METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Multi-Choice Wavelet Thresholding Based Binary Classification Method

Seung Hyun Baek[a], Alberto Garcia-Diaz[b], Yuanshun Dai[c]

[a] *Hanyang University ERICA Campus, South Korea.* **[b]** *University of Tennessee, Knoxville, USA.* **[c]** *University of Electronics Science and Technology of China, Chengdu, China.*

Methodology, 2020, Vol. 16(2), 127–146, https://doi.org/10.5964/meth.2787

Received: 2018-08-29 • **Accepted:** 2019-09-14 • **Published (VoR):** 2020-06-18

Corresponding Author: Seung Hyun Baek, Division of Business Administration, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan Gyeonggi-do, 15588, South Korea. E-mail: sbaek4@hanyang.ac.kr

Abstract

Data mining is one of the most effective statistical methodologies to investigate a variety of problems in areas including pattern recognition, machine learning, bioinformatics, chemometrics, and statistics. In particular, statistically-sophisticated procedures that emphasize on reliability of results and computational efficiency are required for the analysis of high-dimensional data. Optimization principles can play a significant role in the rationalization and validation of specialized data mining procedures. This paper presents a novel methodology which is Multi-Choice Wavelet Thresholding (MCWT) based three-step methodology consists of three processes: perception (dimension reduction), decision (feature ranking), and cognition (model selection). In these steps three concepts known as wavelet thresholding, support vector machines for classification and information complexity are integrated to evaluate learning models. Three published data sets are used to illustrate the proposed methodology. Additionally, performance comparisons with recent and widely applied methods are shown.

Keywords

data mining, search procedures, optimization, classification analysis, multi-choice wavelet thresholding

Data mining procedures are essentially based on statistical principles and machine learning theory both creatively integrated to effect and facilitate the identification of significant informative patterns for a given database. Recurrent strategies used in data mining include preprocessing, data partitioning, machine learning (modeling), and validation. The ultimate goal of these procedures is the disclosure of unknown and valuable information. Adams et al. (2000) have discussed several models and patterns.

As indicated by Meisel and Mattfeld (2010), operations research and data mining are complementary and supportive due to three facts: (i) operations research techniques

expedite efficiency of data mining; (ii) data mining methodologies enlarge the scope of operations research applications; and (iii) integration of both data mining and operations research boost systems performance. Furthermore, the key element that allows effective fusion of both areas is the use of optimization algorithms (with particular emphasis on search procedures) to find an accurate model and development of metaheuristics. An example of such procedures is the search algorithm by Olafsson et al. (2008) to find the best variable subset.

Classification analysis methods, based on several types of different algorithms, have been proposed to find successful models for complicated data in an extensive range of application domains. The objective of classification analysis is to identify groups of observations based on the input variables which minimize the within group-variability and maximize the between group-variability. Recently, not only classification area but also other supervised or unsupervised learning areas have faced two challenging issues: (i) the curse of dimensionality; and (ii) nonlinearity. Several researchers developed new classification analysis techniques for preventing problems of the curse of dimensionality; spectral regression discriminant analysis (Cai et al., 2008), automatic non-parameter uncorrelated discriminant analysis (Yang et al., 2008), high-dimensional discriminant analysis (Bouveyron et al., 2007), and for avoiding problems of nonlinearity; adaptive nonlinear discriminant analysis (Kim et al., 2006), kernel Fisher discriminant analysis (Mika, 2002), support vector machines for classification and regression (Vapnik, 1995).

Variable selection is an important area of research in machine learning, pattern recognition, statistics, and related fields. The key idea of variable selection is to find input variables which have predictive information and to eliminate non-informative variables. The use of variable selection techniques is motivated by three reasons: (i) to improve discriminant power; (ii) to find fast and cost-effective variables; and (iii) to reach a better understanding of the application process (Guyon & Elisseeff, 2003). In the case of high-dimensional data, variable selection plays a crucial role because of four challenges (Theodoridis & Koutroumbas, 2006): (i) large set of variables; (ii) existence of irrelevant variables; (iii) presence of redundant variables; and (iv) data noise.

This article proposes a novel methodology based on an integration of both the multi-choice wavelet thresholding (MCWT) and a variable selection method for classification to perform three steps known as perception, decision, and cognition. The proposed procedure will be referred to as a *Perception-Decision-Cognition Methodology* (PDCM). The main idea of this methodology is to provide better classification power by integrating both optimal search and data mining procedures. The perception step includes five different dimension reduction methods, based on wavelets, to transform original data into a representation form that exhibits orthogonality and de-noising. Multi-choice wavelet thresholding tries to say that the proposed method integrates these five different dimension reduction methods. But in the computational approach, five versions of the proposed procedure are compare and each version uses one of these dimension

reduction methods. The decision step uses information complexity to find informative variables which can be used to identify groups based on prior modeling information. The cognition step recognizes the best model based on the support vector machines for classification, a well-known kernel-based statistical data mining approach. As the optimal way, new methodologies are usually tested to check capabilities of the procedures with simulated datasets which have different characteristics. But, the proposed PDCM is directly applied to three real datasets in this article. Three numerical experiments were run to compare the PDCM to other often-used procedures. The results from the experiments show that the proposed method outperforms all the procedures used in the experiments.

The section "The Proposed PDCM" reviews relevant procedures integrated to design the proposed methodology. These procedures can be classified into the following areas:

a. Wavelet thresholding-based dimension reduction
b. Variable selection (feature ranking)
c. Cognition accuracy (model selection)

The performance of the methodology is tested using three published data sets and the corresponding results are documented in section "Experimental Results". Section "Discussion" concludes this article. The information of three real benchmark datasets is presented in section "Supplementary Materials".

The Proposed PDCM

Perception-Decision-Cognition Methodology (PDCM)

The proposed Perception-Decision-Cognition Methodology (PDCM) for discriminant analysis is conceptually represented in Figure 1. As indicated in this figure, it consists of three steps:

1. Perceive environmental information.
2. Decide on response (actions).
3. Cognize the accuracy of results to adjust the response.

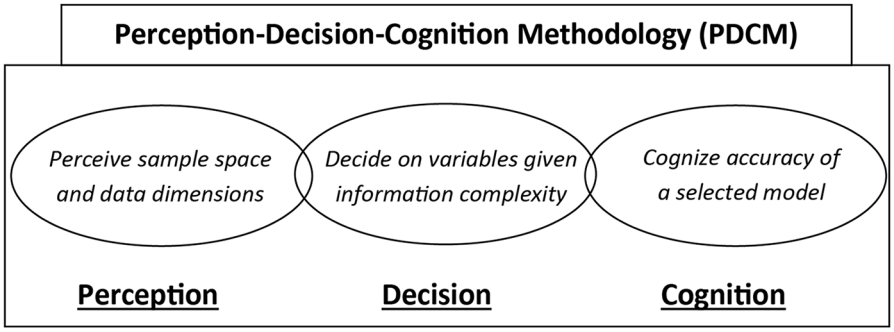
The algorithm used by the PDCM consists of three steps conceptually described below, after assuming that all data have been classified according to three sets: training set, cognition (validation) set, and test set.

Step 1: Perceive Sample Space and Data Dimensions

Let the sample data be $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2,...,\mathbf{x}_q)$,and the corresponding response be $\mathbf{y} = (y_1, y_2,...,y_n)^T$, where q is the dimension of $\mathbf{X}$ and n is the number of samples. Now apply all available MCWT techniques: VisuShrinkUnion, VisuShrinkIntersect, VertiShrink,

Figure 1

Conceptual View of the Perception-Decision-Cognition Methodology (PDCM)



VET, and MSVET. For each dimension reduction technique, generate a new training set $\mathbf{X} = (\mathbf{d}_1, \mathbf{d}_2,...,\mathbf{d}_p)$, where the *reduced* dimension p is the number of coefficients *perceived* by the reduction techniques ($p \leq q$).

Step 2: Decide on Variables Given Information Complexity

The procedure can be described as follows. Remove each of the p variables one at a time, and evaluate the corresponding information complexity measure, $ICOMP_{PERF}$. Once the p - 1 removal procedures are completed, the removed variable resulting in minimum value of $ICOMP_{PERF}$ is identified and assigned the lowest rank (i.e. p). This procedure is repeated for the remaining p - 1 variables for which there is no rank yet. As a result of this, a variable receives rank equal to p - 1. This procedure is repeated until the p variables have been arranged according to their ranks.

Step 3: Cognize Accuracy of Selected Models

Compute the accuracy value of each cognition data set using the SVMC for all possible subsets of the ranked variables selected in Step 2. Specifically, first consider the variable with the highest rank (i.e, rank = 1), and calculate the cognition accuracy value. After this, the two variables with rank = 1 and rank = 2 are considered, and a new cognition accuracy value is calculated. This procedure is repeated until all ranked variables are considered. Finally, the subset of variables resulting in the highest accuracy value is chosen as the best model. The reason of using the two steps for transforming the nonlinear input data with wavelets and for finding informative variables with $ICOMP_{PERF}$ - *RFE* is to find more accurate and faster models.

Step 1: Wavelet Thresholding-Based Dimension Reduction Techniques

Dimension reduction is a preferred strategy in the area of machine learning. As anticipated, there are several approaches to perform dimensional reduction. The following methods are among the most popular: principal component analysis (Jolliffe, 2002), rotational linear discriminant analysis technique (Sharma & Paliwal, 2008), independent component analysis (Stone, 2004), semi-definite embedding (Weinberger & Saul, 2006), multifactor dimensionality reduction (Ritchie & Motsinger, 2005), factor analysis (Basilevsky, 1994), and wavelet-based dimension reduction (Chang & Vidakovic, 2002; Cho et al., 2009; Donoho & Johnstone, 1994; Jung et al., 2006).

The dimension reduction strategy has important benefits that can be measured not only in terms of computational time savings, but also in accuracy improvement. In the novel PDCM, the wavelet-based dimension reduction is applied in Step 1. The wavelets approach was selected because of several attractive attributes, among which the following two are most relevant: (a) wavelets adapt effectively to spatial features of a function such as discontinuities and varying frequency behavior; (b) wavelets have efficient $O(n)$ algorithms to do transformations (Mallot, 1999). Based on perceived knowledge, wavelet-based techniques are applied to obtain a well-fitted reduced-dimension representation of the original data.

Discrete Wavelet Transformation (DWT) is often used for dimension reduction (also known as shrinkage or threshold). The data constructed with the scaling and wavelet functions based on orthogonal base in time domain is as follows:

$$f(t) = \sum_k c_{L,k}\phi_{L,k}(t) + \sum_{j \geq L}\sum_k d_{j,k}\psi_{j,k}(t)$$

$$\phi_{L,k}(t) = 2^{L/2}\phi(2^L t - k), \qquad L, k \in \mathbf{Z}$$

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \qquad j \geq L \text{ and } j, k \in \mathbf{Z}$$

where $\mathbf{Z}$ is the set of all the possible integer values, $c_{L,k}$ is the coarse level coefficient and $d_{L,k}$ is the finer level coefficient. Let $\mathbf{y}_m = [y_{m1}, y_{m2}, \cdots, y_{mN}]^T$ is an m^{th} observed sample. For a single sample, the DWT procedure uses the orthonormal matrix $\mathbf{W}$ of dimension $N \times N$ to find the wavelet coefficient

$$\mathbf{d} = (\mathbf{c}_L, \mathbf{d}_L, \mathbf{d}_{L+1}, \cdots, \mathbf{d}_J)$$

where $\mathcal{J} > L$, L corresponds to the lowest decomposition level.

$$\mathbf{c}_L = (c_{L,0},\ldots,c_{L,2^L-1}), \mathbf{d}_L = (d_{L,0},\ldots,d_{L,2^L-1}),\ldots,\mathbf{d}_J = (d_{J,0},\ldots,d_{J,2^J-1}),$$

through the transformation

$$\mathbf{d} = \mathbf{W}\mathbf{y}.$$

For multiple samples, let vector $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2,...,\mathbf{y}_M]$ be the data set with M observed samples. The wavelet coefficient vector is obtained from the transformation

$$\mathbf{D} = \mathbf{W}\mathbf{Y}$$

where $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2,...,\mathbf{d}_M]$, and $\mathbf{d}_m = (\mathbf{c}_{mL}, \mathbf{d}_{mL}, \mathbf{d}_{mL+1}, \cdots, \mathbf{d}_{mJ})$, $m = 1,2,...,M$.

Small absolute values of wavelet coefficients are undesirable since they may be influenced more by noise than by information. On the other hand, large absolute values are more influenced by information than noise. This observation motivates the development of threshold methods. There are two threshold rules usually referred to as *soft* and *hard* thresholds. The soft rule is a continuous function of the data that shrinks each observation, while the hard rule retains unchanged only large observations (Donoho & Johnstone, 1994). The hard and soft threshold methods are defined as following:

$$D(U,\lambda) = sign(U)\max(0,|U| - \lambda) \qquad \text{(Soft)}$$

$$D(U,\lambda) = \begin{cases} U, & \text{all } |U| > \lambda \\ 0, & \text{otherwise} \end{cases} \qquad \text{(Hard)}$$

here λ is the threshold value. The threshold method can be used not only for data reduction but also for de-noising. One of the 5 different thresholding methods will be selected based on the performance of PDCM (*Multi-Choice*). Also, there are 5 different wavelet-based thresholding methods which will handle high-dimensional data. Therefore, the dimension reduction method is referred as *Multi-Choice Wavelet Thresholding* (MCWT).

VisuShrink (VS)

VisuShrink is a soft thresholding technique that applies a universal threshold proposed by Donoho and Johnstone (1994). The VisuShrink threshold is given by $\sigma\sqrt{2\log N}$, where N is the number of wavelet coefficients, and σ is the standard deviation of the wavelet coefficients (or noise standard deviation). When ε_i is a white noise sequence, independent and identically distributed as $N(0,1)$, then as $N \to \infty$, $P\{\max|\varepsilon_i| > \sqrt{2\log N}\} \to 0$. That is, the maximum of the N values will most likely be smaller than the universal threshold. The VisuShrink guarantees a noise free reconstruction. However, when setting the threshold large, the degree of data fitting may be unsatisfactory. For multiple curves or samples, the VS procedure uses the union (VisuShrinkUnion, VSU) or intersection (VisuShrinkIntersection, VSI) of data sets in the selection of wavelet coefficients (Jung et al., 2006).

VertiShrink (VERTI)

Chang and Vidakovic (2002) developed a Stein-type shrinkage method, known as VertiShrink, to maximize the predictive density under appropriate model assumptions regarding wavelet coefficients. The main goal of VertShrink is the estimation of the baseline curve by using the average of block vertical coefficients. The estimated wavelet coefficients are given by:

$$\widehat{\boldsymbol{\theta}} = \left(1 - \frac{M\sigma^2}{\mathbf{d}^T\mathbf{d}}\right) + \mathbf{d}$$

where, $\mathbf{d}$ is the wavelet coefficient, $\mathbf{d} = (\mathbf{c}_L, \mathbf{d}_L, \mathbf{d}_{L+1}, \cdots, \mathbf{d}_J)$, M is the number of curves and σ is the standard deviation of the wavelet coefficients.

Vertical-Energy-Thresholding (VET)

VET was proposed by Jung et al. (2006). The procedure is based on the concept of *energy of a function* with some smoothness, since it is often concentrated on few coefficients, while the *energy of noise* is still spread over all coefficients in the wavelet domain. The *vertical* energy of wavelet coefficients is defined by

$$\left|\left|\mathbf{d}_{vj}\right|\right|^2 = d_{1j}^2 + d_{2j}^2 + \ldots + d_{Mj}^2$$

where d_{mj} is the wavelet coefficient at the j^{th} wavelet position for the m^{th} data curve, $m = 1, 2, \ldots, M$.

The VET method minimizes the overall relative reconstruction error (*ORRE*), formulated below, to determine a threshold value, namely λ:

$$ORRE(\lambda) = \frac{\sum_{j=1}^{N} E\left[\left|\left|\mathbf{d}_{vj}(1 - I(\left|\left|\mathbf{d}_{vj}\right|\right|^2 > \lambda))\right|\right|^2\right]}{\sum_{j=1}^{N} E\left[\left|\left|\mathbf{d}_{vj}\right|\right|^2\right]} + \frac{\sum_{j=1}^{N} E\left[\left|\left|I(\left|\left|\mathbf{d}_{vj}\right|\right|^2 > \lambda)\right|\right|^2\right]}{N}$$

MultiScale-Vertical-Energy-Thresholding (MSVET)

Since the VET procedure does not consider the scale information of wavelets, an improved procedure proposed by Cho et al. (2009) and known as multi-scale vertical energy thresholding (MSVET) obtains a different optimal thresholding value for each scale by extending the idea of the VET procedure. In the MSVET procedure, the multi-scale overall relative reconstruction error (*MSORRE*) is defined as follows to determine the threshold values, λ_i:

$$MSORRE(\lambda_L, \lambda_{L+1},...,\lambda_J) = \frac{\sum_{i=L}^{J}\sum_{j=1}^{i+j-2L} E\left[\left|\left|\mathbf{d}_{vji}(1 - I(\left|\left|\mathbf{d}_{vji}\right|\right|^2 > \lambda_i))\right|\right|^2\right]}{\sum_{j=1}^{N} E\left[\left|\left|\mathbf{d}_{vji}\right|\right|^2\right]} + \frac{\sum_{i=L}^{J}\sum_{j=1}^{i+j-2L} E\left[I(\left|\left|\mathbf{d}_{vji}\right|\right|^2 > \lambda_i))\right|\right|^2\right]}{N - 2^{J-L}}$$

where, $\mathbf{d}_{vji} = (d_{1ji}, d_{2ji},...,d_{Mji})$, $\left|\left|\mathbf{d}_{vji}\right|\right|^2 = d_{1ji}^2 + d_{2ji}^2 + ... + d_{Mji}^2$; d_{mji} represents the wavelet coefficient at the j^{th} wavelet position of the i^{th} scale for the m^{th} curve, $m = 1, 2, ..., M$.

Step 2: Variable Selection Based on Information Complexity and Recursive Feature Elimination

Once the reduced sample space is determined in Step 1, the decision regarding which of the remaining variables should be selected for ranking is made on the basis of minimal information complexity values, following the Information Complexity Performance Testing with Recursive Feature Elimination ($ICOMP_{PERF}$-RFE) procedure proposed by Bozdogan and Baek (2018). Since this procedure resulted in better performance than other RFE-based methods. Also, this procedure essentially generates a stabilized and smoothed covariance estimator to calculate the information complexity measure, and, finally performs ranking using recursive elimination on the remaining variables.

The development of information complexity for the discriminant analysis is evaluated using the modified maximal entropic complexity C_{1F}

$$C_{1F}(\widehat{\Sigma}) = \frac{1}{4\overline{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \overline{\lambda}_a^2),$$

where s is the rank of $\widehat{\Sigma}$, λ_j is the j^{th} eigenvalue of $\widehat{\Sigma} > 0$, $j = 1, 2, \cdots, s$ and $\overline{\lambda}_a$ is arithmetic means of the eigenvalues

$ICOMP_{PERF}$ can be evaluated as indicated below:

$$ICOMP_{PERF} = n\log 2\pi + n\log(\widehat{\sigma}^2) + n + 2C_{1F}(\widehat{\Sigma}_{STA\_CSE})$$

where lack of fit is assessed by means of the first three terms and complexity by the fourth one. In the above expression, $\widehat{\sigma}^2$ is the estimated mean squared error given by

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2,$$

and $\widehat{\Sigma}_{STA\_CSE}$ is the stabilized and smoothed *convex sum covariance matrix estimator* given by

$$\widehat{\Sigma}_{STA\_CSE} = \frac{n}{n+k}\widehat{\Sigma}_{STA} + \left(1 - \frac{n}{n+k}\right)\left[\frac{trace(\widehat{\Sigma}_{STA})}{h}\right]\mathbf{I}_h,$$

where $\widehat{\Sigma}_{STA}$ is the stabilized covariance matrix proposed by Thomaz (2004), h is the number of variables, $\mathbf{I}_h$ is $h \times h$ identity matrix, and k is chosen such that

$$0 < k < \frac{2[h(1+\beta)-2]}{h-\beta},$$

and

$$\beta = \frac{\left(tr\widehat{\Sigma}_{\mathbf{W}}\right)^2}{tr(\widehat{\Sigma}_{\mathbf{W}}^2)}.$$

Specific details on this procedure are provided by Bozdogan and Baek (2018).

Step 3: Cognition Accuracy of Selected Models

When the ranking decision is finished in Step 2, the corresponding accuracies are determined using the corresponding cognition sets and the support vector machines for classification (SVMC) described below. Once the accuracies are calculated for the selected models the most-accurate one is chosen.

The SVMC find an optimal separating hyperplane that maximizes the margin between the classes (Vapnik, 1995). Consider the case of classifying a set of linearly separating data into two groups. Assume a set of training data is given by $[(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2),...,(\mathbf{x}_n, y_n)]$ where $\mathbf{x}_i \in \Re^p$ is an input vector, $y_i \in \{-1,1\}$ is a binary class index, and n is the size of the training data set. Then, a decision boundary that partitions the underlying vector space into two classes can be represented by the hyperplane

$$\mathbf{w}^T\mathbf{x} + b = 0$$

where $\mathbf{w}$ is the weight vector and b is the bias. The objective of SVMC is to find a maximum margin decision boundary between two parallel hyperplanes, $\mathbf{w}^T\mathbf{x} + b = 1$ and $\mathbf{w}^T\mathbf{x} + b = -1$. The dual model of the corresponding primal model can be formulated as

$$\max\ l(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i^T, \mathbf{x}_j)$$

subject to

$$\sum_{i=1}^{n} \alpha_i \mathbf{x}_i = 0,\ 0 \le \alpha_i \le C,\ i = 1,2,...,n$$

where $K(\mathbf{x}_i^T, \mathbf{x}_j)$ is the kernel function and C is a predefined coefficient. Kernel functions used in the numerical experiments are described in Table 1.

Table 1

Kernel Functions

Function	$K(\mathbf{x}_i^T, \mathbf{x}_j)$	Parameters
Gaussian	$\exp\left[-\left(\frac{1}{a^b}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right)^c\right]$	$a = 2,\ b = c = 1$
Cauchy	$\left(1 + \frac{1}{a}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right)^{-1}$	$a = 1$
Inverse Multi-Quadratic	$\left(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 + a^2\right)^{-1/2}$	$a = 1$

The point $\mathbf{x}^o$ with coordinates corresponding to new data can be classified as indicated below:

Class 1: $\sum_{i=1}^{n} \alpha_i^{ov} y_i K(\mathbf{x}_i^T, \mathbf{x}^o) + b^{ov} < 0$

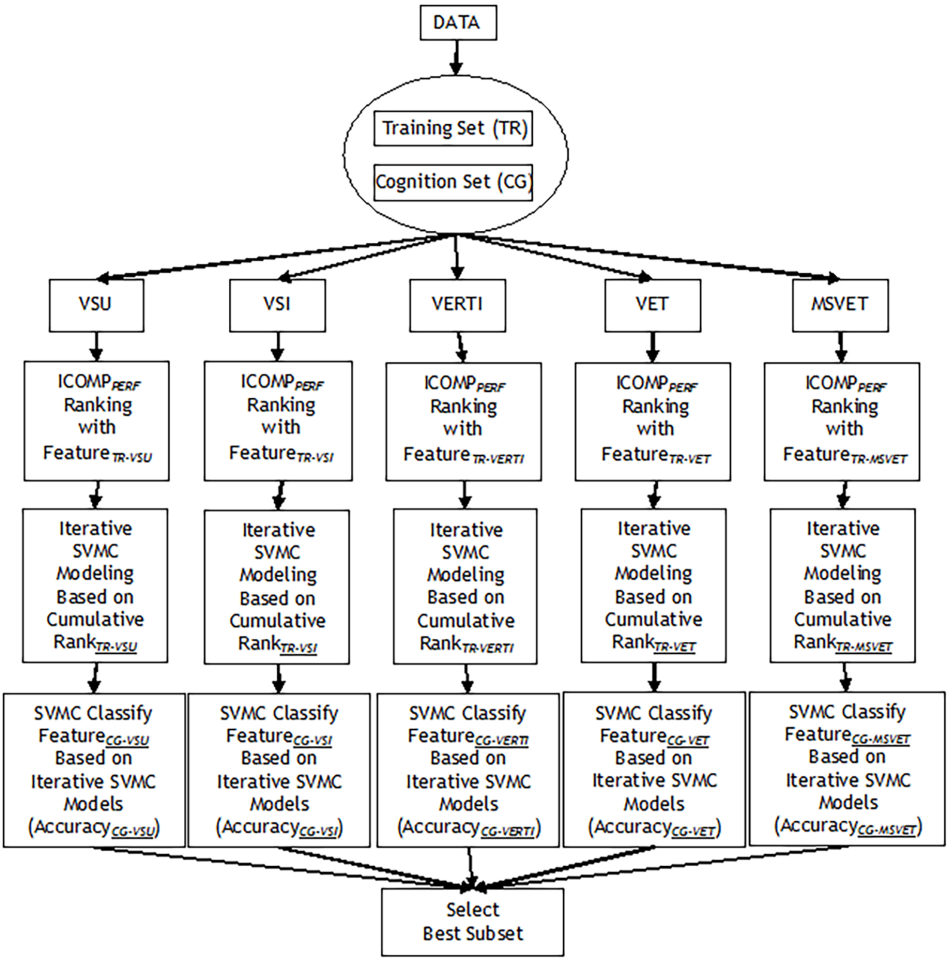
and

Class 2: $\sum_{i=1}^{n} \alpha_i^{ov} y_i K(\mathbf{x}_i^T, \mathbf{x}^o) + b^{ov} > 0$

where α^{ov} and b^{ov} are optimal values found based on the training data. A classification example based on the PDCM is illustrated in Figure 2.

Figure 2

Classification Example Based on PDCM with SVMC



Experimental Results

In order to emphasize the effectiveness of the PDCM, it will be applied to three different data sets (Heart, Fat, & Handwritten) used for experiments with low-dimension high sample size, and high-dimension low sample size. The data sets are divided as 3 sets; training (50% of total set), cognition (10% of total set), and test (40% of total set) sets. For the wavelet transformation of the three data sets, the linear padding suggested by

(Strang & Nguyen, 1997) is applied. Also, well-known MATLAB software was used to get the experimental results. MATLAB software gives several advantages such as easy calculation for data matrix, vector operation, easy plotting, and function operations with MATLAB toolbox. This article documents the comparison of the PDCM to the following procedures:

a. *SVMC recursive feature elimination (SVMC-RFE)* (Guyon et al., 2002): The SVMC-RFE is used weight as a criterion to rank each feature using support vector machines for classification and recursive feature elimination algorithm.
b. *Two-stage method* (Cho et al., 2009): The Two-stage method is used multi-scale vertical energy thresholding (MSVET) to reduced dimension and applied support vector machines for classification and recursive feature elimination to select important wavelet coefficients based on gradient.

Heart Data (44 Variables)

The data set includes 267 samples and 44 variables on cardiac single proton emission computed tomography (SPECT) images with two categories, i.e., normal and abnormal. There are 55 normal and 212 abnormal classes (Cios & Kurgan, 2001). The data set is divided into 134 samples as a training set, 13 samples as a cognition set, and 120 samples as a test set. The training set has 25 normal and 109 abnormal classes. The cognition set has 1 normal and 12 abnormal classes. The test set has 29 normal and 91 abnormal classes[1]. Table 2 and Table 3 show comparison results in terms of the variables selected from $ICOMP_{PERF}$-*RFE*, the cognition accuracy, and the test accuracy. Cauchy and inverse multi-quadratic kernel functions are used in Table 2 and Table 3, respectively.

Table 2

PDCM Versus Various Ranking Based Method Using Cauchy Kernel

Method	Selected Variables from $ICOMP_{PERF}$-*RFE*	Number of Variables	Cognition Accuracy	Test Accuracy
PDCM (MSVET)	20, 28, 30	3	100%	92%
PDCM (VET)	20, 28, 50	3	100%	92%
PDCM (VERTI)	19, 34	2	100%	92%
PDCM (VISU UNION)	9, 20, 44, 53	4	100%	92%
PDCM (VISU INTERSECT)	8, 15	2	100%	92%
SVMC-RFE	6, 16, 17, 18, 26, 32, 35	7	92%	76%
Two-Stage	43	1	92%	78%

1) The data is available at http://archive.ics.uci.edu/ml/datasets/SPECT+Heart?ref=datanews.io.

Table 3

PDCM Versus Various Ranking Based Method Using Inverse Multi-Quadratic Kernel

Method	Selected Variables from $ICOMP_{PERF}$-RFE	Number of Variables	Cognition Accuracy	Test Accuracy
PDCM (MSVET)	6, 10, 18, 29	4	92%	76%
PDCM (VET)	6, 10, 18, 29	4	92%	76%
PDCM (VERTI)	6, 10, 18, 29	4	92%	76%
PDCM (VISU UNION)	9, 10, 16, 29	4	92%	76%
PDCM (VISU INTERSECT)	10, 19, 28	3	92%	78%
SVMC-RFE	22	1	92%	76%
Two-Stage	43	1	92%	78%

As observed in Table 2, PDCM achieves better cognition accuracies comparing to SVMC-RFE and two-stage and yields more accurate results in test. Also, as shown in Table 3, PDCM (visu intersect) and two-stage both reach the highest test accuracy, although two-stage requires fewer variables.

Near Infrared Spectroscopy Data (100 Variables)

These data were collected by a Tecator infratec food and feed analyzer to predict the fat content of a meat sample based on near infrared (NIR) spectroscopy. The data set was divided into two classes defined on the basis of fat content; one class (low-fat) corresponded to 20% or less, and another class (high-fat) to more than this level (Rossi & Villa, 2006). There are 77 low-fat and 138 high-fat classes. The entire data set consists of 215 samples with measured values for each of 100 predictive variables (wavelengths). These samples are divided randomly to configure a training set consisting of 108 samples; a cognition set consisting of 11 samples for cognition; and a test set consisting of the remaining 96 samples. The training set has 39 low-fat and 69 high-fat classes. The cognition set has 4 low-fat and 7 high-fat classes. The test set has 34 low-fat and 62 high-fat classes[2]. Table 4 and Table 5 show comparison results in terms of the variables selected from $ICOMP_{PERF}$-RFE, the cognition accuracy, and the test accuracy. Cauchy and Gaussian kernel functions are used in Table 4 and Table 5, respectively. Although both PDCM and two-stage reach the 91% accuracy level for the cognition set in Table 4, the test accuracy of PDCM is higher than that of SVMC-RFE and two-stage except PDCM (verti). Additionally, Table 5 shows that the cognition accuracy of PDCM is 91% and the test accuracy is higher than that of other methods. Furthermore, PDCM uses only less than 9 variables to reach the accuracy levels previously mentioned.

2) The data is available at http://lib.stat.cmu.edu/datasets/tecator.

Table 4

PDCM Versus Various Ranking Based Method Using Cauchy Kernel

Method	Selected Variables from $ICOMP_{PERF}$-RFE	Number of Variables	Cognition Accuracy	Test Accuracy
PDCM (MSVET)	2, 5	2	91%	90%
PDCM (VET)	5, 32	2	91%	90%
PDCM (VERTI)	6, 32	2	91%	84%
PDCM (VISU UNION)	1, 5	2	91%	90%
PDCM (VISU INTERSECT)	4, 28	2	91%	91%
SVMC-RFE	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100	96	82%	84%
Two-Stage	33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45	13	82%	86%

Table 5

PDCM Versus Various Ranking Based Method Using Gaussian Kernel

Method	Selected Variables from $ICOMP_{PERF}$-RFE	Number of Variables	Cognition Accuracy	Test Accuracy
PDCM (MSVET)	4, 5, 14, 17, 18, 29	6	91%	92%
PDCM (VET)	4, 5, 10, 16, 22, 25, 32	7	91%	91%
PDCM (VERTI)	1, 5, 8, 12, 14, 20, 22, 25, 27	9	91%	86%
PDCM (VISU UNION)	1, 5, 6, 13, 17, 19, 24, 25, 30	9	91%	91%
PDCM (VISU INTERSECT)	5, 9, 10, 16, 19, 28	6	91%	88%
SVMC-RFE	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88,	92	73%	85%

Method	Selected Variables from $ICOMP_{PERF}$-RFE	Number of Variables	Cognition Accuracy	Test Accuracy
	89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100			
Two-Stage	40, 43, 45	3	73%	80%

Handwritten Data (240 Variables)

This data set has features of handwritten numerals from 0 to 9 extracted from a collection of Dutch utility maps. The entire set consists of 200 samples digitized in binary images per numerals and six different variable sets: 76 fourier coefficients, 216 profile correlations, 64 Karhunen-Love coefficients, 240 pixel averages, 47 Zernike moments and 6 morphological features (Van Breukelen et al., 1998). Two of six variable sets (216 profile correlations and 240 pixel averages) are used for the experiment. Two (0 and 9) out of 10 numerals are selected to verify the proposed method. Each numeral has 100 samples which are included in the experimental data set. For the experiment, 100 samples are used as a training set, 10 samples are used as a cognition set and 90 samples are used as a test set. The training set has 45 zero-numeral and 55 nine-numeral classes. The cognition set has 4 zero-numeral and 6 nine-numeral classes. The test set has 51 zero-numeral and 39 nine-numeral classes[3]. Table 6 and Table 7 show comparison results in terms of the selected variables from $ICOMP_{PERF}$-RFE, the cognition accuracy, and the test accuracy.

Table 6

PDCM Versus Various Ranking Based Method Using Cauchy Kernel

Method (Wavelet-Based Method)	Selected Variables from $ICOMP_{PERF}$-RFE	Number of Variables	Cognition Accuracy	Test Accuracy
PDCM (MSVET)	112	1	100%	88%
PDCM (VET)	25	1	100%	93%
PDCM (VERTI)	25	1	100%	93%
PDCM (VISU UNION)	25	1	100%	93%
PDCM (VISU INTERSECT)	25	1	100%	93%
SVMC-RFE	324, 334, 339, 340, 344, 349, 456	7	100%	89%
Two-Stage	8, 10, 11, 15	4	80%	80%

Cauchy and inverse multi-quadratic kernel functions are used in Table 6 and Table 7, respectively. As shown in the tables, PDCM reaches 100% cognition accuracy levels as did

3) The data is available at https://archive.ics.uci.edu/ml/datasets/Multiple+Features.

the SVMC-RFE, except two-stage. Furthermore, PDCM achieves a higher accuracy level than the other methods for the test set, except PDCM (msvet) in Cauchy kernel.

Table 7

PDCM Versus Various Ranking Based Method Using Inverse Multi-Quadratic Kernel

Method	Selected Variables from $ICOMP_{PERF}$-RFE	Number of Variables	Cognition Accuracy	Test Accuracy
PDCM (MSVET)	112, 155	2	100%	97%
PDCM (VET)	25	1	100%	100%
PDCM (VERTI)	25	1	100%	100%
PDCM (VISU UNION)	87, 345	2	100%	98%
PDCM (VISU INTERSECT)	24	1	100%	100%
SVMC-RFE	48, 82, 106, 154, 166, 217, 218, 223, 224, 225, 226, 230, 231, 232, 235, 236, 237, 238, 239, 240, 241, 246, 249, 250, 251, 257, 258, 261, 262, 264, 265, 273, 276, 277, 279, 280, 283, 284, 288, 289, 291, 294, 295, 298, 299, 300, 303, 304, 309, 310, 312, 318, 319, 325, 333, 334, 335, 337, 342, 348, 349, 350, 352, 358, 363, 364, 365, 367, 373, 374, 376, 378, 379, 380, 382, 387, 389, 391, 393, 394, 395, 397, 402, 404, 408, 410, 412, 422, 423, 424, 427, 428, 435, 436, 437, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 454, 455, 456	109	100%	92%
Two-Stage	8, 10, 11, 15	4	70%	80%

Discussion

This article documents the development and application of a novel Perception-Decision-Cognition Methodology (PDCM) for classification analysis based on the MCWT and SVMC with $ICOMP_{PERF}$-RFE. Five different wavelet-based dimension reduction techniques called MCWT are applied in the perception step. It is shown that the procedure yields a good representation of the original data, using only reduced variables. The decision step is performed using a rank-based variable selection approach, using the information complexity criterion. The information complexity based variable selection approach shows a good ability to achieve reasonable variable ranks, which in turn can affect decision making. In the cognition step, the number of variables and accuracy are cognized for further discrimination.

The PDCM is directly applied to three real datasets instead of using simulated datasets having different characteristics in this article. As supported by the numerical experiments documented in this article, the PDCM outperforms the currently available data mining approaches, and, furthermore, shows to be applicable to various areas, such as bioinformatics, chemometrics, pattern recognition, and other data mining fields. The PDCM has three advantages:

1. Dimension simplification.
2. Multiple model choices based on simplified dimension.
3. Novel rank based variable selection: $ICOMP_{PERF}$-*RFE*.

Funding: The authors have no funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Acknowledgments: The authors have no support to report.

Author note: This paper was modified from the Ph.D. dissertation (Baek, 2010) of the first author.

Data Availability: Datasets used for this article are freely available (see Supplementary Materials section).

Supplementary Materials

1. Dataset on cardiac Single Proton Emission Computed Tomography (SPECT) images.
2. Dataset on near infrared absorbance spectrum of fat content in meats.
3. Dataset on features of handwritten numerals (`0'--`9') extracted from a collection of Dutch utility maps.

Index of Supplementary Materials

Cios, K. J., & Kurgan, L. A. (2001). *SPECT heart data set.* UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/SPECT+Heart?ref=datanews.io

Thodberg, H. H. (1995). *Tecator data set.* StatLib - Datasets Archive. http://lib.stat.cmu.edu/datasets/tecator

Duin, R. P. W. (n.d.). *Multiple features data set.* UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Multiple+Features

References

Adams, N., Blunt, G., Hand, D., & Kelly, M. (2000). Data mining for fun and profit. *Statistical Science, 15*(2), 111-131. https://doi.org/10.1214/ss/1009212753

Baek, S. H. (2010). *Kernel-based data mining approach with variable selection for nonlinear high-dimensional data* [Unpublished doctoral dissertation]. University of Tennessee, Knoxville, USA.

Basilevsky, A. (1994). *Statistical factor analysis and related methods: Theory and application*. New York, NY, USA: Wiley.

Bouveyron, C., Girard, S., & Schmid, C. (2007). High-dimensional discriminant analysis. *Communications in Statistics. Theory and Methods, 36*, 2607-2623. https://doi.org/10.1080/03610920701271095

Bozdogan, H., & Baek, S. H. (2018). Hybridized support vector machine and recursive feature elimination with information complexity. *Statistics, Optimization & Information Computing, 6*(2), 159-177. https://doi.org/10.19139/soic.v6i2.327

Cai, D., He, X., & Han, J. (2008). An efficient algorithm for large-scale discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering, 20*(1), 1-12. https://doi.org/10.1109/TKDE.2007.190669

Chang, W., & Vidakovic, B. (2002). Wavelet estimation a baseline signal from repeated noisy measurements by vertical block shrinkage. *Computational Statistics & Data Analysis, 40*, 317-328. https://doi.org/10.1016/S0167-9473(02)00053-1

Cho, H., Baek, S. H., Youn, E., Jeong, M. K., & Taylor, A. (2009). A two-stage classification procedure for near-infrared spectra based on multi-scale vertical energy wavelet thresholding and SVM-based gradient-recursive feature elimination. *The Journal of the Operational Research Society, 60*, 1107-1115. https://doi.org/10.1057/jors.2008.179

Cios, K. J., & Kurgan, L. (2001). Hybrid inductive machine learning: An overview of CLIP algorithms. In L. C. Jain & J. Kacprzyk (Eds.), *New learning paradigms in soft computing* (pp. 276-321). Heidelberg, Germany: Physica-Verlag.

Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika, 81*, 425-455. https://doi.org/10.1093/biomet/81.3.425

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(1), 1157-1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1/3), 389-422. https://doi.org/10.1023/A:1012487302797

Jolliffe, I. T. (2002). *Principal component analysis*. New York, NY, USA: Springer.

Jung, U., Jeong, M. K., & Lu, J. C. (2006). A vertical-energy-thresholding procedure for data reduction with multiple complex curves. *IEEE Transactions on Cybernetics, 36*(5), 1128-1138. https://doi.org/10.1109/TSMCB.2006.874681

Kim, H., Drake, B. L., & Park, H. (2006). Adaptive nonlinear discriminant analysis by regularized minimum squared errors. *IEEE Transactions on Knowledge and Data Engineering, 18*(5), 603-612. https://doi.org/10.1109/TKDE.2006.72

Mallot, S. (1999). *A wavelet tour of signal processing*. San Diego, CA, USA: Academic Press.

Meisel, S., & Mattfeld, D. C. (2010). Synergies of operations research and data mining. *European Journal of Operational Research, 206*(1), 1-10. https://doi.org/10.1016/j.ejor.2009.10.017

Mika, S. (2002). *Kernel fisher discriminants* [Unpublished doctoral dissertation]. Technical University of Berlin, Berlin, Germany.

Olafsson, S., Li, X., & Wu, S. (2008). Operations research and data mining. *European Journal of Operational Research, 187*(3), 1429-1448. https://doi.org/10.1016/j.ejor.2006.09.023

Ritchie, M. D., & Motsinger, A. A. (2005). Multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions in pharmacogenomics studies. *Pharmacogenomics, 6*(8), 823-834. https://doi.org/10.2217/14622416.6.8.823

Rossi, F., & Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing, 69*(7-9), 730-742. https://doi.org/10.1016/j.neucom.2005.12.010

Sharma, A., & Paliwal, K. K. (2008). Rotational linear discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering, 20*(10), 1336-1347. https://doi.org/10.1109/TKDE.2008.101

Stone, J. (2004). *Independent component analysis: A tutorial introduction*. London, United Kingdom: MIT press.

Strang, G., & Nguyen, T. (1997). *Wavelets and filter banks*. Wellesley, MA, USA: Wellesley-Cambridge Press.

Theodoridis, S., & Koutroumbas, K. (2006). *Pattern recognition*. San Diego, CA, USA: Academic Press.

Thomaz, C. (2004). *Maximum entropy covariance estimate for statistical pattern recognition* [Unpublished doctoral dissertation]. University of London, London, United Kingdom.

Van Breukelen, M., Duin, R. P. W., Tax, D. M. J., & Den Hartog, J. E. (1998). Handwritten digit recognition by combined classifiers. *Kybernetika, 34*(4), 381-386.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag.

Weinberger, K. Q., & Saul, L. K. (2006). Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision, 70*(1), 77-90. https://doi.org/10.1007/s11263-005-4939-z

Yang, W. H., Dai, D. Q., & Yan, H. (2008). Feature extraction and uncorrelated discriminant analysis for high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering, 20*(5), 601-614. https://doi.org/10.1109/TKDE.2007.190720

Methodology
2020, Vol.16(2), 127–146
https://doi.org/10.5964/meth.2787