FINANCIAL BIG DATA ANALYSIS BY SPARSE REPRESENTATION CLASSIFIERS

Shian-Chang Huang [a], Nan-Yu Wang [b], Tung-Kuang Wu [c]
[ac]National Changhua University of Education, Taiwan
[b] Ta Hwa University of Science and Technology, Hsinchu, Taiwan
Corresponding email: shhuang@cc.ncue.edu.tw

Abstract

Financial big data analysis has recently become a popular research field. Kernel machines (such as support vector machines, SVM) have demonstrated good performance in many areas of pattern recognition. However, the representation of traditional kernel machines is not sparse. A sparse model representation in machine learning is expected to improve the generalization performance and computational efficiency. Moreover, in big data analysis, high-dimensional and nonlinear distributed data generally degrade the performance of a classifier due to the curse of dimensionality, especially in financial distress predictions. To address these problems, this study proposes a novel system using kernel sparse representation classifiers (KSRC) to discriminate financial statement data. The statement data are first projected to a low-dimensional subspace and is then classified by the KSRC. Compared with other data mining systems, the proposed system performs best.

Keywords: Sparse Representation, Orthonormalized Partial Least Square Analysis, Dimensionality Reduction, Data Mining.

1 Introduction

Mining financial big data is important in many applications. Financial data mining, such as that involved in predicting bankruptcies, has become a popular topic owing to the late-2000s financial crisis. The objective of this paper is to develop a reliable big data analysis system for bankruptcy predictions. Our strategy is to develop a sparse representation based classifiers on some low-dimensional representative space that can effectively discriminate the data. This expert system can prevent banking and investment institutions from investing in a distressed company.

Recently, many approaches from data mining to artificial intelligence have been developed for solving these problems. These approaches include inductive learning (Han et al., 1996; Shaw &Gentry, 1998), case-based reasoning (Buta, 1994; Bryant, 1997), neural networks (Bortiz & Kennedy, 1995; Jo &Han, 1996;Coakley & Brown, 2000; Nasir et al., 2000; Tang &Chi, 2005), rough set theory (Dimitras et al., 1999; Ahn et al., 2000), and support vector machines (SVMs) (Wu et al., 2006; Hua et al., 2007). SVM (Vapnik, 1999; Cristianini &Shawe-Taylor, 2000), a special form of kernel classifiers, has become increasingly popular. The formulation of an SVM simultaneously embodies the structural risk (a maximum margin classifier) and empirical risk minimization principles. SVM has good performance in many areas and has been regarded as the state-of-the-art technique for regression and classification applications. Despite these attractive features and many good empirical results obtained using SVMs, some data modelling participants have begun to realize that the ability of the SVM method to produce sparse models has perhaps been overstated. For example, it has been shown that the standard SVM technique is not always able to construct parsimonious models in system identification (e.g. Drezet &Harrison, 1998).

Sparsity-based representation is a task of reconstructing a given signal by selecting a relatively small subset of dictionary or basis elements from a large dictionary while keeping the reconstruction error as small as possible. A sparse model representation in machine learning is expected to improve the generalization performance and computational efficiency (Floyd &Warmuth, 1995; Graepel et al., 2000; Zhang &Zhou, 2010). Sparsity-based algorithms have been rapidly applied to many practical engineering problems and almost always leads to encouraging results (Mallat, 2008; Baraniuk et al., 2010). This is because most natural signals can be compactly represented by only a few coefficients that carry the most important information in a certain basis or dictionary. Moreover, it is based on the observation that despite the high dimensionality of natural signals, the signals in the same class usually lie in a low-dimensional subspace.

In machine learning, sparsity usually refers to the extent to which a representation model contains null values and can be measured by the number of nonzero coefficients in a decision function. The less the number of nonzero elements is, the better sparsity we get. The mechanism of maximizing the sparsity of a model representation can be regarded as an approximative form of the minimum description length principle which can be used to improve the generalization performance (Duda et al., 2000). Sparse representation based classifier (SRC), a combined result of machine learning and compressed sensing, shows its good classification performance on many data.

Kernel methods are known to significantly improve the performance of classical pattern recognition algorithms, by implicitly exploiting the higher-order structure of the given data that may not be captured by linear models. This is exceedingly true in cases where the background and target subspaces are not linearly separable. Kernel sparse representation based classifier (KSRC) based on SRC and the kernel trick (a technique in machine learning) is a nonlinear extension of SRC, which can remedy the drawback of SRC. Namely, KSRC is based on kernelized sparse representation, where a test sample in the high-dimensional feature space induced by a kernel is assumed to be sparsely represented by a non-linear combination of the training samples in the same feature space. The sparse vectors can be efficiently recovered by kernelized algorithms.

The power of the kernel method lies in the implicit use of a high-dimensional RKHS (reproducing kernel Hilbert space), induced by a positive semidefinite (PSD) kernel (Scholkopf &Smola, 2002). Kernel classifiers map input data into a high-dimensional RKHS, where simple linear classification is performed. However, due to the large amount of data from public financial statements that can be used for bankruptcy predictions, the high dimensional input data makes kernel classifiers infeasible due to the curse of dimensionality (Bellman, 1961). Consequently, one needs to transform the input data space into a suitable low dimensional subspace that optimally represents the data.

Traditional dimensionality reduction methods are unsupervised in nature. They fail to incorporate the label (or class) information so as to guide subspace or manifold learning. To address this problem, this study constructs classifiers on subspace extracted by kernel orthonormalized partial least square (KOPLS, Arenas-Garcia &Camps-Valls, 2008) so as to prevent the curse of dimensionality. The method of partial least squares (PLS) (Rosipal &Kramer, 2006) 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 (or basis vectors) that are good at distinguishing between different output labels. These basis vectors can be used as the dictionary for sparse representation classifiers. 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., 1998).

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

2 The Proposed Methodology

In the first stage, KOPLS optimally project original data space to a low dimensional subspace which has maximum covariance between inputs and outputs. In the second stage, SRC is constructed on the the low dimensional representative space for classifications. For the details of KOPLS, please refer to Arenas-Garcia and Camps-Valls (2008).

2.1 Sparse Representation Based Classifier

The basic problem of classification is to use training samples from different classes of signals to correctly determine as to which class the test signal belongs to. The n_i training samples of a particular i th class are arranged as the columns of a matrix $D_i = [x_{i,1}, x_{i,2},..., x_{i,n_i}] \in R^{m \times n_i}$ where each training sample x is a column vector and $x \in R^m$. The basic assumption in the theoretical development of SRC is that the samples of the matrix D_i lie on linear subspace.

Any test sample lies in the linear span of training samples belonging to the i th class, given as follows (Wright et al.,2009):

$$y = \alpha_{i,1}x_{i,1} + \alpha_{i,2}x_{i,2} + ... + \alpha_{i,n_i}x_{i,n_i} \quad (1)$$

for some scalars $\alpha_{i,j}$. Matrix D, which is also known as the dictionary, is defined to be formed by the entire set of basis vectors (produced by KOPLS) of all the K classes, given as follows:

$$D = [D_1, D_2,..., D_K]$$

here $D \in R^{m \times n}$ where n is the size of the dictionary i.e. total number of basis vectors. Then y may be linearly represented in terms of all the training samples as follows:

$$y = D\alpha \in R^m, \quad (2)$$

where α is known as the coding vector, given as $\alpha = [\alpha_1, \alpha_2,..., \alpha_K]$. Here α_i is the sub-coding vector associated with the matrix D_i. Ideally $\alpha = [0,...,0, \alpha_{i,1}, \alpha_{i,2},..., \alpha_{i,ni}, 0,...0]^T \in R^n$ is a vector whose most entries are zero except the elements that are associated with the i th class. Then one can represent y as $y \approx D_i\alpha_i \in R^m$, which means that, assuming that y belongs to the i th class, in a practical situation, most of the coefficients in $\alpha_k (k \neq i)$ are quite small and only coefficients $\alpha_k (k = i)$ have significant values. This means a test sample may be represented by training samples from the same class. The test signal is more accurately determined depending on the sparsity of recovered α. This leads to the determination of the sparsest solution of $y = D\alpha$ (Huang &Aviyente, 2006) using the l_0-norm minimization problem:

$$\min \| x \|_0 \ \textit{subject to } y = D\alpha. \quad (3)$$

Recent researches reveal the fact that finding the sparsest solution via l_0-minimization has been converted to an l_1-minimization problem, since the combinatorial l_0-minimization is essentially an NP-hard problem and l_1-minimization is known as the closest convex function to l_0-minimization (Donoho, 2006). Consider the case of noisy data, the model can be modified as

$$y = D\alpha + \xi$$

where ξ is a noise vector with bounded energy $\| \xi \| < \varepsilon$. The problem can be modified as

$$\min \| \alpha \|_1 \ \textit{subject to} \ \| y - D\alpha \|_2 \leq \varepsilon, \tag{4}$$

where $\| \cdot \|_2$ denotes the l_2-norm. This model can be written in its equivalent form given as follows:

$$\hat{x} = \arg\min_x \| y - D\alpha \|_2^2 + \lambda \| \alpha \|_1 . \tag{5}$$

The residuals, $r_i = \| y - D_i \hat{\alpha}_i \|_2$ are computed ($\hat{\alpha}_i$: coding sub-vector associated with class i, determined from $\hat{\alpha}$). The output of the classifier is the class with the smallest residual.

For kernel extension of SRC, we map samples from original feature space into a high dimensional feature space H by a nonlinear function $\phi : x \to \phi(x)$. Let $\Phi = [\phi(D_1), \phi(D_2), \ldots, \phi(D_K)]$ represent the matrix composed of all the training samples after the nonlinear mapping ϕ. The problem of sparse representation in H can be described as (Wright et al.,2009)

$$\hat{\beta} \arg\min_\beta \| \beta \|_0 \ \textit{subject to} \ \phi(y) = \Phi\beta, \tag{6}$$

where $\phi(y)$ is any test sample in the high dimensional feature space, which corresponds to y in the original feature space. Similarly, the approximate solution of Eq. (6) can be obtained through the following convex relaxed optimization (Candè & Tao, 2006; Candè et al., 2006; Donoho, 2006)

$$\hat{\beta} \arg\min_\beta \| \beta \|_1 \ \textit{subject to} \ \phi(y) = \Phi\beta. \tag{7}$$

When the observations are not accurate, the constraint in Eq. (7) should be relaxed and the following optimization problem is obtained:

$$\hat{\beta} \arg\min_\beta \| \beta \|_1 \ \textit{subject to} \ \| \phi(y) - \Phi\beta \|_2 \leq \varepsilon. \tag{8}$$

The above equation can be transformed to the kernel form:

$$\hat{\beta} \arg\min_\beta \| \beta \|_1 \ \textit{subject to} \ \| \Phi^T \phi(y) - \Phi^T \Phi\beta \|_2 = \| K_{XY} - K_X \beta \|_2 \leq \delta, \tag{9}$$

where $K_{XY} = \Phi^T \phi(y)$ and $K_X = \Phi^T \Phi$ are the kernel matrixes.

3 Experimental Results and Analysis

This study takes companies listed on the Taiwan Stock Exchanges (TSE) as the samples for analysis. This investigation used publicly disclosed financial information of companies as the model input. Stocks of companies that are bankrupt or de-listed and labeled as full delivery securities on the TSE were selected as the samples in this study. These samples were matched with normal companies for comparison. The sample data covers the period from 1999 to 2010.

On behalf of sample matching, each company experiencing financial failure should be matched against two normal companies in the same year, same industry and running similar business items. Restated, the comparison companies should produce the same products as the failed companies and have similar scale of operations. Generally, the comparison company had similar total assets or the scale of operation income is close to the failed company. As a result, 57 failed firms and 114 non-failed firms were selected in the period between 2005-2010. This study traced data over 5 years, counted backwards from the day a company fell into financial distress for 5 years. The financial reports of the comparison companies are matched (pooled together) with those of the failed companies in the same year. The variables used in this research are selected from the TEJ (Taiwan Economic Journal) financial database, which contains the following eight catalogues of financial indexes: corporate governance, macroeconomic condition, auditor opinion, and auditor quality. Totally, 18 indexes comprise 111 variables.

This study tested traditional and kernel classifiers for bankruptcy predictions, including a decision tree (J48), nearest neighbors (KNN), logistic regressions (LR), Bayesian networks (BN), and SVM. The data set was randomly divided into ten parts, and ten-folds cross validation will be applied to evaluate the model's performance.

Table 1 shows the performance for all the classifiers. On average, their accuracies are about 70% . The performance of SVM, BN, and LR are similar. The accuracy of J48 is slightly better. The performance of KNN is the poorest. All of their performance are not satisfactory.

Table 1: Performance comparison on basic prediction models (accuracy %)

	1st year	2nd year	3rd year	4th year	5th year
SVM	73.10	72.51	70.76	70.18	63.16
BN	70.76	69.01	69.59	65.50	59.65
LR	67.84	73.10	71.93	69.59	76.02
J48	80.70	80.12	76.61	84.21	84.21
KNN	64.91	63.74	60.82	59.06	53.80

The performance of the new system is shown in Table 2. Average performance of all models is shown in Table 3. Figure 1 is the performance comparison. Our new system, KSRC on KOPLS space, significantly outperforms traditional classifiers. KSRC can enhance out-of-sample model generalization and preventing the problem of overfitting.

These results demonstrate that in financial big data mining, the data is not from a linear subspace. Hence, linear algorithms fail to extract key discriminative information for classification. It is more effective to consider nonlinear subspace learning (such as KOPLS) and sparse representation based classifier. The basis vectors found by KOPLS are optimal candidates to serve as the dictionary of the KSRC to improve classification performance.

Table 2: Performance of the proposed system (accuracy %)

	1st year	2nd year	3rd year	4th year	5th year
The system	97.06	94.12	94.12	91.67	88.24

Table 3: Average performance (accuracy %)

	Average
SVM	69.94
BN	66.90
LR	71.70
J48	81.17
KNN	60.47
The system	93.04

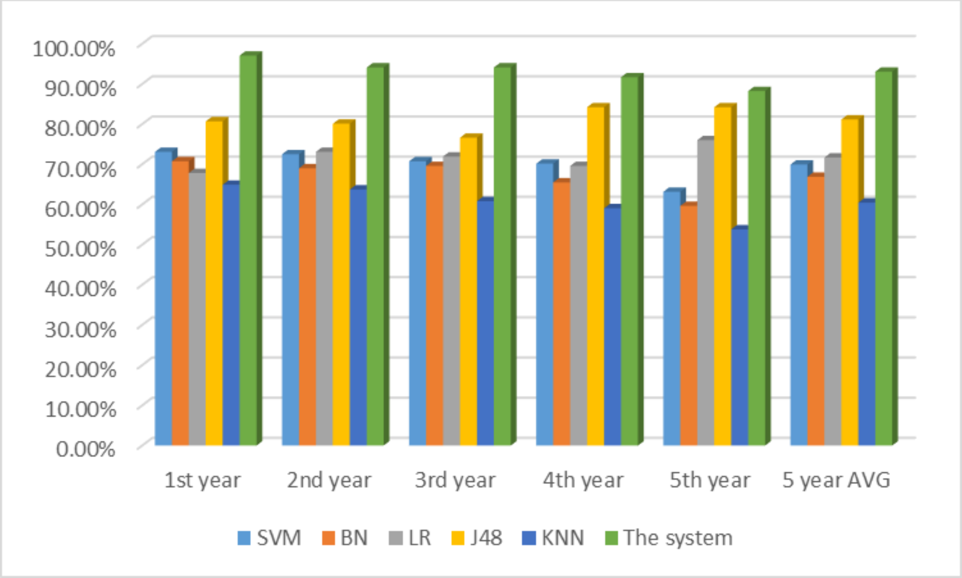

Figure 1: Performance comparision of all models

4 Conclusions

In financial big data analysis, bankruptcy prediction is important for banks or investors to control risk in their investments. Traditional classifiers usually perform poorly when they encounter the high-dimensional and nonlinear-distributed financial input data. This study addresses this problem by constructing a KSRC on subspaces of KOPLS for high-dimensional data mining. KOPLS extracted representative subspaces that optimally discriminate the output labels, significantly reduce the computational loading of KSRC and simultaneously enhance their performance. Empirical results will indicate that, compared to other classifiers, the proposed system performs best and robustly. The proposed method can help financial institutions accurately assess their investment risk and substantially reduce losses.

Future research may include more financial information, such as non-financial and macroeconomic variables. However, high-dimensional data mining remains a great challenge. More effective subspace learning algorithms require further study.

References

i. Ahn, B.S., Cho, S. S. &Kim, C. Y., 2000. The Integrated Methodology of Rough Set Theory and Artificial Neural Network for Business Failure Prediction. *Expert Systems with Applications*, 18(2), pp. 65-74.

ii. Arenas-Garcia, J. & Camps-Valls, G., 2008. Efficient Kernel Orthonormalized PLS for Remote Sensing Applications. *IEEE Transactions on Geoscience and Remote Sensing*,46(10), pp. 2872-2881.

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

iv. Bortiz, J.E. &Kennedy, D. B., 1995. Effectiveness of Neural Network Types for Prediction of Business Failure. *Expert Systems with Application*, 9(4), pp. 503-512.

v. Bryant, S. M., 1997. A Case-Based Reasoning Approach to Bankruptcy Prediction Modeling. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 6(3), pp. 195-214.

vi. Buta, P., 1994. Mining for Financial Knowledge with CBR. *AI Expert*, 9(10), pp. 34-41.

vii. Candè, E., Romberg, J. & Tao, T., 2006. Stable Signal Recovery from Incomplete and Inaccurate Measurements. *Commun. Pure Appl. Math*, 59(8), pp. 1207-1223.

viii. Candè, E. & Tao, T., 2006. Nearest-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies. *IEEE Trans. Inf. Theory*, 52(12), pp. 5406-5425.

ix. Coakley, J. R. & Brown, C. E., 2000. Artificial Neural Networks in Accounting and Finance: Modeling Issues. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 9(2), pp. 119-144.

x. Cristianini, N. & Shawe-Taylor, J., 2000. *An Introduction to Support Vector Machines*. Cambridge University Press.

xi. Dimitras, A. I., Slowinski, R., Susmaga, R. & Zopounidis, C., 1999. Business Failure Prediction Using Rough Sets. *European Journal of Operational Research*, 114, pp. 263-280.

xii. Donoho, D., 2006. For Most Large Underdetermined Systems of Linear EquationsThe Minimal l1-Norm Solution is Also The Sparsest Solution. *Comm. Pure and Applied Math*, 59(6), pp. 797-829.

xiii. Drezet, P.M.L. &Harrison, R. F., 1998. Support Vector Machines for System Identification. In *Proceedings of the UKACC International Conference on Control'98*. Swansea, UK, pp. 688-692.

xiv. Duda, R., Hart, P. & Stork, D., 2000. *Pattern Classification*, 2nd edn. New York: Wiley.

xv. Floyd, S. & Warmuth, M., 1995. Sample Compression Learnability, and Vapnik-Chervonenkis Dimension. *Mach. Learn*, 21(3), pp. 269-304.

xvi. Graepel, T., Herbrich, R. & Shawe-Taylor, J., 2000. Generalisation Error Bounds for Sparse Linear Classifiers. In *Proc. 13th Annu. Conf. Comput. Learn. Theory*. Stanford, CA, pp. 298-303.

xvii. Han, I., Chandler, J. S. &Liang, T. P., 1996. The Impact of Measurement Scale and Correlation Structure on Classification Performance of Inductive Learning and Statistical Methods. *Expert System with Applications*, 10(2), pp. 209-221.

xviii. Hua, Z., Wang, Y., Xu, X., Zhang, B. & Liang, L., 2007. Predicting Corporate FinancialDistress Based on Integration of Support Vector Machine and Logistic Regression. *Expert Systems with Applications*, 33(2), pp. 434-440.

xix. Huang, K. & Aviyente, S., 2006. *Sparse Representation for Signal Classification*. Neural Information Processing Systems.

xx. Jo, H. & Han, I., 1996. Integration of Case-Based Forecasting, Neural Network, and Discriminant Analysis for Bankruptcy Prediction. *Expert Systems with Applications*, 11(4), pp. 415-422.

xxi. Nasir, M.L, John, R.I., Bennett, S. C. &Russell, D. M., 2000. Predicting Corporate BankruptcyUsing

Modular Neural Networks. *Computational Intelligence for Financial Engineering*, 2000, pp. 86-91.

xxii. Rosipal, R. & Kramer, N., 2006. Overview and Recent Advances in Partial Least Squares. In *Subspace, Latent Structure and Feature Selection Techniques*, Saunders, C., Grobelnik, M., Gunn, S. & Shawe-Taylor, J. (eds.). Springer, pp. 34-51.

xxiii. Scholkopf, B. &Smola, A. J., 2002. *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, M.A: MIT Press.

xxiv. Shaw, M. & Gentry, J., 1998. Using and Expert System with Inductive LearningTo Evaluate Business Loans. *Financial Management*, 17(3), pp. 45-56.

xxv. Tang, T.C. &Chi, L. C., 2005. Neural Networks Analysis in Business Failure Prediction of Chinese Importers: A Between-Countries Approach. *Expert Systems with Applications*, 29(2), pp. 244-255.

xxvi. Vapnik, V. N., 1999. *The Nature of Statistical Learning Theory*, 2nd edn. New York: Springer.

xxvii. Worsley, K., Poline, J., Friston, K. & Evans, A., 1998. Characterizing The Response of PET and fMRI Data Using Multivariate Linear Models. *NeuroImage*, 6(4), pp. 305-319.

xxviii. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S. S. & Ma, Y., 2009. Robust Face Recognition Via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2), pp. 210-227.

xxix. Wu, C.H., Fang, W. C. & Goo, Y. J., 2006. *Variable Selection Method Affects SVM-based Models in Bankruptcy Prediction*. 9th Joint International Conference on Information Sciences.

xxx. Zhang, L. &Zhou, W. D., 2010. On the Sparseness of 1-norm Support Vector Machine. *Neural Networks*, 23(3), pp. 373-385.