Investigation of the Impact of Data Comparability on Performance of Support Vector Machine Models for Credit Scoring

Yanwen Dong 1, Xiyong Hao 1, and Hideo Sato 2

1 Cluster of Science and Technology, Fukushima University, Fukushima City, Japan
2 Graduate School of Symbiotic Systems Science and Technology, Fukushima University, Fukushima City, Japan

Abstract: Although most of the studies of support vector machines (SVMs) models focused on the algorithm improvement or parameters tuning, the performance of SVM models also depends on datasets, based on which the models were constructed. This paper investi gates the impact of data comparability on performance of SVM models for credit scoring. After giving several examinations into data comparability and its impairing factors, we collect two practical datasets for credit scoring and then carry out several experiments to construct and test SVM models. According to the experiments’ results, it has been clarified that SVM models can classify training datasets perfectly whatever data comparability may be, if we choose appropriate kernel function and related parameters. However, the performance of SVM models to classify new data depends heavily on data comparability. If data comparability is low, the accuracy for classifying test datasets is proportionally low and fluctuates irregularly. It is obvious that guaranteeing data comparability is more important and effective than improving algorithm or turning parameters of SVM models.

Key Words: Support vector machine, Data comparability, Credit scoring, Class imbalance

1. Introduction

In today’s increasingly competitive business environmental, successful risk management is very important. Credit risk is most simply defined as the potential that counterparty will fail to meet its obligations in accordance with agreed terms. Because there are many types of counterparties – from individuals to sovereign governments – and many different types of obligations – from auto loans to derivative transactions – credit risk takes many forms. Credit risk assessment has attracted significant attention from managers at various organizations around the world. This increased interest has been in no small part caused by the weaknesses of existing risk management techniques that have been revealed by the recent financial crisis such as the U.S. subprime mortgage crisis of 2008-2012 and the growing demand for consumer credit [1].

There has been significant prior research on credit analysis or credit assessment [2]. The models and methodologies published so far for credit risk assessment generally fall into two categories: default models and credit scoring models. Default models assess the likelihood or probability of default (bankruptcy) by an obligor. Credit scoring models are used to assess the credit quality of counterparty. Credit scoring is usually applied to individuals or small businesses. Default models are applied more to larger credits such as corporation or sovereigns.

A variety of traditional statistical models and artificial intelligence approaches have been applied to default prediction and credit scoring, which involve multiple discriminant analysis (MDA) models, logistic regression models, k-nearest neighbor models, decision tree models, neural network models, case-based reasoning (CBR) and genetic programming models [1]-[4]. Recently, many studies tried to propose hybrid approaches to construct more effective default prediction and credit scoring models [5].

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. SVMs were first suggested by Vapnik [6] and have yielded excellent generalization performance on a wide range of problems. SVMs have been applied for bankruptcy prediction and many articles reported that SVMs are competitive and outperform other classifiers (including neural networks and linear discriminant classifier) in terms of generalization performance [7],[8].

Meanwhile, many applications of SVMs to credit scoring have been reported. Van Gestel, et al.[9] used least squares support vector machines (LS-SVMs) to credit rating of banks and reported significantly better results when contrasted with the classical techniques. The bank dataset was retrieved from BankScope and consists of 8 year data (1995-2002) of Moody’s financial strength rating for 831 financial institutions. Lee [10] applied SVMs to the corporate credit rating problem and compared its performance with MDA, CBR, and three-layer fully connected back-propagation neural networks (BPNs). He used a grid-search technique to find out the optimal parameter values of RBF kernel function of SVM and showed that SVM outperforms the other methods.

Some efforts have also been made to propose hybrid models of combining SVMs with other methods [11]-[13]. Huang, et al.[12] used three strategies to construct the hybrid SVM-based credit scoring models to evaluate the applicant’s credit score from the applicant’s input features. Two credit datasets...
in UCI database are selected as the experimental data. Combining genetic algorithms with SVMs classifier, the proposed hybrid GA-SVM strategy can simultaneously perform feature selection task and model parameters optimization. Their experimental results showed that SVM is a promising addition to the existing data mining methods.

Although a great many studies have applied the SVMs to credit scoring and showed that the SVMs achieved better performance than traditional statistical models and artificial intelligence methods, there remains serval issues to be addressed. The most important issue is how to improve the accuracy or hit rate for classifying test datasets. Like other machine learning methods, SVMs can learn well from known data and show very high accuracy for classifying training datasets on which the models were constructed. In contrast to this, SVM models cannot give adequate accuracy for classifying new data or test datasets.

This paper investigates the impact of data comparability on performance of SVM models for credit scoring. We shortly review the recent studies to improve the generalization ability of SVM models, and then give several examinations into data comparability and its impairing factors. We conduct several experiments using two practical datasets for credit scoring to demonstrate how data comparability give a strong impact on the performance of SVM models.

The remainder of this paper is presented as follows. Section 2 gives a brief review of recent literatures concerning the improvement of SVM models for credit scoring, and describes the aim and contributions of this study. In section 3, we introduce the basic aspects of the SVM method. Several examinations into data comparability and its impairing factors are presented in Section 4. The details of two practical datasets for credit scoring are highlighted in section 5. Section 6 describes the experiments and results, as well as some discussions on the results. Section 7 summarize this study and give some concluding remarks.

2. Aim and Contributions of This Study
As described above, though many applications of SVM models for credit scoring have been proposed so far, there is an important problem to be solved that SVM models could only give a comparably lower accuracy for classifying test datasets. It is usual that credit-scoring models result in higher misclassification rates than would normally be considered acceptable. According to the latest results given by Harris [14], he used the clustered support vector machine (CSVM) for credit scorecard development and applied the model to the German credit scoring dataset taken from the UCI Machine Learning Repository. He demonstrated that the CSVM can achieve comparable levels of classification performance to other nonlinear SVM based techniques. However, the mean of accuracies for classifying training datasets and test datasets are 84.5% and 71.1% (CSVM-RBF) or 79.3% and 76.3% (CSVM-linear) respectively.

As same as the results of Harris [14], accuracy of SVM models for classifying test datasets is usually in the range of 70%–80%. In order to improve this accuracy, some studies have been made toward enhancing SVM models. Shi, et al.[15] proposed a novel feature-weighted support vector machine (SVM) credit scoring model, in which an F-score is adopted for feature importance ranking. Focused on the performance of SVM models depends on their parameters’ setting, Zhou, et al.[16] used direct search method to optimize the SVM-based credit scoring model and compared it with other three parameters optimization methods, such as grid search, method based on design of experiment (DOE) and genetic algorithm (GA). They showed that the direct search method can find the effective model with high classification accuracy and good robustness and keep less dependency on the initial search space or point setting. Furthermore, Okada, et al.[17],[18] proposed new approaches for failure discriminant analysis. The basic idea is to separate multi-dimensional financial data corresponding to ongoing and failed companies by an ellipsoidal surface which is estimated by using a maximal margin hyper plane.

Besides the studies described above, many other improvement to SVM models have also been made. However, there is not yet any effective method that can enhance remarkably the performance of SVM models for credit scoring. In fact, most of the studies reported so far focused only on the algorithm or parameters related the SVM models, but the performance of credit scoring models depends also on the datasets, based on which the models were constructed.

This paper aims at investigating the impact of data comparability on performance of SVM models for credit scoring, and making the following contributions:

(1) As there is an absence of literatures to consider the impact of data comparability on performance of SVM models, this study change the viewpoint from algorithm or parameters into data comparability. The basic viewpoint is different from all of previous studies.

(2) We give several examinations into data comparability and discuss its impairing factors. In addition to this, we consider the reason why SVM models for credit scoring could only classify test datasets with comparably low accuracy.

(3) We carry out several experiments using two practical datasets for credit scoring and demonstrate how the data comparability affect the performance of SVM models through the experiments results.

3. Support Vector Machines
Support vector machine (SVM) is a novel learning machine based on statistical learning theory, and its basic concept can be briefly introduced as the following [6].

Given a training set \( D = \{x_i, y_i; i = 1,2,\cdots,N\} \), where \( x_i = (x_{i1}, x_{i2}, \cdots, x_{in}) \in \mathbb{R}^n \) are input vectors and \( y_i \in \{+1, -1\} \) are target labels representing two classes (for example, +1="healthy customer", -1="insolvent customer"), SVM aims at finding an unknown decision function

\[
y = \text{sign} \left( \sum_{j=1}^{N} w_j \phi_j(x) + b \right) \tag{1}
\]

where the nonlinear functions

\[
\phi(x) = [\phi_1(x), \phi_2(x), \cdots, \phi_N(x)]^T \tag{2}
\]

represent the features of the input space. Parameter \( w_j \) denotes a set of linear weights connecting the feature space to the output space, and \( b \) is the bias or threshold.
Let \( f(x) = \sum_{j=1}^{M} w_j \phi_j(x) + b \) be the classifying function, we have to construct a classification hyperplane (or hypersurface) and obtain the classifying function \( f(x) \). If the data set \( D \) is linearly separable, the perfect classification hyperplane exists and can be derived from maximizing the margin \( 2/||u||^2 \), or minimizing \( ||u||^2/2 \).

If there exists no hyperplane that can split \( \{1,-1\} \) classes, the soft margin method will choose a hyperplane that splits the examples as cleanly as possible, while still maximizing the distance to the nearest cleanly split examples. The method introduces non-negative slack variables \( \xi_i \geq 0 \), which measure the degree of misclassification of the data \( x_i \). The classifying function \( f(x) \) can be obtained by solving the following optimization problem:

\[
\begin{align*}
\text{Minimize} & \quad \frac{1}{2} ||u||^2 + C \sum_{i=1}^{N} \xi_i \\
\text{Subject to:} & \quad y_i(w^T \phi(x_i) + b) + \xi_i \geq 1, \quad i = 1, 2, \ldots, N
\end{align*}
\]

This is the primal problem of the SVM and its solution can be easily obtained by solving the following dual problem:

\[
\begin{align*}
\text{Minimize} & \quad \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i \\
\text{Subject to:} & \quad \sum_{i=1}^{N} y_i \alpha_i = 0 \\
& \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, N
\end{align*}
\]

where \( \alpha_i \) is Lagrange multipliers. Then we can use the optimal solution \( \alpha_i^* \) to obtain the nonlinear classifying (score) function:

\[
f(x) = \sum_{j=1}^{N} \alpha_j^* K(x, x_j) + b
\]

where \( K(x, x_j) = \phi^T(x) \phi(x_j) \) is the kernel function, and the following three kernel functions are frequently used.

1. **The Polynomial function kernel**

\[
K(x, x_j) = (x^T x_j + 1)^\rho
\]

where \( \rho \) is the degree of the polynomial kernel.

2. **The Gaussian radial basis function kernel**

\[
K(x, x_j) = \exp(-\sigma^2||x - x_j||^2)
\]

where \( \sigma \) is a free parameter.

3. **The Laplace radial basis kernel**

\[
K(x, x_j) = \exp(-\sigma||x - x_j||)
\]

where \( \sigma \) is a free parameter.

### 4. Data Comparability and Its Deteriorating Factors

#### 4.1 Meaning of Data Comparability

In general, data are facts and statistics collected together for reference or analysis. In the field of credit scoring, data are mappings from economic/social phenomena or objects’ attributes to numbers or characters. Under this view, two objects have comparable data if their mappings are similar. Data comparability is usually described as a quality of the similarity between two or more pieces of data in the situations of comparing data about an entity with similar data about other entities and with similar data about the same entity for another period or another date. It is the typical case that the firms report similar accounting numbers for similar economic events and appropriately different accounting numbers for different economic events.

In order to enhance the comparability of financial statements, the Financial Accounting Standards Board (FASB) has established and improved many standards of financial accounting and reporting. In Statement of Financial Accounting Concepts No. 8 [19], FASB states that the objective of general-purpose financial reporting is to provide the reporting firm’s investors and creditors with information that enables them to assess “the amount, timing, and uncertainty of (the prospects for) future net cash inflows to the entity”. Information that can be compared across firms and time, enabling “users to identify and understand similarities in, and differences among, items,” is more likely to satisfy this objective.

FASB argues also that “If financial information is to be useful, it must be relevant and faithfully represent what it purports to represent. The usefulness of financial information is enhanced if it is comparable, verifiable, timely, and understandable.” For “information to be comparable, like things must look alike and different things must look different. Comparability of financial information is not enhanced by making unlike things look alike any more than it is enhanced by making like things look different”.

#### 4.2 Impairing factors of data comparability

Data comparability can be impaired by many factors. The most important factor is the error or noise existing in data. In many situations, we almost cannot collect correct data. When gathering data by means of personal interview or questionnaires, several error sources exist [20].

1. Both the interviewer and the subject can be misunderstood by each other;
2. In the situation of mutual interaction, one or the other may lead the reactions of the counterpart, probably by nonverbal suggestions;
3. The subject may consciously give wrong information, being dependent on the contents of the questions;
4. Slips of memory;
5. Somebody else to fill in the questionnaire than the respondent to whom the questionnaire was mailed.

Non comparability of financial data can arise from Generally Accepted Accounting Principles (GAAP), firms’ application of GAAP, and transaction structuring. GAAP standards often include bright-line criteria that require different accounting treatments for economically similar transactions. This means that GAAP permits alternative accounting methods for the same economic phenomenon, therefore a single economic phenomenon can be faithfully represented in multiple ways.

Many GAAP standards only cover transactions with specified characteristics, even though the transactions are economically similar to transactions with distinct characteristics covered by other standards. Firms may structure their transactions to fall
on the sides of the bright lines or to exhibit the characteristics that yield the desired accounting treatments. Companies may choose among alternative accounting methods allowed by GAAP or exercise judgment over accrual estimates, and their choices may impair rather than enhance comparability.

Data comparability can also be impaired by the timing difference. In accounting under the traditional historical cost paradigm, historical cost is the original nominal monetary value of an economic item. Assets and liabilities may be shown at their historical cost, as if there had been no change in value since the date of acquisition. Even if some assets or liabilities were virtually identical, their historical costs may differ largely, deepening on the date of acquisition. Therefore the balance sheet value of these assets or liabilities may differ from the real value.

This timing difference also appears in another form when some social or economic indexes change notably as time goes by. For example, if economic situation changes conspicuously, monthly income for the same person with the same job and the same position may change largely. Therefore, the amount of monthly income deepends on the date of data collection [21].

In normal business circumstance, only a few firms are going into bankruptcy or failure. When collecting data for credit scoring, the samples of healthy firms’ class significantly outnumber the numbers of insolvent firms’ class. As the result, we can only obtain a class imbalanced dataset to construct credit scoring models. Since most of the standard classification algorithms assume the underlying distribution of data is identical and are proposed to classify balanced training datasets, class imbalanced datasets cause most learning algorithms perform less optimally. The classification models achieve a good coverage of the majority samples (healthy firms), whereas the minority ones (insolvent firms) are misclassified frequently. Therefore, those algorithms, which obtain a good behavior on balanced datasets, do not necessarily achieve the best performance for imbalanced datasets.

In order to avoid an imbalanced dataset, a lot of prior studies tried to increase the samples of insolvent firms intentionally through collecting data of insolvent firms over a longer period or from different industries. However, the longer the period is, the stronger the effect of timing difference become. If collecting data from different industries, the financial data with the same account title is really mappings of different economic events, because the companies in different industries provide different product/service and their business processes differ from each other.

Data comparability is the basic prerequisite for constructing credit scoring models. If this prerequisite is not guaranteed, credit scoring models cannot achieve satisfactory performance. This is an important reason why SVM models for credit scoring could not achieve the best accuracy for classifying test datasets in most of prior studies.

5. Two Datasets

In order to investigate the impact of data comparability on the performance of SVM models, we prepared two datasets. One is 498 customers’ data which is used to assess their credit in a small company, and this dataset has a very high comparability. The other is 54 companies’ data that were collected from real estate industry and construction industry, and its comparability is very low.

(1) 498 customers of a small company (SC498)

In a small company of Japan, the main business is selling school uniforms and accessories at wholesale. There are 20 employees in the company, and the annual sale is about 600 million Japanese yen. Orders come from about 800 customers that are classified into three types: retailers, schools and others [4].

The customers’ credit is assessed through a four-grade credit score: score of 1 meaning a healthy customer for which all orders are accepted, score of 2 meaning a customer for which orders are accepted and limited to a given amount, score of 3 meaning a customer for which orders are accepted only in cash sale, and score of 4 meaning an insolvent customer for which all of orders are rejected.

It is almost impossible to obtain the customers’ financial data because most of them are minor small businesses without disclosure of financial information. It is also difficult to frequently ask an agency for evaluating customers’ credit due to limited budget. For these reasons, we extracted the following eight features of the customers from the database of the management information system and assess the customers’ credit based on these features data.

- $x_1, x_2$: dummy variables representing the type of customers as shown in table 1.

| Type of customers | Value of $x_1$ | Value of $x_2$ |
|-------------------|----------------|----------------|
| retailer          | 0              | 0              |
| school            | 0              | 1              |
| others            | 1              | 0              |

- $x_3$: average amount of overdue payment in the year considered.
- $x_4$: maximum overdue days for all of overdue payment in the year considered.
- $x_5$: number of times that overdue payment occurs in the year considered.
- $x_6$: total sales in the year considered.
- $x_7$: rate of the average amount of overdue payment to the total sales, i.e. $x_7 = x_3/x_6$.
- $x_8$: number of transaction months that any order from the customer is fulfilled in the year considered.

Furthermore, features $x_3 \sim x_8$ are dimensionless quantities derived by the following standardization equation (13).

$$x_i = \frac{\text{raw value of } x_i - \text{the mean of raw value of } x_i}{\text{the standard deviation of raw value of } x_i}$$

Table 2 shows basic statistics of the dataset SC498, which is the data collected in the 2001 financial year. As these features data were extracted from the daily transaction data of the company according to the identical method and from the same period, its comparability is very high.

(2) 54 companies of real estate and construction industry (REC54)

In Japan, bankruptcy probability of real estate industry and construction industry is relatively higher than other industries.
We collected the data of 27 insolvent companies from real estate/construction industry in Japan, these companies applied bankruptcy law from 2008 to 2009. Then, we selected randomly 27 healthy companies from the same two industries and collected their data within the same period [22]. The dataset consists of the following nine financial items:

- ROA
- ROE
- Current ratio
- Quick ratio or acid test ratio
- Debt-to-equity ratio
- Capital-asset ratio
- Turnover of total capital
- Income growth
- Profit growth

The comparability of this dataset is relatively low because the business systems of real estate industry and construction industry differ from other industries in Japan:

- There are two alternative accounting methods: percentage of completion method and completed contract method, for the companies of these two industries to amount their sales and profits. The sales and profits vary largely according to the selected accounting method.
- In real estate industry and construction industry, prices of products or services depend not only on the market but also on some political factors. There is not fair prices in these two industries and therefore, it is difficult to compare directly sales of two companies.
- Some major companies act as the original contractors and underneath them there are a large number of small and medium sized subcontractors and sub-subcontractors. Major companies usually provide security to liabilities of subcontractors and sub-subcontractors. These liabilities are not reflected in the financial data.

### Table 2: Basic statistics of the dataset SC498

| Feature                          | Mean       | Standard deviation | Number of customers |
|----------------------------------|------------|--------------------|---------------------|
| x₁(amount of overdue payment)    | 196,171    | 762,532            | Credit score 1     |
| x₂(overdue time in days)         | 56.37      | 73.30              | 474                 |
| x₃(number of times overdue)       | 2.92       | 3.06               | 2                   |
| x₄(total sales)                  | 1,201,390  | 3,706,057          | 3                   |
| x₅(overdue rate)                 | 0.22       | 0.37               | 20                  |
| x₆(number of transaction months)  | 5.56       | 4.14               | Total 498           |

Table 3 and table 4 shows the classification performance of SVM models on the dataset REC54 and SC498 respectively, where k is the number of subsets for k-fold cross-validation, and sigma shows the good value of σ for the Gaussian RBF or Laplace kernel, which was decided using the heuristics in sigest function of package kernlab.

From table 3 and table 4, it is obvious that:

1. Training accuracy for classifying the dataset REC54 ranged from 77% to 100%. If appropriate kernel function and related parameters were chosen, SVM models can classify the training dataset perfectly. Training accuracy for classifying the dataset SC498 ranged from 94% to 99%. Although training accuracy on the dataset SC498 could not achieve 100%, it is still a very good and satisfactory because the dataset SC498 is an imbalanced one. According to the research reported by Hao, et al.[23], the class imbalance in dataset leads to a notable deterioration on the performance of classification methods.

2. As it is more important to recognize insolvent customers or companies than healthy ones, the classification error (%) for insolvent customers or companies in training datasets was also showed in table 3 and table 4 (error for insolvent). It is weakly proportional to training accuracy. This error could be decreased into zero for the balanced datasets if we choose appropriate kernel and related parameters of SVM models.

3. The average accuracy (test accuracy) of k-fold cross-validation (k=5 or 10) for the test datasets represents the performance of SVM models to classify new data. As shown in table 3, the test accuracy for classifying the dataset REC54 ranged from 56% to 63%, even though the training accuracy for the same dataset could achieve 100%. There is a big difference between the test accuracy and training accuracy for the dataset REC54. On the other hand, as shown in table 4, the test accuracy for classifying the dataset SC498 ranged from 96% to 97%, there is less difference from the training accuracy.

6. Experiment and Results Discussion

In this study, we applied the SVM model of equations (5)-(7) to the two datasets SC498 and REC54 described above. The Gaussian radial basis function (11) and the Laplace radial basis function (12) were used as kernel. We conducted the SVM analysis in R, concretely used the ksvm function in package kernlab. As k-fold cross-validation of an SVM model is also supported by ksvm function, the dataset is divided into k subsets, and the learning and validation of the SVM model is repeated k times. Each time, one of the k subsets is used as the test set and the other k – 1 subsets are put together to form a training set. Then the average error across all k trials is computed.

As the SVM models are identical for the datasets REC54 and SC498, it is reasonable to consider that the difference of test accuracy between these two dataset was caused by the difference in the property of these two datasets rather than the models. As described above, dataset REC54 is in lower comparability, this is the true reason to bring out the poor performance for classifying test datasets.

Besides the above observations, we can also find out that the SVM models with the Laplace radial basis kernel function outperform the models with the Gaussian radial basis kernel function. The impact of the parameters C and σ on the performance
of SVM models could not be reflected clearly because there was only six results in table 3 and table 4, and \( \sigma \) was calculated automatically and heuristically.

In order to examine the impact of the parameters \( C \) and \( \sigma \), we carried out a grid-search giving \( C = 1, 2, \cdots, 20; \sigma = 0.1, 0.2, \cdots, 1.0. \) 5-fold cross-validation was used and the average of accuracy for training data and test data were calculated.

Figure 1 and figure 2 shows the grid-search results for the dataset REC54 and SC498 respectively. From figure 1 and figure 2, we can observe that:

(1) The accuracies for the training datasets have a clear correlation to the parameters. The training accuracies are proportional to \( C \). Meanwhile, larger values of \( \sigma \) lead to higher training accuracies. However, both \( C \) and \( \sigma \) has a saturated value, over which training accuracies cannot be improved.

(2) The relation between the parameters and the test accuracies is very complex. The test accuracies vary irregularly when the parameters increased. The test accuracy for dataset REC54 ranges from 48.0% to 68.9% and achieves its maximum at \( C = 6 \) and \( \sigma = 0.6 \). On the other hand, the test accuracy for dataset SC498 varies in a very narrow ranges and achieves its maximum 97.4% at \( C = 9 \) and \( \sigma = 0.2 \). As the average of the test accuracy for dataset SC498 is 96.8% and the minimum is 95.6%, it is obvious that the test accuracy for dataset SC498 is not sensitive to the parameters setting.

(3) It is reasonable to consider that data comparability is the main reason that lead to the difference in the relation between test accuracy and parameters. The higher comparability of dataset SC498 can guarantee its sub datasets having the same higher comparability, therefore the test accuracy for dataset SC498 was very high and not sensitive to parameters. In contrast to this, dataset REC54 has a lower comparability and its sub datasets differ from each other. As the result, the test accuracy for dataset REC54 was relatively low and varied irregularly according to the sub datasets generated randomly.

In the above experiments, we used \( k \)-fold cross-validation and generated sub datasets for training and test randomly, based on these sub datasets the SVM models were constructed and tested. As the experiment results depended on the sub datasets generated randomly, we cannot always obtain the same experiment results as described above. However, after carrying out the above experiments repeatedly, it could be confirmed that the main characteristics or points described previously have not changed except for a few difference in the values of accuracies.

7. Concluding Remarks

Most of the studies of SVM models focused on the algorithm improvement or parameters tuning. This paper investigated the impact of data comparability on performance of SVM models for credit scoring. After giving several examinations into data comparability and its impairing factors, two practical datasets: REC54 and SC498 are collected and then several experiments of SVM models were carried out. According to the experiments’ results, it has been clarified that:

(1) If we choose appropriate kernel function and related parameters, SVM models can classify the training dataset perfectly whatever data comparability may be.

(2) The performance of SVM models to classify new data depend heavily on the comparability between the new data and training data. If the comparability is low, the accuracy for classifying test dataset is proportionally low.

(3) If comparability is high, the accuracy for classifying test datasets is not only high but also insensitive to parameters setting. Meanwhile, low comparability leads to irregular fluctuation on the accuracy for test datasets.

From the experiments’ results of this paper, it is obvious that guaranteeing data comparability is more important and effective than improving algorithm or turning parameters of SVM models. However, this important issue has almost been ignored so far. In order to apply SVM models to solve practical problems successfully, we should take more time to examine data comparability. It may be necessary to develop an effective method or tool to check or investigate data comparability analytically.

Acknowledgements

This work was partly supported by Grant-in-Aid for Scientific Research (C) from the Japan Society for the Promotion of Science (JSPS KAKENHI) under Grant No. 25380497.
References

[1] Galindo, J. and Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications, Computational Economics, 15(1-2), 107-143.

[2] Crook, JN., Edelman, DB and Thomas, LC. (2007). Recent developments in consumer credit risk assessment, European Journal of Operational Research, 183 (3), 1447-1465.

[3] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627-635.

[4] Dong, Y. (2008). Development of a Customer Credit Evaluation System via Case-based Reasoning Approach, Asia-Pacific Journal of Industrial Management, 1(1), 1-7.

[5] Hsieh, N.C. (2005). Hybrid mining approach in the design of credit scoring models, Expert Systems with Applications, 28(4), 655-665.

[6] Vapnik, V. (1995). The Nature of Statistical Learning Theory, Springer-Verlag, New York.

[7] Shin, K., Lee, T.S. and Kim, H. (2005). An application of support vector machines in bankruptcy prediction model, Expert Systems with Applications, 28(1), 127-135.

[8] Min, J.H. and Lee, Y.C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters, Expert Systems with Applications, 28(4), 603-614.

[9] Van Gestel, T., Baesens, B., Garcia, J. and Van Dijcke, P. (2003). A Support Vector Machine Approach to Credit Scoring, Bank en Financiewezen, 2, 73-82.

[10] Lee, Y.C. (2007). Application of support vector machines to corporate credit rating prediction, Expert Systems with Applications, 33(1), 67-74.

[11] Huang, Z., Chen, H., Hsu, C.J., Chen, W.H. and Wu, S. (2004). Credit rating analysis with support vector machine and neural networks: A market comparative study. Decision Support Systems, 37(4), 543-558.

[12] Huang, C.L., Chen, M.C. and Wang, C.J. (2007). Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 33(4), 847-856.

[13] Liu, X., Fu, H. and Lin, W. (2010). A Modified Support Vector Machine model for Credit Scoring, International Journal of Computational Intelligence Systems, 3(6), 797-804.

[14] Harris, T. (2015). Credit scoring using the clustered support vector machine, Expert Systems With Applications, 42(2), 741-750.

[15] Shi, J., Zhang, S. and Qiu, L. (2013). Credit scoring by feature-
weighted support vector machines, Journal of Zhejiang University SCIENCE C, 14 (3), 197-204.

[16] Zhou, L., Lai, K. and Yu, L. (2009). Credit scoring using support vector machines with direct search for parameters selection, January, 13(2), 149-155.

[17] Okada, Y. and Konno, H. (2009). Failure Discrimination by Semi-Definite Programming Using a Maximal Margin Ellipsoidal Surface, Journal of Computational Finance, 12(3), 63-78.

[18] Saito, M. and Konno, H. (2013). Classification of Companies Using Maximal Margin Ellipsoidal Surfaces, Computational Optimization and Applications, 55, 469-480.

[19] Financial Accounting Standards Board (FASB) (2010). Statement of financial accounting concepts No. 8: Conceptual framework for financial reporting. Available at http://www.fasb.org/sp/FASB/DocumentC/DocumentPage?cid=1176157498129&acceptedDisclaimer=true (Accessed at April 25, 2015).

[20] Ilhalainen, O. and Vaananen, S. (1975). Comparability of Interview and Questionnaire Data, Acta Psychiatrica Scandinavica, 52(S262), 44-48.

[21] Kim, S., Kraft, P. and Ryan, S. G. (2013). Financial statement comparability and credit risk. Review of Accounting Studies, 18(3), 783-823.

[22] Dong, Y. 2015. Data science for business analysis and credit scoring. Kyoritsu Shuppan, 139-151. (in Japanese)

[23] Hao, X., Dong, Y. and Wu, S. (2014). An Experimental Study on Relationship between Classification Performance and Degree of Class Imbalance, INFORMATION, 17 (6B), 2567-2582.

Yanwen Dong

is a professor in the cluster of science and technology at Fukushima University. He received his bachelor degree in 1982 and a master degree in 1984 from University of Science and Technology Beijing, China. He also received PhD in 1996 from the Osaka Prefecture University, Japan. He worked in University of Science and Technology Beijing as a lecture from 1984 to 1995 and in Faculty of Economics, Fukushima University as an associate professor from 1997 to 2004. His current research interests include cell production system, production scheduling, data mining and management information system.

Xiying Hao

is now a researcher in the cluster of science and technology at Fukushima University. She received her master degree in 2011 from Fukushima University and also received PhD in 2014 from the Graduate School of Symbiotic System Science, Fukushima University. Her research interests include risk management, multivariate statistics, data mining, cell production system and management information system.

Hideo Sato

is now a PhD candidate of Graduate School of Symbiotic System Science, Fukushima University. He is also the Executive Managing Director in Fukushima Shinkin Bank. He received his master degree in economics and business administration from Fukushima University in Japan. His research interests include financial engineering, regional administration, credit assessment and project management.