Expert Systems With Applications 74 (2017) 105–114

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Reject inference in credit scoring using Semi-supervised Support Vector Machines

Zhiyong Li[a,e], Ye Tian[b,e], Ke Li[c], Fanyin Zhou[c], Wei Yang[d,e,*]

[a] School of Finance, Southwestern University of Finance and Economics, China
[b] School of Business Administration, Southwestern University of Finance and Economics, China
[c] Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, China
[d] School of Insurance, Southwestern University of Finance and Economics, Wenjiang, Chengdu 611130, Sichuan, China
[e] Collaborative Innovation Center of Financial Security, Southwestern University of Finance and Economics, China

ARTICLE INFO

Article history:
Received 2 October 2016
Revised 21 December 2016
Accepted 11 January 2017
Available online 12 January 2017

Keywords:
Reject inference
Credit scoring
Semi-supervised Support Vector Machines
Online lending
Predictive accuracy

ABSTRACT

Credit scoring models are commonly built on a sample of accepted applicants whose repayment and behaviour information is observable once the loan has been issued. However in practice these models are regularly applied to new applicants, which may cause sample bias. This bias is even more pronounced in online lending, where over 90% of total loan requests are rejected. Reject inference is a technique to infer the outcomes for rejected applicants and incorporate them in the scoring system, with the expectation that predictive accuracy is improved. This paper extends previous studies in two main ways: firstly, we propose a new method involving machine learning to solve the reject inference problem; secondly, the Semi-supervised Support Vector Machines model is found to improve the performance of scoring models compared to the industrial benchmark of logistic regression, based on 56,626 accepted and 563,215 rejected online consumer loans.

1. Introduction

Credit scoring as an automatic credit assessment tool has been used by lenders such as banks for decades. Various statistical and machine learning techniques (e.g. Bijak and Thomas (2012), Min and Lee (2008)) help creditors identify borrowers who cannot fulfil their financial obligations within a group of loan applicants through a process of application scoring, although other types of implementations such as behaviour scoring and profit scoring are also popular. In credit scoring, loans are normally classified into good ones and bad ones according to their rank orderings of default probabilities using *ex post* methods constructed from historic loan records. Their success in risk control has led to a flourishing within the borrowing and lending market. However, since the beginning of credit scoring, the issue has been that samples of training scoring models differ from those pertaining to applications, because lenders normally only have performance information about accepted borrowers. By predicting the default probability based solely on the accepted borrowers, sample bias may arise. Analysts have tried to make use of information from rejected borrowers to improve the prediction performance of their models. When the actual status of rejected applicants is missing, reject inference is used by credit scorers to assess the default risk of the rejection. Many scholars such as Hand and Henley (1993), Sohn and Shin (2006) and Bücker, van Kampen, and Krämer (2013) have started to pay attention to this issue, and some techniques of reject inference have already been explored by Anderson and Hardin (2013), Banasik and Crook (2007), Chen and Astebro (2012) and others.

In recent years, P2P (Peer-to-Peer) lending as a financial innovation has proved a phenomenal success in the online loan market in countries such as the US and China. A borrower with internet access can easily submit a loan request on a P2P lending platform, so crowds of lenders can thereby take different stakes in the loan based on their own perception of risk. Lenders have always had means to assess the creditworthiness of borrowers at their disposal, for example from their outward appearance (Duarte, Siegel, & Young, 2012) or other soft information (Iyer, Khwaja, Luttmer, & Shue, 2016), but obviously these are not enough to eliminate the high default risk of online loans. Therefore, platforms set up high requirements and use various risk control tools to reject unqualified borrowers. Lending Club, one of the largest P2P lending organisations, has thus far only accepted roughly 9% of all loan requests on their website. The high rate of rejection has provided us with an opportunity to investigate the problem of reject inference,

* Corresponding author.
E-mail addresses: liz@swufe.edu.cn (Z. Li), tianye@swufe.edu.cn (Y. Tian), likec@swufe.edu.cn (K. Li), zfy@swufe.edu.cn (F. Zhou), yangw@swufe.edu.cn (W. Yang).

http://dx.doi.org/10.1016/j.eswa.2017.01.011
0957-4174/

which is obviously of importance to both the organisation and the individual lenders.

This paper addresses the issue of reject inference in credit scoring in two ways. First, we propose a new method in reject inference using the machine learning technique of Semi-supervised Support Vector Machines (SSVM) to classify the status of rejected borrowers. Support Vector Machines (SVM) have received much attention in credit scoring in recent years for their use in classifying good and bad accounts (Bellotti & Crook, 2009), and its application to reject inference is also promising. Second, we have used a large number of real consumer loans, consisting of 56,626 accepted loans and 563,215 rejected loans, to assess the performance of the new method. This data basically consists of all loan requests made on the online lending platform by September 2012, so it constitutes the total population of loan applicants, rather than a limited sample of them, as has appeared in previous literature. The availability of such a wide range of data makes the investigation possible for the whole population, which we believe to be advantageous in resolving the sample bias problem.

The rest of paper is organised as follows: first a group of reject inference studies in credit scoring has been identified and summarised in the literature review. The SSVM algorithm in the reject inference context is introduced in the methodology section, and a simulation approach is involved in the analysis. Then from the evidence of real consumer loans, we compare and discuss the results from logistic regression and traditional supervised SVM. Our findings are summarised in the conclusion.

2. Literature review

One of the major concerns in credit risk modelling is the sample bias when we apply the credit model to new applicants, since the model has been developed based only on applicants who are offered a loan. This not only causes possible bias in parameter estimates, but also results in an unstable predictive performance for the application population (Crook, Edelman, & Thomas, 2007). Ideally, credit models should be built based on the total population of applicants and applied to new applicants in the real lending process. However in practice, credit modellers only have records of those applicants who pass the credit assessment, and for obvious reasons, the repayment performance of the rejected group is missing. The sample bias in this circumstance essentially becomes a missing data problem. As summarised by Feelders (1999), there are three types of 'missing' in the context of credit scoring, namely Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). Various techniques of reject inference have been developed to infer the missing status of the rejects and mitigate this bias according to different types of missing data. Although the missing data problem has drawn attention from both academics and practitioners, there has been no thorough survey on the literature in this field. We therefore discuss the problem separately according to the type of missing.

MCAR indicates that the class of the outcome is completely missing at random and is independent of both the characteristics of applicants**x**and the class label*y*. It simply means that credit is randomly offered to applicants no matter what their attributes are or how they behave in repayment. Under the MCAR condition, no selection mechanism is in execution, so no sample bias exists in the lending process. Most lenders are believed to not implement this as a long-term policy, as their businesses are unlikely to become profitable in that way.

MNAR represents another type of missing data, wherein the class of the outcome depends on not only**x**but also on*y*, given*y*is influenced by some unobserved variables. This is discussed in Heckman (1979). Manual overrides of the outcome of the model decision are typically MNAR. Loan officers have the right to alter the decision according to their overall impression of an applicant, based on personal experience or business strategy. In countries such as Germany, for instance, decisions relied purely on credit models are not in compliance with the law, but should instead be combined with human judgement. Bücker et al. (2013) represents one of the very few empirical studies that investigate the MNAR reject inference problem due to difficulties caused by unobserved variables. By adding the information of the rejected group to the predictive model, they found that parameters were significantly different and the predictions improved, which shows that ignorance of the rejected group in the scorecard building process is not appropriate when considering statistical and economic concerns.

In credit scoring practice, not surprisingly, most rejected cases are MAR, where acceptance is based on the values of characteristics **x**, as well as some arbitrary cut-offs. If the probability distributions of characteristics of both the accepted and rejected applicants are observed, Feelders (1999) proposes the Expectation-Maximisation (EM) algorithm to solve the reject inference problem. Rather than applying discriminant analysis and logistic regression based on artificial data as in Anderson and Hardin (2013), Feelders (1999) tested the EM logistic model on real consumer loan data and found that on average it outperformed augmentation alone.

If looking at the selection mechanism and the outcome mechanism separately, we have the Accept-Reject (AR) decision depending on a set of characteristics X_{AR}, and the Good-Bad (GB) decision depending on another set of characteristics X_{GB}. If X_{AR} is a subset of X_{GB}, a class can be assigned to each of the rejected applicants given a posterior probability based on X_{GB} and a cut-off. A new Good-Bad is then constructed when all cases have their own classes. This extrapolation method was used by Crook and Banasik (2004) but showed limited improvement in the predictive performance compared to the model based on the accepts alone. If no set of characteristics can determine both the AR and the GB decision, *i.e.* X_{AR} is not a subset of X_{GB}, a similar method in reject reference, that of augmentation, is frequently used. Empirical tests have been carried out by Banasik, Crook, and Thomas (2003) and Banasik and Crook (2005), using various models including logit and probit on the same dataset provided by a credit grantor where the status of the rejects were available. However, their results showed no positive evidence of augmentation in increasing the overall predictive accuracy. On the corporate credit side, Kim and Sohn (2007) found that predictive accuracy is improved by combining samples of the accepts and the rejects.

Survival analysis, another popular credit scoring method in recent years, can deal with censoring where the status of performance is not observed, as well as model the time to default, which is critical to the profitability of a loan. Sohn and Shin (2006) tried introducing the time in late repayment as a measure of default in the reject inference problem. In their model, the class of rejected applicants depends on their position on the distribution of the survival time. Banasik and Crook (2010) also tried survival analysis on multi-period consumer loan data over 40 months. However, their empirical analysis suggests this method of reject inference has a negative bearing on results.

Apart from extrapolation and augmentation, there are also other techniques such as reweighting (Bücker et al., 2013) and reclassification, where Joanes (1993) suggested reclassification be combined with augmentation to amplify their mutual strengths. Other attempts are Verstraeten and Van den Poel (2004) who found neither performance nor profitability was significantly improved by the inclusion of the rejected sample. Chen and Astebro (2012) introduced the Bayesian theory into reject inference and found that classification power was improved under the MNAR as-

Table 1
Summary of the reject inference literature.

Author	Data type	Status of rejects	No. of accepts	No. of rejects	Reject Inference approach	Classification method
Joanes (1993)	Artificial	Unknown	75	12	Reclassification	Logistic
Feelders (1999)	Artificial	Unknown	Varying	Varying	EM	QDA, Logistic
Banasik et al. (2003)	Consumer	Known	8168	4040	Augmentation	Logistic, Probit
Crook and Banasik (2004)	Consumer	Known	8168	4040	Reweighting, Extrapolation	Logistic
Verstraeten and Van den Poel (2004)	Consumer	Partially known	38,048	6306	Augmentation	Logistic
Banasik and Crook (2005)	Consumer	Known	8168	4040	Augmentation	Logistic
Sohn and Shin (2006)	Consumer	Unknown	759	10	Reclassification	Survival analysis
Banasik and Crook (2007)	Consumer	Known	8168	4040	Augmentation	Logistic, Probit
Kim and Sohn (2007)	Corporate	Known	4298	689	Extrapolation	Bivariate Probit
Banasik and Crook (2010)	Consumer	Known	147,179	Varying	Augmentation	Survival analysis
Maldonado and Paredes (2010)	Consumer	Known	800	200	Extrapolation	SVM
Chen and Astebro (2012)	Corporate	Known	4589	Varying	Bound and Collapse	Bayesian
Bücker et al. (2013)	Consumer	Unknown	3984	5667	Reweighting	Logistic
Anderson and Hardin (2013)	Consumer	Unknown	3000	1500	Augmentation, EM	Logistic

sumption. Chronologically, Table 1 gives an overview of the empirical studies of reject inference up to date.

In a more general sense, as described by Hand and Henley (1993), reject inference can only be effective if the methods are based either on extrapolation of the acceptance model to the rejects, or on the distribution of rejected applicants. Otherwise, supplementary information (*i.e.* actual status of the rejects) should be available from other sources, as in Banasik and Crook (2010), Crook and Banasik (2004), Kim and Sohn (2007) and Chen and Astebro (2012), although acquiring supplementary information can at times be costly.

In contrast to those statistical methods (Discriminant Analysis, Logistic Regression, Survival Analysis) mentioned above, machine learning techniques as a main stream of classifiers are also very popular in both consumer and corporate credit modelling. Among them, Neural Networks and Genetic Algorithm are frequently compared to conventional statistical methods, and Crook et al. (2007) and Ravi Kumar and Ravi (2007) provide a comprehensive review of them and their extensions. Among various intelligent algorithms, SVM has recently received much attention for its mathematical simplicity and least restriction, and has been applied to credit assessment in the literature, *e.g.* Bellotti and Crook (2009), Boyacioglu, Kara, and Baykan (2009), Hua, Wang, Xu, Zhang, and Liang (2007), Yang, You, and Ji (2011), Yeh, Chi, and Hsu (2010) and so on. Particularly Huang, Tang, Lee, and Chang (2012) used semi-supervised SVM to predict the probability of financial distress.

Maldonado and Paredes (2010) provide the idea of extrapolating the distance measurement in SVM between the two classes of accepted to the class of rejected. They did not, however, make use of the information of the rejects and only applied linear surface SVMs in classification, which is not suitable for most real-life loan data. On the contrary, this paper proposes a real Semi-supervised SVM model to infer the class of the rejects, which fully and directly incorporates the information of the rejects into modelling and applies it to a large real consumer loan dataset. Furthermore, we apply a kernel function to the SSVM iterative process, which is believed to be superior in inferring the results. In the next part, we introduce the details of the model based on this SSVM algorithm.

3. Methodology

3.1. Semi-supervised SVM for reject inference

A number of classification tools have been introduced in the field of consumer credit risk assessment. Such tools can help a lender to discriminate between loan applicants who are believed to be 'good' ones and those in whom the lender is insufficiently confident. The most common method used to build a classification rule is logistic regression. For a borrower i, good/bad can be denoted by a label y_i, where $y_i = 1$ means default or bad and $y_i = 0$ means good, and the characteristics of the borrower are denoted by $\mathbf{x}_i$ which consists of several attributes. Then the probability of default can be expressed by

$$P(y_i = 1 \,|\, \mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}^T)}{1 + \exp(\mathbf{x}_i \boldsymbol{\beta}^T)}, \tag{1}$$

where the parameter vector can be estimated by the maximum likelihood estimation:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \mathbf{x}_i \boldsymbol{\beta}^T - \ln[1 + \exp(\mathbf{x}_i \boldsymbol{\beta}^T)]. \tag{2}$$

Besides logistic regression, a recently popular method used for credit classification is SVM, which originally comes from the field of machine learning, first proposed by Vapnik (1998). The idea is to find two parallel hyperplanes to separate two groups of data in a multi-dimensional space and to maximise the margin between the hyperplanes. Given a training data set within which each applicant is either labelled (with known good/bad status) or unlabelled (with unobservable good/bad status), we can get a hyperplane (linear classifier) which separates all the points into two classes (good or bad) by solving the SVM mathematical programming problem. Semi-supervised SVM or transductive SVM extend basic SVM to take use of additional information in unlabelled data (Joachims, 1999). The boundary lies in the low density regions of unlabelled data and makes reference to the labelled data, though this may cause computational difficulties (Chapelle, Sindhwani, & Keerthi, 2008). A detailed discussion of SSVM issues can be found in Chapelle, Schölkopf, and Zien (2006) and Zhu and Goldberg (2009). In the context of machine learning and credit scoring, the learning problems are mainly categorised as follows:

(a) Supervised learning: In the training data set, all applicants are labelled. This actually describes a situation where the model is built based on accepted applicants whose good/bad statuses are known. In this case, a traditional supervised SVM can be built and applied to new applicants for their credit risk assessment.
(b) Unsupervised learning: In the training data set, all applicants are unlabelled. We can see the characteristics of applicants, but never know their eventual good/bad statuses. In such a case, only some unsupervised machine learning methods (*e.g.* clustering) can be used to summarise the characteristics of applicants.
(c) Semi-supervised learning: In the training data set, a part of the applicants are labelled and the rest are unlabelled. In other words, some applicants are known to be good/bad, but for the rest only the characteristics are observed. For the SSVM, structural information of both labelled and unlabelled points can be used in the optimisation process. In this way, the valuable

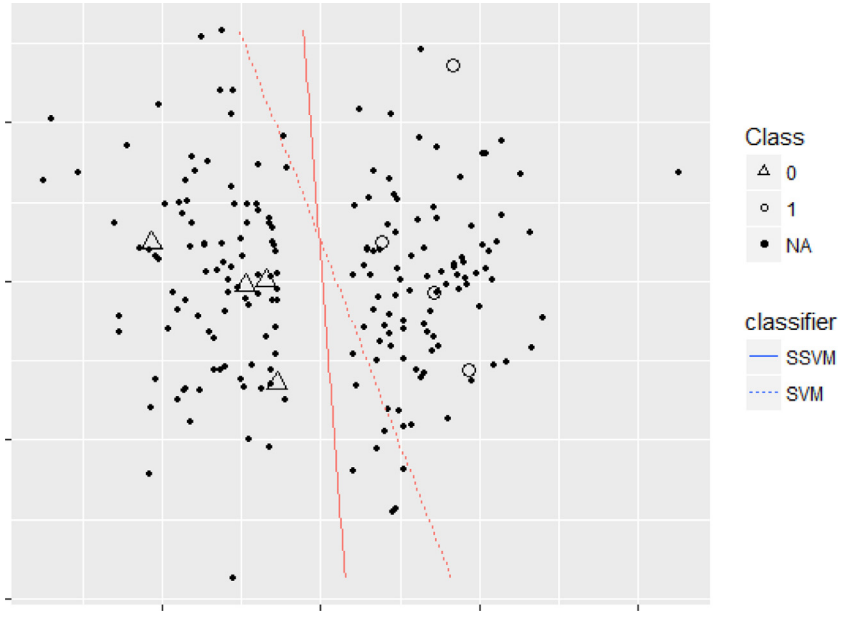

Fig. 1. Illustration of SSVM.

hidden information behind the rejected applicants will be considered. Thus, the new separation plane will help us improve decision criteria. It is worth pointing out that a learning process containing both labelled and unlabelled points is a common phenomenon in many real-world applications. This is because labelled instances are usually scarce and often difficult, expensive, or time-consuming to obtain. As denoted in Fig. 1, triangles and circles are known goods and bads while dots are rejected applicants with unknown outcomes. In such a case, a traditional supervised SVM cannot work well (dashed plane), but advanced tools such as SSVM can build up a better classification rule (full plane), which is a promising method for reject inference.

The major innovation of this paper is to propose a semi-supervised SVM approach (S3VM) (Tian & Luo, 2016), in order to conduct reject inference. Apart from a traditional logit model or supervised SVM approach, the S3VM method builds a model based directly on information of both accepted/labelled and rejected/unlabelled borrowers, which means it provides an effective way to accomplish semi-supervised learning, or more attractively to us, a natural solution for problems involving reject inference.

The main idea of an S3VM model is to find a hyperplane to maximise the margin between the two classes and keep the boundary traversing through low density regions of unlabelled data points, while still respecting labels in the input space. Compared to traditional SVM models, the S3VM model takes additional information provided by the unlabelled patterns into account to reveal more about the data structure at hand (Gieseke, Airola, Pahikkala, & Kramer, 2014). It is thus expected to perform better than traditional SVM when the data set contains unlabelled points, especially for the case which only contains a small proportion of labelled points in the training sample. For reject inference, since most rejected applicants are unknown as real good or bad customers, it is very suitable for the semi-supervised learning model.

Here we give some details of the methodology used in this paper. Firstly, the notations are introduced. Given a training data set $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, where each point $(\mathbf{x}_i, y_i) \in \mathrm{R}^m \times \{-1, 1\}$, let A_l denote the index set of the l labelled training points and $A_u = \{1, ..., n\} \backslash A_l$ be the index set of all unlabelled training points. Thus $\mathbf{y} = (\mathbf{y}_l^T, \mathbf{y}_{n-l}^T)^T$ are the corresponding label vectors, where $\mathbf{y}_l = (y_1, \ldots, y_l)^T \in \{-1, 1\}^l$ is known, but $\mathbf{y}_{n-l} = (y_{l+1}, \ldots, y_n)^T \in \{-1, 1\}^{n-l}$ is unknown. To implement non-linear models of the data, a kernel function $\phi(x) : \mathrm{R}^m \to \mathrm{R}^d$, where d is the dimension of the feature space, is used to project the original data points to a higher dimensional feature space Schölkopf & Smola, 2002). Like the traditional SVM model, positive slack variables $\boldsymbol{\eta} = (\eta_1, ..., \eta_n) \in \mathrm{R}_+^n$ are also introduced to allow for the misclassification of each point (Cortes & Vapnik, 1995), and the misclassification errors $\sum_{i=1}^n \eta_i^2$ is then penalised by $C > 0$ in the objective function in problem ((3). To maximise the margin between the two supporting hyperplanes $\boldsymbol{\omega}^T\phi(\mathbf{x}) + b = -1$ and $\boldsymbol{\omega}^T\phi(\mathbf{x}) + b = 1$, the 2-norm soft-margin S3VM model can be written as:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\boldsymbol{\omega}\|_2^2 + C\sum_{i=1}^n \eta_i^2 \\ s.t. \quad & y_i(\boldsymbol{\omega}^T\phi(\mathbf{x}_i) + b) \geq 1 - \eta_i, i = 1, \ldots, n, \\ & y_i \in \{-1, 1\}, i \in A_u. \end{aligned} \tag{3}$$

Note that a S3VM model differs from the standard SVM in terms of the last constraint in problem (3), which reflects the uncertainty of unlabelled training points when making use of their information for modelling. However, with the extra constraint, problem (3) becomes a mixed integer constrained quadratic programming (MIQP) problem and thus cannot be solved using the standard dual-problem methods for the SVM model. A number of methods have been proposed for solving the non-convex problems associated with S3VM (for example, see Bennett and Demiriz (1999) for an early solution, and Blum and Chawla (2001), Reddy, Shevade, and Murty (2011) for more recent developments). However, most of proposed methods are only efficient for small-sized problems. In order to conduct reject inference on a large scale of rejected applicants, here we consider to use the branch-and-band algorithm proposed in Tian and Luo (2016), which was shown to be more efficient for large-sized problems.

To solve for the S3VM model, the bias term b is dropped first to avoid the non-convexity in the dual problem (Bai, Niu, & Chen, 2013). Let $\alpha_i \in \mathrm{R}_+$ be the dual variable for the constraint $y_i\boldsymbol{\omega}^T\phi(\mathbf{x}_i) \geq 1 - \eta_i, i = 1, \ldots, n$. Then the Lagrangian function with respect to $\boldsymbol{\omega}$ and $\boldsymbol{\eta}$ can be written as

$$L(\boldsymbol{\omega}, \boldsymbol{\eta}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\omega}\|_2^2 + C\|\boldsymbol{\eta}\|_2^2 + \sum_{i=1}^n \alpha_i(1 - \eta_i - y_i\boldsymbol{\omega}^T\phi(\mathbf{x}_i)). \tag{4}$$

Table 2
Yearly statistics 2007–2012.

	2007	2008	2009	2010	2011	2012	Total
Accepted	603	2393	4716	8211	13,546	27,157	56,626
Default Rate	17.67%	15.91%	12.60%	9.61%	10.32%	13.67%	12.24%
Rejected	4725	21,752	46,756	92,322	191,499	206,161	563,215
Acceptance Rate	11.32%	9.91%	9.16%	8.17%	6.61%	11.64%	9.14%

We first solve the problem of $\min_{\boldsymbol{\omega},\boldsymbol{\eta}} L(\boldsymbol{\omega},\boldsymbol{\eta},\boldsymbol{\alpha})$ by using the Karush–Kuhn–Tucker (KKT) conditions of regularity and assuming $\mathbf{y}$ and $\boldsymbol{\alpha}$ as constants, we then get

$$L(\boldsymbol{\alpha}|\hat{\boldsymbol{\omega}},\hat{\boldsymbol{\eta}}) = (\mathbf{e}_n)^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T(K^* \circ \mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha} \tag{5}$$

where $K^* = K + \frac{I_n}{2C}$ Lanckriet, Cristianini, Bartlett, Ghaoui, & Jordan, 2004), $K$is a kernel matrix with elements $K_{ij} = K(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ and $\mathbf{e}_n = (1,\ldots,1)$. Based on these results, we can write the dual problem of problem ((3) as

$$\begin{aligned} &\min_{\mathbf{y}_{n-l}\in\{-1,1\}^{n-l}} \max_{\boldsymbol{\alpha}} (\mathbf{e}_n)^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T(K^* \circ \mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha} \\ &s.t.\ \boldsymbol{\alpha} \geq \mathbf{0}^n \end{aligned} \tag{6}$$

Since both K^* and $(K^*)^{-1}$ are positive definite and invertible, $K^*\circ\mathbf{y}\mathbf{y}^T$ is also positive semidefinite. Therefore, the 'max' part in problem (6) is a convex problem and the optimal solution can be derived as $\boldsymbol{\alpha}^* = (K^* \circ \mathbf{y}\mathbf{y}^T)^{-1}(\mathbf{e}_n + \boldsymbol{\mu})$, for $\boldsymbol{\mu} = (\mu_1,\ldots\mu_n)^T \in \mathrm{R}_+^n$ being the dual variable for the constraint $\boldsymbol{\alpha} \geq \mathbf{0}^n$. Replacing $\boldsymbol{\alpha}$ in problem (6) and rearranging the problem, we have

$$\begin{aligned} \min\quad & \frac{1}{2}\mathbf{u}^T Q\mathbf{u} \\ s.t.\quad & y_i u_i \geq 1,\ i = 1,\cdots,l, \\ & 1 - u_i^2 \leq 0, i = l+1,\cdots,n. \end{aligned} \tag{7}$$

where $Q = (K^*)^{-1}, \boldsymbol{\delta} = \mathbf{e}_n + \boldsymbol{\mu}$ and $\mathbf{u} = \boldsymbol{\delta} \circ \mathbf{y}$. Then the objective function of problem (7) becomes a normal convex quadratic function $\mathbf{u}^TQ\mathbf{u}$ with non-convex constrains, which can be solved by the branch-and-band algorithm proposed in Tian and Luo (2016). For a feasible solution $\boldsymbol{u}$ of problem (7), a feasible solution y can be obtained by the following equation:

$$y_i = \mathrm{sign}(u_i),\ i = 1,\cdots,n. \tag{8}$$

After obtaining the labels of all unknown points, we can solve the corresponding supervised problem to get the optimal solutions $\boldsymbol{\alpha}^*$ and b^*. Then, the decision function is given by

$$f(\mathbf{x}) = \mathrm{sign}\left(\sum_{i=1}^{n} y_i\alpha_i^* K(\mathbf{x}_i,\mathbf{x}) + b^*\right) \tag{9}$$

It is worth pointing out that this function is used for further classification on any unknown point $\mathbf{x}$. By solving the S3VM problems above, we can reach an optimised solution to discriminate new applicants into the acceptance/rejection sets. In the rest of this paper, we are going to conduct numerical tests to compare the performance of traditional logistic regression, supervised SVM and S3VM based on real-world data.

3.2. Data and variables

Lending Club in the United States is currently the world's largest online credit marketplace offering P2P lending. Based on the public data provided by Lending Club until September 2015, it has to date issued 820,398 loans with a total value over 13 billion USD, while it has rejected other 5,398,760 loan requests. The rejected group is much larger than the accepted group, so reject inference is of considerable importance under these circumstances. Our empirical data is publicly sourced from Lending Club, where characteristics of both accepts and rejects are available.

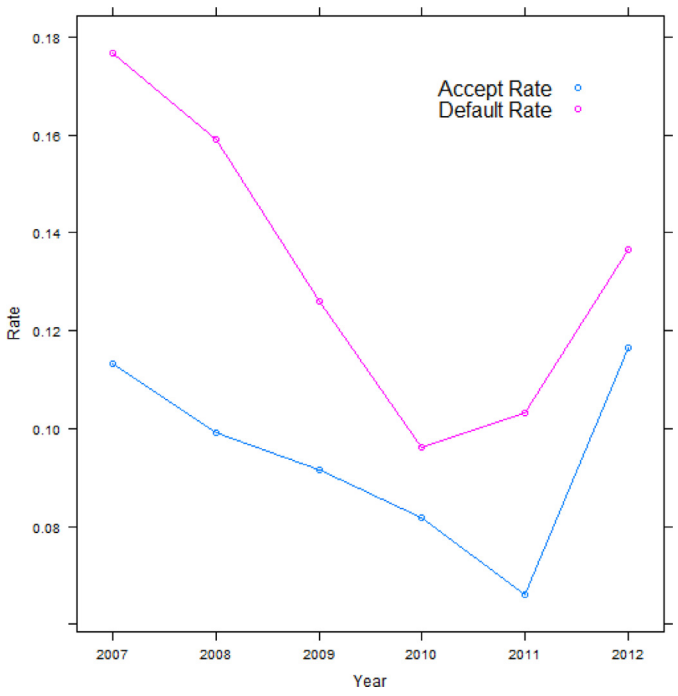

Fig. 2. Accept and default rates 2007–2012.

As we need a definite result for accepted loans, we have considered all 36-month loans issued by September 2012, so that their actual good/bad results were visible in September 2015, when the data was extracted. Excluding records containing obvious errors, this dataset consists of 56,626 issued loans including 6931 defaults, and 563,215 rejected requests.

Table 2 presents yearly statistics of the data in analysis from 2007 to 2012. Since the start of Lending Club, the default rates have varied between 9.61% and 17.67%, while the accepted rates have taken different values, from 6.61% to 11.64%. The reasons for the variation are complicated, possibly in part due to loan demand trends, the company's loan policy, the risk management system and the general economic environment. In Fig. 2, the acceptance rate and the default rate appear to be correlated, as the default rate goes down along with the acceptance rate. It is possibly because the quality of borrowers improves with the decrease of the acceptance rate, and *vice versa*. In order to verify that our proposed reject inference technique can be widely applied, we analyse the data by year to ensure that our method has been adequately exposed to all possible scenarios. In the years 2007 and 2008, the lending business as an innovative lending product was not stable and transaction volumes were small. Therefore, we exclude those two years from the following analysis to keep our models robust.

For each of the accepted records, there are both good/bad statuses and over fifty observed characteristics including personal information and interest/repayment information; while for the re-

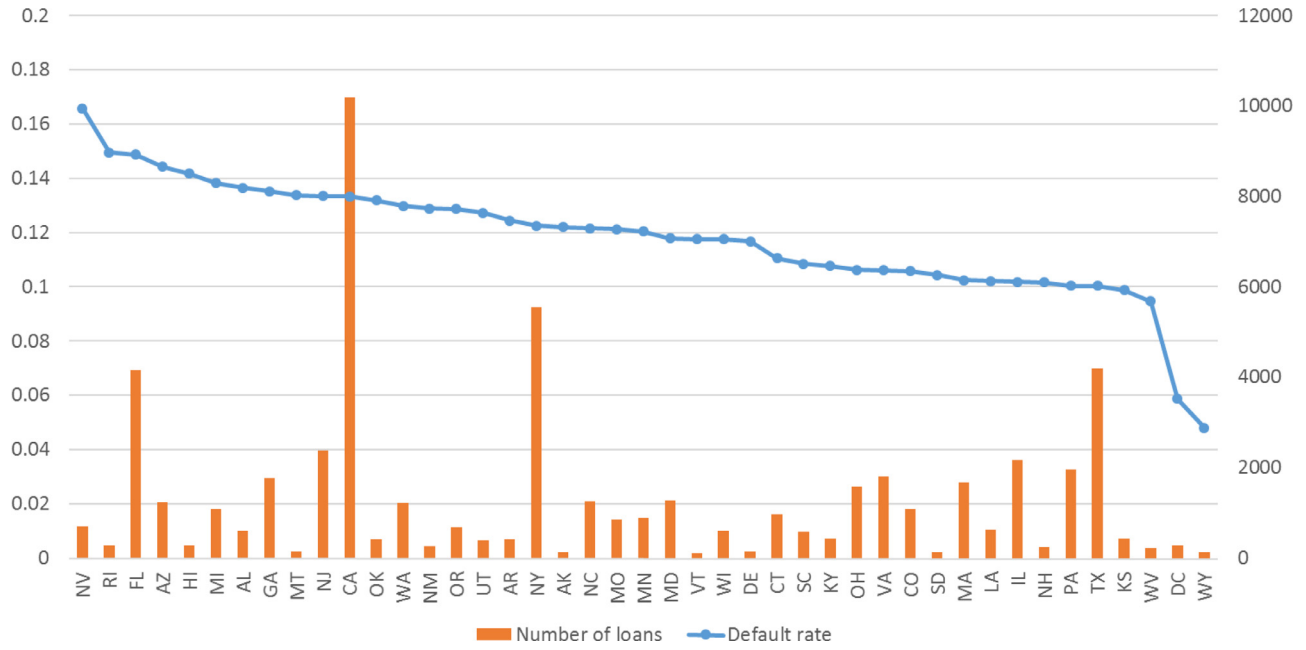

Fig. 3. Number of loans and the default rate across states.

jected ones, there are 9 characteristics of personal information, but neither good/bad status nor interest/repayment data is available. The common characteristics, with comparable and useful information, for the accepted and the rejected records are loan amount, FICO score, address state, debt to income ratio and employment length, which have been used as variables in our modelling.

The loan amount represents the listed amount of the loan applied for by the borrower, ranging from $1000 to $35,000. The FICO score ranges from 300 to 850, which is consistent with its definition. The debt to income ratio is the borrower's total monthly debt excluding mortgage and the requested loan over the borrower's monthly income. The employment length takes values between 0 and 10, where 0 means the length of employment is less than one year, and 10 means 10 or more years of employment.

We have also included the address state to show the location of borrowers in terms of 50 states, plus the District of Columbia. Geographic location has potential predictive power for default risk, as we find default risk spreads unevenly across different areas indicated either by State or by 3-digit ZIP code in the US, shown in Fig. 3. Elliehausen, Christopher Lundquist, and Staten (2007) included geographic location as part of the demographic information for borrowers in their credit scoring model, and dummies were used for nine states. However, in our data, borrowers reside in all 51 states/districts, and it is therefore not convenient to use dummies for all of them. Instead, we transform all regions into four categories with reference to the default rates in Fig. 3, namely states with rates above 0.14, from 0.12 to 0.14, from 0.1 to 0.12 and below 0.1. Three dummies (State d1, d2 and d3) are used to denote the first three categories and we leave the last group (States KS, WV, DC and WY) as the reference set. Any applicant who falls in a specific group of states is given a value 1 (0 otherwise) and only the averages of dummies are calculated in Table 3.

Descriptive statistics of these five variables for the accepted and rejected are presented in Table 3. The distributions of variables are obviously different in two groups.

3.3. Research design

The dependent variable being modelled is the good or bad indicator variable, where $y_i = 1$ if the applicant i has defaulted and $y_i = 0$ if the corresponding applicant has not defaulted. We firstly assign a good/bad label to each of the rejected applicants, assuming these are the truth, although we have no supplementary information regarding it. Based on the characteristics of applicants and corresponding labels, we compare the result from the reject inference method S3VM with that of standard classification methods, logistic regression and supervised SVM, in order to see if our new reject inference technique improves predictive accuracy. We conducted a simulation study following the standard steps proposed in Anderson and Hardin (2013) as follows:

Step 1: Build separate models of Logit and supervised SVM using the sample of all the accepted applicants, and S3VM using the sample of all accepted and rejected applicants.
Step 2: Use models from Step 1 to assign the good/bad label to the rejected applicants, seen as the truth.
Step 3: Sample equal numbers of accepts and rejects in each year.
Step 4: Split the pooled data into a training set and a test set with a ratio 7:3.
Step 5: Respectively build models of Logit, supervised SVM and proposed S3VM on the training set. Note that the Logit or supervised SVM models can only be trained using the accepted samples, whereas the S3VM model are trained on both the accepted with labels and the rejected without labels.
Step 6: Apply the classification rules derived from Step 5 to the test set.
Step 7: Test the performance on the test set.

Please note that for any SVM modelling, a hyperparameter tuning process is required. The target of tuning is to maximise the performance in the test set. SVM is tuned using grid searching for a range of penalty parameter C in problem (3), kernel degrees and parameters as such applied in Bellotti and Crook (2009).

Once the models in Steps 1 and 2 are used to determine the good/bad loan status of the rejected applicants, Steps 3–7 will be repeated 50 times in order to check whether the SSVM inference model performs better than other models without inference (Fig. 4). To eliminate the influence of the sample-size differences

Table 3
Variable statistics.

	Variable	Min	1st Quantile	Median	3rd Quantile	Max	Mean	S.D.
Accepts	Loan amount	1000	6300	10,000	15,000	35,000	11,490	6969
	Fico score	662	682	697	722	848	705.7	33.01
	Debt to income	0	10.36	15.77	21.21	34.99	15.91	7.397
	Employment length	0	2	5	10	10	5.523	3.532
	State d1						0.1202	
	State d2						0.5081	
	State d3						0.3525	
Rejects	Loan amount	1000	5000	10,000	25,000	35,000	14,900	11,057
	Fico score	385	608	652	687	850	642.7	65.05
	Debt to income	0	8.70	18.42	31.43	419.4	23.95	28.32
	Employment length	0	0	0	0	10	0.7696	2.271
	State d1						0.1219	
	State d2						0.4600	
	State d3						0.3970	

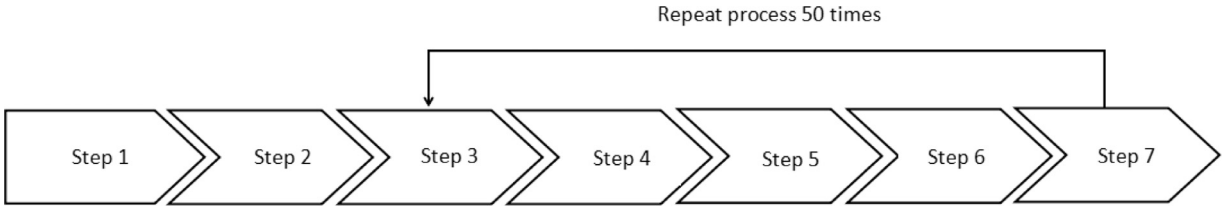

Fig. 4. Simulation process.

Table 4
A confusion matrix.

		Observation		
		Good	Bad	Total
Prediction	Good	Correct goods (True negative)	Type II error (false negative)	Goods predicted
	Bad	Type I error (false positive)	Correct bads (true positive)	Bads predicted
	Total	Goods observed	Bads observed	Sample size

Table 5
Accuracy results.

		Overall accuracy			Type I errors			Type II errors		
		Logit	*SVM*	*SSVM*	*Logit*	*SVM*	*SSVM*	*Logit*	*SVM*	*SSVM*
2009	Accepted	0.754	0.736	**0.857**	0.169	0.248	*0.032*	0.769	*0.270*	0.912
	Rejected	0.780	0.958	**0.969**	0.231	0.075	*0.025*	0.215	0.093	*0.040*
	Both	0.767	0.890	**0.913**	0.185	0.104	*0.029*	0.300	*0.215*	0.227
2010	Accepted	0.761	0.726	**0.881**	0.178	0.328	*0.027*	0.779	*0.258*	0.956
	Rejected	0.819	0.968	**0.976**	0.216	0.069	*0.029*	0.159	0.089	*0.020*
	Both	0.790	0.881	**0.929**	0.189	0.102	*0.028*	0.244	0.201	*0.183*
2011	Accepted	0.811	0.755	**0.883**	0.126	0.240	*0.025*	0.781	*0.202*	0.973
	Rejected	0.820	**0.978**	0.968	0.268	0.070	*0.036*	0.137	0.082	*0.026*
	Both	0.815	0.913	**0.926**	0.163	0.104	*0.029*	0.218	0.200	*0.196*
2012	Accepted	0.762	0.712	**0.827**	0.164	0.316	*0.057*	0.705	*0.278*	0.910
	Rejected	0.787	**0.971**	0.949	0.195	0.065	*0.042*	0.222	0.087	*0.065*
	Both	0.775	0.881	**0.888**	0.169	0.109	*0.051*	0.298	*0.192*	0.258
All	Accepted	0.772	0.732	**0.862**	0.159	0.283	*0.035*	0.759	*0.252*	0.938
	Rejected	0.802	0.969	**0.966**	0.228	0.070	*0.033*	0.183	0.088	*0.038*
	Both	0.787	0.891	**0.914**	0.177	0.105	*0.034*	0.265	*0.202*	0.216

NB: The model with highest accuracy is highlighted in bold and the one with the smallest error rates was highlighted in italics.

by year, we set up a fixed 1000 sample size for each year, 70% from the accepts and the rejects each as training, 30% of them as testing, taking 50 times resampling with replacement. The reject inference model will be built upon the training sets and then evaluated on the test sets.

3.4. Measurement of model performance

The performance of a model can be measured in many ways, but we can generally evaluate the effectiveness of a model by its classification accuracy and discriminant power. Crook et al. (2007) suggest that the confusion matrix, which compares the number of true goods and bads against the number of predicted goods and bads, is a useful tool to gauge the level of misclassification or classification accuracy. Taking the imbalanced into account, a cut-off that fits the empirical default rate is assigned to probabilities of default predicted by a model, where applicants with default probabilities above the cut-off are classified as 'bad' and those with probabilities below the cut-off as 'good', represent-

ing a lower chance of default. A typical example of the confusion matrix is shown in Table 4. Note that the overall accuracy of the modelling procedure is given by the percentage of goods and bads that the model correctly classifies, denoted as the sum of Correct Goods and Correct Bads divided by Sample Size. Type I errors (true goods classified as bads) and Type II errors (true bads classified as goods) are also reported.

The Receiver Operation Curve (ROC) is a good measure for discriminant power (Crook et al., 2007). The ROC curve is a graph of the true positive rate against the false positive rate at all values of cut-offs. The Area Under ROC (AUC) ranges between 0.5 and 1 is an indicator of how good a model is. If two groups of cases can be completely separated by a model, the ROC curve would go along the edges of the square (AUC equals to 1), whereas if a model performs just as well as a random guess, the curve would be a diagonal line (AUC equals to 0.5).

4. Results

Table 5 shows the summary of accuracy of the three models, respectively for the accepted, the rejected, the good and the bad (Type I & II errors). The rank of overall accuracy from best to worst is SSVM, SVM and Logit, across all years. Aggregated all together, the overall accuracy for SSVM, SVM and Logit is 91.4%, 89.1% and 78.7% respectively. We found that the advantages of SSVM over the other two mainly comes from the accepted applications with true labels, which means that SSVM can make use of the information in the rejected to improve the classification in the accepted. In

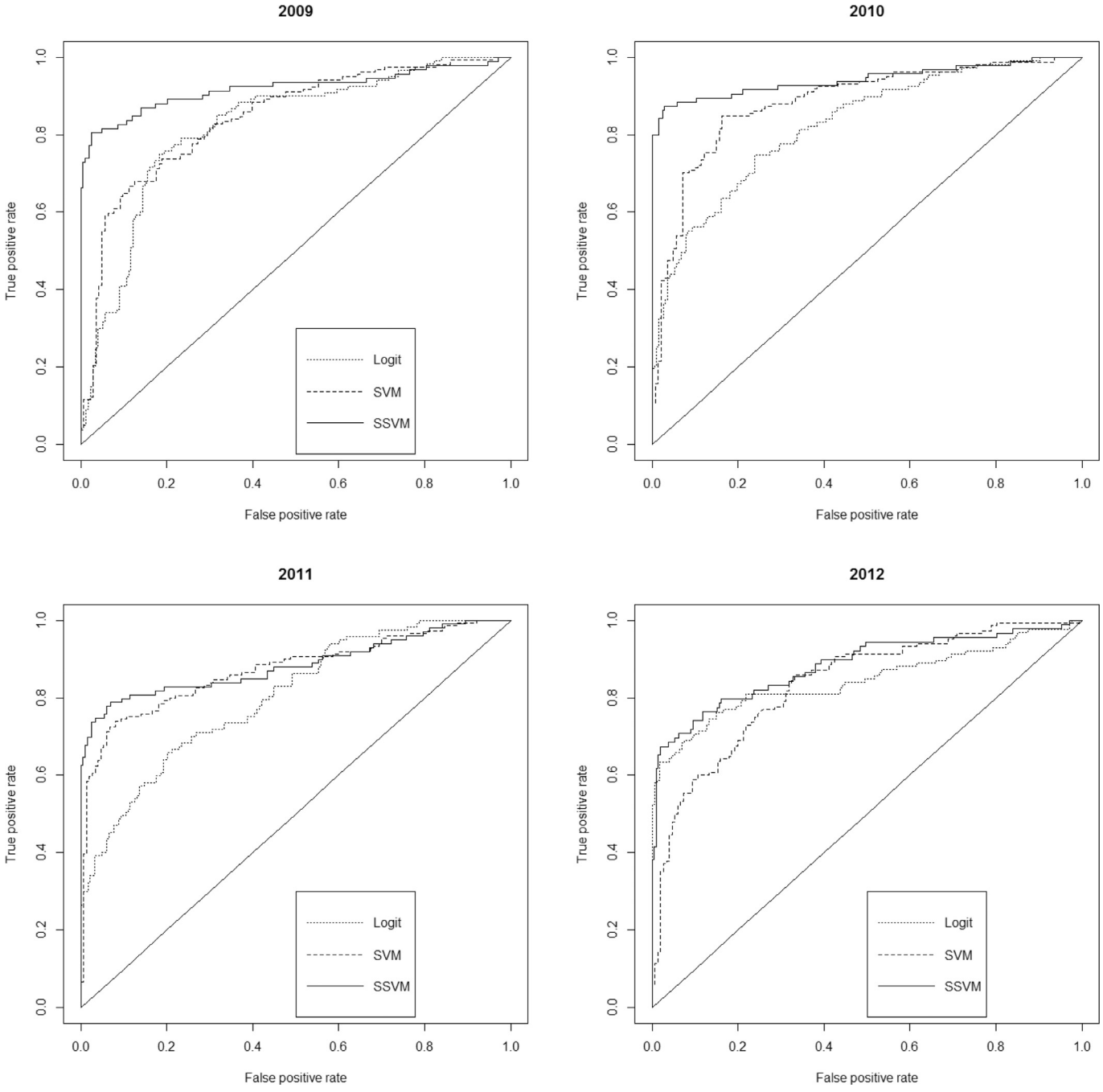

Fig. 5. ROC curves.

Table 6
AUC results.

	Logit	SVM	SSVM
2009	0.825	0.859	**0.872**
2010	0.852	0.883	**0.895**
2011	0.881	0.872	**0.887**
2012	0.841	0.828	**0.845**
Average	0.850	0.861	**0.875**

NB: The model with highest AUC is highlighted in bold

fact, for accepted applications all three methods achieved a predictive accuracy ranging between 73% and 86%. On average, SSVM and SVM perform equally well in the rejected group and are better than Logit.

In terms of Type I errors where good applicants are predicted as bads, SSVM is consistently better than the other two models across all years, with only 3.4% Type I error rate in all. If we look at the average Type II errors, SVM slightly outperforms SSVM by 1 percentage point and SSVM outperforms Logit by 5 percentage points. It is normally the case that credit scoring models can identify true goods better than true bads. However we find that SSVM in particular has a very low Type II error rate in the rejected group, which indicates that it is unlikely to treat a bad applicant as a good one in the group of applicants the credit grantor turns down. In practice, this helps reduce loss in the business.

For discriminant power between the three models, we calculated their average values of AUC, based on 50 simulations in Table 6. It shows that using SSVM as a reject inference technique is generally better than Logit and SVM. The improvement of AUC from logistic regression to SSVM is evident at all cut-offs, from 0.850 to 0.875 averagely. The SSVM model apparently shows a significant improvement to both traditional classification methods and makes the best use of the position information of unlabelled points.

We randomly extracted the result of one test from the repeated sampling/modelling and drew the ROC curve for three models in Fig. 5. Overall, the curves of SSVM are generally above those of SVM and the curves of SVM are above those of logistic regression, though the margins vary across all years. It is noted that in the years 2009 and 2012, the performance of Logit and SVM is very close, with the difference between them being 0.822 to 0.842 in 2009, and 0.845 to 0.835 in 2012.

5. Concluding remarks

The problem of reject inference has a long history in credit scoring, but so far it has not been adequately resolved. It can be seen as a statistical problem with missing data, given that the repayment behaviours of rejected applicants are unavailable. Depending on the type of missing, *i.e.* whether it is missing at random or missing not at random, various statistical techniques are used by credit scorers. From another perspective, reject inference can also be considered as a machine learning problem where algorithms learn to use information from the rejected group and optimise the objective gradually. Given the wide applications of machine learning techniques and the rising influence of SVM as an efficient classification algorithm, this paper proposes a semi-supervised SVM, with significant improvement to the previous reject inference methods, as an informative solution to the reject inference problem.

In this study, we have tested the predictive performance on a large sample of online lending data between 2007 and 2012. Compared to logistic regression and supervised SVM without the use of the rejected, the new semi-supervised S3VM built on the accepted and rejected groups shows better performance. Through simulation, we have proved that making use of the rejected applicants' information is of value in practice. The semi-supervised machine learning method can be used as an effective reject inference technique for credit scoring applications. In future studies, we are also interested in introducing more SVM algorithms such as cS3VM and S3VM light to reject inference.

Acknowledgements

This research is supported in part by National Natural Science Foundation of China (No. 11501463) and the Fundamental Research Funds for the Central Universities (No. JBK120509 and JBK140507).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2017.01.011.

References

Anderson, B., & Hardin, J. M. (2013). Modified logistic regression using the EM algorithm for reject inference. *International Journal of Data Analysis Techniques and Strategies, 5*, 359–373.
Bücker, M., van Kampen, M., & Krämer, W. (2013). Reject inference in consumer credit scoring with nonignorable missing data. *Journal of Banking & Finance, 37*, 1040–1045.
Bai, Y. Q., Niu, B. L., & Chen, Y. (2013). New SDP models for protein homology detection with semi-supervised SVM. *Optimization, 62*, 561–572.
Banasik, J., & Crook, J. (2005). Credit scoring, augmentation and lean models. *Journal of the Operational Research Society, 56*, 1072–1081.
Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research, 183*, 1582–1594.
Banasik, J., & Crook, J. (2010). Reject inference in survival analysis by augmentation. *Journal of the Operational Research Society, 61*, 473–485.
Banasik, J., Crook, J., & Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society, 54*, 822–832.
Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications, 36*, 3302–3308.
Bennett, K., & Demiriz, A. (1999). Semi-supervised support vector machines. In *Advances in neural information processing systems* (pp. 368–374).
Bijak, K., & Thomas, L. C. (2012). Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications, 39*, 2433–2442.
Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the eighteenth international conference on machine learning* (pp. 19–26).
Boyacioglu, M. A., Kara, Y., & Baykan, Ö. K. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. *Expert Systems with Applications, 36*, 3355–3366.
Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning*. Cambridge: MIT Press.
Chapelle, O., Sindhwani, V., & Keerthi, S. S. (2008). Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research, 9*, 203–233.
Chen, G. G., & Astebro, T. (2012). Bound and collapse Bayesian reject inference for credit scoring. *Journal of the Operational Research Society, 63*, 1374–1387.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.
Crook, J., & Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance, 28*, 857–874.
Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*, 1447–1465.
Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *Review of Financial Studies, 25*, 2455–2483.
Elliehausen, G., Christopher Lundquist, E., & Staten, M. E. (2007). The impact of credit counseling on subsequent borrower behavior. *Journal of Consumer Affairs, 41*, 1–28.
Feelders, A. J. (1999). Credit scoring and reject inference with mixture models. *Intelligent Systems in Accounting, Finance & Management, 8*, 271–279.
Gieseke, F., Airola, A., Pahikkala, T., & Kramer, O. (2014). Fast and simple gradient-based optimization for semi-supervised support vector machines. *Neurocomputing, 123*, 23–32.
Hand, D. J., & Henley, W. E. (1993). Can reject inference ever work? *IMA Journal of Management Mathematics, 5*, 45–55.
Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*, 153–161.

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*, 434–440.

Huang, S. C., Tang, Y. C., Lee, C. W., & Chang, M. J. (2012). Kernel local Fisher discriminant analysis based manifold-regularized SVM model for financial distress predictions. *Expert Systems with Applications, 39*, 3855–3861.

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening peers softly: Inferring the quality of small borrowers. *Management Science, 62*, 1554–1577.

Joachims, T. (1999). Transductive inference for text classifcation using support vector machines. In *ICML: Vol. 99* (pp. 200–209).

Joanes, D. N. (1993). Reject inference applied to logistic regression for credit scoring. *IMA Journal of Management Mathematics, 5*, 35–43.

Kim, Y., & Sohn, S. Y. (2007). Technology scoring model considering rejected applicants and effect of reject inference. *Journal of the Operational Research Society, 58*, 1341–1347.

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

Maldonado, S., & Paredes, G. (2010). A Semi-supervised approach for reject inference in credit scoring using SVMs. In P. Perner (Ed.). In *Advances in data Mining. Applications and theoretical aspects: Vol. 6171* (pp. 558–571). Springer Berlin Heidelberg.

Min, J. H., & Lee, Y.-C. (2008). A practical approach to credit scoring. *Expert Systems with Applications, 35*, 1762–1770.

Ravi Kumar, P., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research, 180*, 1–28.

Reddy, I. S., Shevade, S., & Murty, M. N. (2011). A fast quasi-Newton method for semi-supervised SVM. *Pattern Recognition, 44*, 2305–2313.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge: MIT press.

Sohn, S. Y., & Shin, H. W. (2006). Reject inference in credit operations based on survival analysis. *Expert Systems with Applications, 31*, 26–29.

Tian, Y., & Luo, J. (2016). A new branch-and-bound approach to semi-supervised support vector machine. In *Soft computing* (pp. 1–10).

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley-Interscience.

Verstraeten, G., & Van den Poel, D. (2004). The impact of sample bias on consumer credit scoring performance and profitability. *Journal of the Operational Research Society, 56*, 981–992.

Yang, Z., You, W., & Ji, G. (2011). Using partial least squares and support vector machines for bankruptcy prediction. *Expert Systems with Applications, 38*, 8336–8342.

Yeh, C.-C., Chi, D.-J., & Hsu, M.-F. (2010). A hybrid approach of DEA, rough set and support vector machines for business failure prediction. *Expert Systems with Applications, 37*, 1535–1541.

Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 3*, 1–130.