the same on the dataset Blood, and the latter performs better than the former on most datasets. To visualize the impact of CCA- and KCCA-extracted features on the performance in this paper, we use Principal Component Analysis (PCA) (Jolliffe et al., 2016) to derive the first 3 principle components of each feature dataset. Fig. 5 presents the scatter plot of each feature set after being applied PCA. From Fig. 5(a) presenting the dataset Y-HAR, we can observe that it is difficult to see the difference of the two expressions since the sample size is too large (32,844) even we can see the KCCA performs better in Tables 3 and 4 on the dataset. Fig. 5(b) is the scatter plot of dataset Blood, which appears the same for the CCA and KCCA feature selection. This implies that features in dataset Blood may not contain nonlinear correlation. From Fig. 5(c) which presents the dataset Wine, we can see that the results of mRMJR-KCCA may be better since some dots from class 3 are mixed with class 1 in mRMR-CCA.

5. Conclusions

This paper presents a feature selection method, named mRJMR-KCCA, which replaces the correlation measure of the MI in mRMR with the KCCA. Experimental results demonstrate the superior performance of mRMJR-KCCA on the 13 classification associated datasets used in this paper especially on the larger-dimensionality dadasets (such as Y_HAR and Z_HAR in Tables 3 and 4), compared with the other eight benchmark feature selection methods. The mRMJR-KCCA ranks first regarding the times and it is better than the other FS methods with both SVM and RF classification in Tables 3 and 4. From the mRMR to the mRMJR-KCCA, the FS measure changes from the entropy to the KCCA. The mRMR gives an entropy-based score between two variables and utilizes a sum approximation to measure the correlation between a variable and a set of variables. Instead, the KCCA searches for the nonlinear correlation between two sets of variables in mRMJR-KCCA. The mRMJR-KCCA can avoid the sum approximation in mRMR when measuring the joint redundancy between the feature candidate and the already selected features, which somehow considers the complementarity between the already selected features in the view of KCCA. Whilst, both mRMR and mRJMR-KCCA cannot completely remove the dependencies and redundancies among



(a) Dataset Y_HAR

(b) Dataset Blood

(c) Dataset Wine

Fig. 5. Scatter plot of the principle components of the feature sets selected by CCA and KCCA.

features since the two methods rely on a same selection criteria structure as shown in Eqs. (2) and (10). Meanwhile, from the results in Tables 3 and 4, we can also see that Autoencoder performs best on Wine_red and Wpbc with RF and the other FS methods can also yield comparable or similar results on the smaller-dimensionality datasets. The mRMJR-KCCA do not always beat other FS methods; however, it performs much better on the datasets with larger dimensionality since these datasets may contain nonlinear correlations with another set of variables. The results further prove that there is not a “best method” for all tasks. The choice of the best feature set is usually with the aid of FS methods or empirical evaluation of different combinations of features. As previously mentioned, the optimized parameters in SVM (c, gamma) or RF (the number of trees) classification in the paper are achieved by searching in the preset ranges during 10-fold cross validation. The parameters involved in the classification in Tables 3 and 4 and Fig. 3 can refer to the supplemental document. The number of the parameters in Fig. 2 are too big to be included. It is worth mentioning that the parameters shown in the document are not the only ones to yield the corresponding results. This means different parameters or parameter combinations may produce the similar results in classification.

For the future work, we have the following issues remained to be investigated.

(1) The further work can be carried out to discover different kernels in KCCA measurement.
(2) The computational cost in the KCCA-based feature selection methods can be further reduced especially for larger datasets. The further study can consider employing other state-of-art matrix approximation methods to improve efficiency and accuracy.

(3) The performance of the KCCA-based feature selection is affected by the kernel parameters, and other associated CCA-based selection criteria can be explored to apply on larger datasets, such as sparse KCCA, group sparse KCCA, or deep CCA.

Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

CRediT authorship contribution statement

Yan Wang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing - original draft. **Shuang Cang:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Supervision, Validation, Writing - review & editing. **Hongnian Yu:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Validation, Writing - review & editing.

Acknowledgments

This work was supported by the Erasmus Mundus Fusion Project (grant numbers 545831-EM-1-2013-1-IT-ERAMUNDUSEMA21), in part by the European Commission Marie Skłodowska-Curie SMOOTH (Smart Robots for Firefighting) Project under Grant H2020-MSCA-RISE-2016-734875, CHARMED (Characterisation of a green microenvironment and to study its impact upon health and well-being in the elderly as a way forward for health tourism) Project under Grant H2020-MSCA-RISE-2016-734684, and in part by the Royal Society International Exchanges Scheme (Adaptive Learning Control of a Cardiovascular Robot Using Expert Surgeon Techniques) Project under Grant IE151224. The authors would like to thank the volunteers who involved in this work for data collection.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswax.2019.100014.

References

Alickovic, E., Kevric, J., & Subasi, A. (2018). Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction. *Biomedical Signal Processing and Control, 39*, 94–102.

Ashad Alam, M., & Fukumizu, K. (2015). Higher-order regularized kernel canonical correlation analysis. *International Journal of Pattern Recognition and Artificial Intelligence, 29*, 1551005.

Bach, F. R., & Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research, 3*, 1–48.

Bennasar, M., Hicks, Y., & Setchi, R. (2015). Feature selection using joint mutual information maximisation. *Expert Systems with Applications, 42*, 8520–8532.

Bilenko, N. Y., & Gallant, J. L. (2016). Pyrcca: Regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Frontiers in neuroinformatics, 10*, 49.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems, 34*, 483–519.

Brown, G., Pocock, A., Zhao, M.-J., & Luján, M. (2012). Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research, 13*, 27–66.

Chakraborty, S., Chatterjee, S., Dey, N., Ashour, A. S., & Hassanien, A. E. (2017). Comparative approach between singular value decomposition and randomized singular value decomposition-based watermarking. In *Intelligent techniques in signal processing for multimedia security* (pp. 133–149). Springer.

Chernbumroong, S., Cang, S., & Yu, H. (2014). A practical multi-sensor activity recognition system for home-based care. *Decision Support Systems, 66*, 61–70.

Chu, D., Liao, L., Ng, M., & Zhang, X. (2013). Sparse kernel canonical correlation analysis. In *Proceedings of International Multiconference of Engineers and Computer Scientists*.

Dessì, N., & Pes, B. (2015). Similarity of feature selection methods: An empirical study across data intensive classification tasks. *Expert Systems with Applications, 42*, 4632–4642.

Gao, S., Ver Steeg, G., & Galstyan, A. (2016). Variational information maximization for feature selection. In *Advances in Neural Information Processing Systems* (pp. 487–495).

Gheid, Z., & Challal, Y. (2016). Novel efficient and privacy-preserving protocols for sensor-based human activity recognition. In *Ubiquitous intelligence & computing, advanced and trusted computing, scalable computing and communications, cloud and big data computing, internet of people, and smart world congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016 intl IEEE conferences* (pp. 301–308). IEEE.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation, 16*, 2639–2664.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika, 28*, 321–377.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*, 20150202.

Kaya, H., Eyben, F., Salah, A. A., & Schuller, B. (2014). CCA based feature selection with application to continuous depression recognition from acoustic speech features. In *Acoustics, speech and signal processing (ICASSP), 2014 IEEE international conference on* (pp. 3729–3733). IEEE.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., et al. (2017). Feature selection: A data perspective. *ACM Computing Surveys (CSUR), 50*, 94.

Li, M., Bi, W., Kwok, J. T., & Lu, B.-L. (2015). Large-scale Nyström kernel matrix approximation using randomized SVD. *IEEE Transactions on Neural Networks and Learning Systems, 26*, 152–164.

Li, Q., Zhu, D., Zhang, J., Hibar, D.P., Jahanshad, N., Wang, Y. et al. (2017). Large-scale feature selection of risk genetic factors for Alzheimer's disease via distributed group lasso regression. arXiv:1704.08383.

Lisanti, G., Masi, I., & Del Bimbo, A. (2014). Matching people across camera views using kernel canonical correlation analysis. In *Proceedings of the international conference on distributed smart cameras* (p. 10). ACM.

Lopez-Paz, D., Hennig, P., & Schölkopf, B. (2013). The randomized dependence coefficient. *Advances in Neural Information Processing Systems*, 1–9.

Machado, I. P., Gomes, A. L., Gamboa, H., Paixão, V., & Costa, R. M. (2015). Human activity data discovery from triaxial accelerometer sensor: Non-supervised learning sensitivity to feature extraction parametrization. *Information Processing & Management, 51*, 204–214.

Mehrkanoon, S., & Suykens, J. A. (2017). Regularized semipaired kernel CCA for domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems, 29*, 3199–3213.

Montalto, F., Guerra, C., Bianchi, V., De Munari, I., & Ciampolini, P. (2015). MuSA: Wearable multi sensor assistant for human activity recognition and indoor localization. In *Ambient assisted living* (pp. 81–92). Springer.

Ngiam, J., Chen, Z., Bhaskar, S. A., Koh, P. W., & Ng, A. Y. (2011). Sparse filtering. *Advances in Neural Information Processing Systems*, 1125–1133.

Patel, R., Goldstein, T., Dyer, E., Mirhoseini, A., & Baraniuk, R. (2016). Deterministic column sampling for low-rank matrix approximation: Nyström vs. incomplete Cholesky decomposition. In *Proceedings of the 2016 SIAM international conference on data mining* (pp. 594–602). SIAM.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 1226–1238.

Sakar, C. O., Kursun, O., & Gurgen, F. (2012). A feature selection method based on kernel canonical correlation analysis and the minimum redundancy–maximum relevance filter method. *Expert Systems with Applications, 39*, 3432–3437.

Sani, S., Massie, S., Wiratunga, N., & Cooper, K. (2017). Learning deep and shallow features for human activity recognition. In *International conference on knowledge science, engineering and management* (pp. 469–482). Springer.

Suto, J., Oniga, S., & Sitar, P. P. (2016). Feature analysis to human activity recognition. *International Journal of Computers Communications & Control, 12*, 116–130.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications, 37*.

Urbanowicz, R.J., Meeker, M., LaCava, W., Olson, R.S., & Moore, J.H. (2017). Relief-based feature selection: Introduction and review. arXiv:1711.08421.

Wang, L. (2016). Recognition of human activities using continuous autoencoders with wearable sensors. *Sensors, 16*, 189.

Wang, W., & Livescu, K. (2015). Large-scale approximate kernel canonical correlation analysis. arXiv:1511.04773.

Yoshida, K., Yoshimoto, J., & Doya, K. (2017). Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data. *BMC Bioinformatics, 18*, 108.

Zou, Q., Zeng, J., Cao, L., & Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing, 173*, 346–354.