# High Dimensional Data Mining Systems by Kernel Orthonormalized Partial Least Square Analysis

Shian-Chang Huang, Nan-Yu Wang, and Tung-Kuang Wu

***Abstract*—Mining high-dimensional business data is a popular and important problem. However, there are two challenges for mining such data, including 1) the curse of dimensionality and 2) the meaningfulness of the similarity measure in the high dimension space. This paper proposes a novel approach to overcome the problems, which builds a generalized multiple kernel machine (GMKM) on a special subspace created by the kernel orthonormalized partial least square (KOPLS). GMKM takes products of kernels-corresponding to a tensor product of feature spaces. This leads to a richer and much higher dimensional feature representation. Therefore, GMKM is powerful in identifying relevant features and their apposite kernel representation. KOPLS finds a low dimensional representation of data, which uncovers the hidden information and simultaneously respects the intrinsic geometry of data manifold. Our new system robustly overcomes the weakness of traditional multiple kernel machines, and outperforms traditional classification systems.**

***Index Terms*—Data mining, multiple kernel learning, dimensionality reduction, support vector machine.**

## I. Introduction

Mining high-dimensional data is a great challenge for most existing data mining algorithms. There are two common challenges for analyzing high-dimensional data (Wang and Yang [1]). The first one is the curse of dimensionality. The complexity of many existing data mining algorithms is exponential with respect to the number of dimensions. With increasing dimensionality, these algorithms soon become computationally intractable and therefore inapplicable in many real applications. Secondly, the specificity of similarities between points in a high dimensional space diminishes. For any point in a high dimensional space, the expected gap between the Euclidean distance to the closest neighbor and that to the farthest point shrinks as the dimensionality grows. This phenomenon may render many data mining tasks (e.g., clustering) ineffective and fragile because the model becomes vulnerable to the presence of noise. The objective of this paper is to overcome the above problems by a novel embedding, which fully respects the data manifold and maps data to a low-dimensional subspace for further handling.

Reviewing recent literature, many advanced approaches from data mining or artificial intelligence were developed to solve the problems as mentioned above. These methods (Witten and Frank [2]) include inductive learning, case-based reasoning, neural networks, rough set theory (Ahn *et al*. [3]), and support vector machines (SVM) (Wu *et al*. [4]; Hua *et al*. [5]). SVM, a special form of kernel classifiers, has become increasingly popular. SVM considers the structural risk in system modeling, and regularizes the model for good generalization and sparse representation. SVMs are successful in many applications. They outperform typical methods in classifications. However, the success of SVM depends on the good choice of model parameters and the kernel function, (namely, the data representation). In kernel methods, the data representation is implicitly chosen through the so-called kernel. This kernel actually plays two important roles: it defines the similarity between two examples, while defining an appropriate regularization term for the learning problem.

The success of SVMs is often dependent on the choice of kernel and features are typically hand-crafted and fixed in advance. However, hand-tuning kernel parameters can be difficult as can selecting and combining appropriate sets of features. Recent applications have also shown that using multiple kernels instead of a single one can enhance the interpretability of the decision function and improve performances. Multiple Kernel Learning (MKL) seeks to address this issue by learning the kernel from training data. In particular, it focuses on how the kernel can be learnt as a linear combination of given base kernels.

Traditional multiple kernel learning (MKL) approaches are limited in that they focus on learning linear combinations of base kernels-corresponding to the concatenation of individual kernel feature spaces. Conventional MKL formulations can be easily extended to learn general kernel combinations subject to general regularization on the kernel parameters (Varma and Babu [6] and Varma and Ray [7]). Far richer representations, this paper took products of kernels-corresponding to a tensor product of their feature spaces. This leads to a much higher dimensional feature representation as compared to feature concatenation. The generalized multiple kernel machine (GMKM) based on products of kernels gives good results for feature selection problems. The advantages of GMKM is two folds: 1) it can learn to achieve the same classification accuracy but using far fewer features. 2) the model learning can also be achieved very efficiently based on gradient descent optimization and existing large scale SVM solvers.

In financial data mining, high dimensional data from public financial statements and stock markets can be used for bankruptcy predictions. However, the high dimensional data make kernel classifiers infeasible due to the curse of

Manuscript received March 6, 2015; revised September 15, 2015.

Shian-Chang Huang and Tung-Kuang Wu are with the National Changhua University of Education, Changhua, Taiwan (e-mail: shhuang@cc.ncue.edu.tw, tkwu@im.ncue.edu.tw).

Nan-Yu Wang is with Ta Hwa University of Science and Technology, Hsinchu, Taiwan (e-mail: nanyu@tust.edu.tw).

doi: 10.18178/ijfcc.2015.4.5.417

dimensionality (Bellman [8]). Regarding dimensionality reduction, linear algorithms such as principal component analysis (PCA) and linear discriminant analysis (LDA) are the two most widely used methods due to their relative simplicity and effectiveness. However, classical techniques for manifold learning are designed to operate when the submanifold is embedded linearly, or almost linearly, in the observation space. Such algorithms often fail when nonlinear data structure cannot simply be regarded as a perturbation from a linear approximation. The task of nonlinear dimensionality reduction (NLDR) is to recover meaningful low-dimensional structures hidden in high dimensional data.

The method of partial least squares (PLS) (Rosipal and Kramer [9]) creates score vectors of inputs and outputs, which have a maximum covariance with each other. PLS could be thought of as a method for finding directions that are good at distinguishing between different output labels. However PLS is not invariant to linear transformations. This means that the analysis will be different depending on how the inputs are scaled. For example, doubling a input, or choosing different inputs within the same space, will give different answers. We could overcome the lack of invariance by simply orthonormalizing the inputs first. This is the orthonormalized PLS (OPLS, Worsley *et al*. [10]). OPLS can also be seen as a form of penalized canonical correlation analysis (CCA), which produces more compact data representation and is useful in high-dimensional data with heavy multicollinearity.

The remainder of this paper is organized as follows: Section 2 introduces the KOPLS. Subsequently, Section 3 describes the study data and discusses the empirical findings. Conclusions are given in Section 4.

## II. Kernel Orthonormalized Partial Least Square Analysis

Notationally, the basic PLS algorithm consider that we are given a set of pairs $\{x_i, y_i\}_{i=1}^{l}$, with $x_i \in R^N, y_i \in R^M$. Let us now introduce matrices $X = [x_1,...,x_l]^T$ and $Y = [y_1,...,y_l]^T$, where the $T$ superscript denotes matrix or vector transposition. Let us also denote by $X' = XU$ and $Y' = YV$ two matrices, each one containing $n_p$ projections of the original input and output data, $U$ and $V$ being the projection matrices of sizes $N \times n_p$ and $M \times n_p$, respectively. The goal of PLS is to find the directions $(U, V)$ of maximum covariance between the projected input and output data:

$$\max: Tr\{U^T C_{XY} V\} \tag{1}$$

$$\text{subject to}: U^T U = V^T V = I, \tag{2}$$

where $I$ is the identity matrix of size $n_p$, and $C_{XY}$ represents the covariance between the input and output datasets; namely, $C_{XY} = \frac{1}{l}\tilde{X}^T\tilde{Y}$, where $\tilde{X}, \tilde{Y}$ are the centered versions of $X$ and $Y$.

Traditional PLS is not invariant to linear transformations. To overcome the lack of invariance one can simply orthonormalize the inputs first. This is the orthonormalized PLS (OPLS, Worsley *et al*. [11]), which tackles the following maximization problem:

$$\max: Tr\{U^T C_{XY} C_{XY}^T U\} \tag{3}$$

$$\text{subject to}: U^T C_{XX} U = I. \tag{4}$$

All previous methods assume that there exists a linear relation between the latent variables of $X$ and of $Y$. However, this might not necessarily hold, and thus non-linear versions have become necessary to solve this problem. Kernel methods are a promising approach to formulate non-linear versions from linear algorithms. Notationally, consider $\phi(x): R^N \to \mathrm{H}$ a function that maps the input data into some Reproducing Kernel Hilbert Space (RKHS), usually referred to as feature space, of very large or even infinite dimension. Let $\Phi = [\phi(x_1),...,\phi(x_l)]^T$ and $Y = [y_1,...,y_l]^T$, and denote by $\Phi' = \Phi U$ the projection containing $n_p$ features of the original input data, $U$ being a projection matrix. The kernel OPLS cab be stated as:

$$\max: Tr\{U^T \tilde{\Phi}^T \tilde{Y} \tilde{Y}^T \tilde{\Phi} U\} \tag{5}$$

$$\text{subject to}: U^T \tilde{\Phi}^T \tilde{\Phi} U = I, \tag{6}$$

where $\tilde{\Phi}$ and $\tilde{Y}$ are centered versions of $\Phi$ and $Y$, respectively.

Making use of the Representer's Theorem, which states that all projection vectors $U$ can be expressed as a linear combination of the training data; namely, $U = \tilde{\Phi}^T A$, where $A = [\alpha_1,...,\alpha_{np}]$ and $\alpha_i$ is an column vector containing the coefficients for the $i$ th projection vector, and the maximization problem of KOPLS can be reformulated as follows:

$$\max: Tr\{A^T \mathbf{K}_X \mathbf{K}_Y \mathbf{K}_X A\} \tag{7}$$

$$\text{subject to}: A^T \mathbf{K}_X \mathbf{K}_X A = I, \tag{8}$$

where $\mathbf{K}_X = \tilde{\Phi}\tilde{\Phi}^T$ is the centered kernel matrices, and $\mathbf{K}_Y = \tilde{Y}\tilde{Y}^T$.

## III. Experimental Results And Analysis

This study used bankrupt companies listed in the Taiwan Stock Exchange (TSE) for analysis. Their public financial information is used for the model input. These bankrupt

companies were matched with normal companies for comparison. The sample data covers the period from 2000 to 2007.

For the balance of positive and negative samples, one company in financial crisis should be matched with one or two normal companies in the same year, in the same industry, running similar business items. Namely, they should produce the same products with the failed company and have similar scale of operation. Additionally, the normal company whose total asset or the scale of operation income should be close to the failed company. In our samples, 50 failed firms and 100 non-failed firms were selected. The study traced the data up to 5 years, which started from the day a respective company falls into financial distress backward up to a period of 5 years. The financial reports of the non-failed companies will be matched (pooled together) with the failed company in the same year. For example, company A failed in 2005 and company B failed in 2007. We will pool them and their matched companies A', B' in the same file labeled C0 representing their financial status in the year of bankruptcy. Companies A and A' (or company B and B') will be traced backward up to five years. These data were put in separate files labeled C0, C1, C2, C3, and C4 respectively for classification.

The variables of this research are selected from the TEJ (Taiwan Economic Journal) financial database, which contains the following five financial indexes: profitability index, per share rates index, growth rates index, debt-paying ability index, management ability index. Altogether, there are 54 financial ratios covered by the five indexes. If some values of a ratio lost on some firms, this ratio was deleted. As a result, overall 48 financial ratios were obtained for analysis.

This study tested five conventional classifiers and a kernel classifier (SVM) for bankruptcy predictions, including decision tree (J48), nearest neighbors with three neighbors (KNN), logistic regressions, Bayesian networks (BayesianNet), radial basis neural network (RBFNetwork), and SVM (with gaussian kernels). The data set was randomly divided into ten parts, and ten-folds cross validation was applied to evaluate the model performance.

Table I shows that SVM outperforms other classifiers. Namely, kernel classifiers outperform traditional classifiers due to their flexibility in dealing nonlinear and high-dimensional data. Consequently, this study implemented an advanced kernel classifier, the GMKM, for subsequent classifications.

TABLE I: PERFORMANCE COMPARISON ON BASIC PREDICTION MODELS (ACCURACY %)

| | Sample C0 | Sample C1 | Sample C2 | Sample C3 | Sample C4 |
|---|---|---|---|---|---|
| **J48** | 90.7007 | 87.5912 | 84.6715 | 81.3433 | 71.8750 |
| **KNN** | 88.3212 | 90.5109 | 81.7518 | 76.1194 | 73.4375 |
| **BayesianNet** | 90.2409 | 90.5109 | 85.4015 | 84.3284 | 80.0313 |
| **Logistic** | 88.4058 | 84.0580 | 75.3623 | 71.1111 | 70.5426 |
| **RBFNetwork** | 89.1304 | 90.5797 | 85.5072 | 74.8148 | 78.2946 |
| **SVM** | 91.2409 | 91.3043 | 86.2319 | 83.5821 | 80.6875 |

TABLE II: PERFORMANCE IMPROVEMENTS BY DIMENSIONALITY REDUCTIONS

| | Sample C0 | Sample C1 | Sample C2 | Sample C3 | Sample C4 |
|---|---|---|---|---|---|
| **ICA+SVM** | 74.8350 | 65.2200 | 72.1980 | 65.2750 | 65.3210 |
| **PCA+SVM** | 82.6400 | 84.0700 | 76.0990 | 78.6810 | 78.4620 |
| **LDA+SVM** | 86.2091 | 81.0989 | 81.0990 | 80.0550 | 77.5640 |
| **KPCA+SVM** | 88.8680 | 87.0330 | 80.9340 | 81.5930 | 79.2310 |
| **Isomap+SVM** | 83.2970 | 86.2641 | 82.6919 | 83.0770 | 76.0260 |
| **KOPLS+GMKM** | 96.0600 | 95.5900 | 94.1200 | 93.1200 | 90.1800 |

Next, we compare our method (GMKM on KOPLS) with other dimensionality reduction methods. We compared our system with other famous subspace or manifold learning algorithms such as the PCA, ICA (Independent Component Analysis, Hyvärinen *et al.* [11]), LDA, kernel PCA (KPCA), and Isomap (Tenenbaum *et al.* [12]). The dimension of subspace was set to five for all algorithms.

Table II shows that GMKM on KOPLS significantly outperform other classifiers. It achieved the highest accuracy. This results fully demonstrate that financial data are not sampled from a linear manifold. Hence, linear algorithms such as PCA ICA, and LDA fail to extract discriminative information from data manifold. Considering nonlinear dimensionality reduction algorithms (KOPLS) are more effective. On the other hand, our data come from diverse sources, only multiple kernel machines such as GMKM are powerful enough to handle the complex data. We also find in Table II that supervised algorithms are better than unsupervised ones, and nonlinear dimensionality reduction methods (such as kernel PCA) is not always better than linear algorithms. KPCA works in an unsupervised manner which lacks information to guide the mapping learning that could maintain most discriminant power. However, KOPLS is a supervised algorithm which nonlinearly forms a manifold not only preserving local geometry of the data samples, but also contains label information to discriminate the data.

## IV. CONCLUSIONS

Multiple kernel learning approaches are limited in that they focus on learning linear combinations of base kernels-corresponding to the concatenation of individual kernel feature spaces. Far richer representations, this paper took products of kernels-corresponding to a tensor product of their feature spaces. This leads to a much higher dimensional feature representation as compared to feature concatenation. The advantage of GMKM is that it can learn to achieve the same classification accuracy but using far fewer features. This study developed KOPLS to find a good low dimensional projection that respected the discriminant structure inferred from the output. KOPLS maximizes the covariance between inputs and outputs. Constructing classifiers on KOPLS

reduces the computational loading and simultaneously enhances performance. The success of our hybrid classifier is attributed to the combination of these two techniques. The empirical results confirmed the superiority of the proposed system.

Future research may consider semi-supervised subspace or manifold learning algorithms to enhance system performance, or to include more variables such as non-financial and macroeconomic variables to improve accuracy.

## REFERENCES


[1] W. Wang and J. Yang, "Mining high-dimensional data," *Data Mining and Knowledge Discovery Handbook*, pp. 793-799, 2005.

[2] I. H. Witten, and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2005.

[3] B. S. Ahn, S. S. Cho, and C. Y Kim, "The integrated methodology of rough set theory and artificial neural network for business failure prediction," *Expert Systems with Applications*, vol. 18, no. 2, pp. 65-74, 2000.

[4] C. H. Wu, W. C. Fang, and Y. J. Goo, "Variable selection method affects SVM-based models in bankruptcy prediction," presented at 9th Joint International Conference on Information Sciences, 2006.

[5] Z. Hua, Y. Wang, X. Xu, B. Zhang, and L. Liang, "Predicting corporate financial distress based on integration of support vector machine and logistic regression," *Expert Systems with Applications*, vol. 3, no. 2, pp. 434-440, 2007,

[6] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proc. the International Conference on Machine Learning*, Montreal, Canada, June 2009, pp. 1065-1072.

[7] M. Varma and D. Ray, "ng the discriminative power-invariance trade-off," presented at International Conference on Computer Vision, October 2007.

[8] R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.

[9] R. Rosipal and N. Kramer, "Overview and recent advances in partial least squares," *Subspace, Latent Structure and Feature Selection Techniques*, Springer, 2006, pp. 34-51.

[10] K. Worsley, J. Poline, K. Friston, and A. Evans, "Characterizing the response of PET and fMRI data using multivariate linear models," *Neuro Image*, vol. 6, no. 4, pp. 305-319, 1998.

[11] A. Hyvärinen, J. Karhunen, and E. Oja, "Independent component analysis," *Wiley Interscience*, 2001.

[12] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, 2000.


**Shian-Chang Huang** received his MS degree in electric engineering from National Tsing Hwa University, and his PhD degree in financial engineering from National Taiwan University. He is currently a professor at the Department of Business Administration, National Changhua University of Education, Taiwan. His research interests include machine learning, soft computing, signal processing, data mining, big data, computational intelligence, financial engineering, e-commerce, econometrics.

**Nan-Yu, Wang** was born in Taiwan, Changhua. She completed his Ph. D. degree from the Department of Business Education, National Changhua University of Education, Focusing on Finance. She is currently an associate professor of the Department of Business and Tourism Planning, Ta Hwa University of Science and Technology. Her research areas include investment, risk management, corporate finance.

**Tung-Kuang Wu** received his B.S. degree in electrical engineering from National Taiwan University in 1984, and the M.S. and Ph.D. degrees in computer engineering from the Pennsylvania State University in 1991 and 1995, respectively. He is currently a professor at the Department of Information Management of National Changhua University of Education, Changhua, Taiwan. His research interests include parallel processing, artificial intelligence and special education technologies.