Expert Systems with Applications 42 (2015) 3194–3204

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Selection of Support Vector Machines based classifiers for credit risk domain

Paulius Danenas [a,b,*], Gintautas Garsva [a]

[a] Department of Informatics, Kaunas Faculty, Vilnius University, Muitines Str. 8, Kaunas, Lithuania
[b] Currently works in: Center of Information Systems Design Technologies, Department of Information Systems, Kaunas University of Technology, Studentu Str. 50-313a, Kaunas, Lithuania

ARTICLE INFO

Article history:
Available online 10 December 2014

Keywords:
Support Vector Machines
SVM
Particle swarm optimization
Credit risk
Default assessment
Classification

ABSTRACT

This paper describes an approach for credit risk evaluation based on linear Support Vector Machines classifiers, combined with external evaluation and sliding window testing, with focus on application on larger datasets. It presents a technique for optimal linear SVM classifier selection based on particle swarm optimization technique, providing significant amount of focus on imbalanced learning issue. It is compared to other classifiers in terms of accuracy and identification of each class. Experimental classification performance results, obtained using real world financial dataset from SEC EDGAR database, lead to conclusion that proposed technique is capable to produce results, comparable to other classifiers, such as logistic regression and RBF network, and thus be can be an appealing option for future development of real credit risk evaluation models.

1. Introduction

One of the most important research issues in financial domain is development of working credit risk evaluation and bankruptcy prediction models. Credit risk is one of frequently faced financial risks, which can be defined as the possibility that counterparty will fail to meet its obligations by agreed terms that will cost invested money for the lender. Minimization of such debts is critical for managing risk in financial institutions as Basel II capital accord defines new standards for capital adequacy in banks thus optimal capital allocation is essential for financial institutions. Thus proper, efficient and effective credit risk evaluation tools for and credit risk, such as highly discriminative credit scoring models, are obligatory for every financial institution. A credit score is primarily based on various financial, social, demographic and other data provided of the customers and about the customer, such as credit reports and information obtained from external evaluators and auditors such as major credit reporting agencies. Credit scores are often used to determine the amount of loan or interest rate that particular customer qualifies for.

Various machine learning techniques, such as artificial neural networks (abbr. ANN), have also gained a lot of attention from various researchers which are working in credit risk domain. ANN is understood as a computing model with a graph, that defines data structure for neural network, and interconnection pattern, which describes its architecture. This technique is well suited for developing accurate credit scoring systems and can perform competitively when compared to other classification techniques, such as logistic regression, MDA, decision trees. However, Support Vector Machines (abbr. SVM) technique has recently become one of the most widely research and applied techniques in this field. This technique offers several advantages compared to ANN such as absence of local minimas and relatively simple architecture. Many works in credit risk evaluation domain showed that they can show performance comparable to ANN or to outperform them (Danenas & Garsva, 2010; Kim & Ahn, 2012; Yu, Yao, Wang, & Lai, 2011).

Linear SVM are not widely explored in this domain, mainly because of its reduced flexibility, related to absence of kernel function and nonlinear mappings. However, state-of-the art linear SVM implementations show much faster performance than nonlinear SVM, leading to their suitability for large-scale SVM classification and regression problems. Nonlinear SVMs are not efficient on larger scale learning and also suffer from imbalanced learning problem (Batuwita & Palade, 2013). To the knowledge of the authors, this problem is not yet addressed in popular SVM packages (LibSVM, SVM^{Light}, etc.), except weight assignment to different classes (a variation of cost-sensitive learning). Large scale learning is important in our context as our research framework involves

* Corresponding author.
E-mail addresses: danpaulius@gmail.com (P. Danenas), gintautas.garsva@khf.vu.lt (G. Garsva).

http://dx.doi.org/10.1016/j.eswa.2014.12.001
0957-4174/

obtaining and preprocessing significant amounts of Extensible Business Reporting Language (abbr. XBRL)[1] documents from external datasources (we used SEC datasource as basis for our framework) after they are issued, as well as mining obtained data using automatic label identification and classifier training. Several linear SVM classifiers from LIBLINEAR package (Fan, Chang, Hsieh, Wang, & Lin, 2008) were chosen for development of classification functionality in our system. Promising results, obtained previously (Danenas & Garsva, 2010; Danenas, Garsva, & Gudas, 2011), as well as results in (Chang, Hsieh, Chang, Ringgaard, & Lin, 2010) motivate the research of linear SVM classifiers in parallel with nonlinear state-of-the-art SVM modeling techniques. In particular, our research seeks to explore the potential of this technique against medium or larger datasets (in this context, "larger" is defined as "having 2000 or more instances") in credit risk domain, while combining it with "sliding window" approach for training and testing. The classifier selection is optimized using swarm intelligence metaheuristic (particularly, particle swarm optimization); this approach has gained significant amount of attention from the research community because of its conceptual simplicity and ability to balance both exploration and exploitation (Thangaraj, Pant, Abraham, & Bouvry, 2011).

The remainder of the paper is organized as follows. In Section 2, the key points on relevant credit risk related research is presented. Section 3 briefly describes Support Vector Machines, the classification technique used to develop our approach; additionally, it presents linear SVM algorithms, used in this research, together with motivation to use them. Particle swarm optimization is described in Section 4, together with necessary improvements. Section 5 presents the whole research methodology, together with metrics used for evaluation, while Section 6 gives a brief description of the data used in experiment, describes the experiment configuration and discusses the obtained results. Finally, Section 7 highlights the conclusions and directions for future research.

2. Earlier works

The earliest works in research of credit risk date to 1968, when Altman applied multiple discriminant analysis (MDA) to develop his Z-Score model (Altman, 1968), using two different samples and obtaining accuracy of 96% and 79%, respectively. MDA was also applied by other researchers to develop their own models (Deakin, 1972; Taffler, 1982) or to improve and analyze existing ones (Grice & Dugan, 2001; Grice & Ingram, 2001). Another well-known early development (Springate, 1978) was also based on stepwise MDA and four ratios, resulting in accuracy rate of 92.5%; 83.3% and 88% accuracy rates later were reported after testing the developed model with other samples (Sands, Springate, & Var, 1983). (Ohlson, 1980) applied logit analysis reporting accuracy of 96.12%, 95.55% and 92.84% for prediction within one year, two years and one or two years respectively. While (Begley, Ming, & Watts, 1996) showed that Ohlson's model might perform better than Altman original and improved Z-Scores, their evaluation has also been criticized (Grice & Dugan, 2001). Zmijewski (1984) used two samples of 840 companies (40 of them were bankrupt companies) for training and prediction purpose, using probit and maximum likelihood techniques, and obtained 72% accuracy. Another known credit risk model (Shumway, 2001) was developed using hazard analysis and the same predictors as in original Altman model.

Different machine learning techniques became an object of interest for solutions in financial domain soon after they were discovered to show their potential in solving different problems. Artificial neural network based techniques (abbr. ANN) were the first techniques to be successfully applied in this field. An early survey (Vellido, Lisboa, & Vaughan, 1999) indicated that backpropagation neural networks (abbr. BPNN) were the most popular machine learning technique among researchers in credit risk domain during 1992–1998; this is also confirmed for both cases of finance and business domain in general (Wong, Lai, & Lam, 2000; Wong & Selvi, 1998). Recent research proposed a lot of state-of-the-art ANN-based hybrid models; fuzzy ANN with particle swarm optimization (abbr. PSO) for parameter selection (Huang, 2008), wavelet neural networks with differential evolution applied for their training (Chauhan, Ravi, & Karthik, 2009), knowledge-based artificial neural network (abbr. KBANN) with rule extraction from trained neural networks (Bae & Kim, 2011), neurofuzzy systems (Chen, Huang, & Lin, 2009), ensembles of ANN (Tsai & Wu, 2008; Yu, Wang, & Lai, 2008) are only a few examples. Other important techniques in the domain of credit risk evaluation and bankruptcy prediction include decision trees (Duman, Ekinci, & Tanrıverdi, 2012; Khandani, Kim, & Lo, 2010) and their ensemble variations, particularly random forests (Fantazzini & Figini, 2008; Kruppa, Schwarz, Arminger, & Ziegler, 2013) or other (Zhang, Zhou, Leung, & Zheng, 2010).

Support Vector Machines (abbr. SVM) are another type of learning machines, which are able to perform comparably to ANN, while overcoming their problems of architectural complexity and entrapment in local minimas. One of the most actively researched and discussed problems, related to SVM, is parameter selection for kernel function and cost/complexity parameter; it is pointed out in the relevant literature that it should be set by the expert. Yet, a lot of work has been done in order to simplify this problem using various heuristic techniques, such as genetic algorithm (Cao, Lu, Wang, & Wang, 2012; Wu, Tzeng, Goo, & Fang, 2007) or swarm intelligence (Yun, Cao, & Zhang, 2011; Zhou, Bai, Tian, & Zhang, 2008). A survey of SVM-based methods in credit risk domain (Danenas & Garsva, 2009) also indicated that evolutionary or swarm intelligence techniques for SVM parameter selection or fuzzy logic/rough sets integration usually helps to improve classifier performance.

Recent SVM technique, Least Squares SVM, abbr. as LS-SVM (Suykens & Vandewalle, 1999), gained a lot of attention from different researchers, as its applications identified the efficiency in performance, while at the same time simplifying SVM computing using to a set of linear equations. LS-SVM has been applied as standalone or part of hybrid technique in credit risk domain by several authors (Cao et al., 2012; Lai, Yu, Zhou, & Wang, 2006; Li, Song, & Li, 2012). Ensemble learning, another trend of soft computing, has also been widely researched in the context of credit risk, as different authors prove empirically the capability of classifier ensembles to obtain better classification performance by stabilizing the classification results by reflecting variation within a data set (Hsieh & Hung, 2010). Recent developments of SVM ensemble models include reliability-based and weight-based strategies (Zhou, Lai, & Yu, 2010), adaptive linear ANN (Yu, Yue, Wang, & Lai, 2010), bagging or boosting procedures (Ghodselahi, 2011; Wang & Ma, 2012).

However; while higher accuracy is mostly obtained using novel nonlinear SVM methods on small amounts of data, performance of such techniques often suffers on real-world larger datasets. Our main focus lies on research which is performed on such datasets. The amount of it is not large, which may be influenced by the limitations of availability of the necessary financial/bankruptcy data (although the number of open financial datasources seems to be rising). (Harris, 2015) used a dataset of over 20,000 entries from Barbados credit unions for model development to develop SVM linear and nonlinear classifier together with clustered SVM; the results indicated that performance of linear SVM did not significantly differ from SVM using RBF kernel; similar conclusion can be drawn from the results in (Niklis, Doumpos, & Zopounidis,

[1] https://www.xbrl.org/