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

Article

An intelligent system for business data mining

Global Business & Finance Review (GBFR)

Provided in Cooperation with:
People & Global Business Association (P&GBA), Seoul

Suggested Citation: Huang, Shian-Chang; Wu, Tung-Kuang; Wang, Nan-Yu (2017) : An intelligent system for business data mining, Global Business & Finance Review (GBFR), ISSN 2384-1648, People & Global Business Association (P&GBA), Seoul, Vol. 22, Iss. 2, pp. 1-7, https://doi.org/10.17549/gbfr.2017.22.2.1

This Version is available at:
https://hdl.handle.net/10419/224366

GLOBAL BUSINESS & FINANCE REVIEW, Volume. 22 Issue. 2 (SUMMER 2017), 1-7
pISSN 1088-6931 / eISSN 2384-1648 | Http://dx.doi.org/10.17549/gbfr.2017.22.2.1

GLOBAL BUSINESS & FINANCE REVIEW

www.gbfrjournal.org

An Intelligent System for Business Data Mining

Shian-Chang Huang[a], Tung-Kuang Wu[b], and Nan-Yu Wang[c]

[a]National Changhua University of Education, Taiwan
[b]National Changhua University of Education, Taiwan
[c]Ta Hwa University of Science and Technology, Taiwan

A B S T R A C T

Mining high-dimensional business data is a challenging problem. Particularly in bankruptcy predictions, we need to analyze large amounts of information from financial statements and stock markets. This paper proposes a new strategy to deal with the problem. Because of the highly correlation among financial information, this study employed a technique called generalized discriminant analysis (GDA) to identify important features and reduce the data dimension. GDA is a nonlinear discriminant analysis using kernel function operator. It's easy to deal with a wide class of nonlinearity in financial data, and can reduce the computational loading of subsequent prediction classifier. Due to the promising success of kernel machines in many applications, this study utilized a generalized multiple kernel machine (GMKM) to serve as the predictor. Combining the strengths of GDA and GMKM, our system robustly outperforms traditional prediction systems.

Keywords: *Business Data Mining, Generalized Discriminant Analysis, Financial Statements, Multiple Kernel Machine, Support Vector Machine*

Ⅰ. Introduction

Big data analysis becomes very popular recently. Reviewing recent literature, many advanced approaches from data mining or artificial intelligence were developed for big data analysis. These methods (Witten and Frank, 2005) include inductive learning, case-based reasoning, neural networks, rough set theory (Ahn et al., 2000), and support vector machines (SVM) (Vapnik, 1999; Wu et al., 2006; Hua et al., 2007). Mining big data is a great challenge, especially for high dimensional data (Wang and Yang, 2005). The first challenge 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. The second challenge is 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

Received: April 21, 2016; Revised: Oct 31, 2016; Accepted: Feb 1, 2017

† Nan-Yu Wang
Ta Hwa University of Science and Technology, Taiwan
Tel. +886-922068381 E-mail: nanyu@tust.edu.tw

of this paper is to overcome the above problems and develop a novel classification system.

SVM, a special form of kernel classifiers (Schoelkopf et al., 1999), 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 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 (Lanckriet et al., 2004). 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 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, 2009 and Varma and Ray, 2007). 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 big data analysis, 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 dimensionality (Bellman, 1961). Regarding dimensionality reduction, linear algorithms such as principal component analysis (PCA) and linear discriminant analysis (LDA, Fukunaga, 1990) are the two most widely used methods due to their relative simplicity and effectiveness. However, such algorithms often fail when nonlinear data distribution cannot simply be regarded as a perturbation from a linear approximation.

Generalized discriminant analysis (GDA, Baudat and Anouar, 2000) is a nonlinear extension of LDA using kernel function operator. GDA overcomes the limitations of LDA nonlinearly finding a good low dimensional projection which respects the discriminant structure inferred from data. GDA method provides a mapping of the input vectors into high dimensional feature space. In the transformed space, linear properties make it easy to extend and generalize the classical LDA to non-linear discriminant analysis. The formulation is expressed as an eigenvalue problem resolution. Using a different kernel, one can cover a wide class of nonlinearities.

The remainder of this paper is organized as follows: Section 2 introduces related works. Section 3 describes the proposed system. Section 4 discusses the empirical findings. Conclusions are given in Section 5.

Ⅱ. Related Works

A. Support Vector Machines

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. SVMs were proposed by Vapnik

(1999). By viewing input data as two sets of vectors (two class classification) in a high-dimensional transformed space, an SVM seeks to construct a separating hyperplane in that space, which maximizes the margin between the two data sets. Based on the structured risk minimization principle, SVMs seek to minimize the upper bound of the generalization error instead of the empirical error as with neural networks. The SVM classification function is formulated as follows:

$$y = sign(\mathbf{w}^T \varphi(\mathbf{x}) + b), y \in \{-1,1\}, \quad (1)$$

where y is the output (1 for type A, -1 for type B); $\varphi(\mathbf{x})$ is a nonlinear mapping from the input space to the high-dimensional transformed space. SVMs exploit the idea of mapping input data into a high-dimensional reproducing kernel Hilbert space (RKHS) where classification can easily be performed. Coefficients $\mathbf{w}$ and b are estimated by the following optimization problem:

$$\min_{\mathbf{w},b} R(\mathbf{w},\xi) = \frac{1}{2}\| \mathbf{w} \|^2 + C\sum_{i=1}^{m}\xi_i \quad (2)$$

with

$$y_i(\mathbf{w}^T\varphi(\mathbf{x})+b) \geq 1-\xi_i, \quad i=1,...,m.$$
$$\xi_i \geq 0,$$

where C is a prescribed parameter to evaluates the trade-off between the empirical risk and the smoothness of the model.

The value of the kernel is equal to the inner product of the two vectors $\mathbf{x}$ and $\mathbf{x}_i$ in the feature space, so that $K(\mathbf{x},\mathbf{x}_i) = \phi(\mathbf{x})\phi(\mathbf{x}_i)$. Any function that satisfies Mercer's condition (Vapnik, 1999) can be used as the Kernel function.

Ⅲ. The Proposed System

To reduce the computational loading of kernel machines and simultaneously enhance their performance. This study constructs GMKMs on the subspace created by the GDA.

A. Generalized Multiple Kernel Machine

In multiple kernel learning (MKL), we start with N_k base kernels, $\mathbf{K}_1,...,\mathbf{K}_{N_k}$, where $\mathbf{K}_k(\mathbf{x},\mathbf{y}) = \exp(\gamma_k f_k(\mathbf{x},\mathbf{y}))$. f_k is the distance function. Given the base kernels, the optimal data descriptor's kernel is approximated as $\mathbf{K}_{opt} = \sum_k d_k \mathbf{K}_k$ where the weights $\mathbf{d}$ correspond to the trade-off level. The optimisation is carried out in an SVM framework so as to achieve the best classification on the training set, subject to regularisation. We set up the following primal cost function

$$\min_{\mathbf{w},\mathbf{d},\mathbf{x}} \frac{1}{2}\mathbf{w}^t\mathbf{w} + C\mathbf{1}^t\xi + \sigma^t\mathbf{d} \quad (3)$$

$$subject\ to \quad y_i(\mathbf{w}^t\phi(\mathbf{x}_i)+b) \geq 1-\xi_i \quad (4)$$

$$\xi \geq 0, \mathbf{d} \geq 0 \quad (5)$$

$$where \quad \phi^t(\mathbf{x}_i)\phi(\mathbf{x}_j) = \Sigma_k d_k \phi_k(\mathbf{x}_i)\phi_k(\mathbf{x}_j). \quad (6)$$

The objective function is similar to the standard l_1 C-SVM objective. Given the misclassification penalty C, it maximises the margin while minimising the hinge loss on the training set $(\mathbf{x}_i, y_i)$. The only addition is an l_1 regularisation on the weights $\mathbf{d}$ since we would like to discover a minimal set of invariances. Thus, most of the weights will be set to zero depending on the parameters which encode our prior preferences for descriptors. The l_1 regularisation thus prevents overfitting if many base kernels are included since only a few will end up being used. The constraints are also similar to the standard SVM formulation. Two additional constraints have been incorporated. $\mathbf{d} \geq 0$, ensures that the weights are interpretable and also leads to a much more efficient optimisation problem.

In addition, one can also tune kernel parameters in general kernels such as

$$\mathbf{K}_d(\mathbf{x}_i,\mathbf{x}_j)=(d_0+\sum_m d_m\mathbf{x}_i^t\mathbf{A}_m\mathbf{x}_j)^n \tag{7}$$

$$\mathbf{K}_d(\mathbf{x}_i,\mathbf{x}_j)=e^{\sum_m d_m\mathbf{x}_i^t\mathbf{A}_m\mathbf{x}_j}. \tag{8}$$

In this paper, we used the second setting (product kernel) for our GMKM. Combined with a sparsity promoting regularizer on $\mathbf{d}$, this can be used for non-linear dimensionality reduction and feature selection for appropriate choices of $\mathbf{A}$.

In order to leverage existing large scale optimizers, we follow the standard procedure (Chapelle et al., 2002) of reformulating the primal as a nested two step optimization. In the outer loop, the kernel is learnt by optimizing over $\mathbf{d}$ while, in the inner loop, the kernel is held fixed and the SVM parameters are learnt.

B. Generalized Discriminant Analysis

Linear discriminant Analysis (LDA) seeks directions on which the data points of different classes are far from each other while requiring data points of the same class to be close to each other. Suppose we have a set of l samples $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_l\in\mathbf{R}^n$, belonging to c classes. The objective function of LDA is as follows:

$$\mathbf{a}_{opt}=\arg\max_{\mathbf{a}}\frac{\mathbf{a}^T S_b\mathbf{a}}{\mathbf{a}^T S_w\mathbf{a}}, \tag{9}$$

$$S_b=\sum_{k=1}^{c} l_k(\mu^{(k)}-\mu)(\mu^{(k)}-\mu)^T, \tag{10}$$

$$S_w=\sum_{k=1}^{c}\sum_{i=1}^{l_k}(\mathbf{x}_i^{(k)}-\mu^{(k)})(\mathbf{x}_i^{(k)}-\mu^{(k)})^T, \tag{11}$$

where S_w stands for the within-class scatter matrix and S_b the between-class scatter matrix. In Eqn. (10) and (11), μ is the total sample mean vector, l_k is the number of samples in the $k-th$ class, $\mu^{(k)}$ is the average vector of the $k-th$ class, and $\mathbf{x}_i^{(k)}$ is the $i-th$ sample in the $k-th$ class. Define the total scatter matrix $S_t=\sum_{i=1}^{l}(\mathbf{x}_i-\mu)(\mathbf{x}_i-\mu)^T$, we have $S_t=S_b+S_w$ (Fukunaga, 1990). The objective function of LDA in Eqn. (9) is equivalent to

$$\mathbf{a}_{opt}=\arg\max_{\mathbf{a}}\frac{\mathbf{a}^T S_b\mathbf{a}}{\mathbf{a}^T S_t\mathbf{a}}. \tag{12}$$

The optimal $\mathbf{a}'s$ are the eigenvectors corresponding to the non-zero eigenvalue of eigen-problem:

$$S_b\mathbf{a}=\lambda S_t\mathbf{a}. \tag{13}$$

GDA (Baudat and Anouar, 2000) extends LDA to non-linear mappings. The data, given as the points $\mathbf{x}_i$, can be mapped to a new feature space, F, via some function ϕ. In this new feature space, the function that needs to be maximized is

$$J(\mathbf{a})=\frac{\mathbf{a}^{\mathrm{T}}\mathbf{S}_b^{\phi}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{S}_w^{\phi}\mathbf{a}}, \tag{14}$$

where

$$\mathbf{S}_b^{\phi}=\sum_{k=1}^{c} l_k(\phi(\mu^{(k)})-\phi(\mu))(\phi(\mu^{(k)})-\phi(\mu))^T, \tag{15}$$

$$\mathbf{S}_w^{\phi}=\sum_{k=1}^{c}\sum_{i=1}^{l_k}(\phi(\mathbf{x}_i^{(k)})-\phi(\mu^{(k)}))(\phi(\mathbf{x}_i^{(k)})-\phi(\mu^{(k)}))^T. \tag{16}$$

Ⅳ. 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. For example, company A failed in 2007 and company B failed in 2009. These two companies are pooled with their matched companies A', B' in a single file labeled F0 representing their financial status in the year of bankruptcy (annual financial reports served as the input data). Companies A and A' (or companies B and B') are traced backward for five years, namely, through years 2006, 2005, 2004, 2003, and 2002 (for companies B and B', we traced the years 2008, 2007, 2006, 2005, 2004). These data were put in separate files labeled F1, F2, F3, F4, and F5 respectively for classification. Where file F1 pooled the 2006 annual reports of companies A and A', and the 2008 annual reports of B and B'. Similarly, file F2 pooled the 2005 annual reports of companies A and A', and the 2007 annual reports of companies B and B'. 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 decision trees (J48), nearest neighbors (KNN), logistic regressions (LR), Bayesian networks (BN), neural networks (NN) 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
NN	70.76	69.01	69.59	65.50	59.65
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. The average performance of every methods is listed in Table 3. Figure 1 is the performance comparison. From Table 3 and Figure 1, it's clear that our new system, GMKM on GDA subspace, significantly outperforms traditional classifiers.

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

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

	1st year	2nd year	3rd year	4th year	5th year
The system	100	98.25	96.49	94.74	95.24

Table 3. Average performance of every methods (accuracy %)

	Average
SVM	69.94
NN	66.90
BN	66.90
LR	71.70
J48	81.17
KNN	60.47
The system	96.94

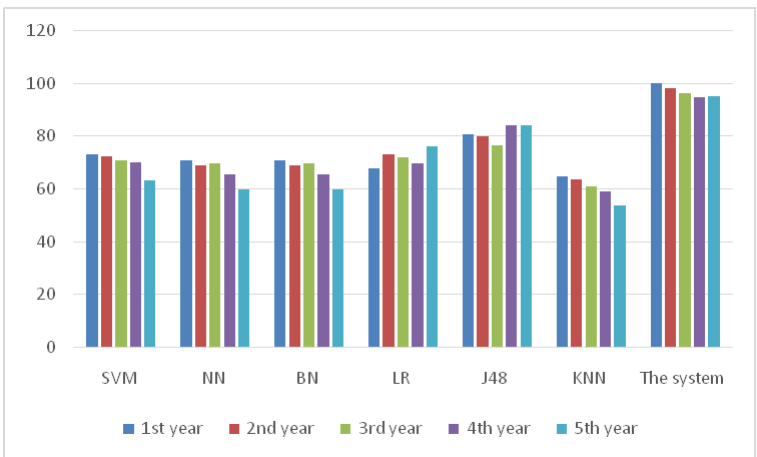

Figure 1. Performance comparision of all models (accuracy %)

information for classification. It is more effective to consider nonlinear subspace learning (such as GDA) and multiple kernel classifiers. The basis vectors found by GDA are optimal for GMKM and significantly improves its performance.

Ⅴ. 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 GMKM on the subspace of GDA for high-dimensional data mining. GDA extracted representative subspaces that optimally discriminate the output labels, significantly reduce the computational loading of GMKM, and simultaneously enhance its performance. Empirical results indicate that, compared to other classifiers, the proposed system performs best, and are more robust. 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

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), 65-74.

Baudat, G., & Anouar, F., (2000). Generalized Discriminant Analysis Using a Kernel Approach. *Neural Computation, 12*(10), 2385-2404.

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

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S., (2002). Choosing multiple parameters for Support Vector Machines. *Machine Learning, 46*, 131-159.

Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition. Academic Press, 2nd edition.

Hua, Z., Wang, Y., Xu, X., Zhang, B., Liang, L., (2007). Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Systems with Applications, 33*(2), 434-440.

Lanckriet, G.R.G., Cristianini, N., Ghaoui, L.E., Bartlett, P., & Jordan, M.I., (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research, 5*, 27-72.

Schoelkopf, B., Burges, C.J.C., & Smola, A.J., (1999). *Advances in kernel methods - Press*, Cambridge, MA.

Vapnik, V.N., (1999). *The nature of statistical learning theory*.

second Edition, New York, Springer.

Varma, M., & Babu, B.R., (2009). *More generality in efficient multiple kernel learning*. In Proceedings of the International Conference on Machine Learning, Montreal, Canada, pages 1065-1072, June 2009.

Varma, M., & Ray, D., (2007). Learning the discriminative power-invariance trade-off. International Conference on Computer Vision, October 2007.

Wang, W., & Yang, J., (2005). Mining High-Dimensional Data, Data Mining and Knowledge Discovery Handbook, 793-799.

Witten, I.H., & Frank, E., (2005). Data Mining: Practical Machine Learning Tools and Techniques (Second Edition).

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.