Advances in Smart Systems Research ISSN 2050-8662
Vol. 6. No. 2 : pp.28-40 : k17is-222

Using Semi-Supervised Learning Methods for Credit Score Problem

Vasileios Papastefanopoulos[a] , Stamatis Karlos[a,b], Sotiris Kotsiantis[a]

[a]Educational Software Development Laboratory (ESDLab), Department of Mathematics, University of Patras, Greece
[b]Technical Educational Institute of Western Greece, Department of Computer Engineering Informatics, Antirrio 30200, Greece

Abstract

Data have always played a cardinal role to financial applications, since the power that pumps out of possessing them is translated to safer and more profitable decisions for the corresponding organizations. However, shortage of such kind of data or the inability to access large amounts of them, especially for organizations of smaller range, settles supervised methods as non-auxiliary predictive tools. Thus, techniques that exploit unlabeled data offer the chance of applying even the most advanced strategies for both mining useful patterns from datasets that stem from the collection of sensitive personal information and analyzing their characteristics. In this work, comparisons between several algorithms of Semi-supervised learning and their supervised variants were conducted for examining the applicability of the first category over credit score problem. The contained datasets come from two different nations (Australia and Germany), concern credit card applications and are publicly available at UCI Machine Learning repository, thus favoring the reproducibility of our results.

1. Introduction

Nowadays, more and more services offered to both individuals and companies are upgraded, or are getting even more automated, assisted by factors such as the digitization of data, evolvement of efficient warehousing systems and integration of Machine Learning (ML) tools on the majority of the known applications. Although the ML term may seem too specialized for anyone non-related with computer science field, its applicability over the last years has exceeded the anticipations of the research community. More specifically, ML is a term that successfully combines several fields, such as Statistical Learning, Data Mining, Pattern Recognition and Predictive Analysis [1].

Such explorative tools could not be neglected by financial applications, especially if we consider the consequences that may occur by a prediction that was based on non-consistent or poor recordings, or even by a non-in-depth analysis of the available solutions. Analogically to the size of the company and the nature of the examined task, any misleading decision may induce various results, ranging from loss of small amounts of money to loss of larger capitals, long term defamation and financial crisis. The most well-known tasks that ML techniques have dealt with are financial credit risk analysis, prediction of credit score/rating, forecasting of bankruptcy phenomena and assessment of fraud risk [2],[3].

In this work, credit score problem is scrutinized under a ML view, which is probably the accounting task that mostly affects directly the bodies of consumers and investors as individual entities and, at the same time, is considered to be one of the most popular application fields for both data mining and operational research techniques [4]. In particular, a credit score is a numerical expression generated by a mathematical expression which is based on an analysis of a person's credit behavior. It represents the creditworthiness of a customer and is designed to predict risk, specifically, the likelihood that a customer will default on his or her credit obligations. Lenders, such as banks and credit card companies, use credit scores to evaluate the potential risk of lending money to any kinds of consumers. The most commonly used credit score model has been introduced by the Fairy Isaac Corporation (FICO) and is highly respected by many financial institutions [5].

Besides the fact that the most variants of accounting problems have been presented and studied since ‘70s, the majority of the published approaches concern construction of supervised models. This assumes the availability of large volumes of data for scoring efficient classification accuracy. In recent years, due to financial distress expanding worldwide and the competitive structure of current markets, rarely are recordings of companies revealed to the public. In order to deal with this shortage of data, Semi-supervised Learning (SSL) schemes have been introduced in the research community and described efficiently using an in-depth taxonomy by Triguero et al. [6]. The ambition of SSL schemes is to exploit unlabeled data, which are often in abundance compared with labeled since the outcome of an instance is the harder to collect information against the usually measured financial attributes, and increase the generalization ability of the contained each time learner. As this property fits with credit score problem in present financial environment, we measure the SSL schemes’ applicability using two separate real-world datasets. To the best of our knowledge, no other approach has been recorded in the literature for examining the efficacy of SSL over the credit problem regarding the classification accuracy.

Consequently, our ambition is to examine possible improvements of predictive accuracy of SSL schemes over the credit score task using a few instances against the corresponding supervised variants. The rest of this paper is organized as follows: Section 2 includes related supervised and SSL works, mainly implemented on other accounting problems. In Section 3, a review of the most known SSL schemes is provided. Section 4 describes both the variables of the two examined datasets and the experimental procedure that was followed together with the produced results. Finally, all the drawn conclusions are provided in the last Section, where also future ambitions for upgrading the compatibility of SSL schemes along with credit score problem are discussed.

2. Related Work

Since there are only two different outcomes of each examined recording, credit score belongs to binary classification problems. Moreover, the produced credit score models can be divided into two different categories: application and behavioral scoring. The former type attempts to predict a customer's probability of default at the time an application for credit is made based on information such as applicant characteristics and financial records, while the latter estimates the risk of an existing customer based on their recent accounting activity.

Several approaches have been recorded in the literature for dealing with this problem, varying from simple statistical models to more recent classification concepts, such as non-parametric and artificial intelligent methods. Avery et al. [7] made one of the first attempts to build credit-history scoring predictive models based on geographically stratified samples from 994 U.S. regions, questioning about several occurred issues (e.g. underrepresentation, omitted variables). More advanced learning techniques have been used over subsequent years, such as recently developed Decision Trees (DTs), Neural Networks (NNs), Artificial Immune Systems (AIS), Evolutionary algorithms such as Grammatical Evolution (GE) and Case-Based Reasoning (CBR) mainly oriented towards forecasting corporate credit rating modelling [8],[9]. Many of these approaches have been also applied to credit score, for instance the AIS system implemented in [10] that exploits the theory of Negative selection algorithms, which are trained on one class instances and try to discriminate any non-suitable example. Two recorded hybrid methods also make use of former ML methods: maximum weighted voting strategy that combines optimized NNs and DTs architectures has been performed very robust classification behavior and presented in [11], while a mixture of algorithms coming from heterogeneous learning categories has been used to construct an ensemble classifier that is ruled by a modified voting scheme (Global and Local Voting – GlLocVot) and managed to outperform several other ensemble methodologies [12].

Support Vector Machines (SVMs) are a category of learners that have been met with great acceptance over several accounting problems. Various comparisons with other learning approaches have been demonstrated, leading the researches to include them also on current scientific works. Multiple Discriminant Analysis (MDA), CBR and three-layer fully connected back-propagation neural networks (BPNs) were outperformed by optimized SVMs with RBF kernel over credit rating prediction in [13] and SVMs with an integrated binary discriminant rule (IBDR) proved superior to conventional structural risk minimization principle that is respected by default SVMs theory over corporate financial distress in [14].

3. Semi-supervised Learning schemes

The main ambition of SSL schemes is the exploitation of one or more learners over a provided dataset and the combination of their decisions according to specified rules under the existence of both unlabeled (*U*) and labeled (*L*) data. Subset *U* is differentiated by *L* because the class variable of its contained instances is not known. Thus, judging by the given *L* subset, various learning strategies are followed for assessing the instances of *U* and assigning the most informative of them for augmenting during the next steps the cardinality of the labeled data (*L'*) and converge to more robust outputs. This property of mining patterns from the unlabeled data renders SSL schemes exceptional tools to scientific community for dealing with real-life scenarios. The parameter that defines the ratio between the size of *L* and the total amount of instances will be called Labeled Ratio (*R*). This could be really helpful to any researcher who wants to simulate his own experiment examining its desired R values.

According to the most usual criterion for studying SSL schemes, there exist 3 distinct categories: single-view, multi-view and hybrid schemes [15]. The self-training scheme is theorized as the most representative algorithm of the single-view category and was initially introduced in [16]. Its simplicity has favored its fast spread to a lot of scientific fields and constitutes the original scheme which other researchers modified for producing innovative SSL algorithms or other SSL schemes, such as Tri-training [17], which encompasses three classifiers for rejecting or approving an instance as useful. During the learning phase, each unlabeled instance – examined with all its variables – whose estimated confidence exceeds a pre-defined trust-threshold according to the temporary built classification model, is added to the existing *L* subset along with its predicted class value. This stage is repeated until *U* subset is empty or other stopping criterion is verified. Intermediate filters could be placed on default scheme for better scrutinizing prospective incoming instances of *U* subset to the training set in order to avoid deterioration of learning ability. A variant of Tri-train scheme is De-Tri-Train algorithm [18] that adopts this strategy by exploiting Depuration data editing technique.

On the other hand, multi-view methods split the vector of variables of each given dataset to two or more views and forward the formatted subsets into different learners. While Co-training [19] – a popular SSL scheme with great success to real-world applications – operates under the hypothesis that two disjoint subsets of variables may contribute to better classification performance by exchanging the mined knowledge between the two different classifiers, other multi-view schemes produce a number of learners over subsets of initial *L*, generated by processes of either randomly selected variables [20] or weighted random selection [21] with replacement, and combine them with voting theory. A hybrid scheme between single and multi-view methods is the Co-training by Committee (CoBC) [22]. Under this scheme, an amount of diverse base learners are built using an Ensemble Learning (Bagging-Boosting-Random Subspace Method) algorithm and their predictions over randomly chosen subsets of *U* without replacement gradually format the final training set for building the appropriate classification model.

Semi-supervised learning methodology has actually been applied to credit score or credit rating problems, but not for identifying the applicability of various methods to this field [23],[24]. On the contrary, investigation of the problem of sample bias in credit score/rating modelling that is caused by underrepresentation or even absence of the rejected applicants in the training set has been conducted. Addressing of this issue is facilitated by reject inference, a method that aims to predict the behavior of the applicants who were not granted a loan. Maldonado and Paredes [25] presented a novel SSL method combined with SVM that labels the rejected loans by using self-training scheme and managed to perform better results compared to other reject inference solutions.

Exploitation of one-class classification (OOC) algorithms, which are constructed based solely on the distribution of given credit worthy applicants, whose cardinality is highly larger than the minority class of non-credit worthy is examined in [26]. The improvement that OOC strategy offers against two-class classifiers, when such conditions hold, constitutes an important indicator, while extensions with oversampling methods over the minority class could indicate even better results when dealing with the low-default portfolio (LDP) problem.

4. Data Description and Experimental Procedure

Two publicly available credit datasets were used in our work, mined from the well-known UCI data repository [27]: Australian Credit and German Credit Dataset. The first of them contains 690 instances, each of which is described through 8 nominal and 6 continuous variables and they all refer to credit card facilities, while the second includes 1000 recordings which, in this experiment, are represented with 20 variables (13 nominal and 7 continuous) used to assess the credit class of the applicants. The titles of the various attributes for both datasets are shown in Table 1 and 2 together with their types, expressing several financial variables.

With regards to the cardinality of each class, in Australian dataset there is a small imbalance of rejected and accepted instances – 383 and 307 respectively – while in German dataset a sharper imbalance is observed, with 300 negative decisions and 700 positive. Furthermore, a few missing values exist inside the Australian dataset, but have been replaced with the modes and means of all the train data, according to the corresponding pre-processing filter of WEKA tool[28].

All the 6 performed SSL schemes and supervised algorithms that are included on the experimental procedure were extracted from the KEEL tool[29]. For acquiring a more equitable view of the examined applicability of SSL schemes, no tuning procedure or parameter choice through possible validation subsets was inserted. Some of the most representative learners of DTs, Instance Based Learning (IBL) methods and SVMs were used as base classifiers both inside each SSL method and individually. In particular, C4.5, K–Nearest Neighbors (KNN) and Sequential Minimal

Table 1. Australian Credit Dataset – Type of Attributes

Attribute	Type
Sex	Nominal
Age	Continuous
Mean time at addresses	Continuous
Home status	Nominal
Current occupation	Nominal
Current job status	Nominal
Mean time with employers	Continuous
Other investments	Nominal
Bank account	Nominal
Time with bank	Continuous
Liability reference	Nominal
Account reference	Nominal
Monthly housing expense	Continuous
Savings account balance	Continuous
Class (Reject / Accept)	Nominal

Optimization (SMO) were exploited as base classifiers. Acquisition of results that come from the use of 4 different *R* values (10%, 20%, 30% and 40%) has been preferred for investigating the influence of the size of the initial *L* subset examples during the training phase of SSL schemes.

A short description of the basic properties of the selected SSL schemes follows:

1. *Self-training* [16]: Parameter of Max iterations equals to 40.
2. *DE-Tri-training* [18]: The number of examined neighbors equals to 3 and the majority of them has to agree on the tested examples.
3. *Co-training* [19]: Parameter of Max iterations equals to 40, while the initial pool from which the possible unlabeled examples are extracted equals to 75.
4. *RASCO* [20] and *Rel–RASCO* [21]: Parameter of Max iterations equals to 10 and the number of views equals to 30.
5. *CoBC (Co–Bagging)* [22]: Parameter of Max iterations equals to 40, while size of *U* equals to 100. In our experiments, only the Bagging was chosen as the default Ensemble Learning strategy. Moreover, the number of the constructed committees equals to 3.

6.*KNN*: Number of neighbors (*K*) is equal to 3.

To more technical information, all the datasets have been partitioned and assessed by using the 10-cross validation technique. According to this technique, the full dataset is split into 10 non-overlapping folds, one of which is kept for the testing process and the rest are used for building the training model. Each fold that is used for the training process is being processed by an unlabelizing stage. More

Table 2. German Credit Dataset – Kind of Attributes

Attribute	Type
Checking account status	Nominal
Duration of credit in months	Continuous
Credit history	Nominal
Purpose of credit	Nominal
Credit amount	Continuous
Average balance in savings account	Nominal
Present employment	Nominal
Installment rate as % of disposable income	Continuous
Personal status	Nominal
Other parties	Nominal
Present resident since – years	Continuous
Property magnitude	Nominal
Age in years	Continuous
Other payment plans	Nominal
Housing	Nominal
Number of existing credits at this bank	Continuous
Nature of job	Nominal
Number of people for whom liable to provide maintenance	Continuous
Applicant has phone in his or her name	Nominal
Foreign worker	Nominal
Class (Reject / Accept)	Nominal

specifically, only a part of the contained examples keep their label, while the labels of the rest are handled as unknown. Regarding the manipulation of nominal variables from SMO algorithm, an appropriate filter was used (RenameNominalValues) through WEKA tool [28] for converting all such values to numeric.

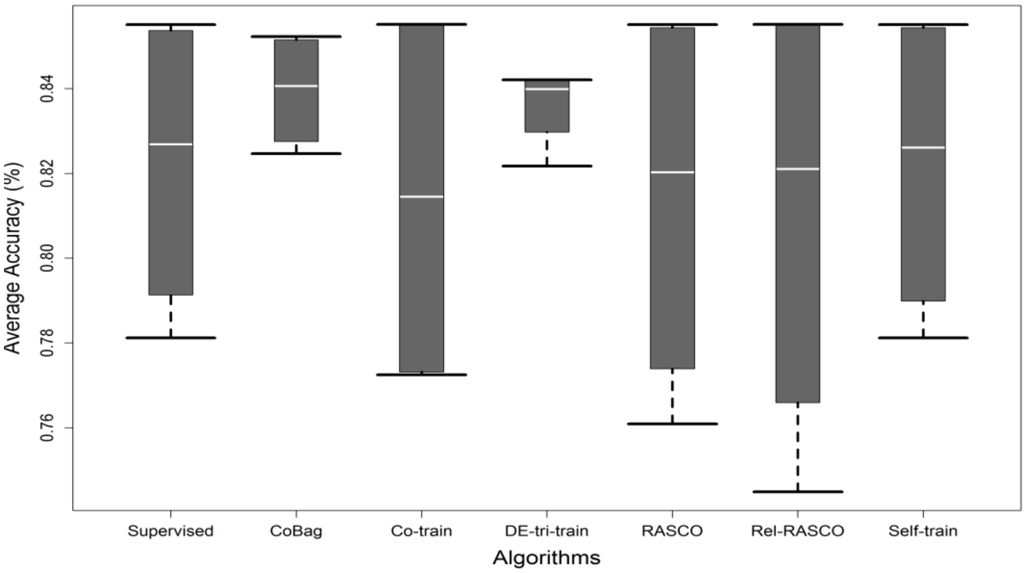

Fig. 1. Boxplot of accuracies of various SSL schemes against Supervised for Australian dataset.

The obtained results concerning the performance of the SSL schemes when low values of *R* parameter were used (10% and 20%) show that their behavior against supervised algorithms was unstable. Too many fluctuations for the majority of the algorithms were detected and as a result a safe generalized conclusion could not be reached. However, when the value of parameter *R* was set equal to 30% and 40%, a more clear learning behavior could be obtained. Furthermore, the use of DTs did not favor the SSL schemes, since not any important accuracy improvement was reported. On the other hand, in the cases where NN and SMO were used as base learners, CoBC scheme achieved the best overall results during our experimental procedure for both datasets, performing about a 2% relative improvement against the supervised variant.

Although De-tri-train scheme presented a similar classification accuracy with CoBC over the Australian dataset, its performance was not proven equivalent to a larger dataset, affected probably more heavily than it should by the quality of the neighbors that were examined by its data editing technique. This is highlighted also by the difference over the bandwidth of the corresponding boxplots. Self-training scheme also managed to perform a slight better accuracy rate than its supervised rival. This means that a more sophisticated filter could boost its accuracy to even higher levels. The two presented figures depict the average accuracies of the referred methods for both datasets, only for 30% and 40% values of *R* parameter and for the two best base learners that were previously mentioned.

5. Conclusions

In credit scoring task, the objective is to assign borrowers to one of two groups: good or bad, depending on their probability of default using methods which are based on the creditor's loan history. A member of the good group is likely to commit to their financial obligations while a member of the bad group is likely to default on them. Various statistical and machine learning techniques help creditors identify borrowers who cannot fulfill their financial obligations within a group of loan applicants. Instead of using supervised theory, as in the majority of the recorded works, SSL strategy seems to fit better with the nature of financial problems, such as the credit score. Thus, experiments for identifying the efficacy of 6 SSL schemes were conducted, resulting in learning behaviors that are improved against supervised and clearly closer to the real-life scenarios of having access to a few data for predicting the outcome of applicants.

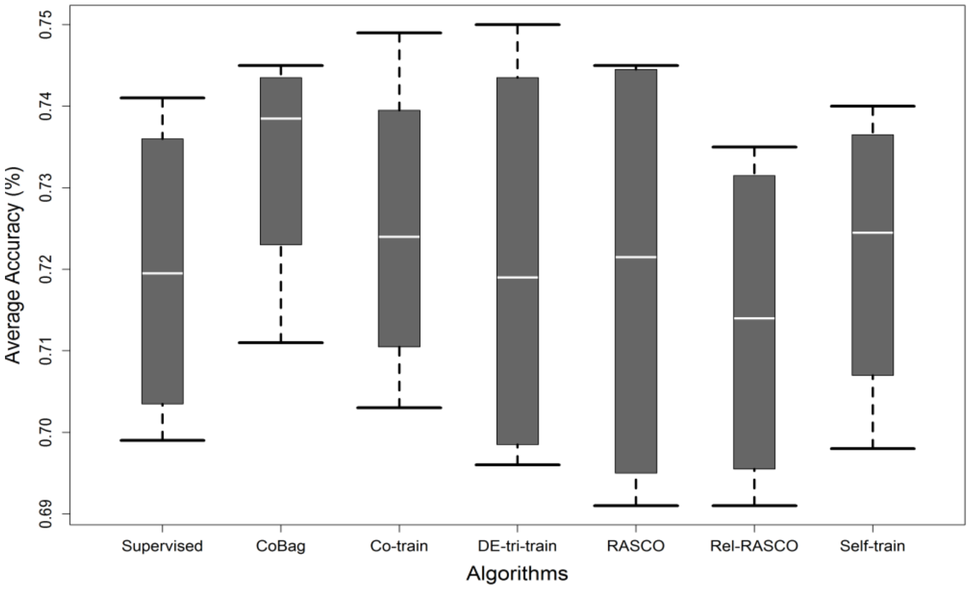

Fig. 2. Boxplot of accuracies of various SSL schemes against Supervised for German dataset.

No tuning stages were added to our experiments, shifting our interests towards combining the most accurate SSL schemes according to our results (CoBag, De-tri-train and Self-train) with data editing techniques or filters that do not permit misclassified instances to be inserted in the upcoming classification models inside the learning stages of SSL schemes. This was initially our ambition, to examine the compatability of several SSL schemes with the credit score problem and present some novel results about their learning ability over different *R* values.

For future work, we refer some pre-processing techniques that have been applied to financial problems and have been proven to boost the performance of supervised learners, so as to integrate them into SSL schemes: use of SMVs as a feature selection method for recognizing the most informative features over the prediction of the likelihood of default [30], the combination of learners with Rough Set Theory (RST) for achieving similar performance by including less variables during the construction of learning model and the inclusion of Data Envelopment Analysis (DEA) that may offer additional insights of the examined data [31]. Finally, Active Learning (AL) theory could be applied to credit score problem, where the most difficult to discriminate instances are added gradually to the learning model under an iterative scheme [32].

Acknowledgements

Stamatis Karlos gratefully acknowledges the support of his work by both the General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI).

References

[1] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*. 2015.

[2] N. Chen, B. Ribeiro, and A. Chen, “Financial credit risk assessment: a recent review,” *Artif. Intell. Rev.*, vol. 45, no. 1, pp. 1–23, 2016.

[3] M. Krivko, “A hybrid model for plastic card fraud detection systems,” *Expert Syst. Appl.*, vol. 37, no. 8, pp. 6070–6076, 2010.

[4] B. Baesens, C. Mues, D. Martens, and J. Vanthienen, “50 Years of Data Mining and OR: Upcoming Trends and Challenges,” *J. Oper. Res. Soc.*, vol. 60, pp. s16–s23, 2009.

[5] “Investopedia - Sharper Insight. Smarter Investing.” [Online]. Available: http://www.investopedia.com/. [Accessed: 12-May-2017].

[6] I. Triguero, S. García, and F. Herrera, “Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study,” *Knowl. Inf. Syst.*, vol. 42, no. 2, pp. 245–284, 2013.

[7] R. B. Avery, R. W. Bostic, P. S. Calem, and G. B. Canner, “Credit Scoring: Statistical Issues and Evidence from Credit-Bureau Files,” *Real Estate Econ.*, vol. 28, no. 3, pp. 523–547, Sep. 2000.

[8] D. Delen, C. Kuzey, and A. Uyar, “Measuring firm performance using financial ratios: A decision tree approach,” *Expert Syst. Appl.*, vol. 40, no. 10, pp. 3970–3983, 2013.

[9] A. Brabazon and M. O’Neill, “Credit classification using grammatical evolution,” *Inform.*, vol. 30, no. 3, pp. 325–335, 2006.

[10] T. P. Mpofu and V. R. Reddy, “Artificial Immune Systems: A Predictive Model for credit scoring,” *Int. J. Sci. Eng. Res.*, vol. 5, no. 8, pp. 113–117, 2014.

[11] A. Arzy Soltan and M. Mehrabioun Mohammadi, “A hybrid model using decision tree and neural network for credit scoring problem,” *Manag. Sci. Lett.*, vol. 2, no. 5, pp. 1683–1688, 2012.

[12] E. Kamos, F. Matthaiou, and S. Kotsiantis, “Credit rating using a hybrid voting ensemble,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7297 LNCS, pp. 165–173, 2012.

[13] Y.-C. Lee, “Application of support vector machines to corporate credit rating prediction,” *Expert Syst. Appl.*, vol. 33, no. 1, pp. 67–74, 2007.

[14] Z. Hua, Y. Wang, X. Xu, B. Zhang, and L. Liang, “Predicting corporate financial distress based on integration of support vector machine and logistic regression,” *Expert Syst. Appl.*, vol. 33, no. 2, pp. 434–440, 2007.

[15] C. Xu, D. Tao, and C. Xu, “A Survey on Multi-view Learning,” *Cvpr*, vol. 36, no. 8, p. 300072, 2015.

[16] D. Yarowsky, “Unsupervised Word Sense Disambiguation Rivaling Supervised Methods,” *Proc. 33rd Annu. Meet. Assoc. Comput. Linguist.*, pp. 189–196, 1995.

[17] Z.Zhou and M.Li, “Tri-Training: Exploiting Unlabled Data Using Three Classifiers,” *IEEE Trans.Data Eng.*, vol. 17, no. 11, pp. 1529–1541, 2005.

[18] C. Deng and M. Z. Guo, “Tri-training and data editing based semi-supervised clustering algorithm,” *Micai 2006 Adv. Artif. Intell. Proc.*, vol. 4293, pp. 641–651, 2006.

[19] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proceedings of the eleventh annual conference on Computational learning theory - COLT’ 98*, 1998, pp. 92–100.

[20] J. Wang, S. Luo, and X. Zeng, “A random subspace method for co-training,” in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 195–200.

[21] Y. Yaslan and Z. Cataltepe, “Co-training with relevant random subspaces,” *Neurocomputing*, vol. 73, no. 10–12, pp. 1652–1661, Jun. 2010.

[22] M. Hady and F. Schwenker, “Co-Training by Committee: A Generalized Framework for Semi-Supervised Learning with Committees,” *Int J Softw. Informatics*, vol. 2, no. 2, pp. 95–124, 2008.

[23] P. Hajek and V. Olej, “Credit rating modelling by kernel-based approaches with supervised and semi-supervised learning,” *Neural Comput. Appl.*, vol. 20, no. 6, pp. 761–773, 2011.

[24] Z. Li, Y. Tian, K. Li, and W. Yang, “Reject inference in credit scoring using Support Vector Machines,” 2017.

[25] S. Maldonado and G. Paredes, “A semi-supervised approach for reject inference in credit scoring using SVMs,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6171 LNAI, pp. 558–571, 2010.

[26] K. Kennedy, B. Mac Namee, and S. J. Delany, “Using semi-supervised classifiers for credit scoring,” *J. Oper. Res. Soc.*, vol. 64, no. 4, pp. 513–529, 2012.

[27] M. Linchman, “UCI Machine Learning Repository,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml/.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, p. 10, 2009.

[29] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, “KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework,” *J. Mult. Log. Soft Comput.*, vol. 17, no. 2–3, pp. 255–287, 2011.

[30] T. Bellotti and J. Crook, “Support vector machines for credit scoring and discovery of significant features,” *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 3302–3308, 2009.

[31] J. H. Min and Y.-C. Lee, “A practical approach to credit scoring,” *Expert Syst. Appl.*, vol. 35, no. 4, pp. 1762–1770, 2008.

[32] B. Settles, “Active Learning,” *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 6, no. 1, pp. 1–114, Jun. 2012.