Effectiveness of semi-supervised learning in bankruptcy prediction

Stamatis Karlos
Department of Mathematics, University of Patras
Patras, Greece
stkarlos@upatras.gr

Nikos Fazakis
Department of Electrical Engineering, University of Patras
fazakis@ece.upatras.gr

Sotiris Kotsiantis
Department of Mathematics, University of Patras
sotos@math.upatras.gr

Kyrgiakos Sgarbas
Department of Electrical Engineering, University of Patras
sgarbas@upatras.gr

***Abstract*—Adoption of techniques from fields related with Data Science, such as Machine Learning, Data Mining and Predictive Analysis, in the task of bankruptcy prediction can produce useful knowledge for both the policy makers and the organizations that are already funding or are interested in acting towards this direction in the near future. The nature of this task prevents analysts from collecting large amount of data for building accurate predictive models. Semi-supervised algorithms overcome this phenomenon and can perform robust behavior based on a few data. Experiments using data from Greek firms have been made in this work, comparing many semi-supervised schemes against well-known supervised algorithms and the results are promising.**

Keywords—bankruptcy prediction; semi-supervised schemes; predictive analysis; labeled ratio

I. INTRODUCTION

Although during the current decade many scientific fields that are related with Data Science (DS) are in bloom and the amount of produced data is increasing exponentially, there are still some services for which the provided information is highly restricted. The main reasons why this phenomenon holds may be the nature of the examined task – sensitive information or data that pose a risk to the personal privacy of any customer are not accessible – bureaucracy or even the inability of recording data in digital format. Bankruptcy prediction [1] belongs to this category, since the owners of the available data are usually private banks that are not willing to share such information because of competitiveness. Moreover, the validation of the gathered data by any private entity could not be safe enough and might lead to distorted decisions or results.

Since bankruptcy prediction does not constitute neither an academic project nor a simulation task that can be redefined and tested as many times the user wants, the cost of any erroneous prediction would induce serious impacts in real-life, such as money loss, financial crises or reduction of labor force. The expected benefits of collecting data from several firms or individuals that describe their profile through financial rates are the following:

- Avoiding granting loans to any enterprise whose financial behavior denotes that a bankruptcy is more possible than not,
- Characterizing with better accuracy the credit behavior of any applicant, so as to propose a more profitable contract,
- Modification of terms of existing collaborations according to more successful recorded strategies for leading to more profitable relationships,
- Capability of predicting in time possible business failure and preventing the increase of the total expenditure.

Judging by the importance of the effects that are caused by the predictions of business bankruptcy, many sophisticated approaches have been implemented both in financial and computer science literature for similar purposes [2], [3]. Chuang [4] refers that the most statistical methods that have already been applied in bankruptcy prediction, such as univariate statistical models, Logistic Regression or Probit analysis may achieve good performance. However, their requirement for existence of data linearity do not allow them to harmonize with the generalized nature of the examined datasets, which is characterized by more complicated relationships without any easily observable property been hold. Inversely, many classical default Machine Learning (ML) and Data Mining (DM) learners have been proven more efficient in these tasks. The most recent studies have been oriented towards constructing ensemble classifiers or using methods that inject diversity into the collected data for tackling with this task. The former techniques are based on encapsulating some "weak" learners and exploiting their decisions under a combinatory structure, while the latter split either the feature set of the dataset or resample its contained examples for forming new subsets of the initial data. These may now be exploited, since they are theorized as totally new views that could reveal subsequent relationships among the tested examples.

Besides the fact that bankruptcy prediction is under research by data scientists more than three decades, all the published works that have examined this subject follow supervised approaches [5], [6], [7]. According to this strategy, all the collected data are used for constructing an appropriate model that will be responsible for the prediction of any incoming example. Despite the good performance and the deep analysis that has been made on several works about the number of the used classifiers on ensembles, or even the exploitation of more sophisticated methods, like Partial Least Square Discriminant Analysis (PLS-DA) [8], which also faces the multicollinearity phenomenon that is usually met on Econometrics and generally on datasets that are related with financial terms, the problem of the shortage of available data is not taken into consideration and consequently it cannot be tackled by them.

Therefore, use of Semi-Supervised Learning (SSL) algorithms could match with the aforementioned situations, something that has not been researched in-depth yet [9]. This kind of iterative methods demand just a few examples, whose outcome is known, for being initialized. Their main asset is to use examples for which the prediction is still unknown, so as to extract useful information both from them and from the gathered. Although the finding of examples of the first category is relatively easy and cheap, their contribution to the final formatted rules about the relationships among the features of general datasets has been proven really efficient. Searching for such automated algorithms reduce the needed time for collecting vast amounts of data and may also achieve improved learning abilities. The rest of this article is organized as follows: Section 2 consists of a review of the most important semi-supervised schemes. A description of the experimental procedure is given in Section 3, while the produced results are commented in Section 4. Conclusions and proposals of future are provided in Section 5.

II. Semi-supervised Schemes

Absence of capability to collect a lot of data for scenarios like bankruptcy prediction is the basic reason why SSL schemes seem promising to be applied. As it concerns the kind of data that are used here, there exist two different categories: Labeled (*L*) and Unlabeled (*U*). The criterion according which this split is made is the awareness or not of the final prediction of each example, respectively. Generally, the size of the data of the first category is quite smaller than the rest. The formula that describes the relationship between these two parameters is called Labeled Ratio (*R*) and is computed as follows:

$$R\ (\%) = \text{size}\ (L) / (\text{size}\ (L) + \text{size}\ (U)) \quad (1)$$

The most known semi-supervised schemes are being discussed in [10] along with their mathematical formulation. Moreover, description of the procedures that permit to each scheme to combine both *L* and *U* subsets is presented. Another great work that also establishes a taxonomy of such techniques has been demonstrated in [11]. The terminology of "self-labeled" techniques has been preferred by Triguero et al. because of the property to expand their initial given examples – mostly known as training set but in these cases this coincides with the *L* subset – with instances that come from the *U* subset and satisfy some kind of metrics or other prerequisites.

To be more specific, SSL schemes can be divided analog to the number of views that they need to single-view and multi-view. Each view consists of a number of features that are related somehow and are more likely to lead to better results in case that are exploited separately from the rest that are not conceptually related with them. If no criterion justifies such a division, the alternative of stacking all the features on a compact dataset is acceptable.

Self-training scheme is the most representative of the single-view methods. Its simplicity has boosted its applicability to many domains. Thus, after having chosen either a weak or a more complex learner, corresponding to the examined task – it could be one of classification, regression, clustering etc. – an appropriate model is being constructed based exclusively on *L*. Then, an evaluation stage follows. During this, each example that belongs to the *U* subset is annotated with a probability class value for each different class. These values express the certainty level that each specific example can be classified to the tested category, respectively. At the end of this phase, only these examples that achieved a class probability larger than a predefined threshold are removed from the *U* and are added to the *L* subset, enriching in this way the source data. These steps are repeated until a stopping criterion to be satisfied. Similar strategy, as it concerns the number of views, is being followed by the Tri-training scheme [12]. Instead of using one learner, three learners are trained by Bootstrap sampling of the *L*. When the decisions two of them agree on an example of the *U*, this is getting labelized and is provided to the third of them for increasing its training data. Both these schemes do not assume any strict assumption about neither the provided data nor during the evaluation stage. However, integration of data editing techniques is able to increase the performance of such schemes by reducing the rate of possible incoming misclassified instances that distort the final hypothesis [13]. Tri-training with Data Editing (DE-Tri-training) [14] uses a Nearest Neighbor Rule based data editing technique named Depuration for achieving a more generalized learning behavior.

The alternative family of semi-supervised schemes respects the multi-view theory [15]. Co-training algorithm, which has been proposed by Blum and Mitchell [16], requires two sufficient and redundant views for training two learners. Both of them follow similar steps with the single-view methods and the most confident predictions over the *U* are used to expand the training set of the other learner. Then they are refined using their updated *L* subset. Its effectiveness under several assumptions and in real-world datasets is discussed in [17]. An extension of Co-training scheme is Random Subspace Method for Co-training (RASCO) [18]. A number of learners is trained over subsets of the initial feature set that have been formatted by random splits. The decisions from all the learners are finally combined for mining knowledge through the unlabeled instances. Rel – RASCO [19] enhances the original scheme by setting some rules according which the "random" splits should be made, especially in occasions that some of the existing features are intensively irrelevant. Relevance scores are computed for each feature and play a cardinal role during the weighted random splits.

A hybrid scheme between single and multi-view methods is the Co-training by Committee (CoBC) [20]. 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 the *U* without replacement gradually format the final *L* subset.

III. Data Description

The framework of this work is set by data related with Greek businesses. The strategy to gather data in national level has also been respected in [8], [21], [22] for reassuring an homogenous financial environment. The period that covers the collected data equals to three years before the bankruptcy filings in the years of 2003 and 2004. The sources that provided them were the National Bank of Greece directories and the business database of the financial information services company (ICAP). The same dataset has also been used in [23], where a novel supervised ensemble classifier was constructed for forecasting corporate bankruptcy having filtered the data with a specific cost matrix that compensates the imbalanced dataset.

The whole data have been partitioned into three distinct datasets, each one for the years before the bankruptcy filing. The strategy that was adopted for collecting the various examples is that for each selected bankrupt firm another two non-bankrupt were chosen. Furthermore, it was mandatory this couple of firms to belong to the same industry without the number of their employees to diverge from the initial. Finally, the dataset of each examined year contains 145 examples, 49 that belong to Bankrupt class and 96 to the Non-Bankrupt. Little modifications of the original dataset has been conducted so as to remove missing values or useless attributes.

The feature set of the dataset that describes the selected firms three years before the bankruptcy filing contains 10 attributes, except for the class attribute, while the rest consist of 13. A short description of them is given in Table I.

Table I. Description of contained features

Feature Abbreviation	Description of features
GRTA	Growth rate of total assets $(TA_t - TA_{t-1})/(ABS(TA_t)+ABS(TA_{t-1})$
SIZE	Size of the firms: ln(Total Assets / GDP price index)
GRNI	Growth rate of net income
GIMAR	Gross income divided by sales
S/CE	Sales divided by capital employed
S/EQ	Sales divided by Shareholder's equity capital
CE/NFA	Capital Employed To Net Fixed Assets
TD/EQ	Total Debt To Shareholder's Equality Capital
EQ/CE	Shareholder's Equity To Capital Employed
WC/TA	Working Capital Divided By Total Assets
COLPER	Average Collection Period For Receivables
PAYPER	Average Payment Period To Creditors
INVTURN	Average Turnover Period For Inventories

For better depicting the chronological order of the datasets, the abbreviation that describes the datasets will be symbolized from this point and after as "1year", "2years" and "3years", where the initial year of collecting the data is three years before the bankruptcy filing. Consequenty, the "1year" dataset contains the examples that were gathered during only the first year, while the "3years" the examples from all the three years. Moreover, the values of each feature have been divided into three intervals. Thus, for each feature exist two values, for instance *a* and *b* with *a* being smaller than *b*. The first interval is defined as the values that are smaller to *a*, the second contains the values that are larger to *b* and the intermediate values set the third interval. As it concerns the "1year" dataset, the features that have not been included are the GRTA and the GRNI.

IV. Experimental Procedure and Results

In order to examine the efficiency of semi-supervised algorithms when they are applied on bankruptcy forecasting scenarios the KEEL tool was used [24]. The learners that were used both separately as supervised algorithms and as base classifiers into the SSL schemes are C4.5, K–Nearest Neighbors (KNN) and Sequential Minimal Optimization (SMO). The default parameters of KEEL were maintained during the experiments. To be more specific:

- Self-training: Parameter of Max iterations equals to 40,
- DE-Tri-training: Number of examined neighbors equals to 3 and the majority of them has to agree on the tested examples,
- Co-training: Parameter of Max iterations equals to 40, while the initial pool from which the possible unlabeled examples are extracted equals to 75,
- RASCO and Rel–RASCO: Parameter of Max iterations equals to 10 and the number of views equals to 30.
- CoBC: Parameter of Max iterations equals to 40, while three committees are being formatted. The Ensemble Learning method that was chosen is Bagging. Therefore, the term Co-Bagging will be used later,
- KNN: K is equal to 3.

All the datasets have been partitioned and assessed by using the 10-cross validation technique. According to this, the full dataset is split into 10 non-overlapping folds where the one 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 filtered through an unlabelezing stage. In other words, only a part of the contained examples keep their label, while the labels of the rest are handled as unknown. The number of the examples that are going through this is defined by the Labeled Ratio (*R*) value. For examining the influence that the parameter *R* induces to the learning behavior of the SSL algorithms, three different ratios were selected: 20%, 30% and 40%. Figure 1 depicts the flowchart that was followed for executing the corresponding experiments.

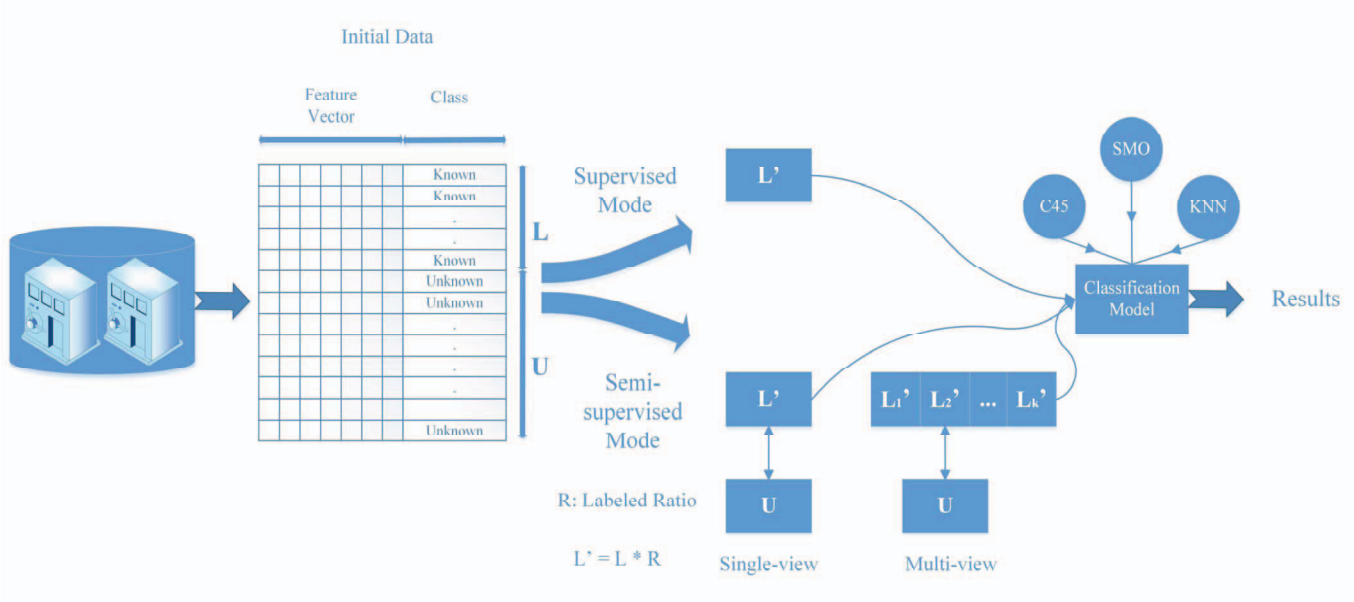

Fig. 1. Flowchart of the experimental procedure

Although the most combinations of the selected SSL algorithms with C4.5 and SMO performed an improved behavior compared with their corresponding supervised algorithm, the combinations that were produced with 3NN did not lead to similar results, except for Self-training (3NN). However, only these algorithms that managed to achieve at least one percent higher classification rate against the abovementioned supervised methods are presented in Table 2.

An interesting point being excluded from the results is that Rel-RASCO has been proved to be more robust in all the tested cases against simple RASCO. Especially, when C4.5 was used as the base classifier, the average accuracy over all the three datasets and the *R* values of Rel-RASCO (C4.5) was improved by 2% against this of RASCO (C4.5). This fact seems to favor the use of compatible criteria that can assign an importance weight over each feature, before their random selection during the creation of diverse views, even in our case where 13 and 11 features are used. DE-Tri-training also performed slight improvement along with C4.5 and SMO as base classifiers. Judging by the poor behavior of KNN methodology in the specific dataset, an alternative editing technique during the assessment of unlabeled instances from the three different classifiers could boost its performance. Similar modification should be inserted into Self-training scheme so as to avoid adding noisy examples into the *L* subset.

Table II. Classification accuracy of Supervised and Semi-Supervised algorithms

Learning Algorithms		**Examined Dataset**								
		1year			**2years**			**3years**		
R		*20%*	*30%*	*40%*	*20%*	*30%*	*40%*	*20%*	*30%*	*40%*
C4.5	*Supervised*	53.66	60.71	65.62	58.67	56.67	60.09	60	58.62	58.04
	Co-train	56.29	59.81	68.1	57.24	55.19	63.48	64.9	61.52	64.81
	RASCO	62.1	58.67	64.76	57.24	60.38	57.9	66.19	61.9	59.95
	Rel-RASCO	59.38	62.91	67.47	58.24	61.14	62.1	66.19	62.62	66.86
SMO	*Supervised*	49.52	54.48	53.1	59.19	48.38	51.05	60.71	58.43	61.2
	Co-Bagging	52.86	51.62	57.86	51.57	61.48	55.95	59.29	57.67	66.67
	Co-train	49.76	58.67	57.19	55.19	53.38	52.43	59.19	62.53	62.67
	DE-Tri-training	55.81	53.14	53.86	58.1	62.15	60.05	62.1	62.81	61.43
	RASCO	51.81	52.33	56.62	55.19	53.76	56	65.62	63.95	69.76
	Rel-RASCO	54.48	54.43	53.62	61.24	56.86	60.1	64	63.38	64.62

For assessing the behavior of the algorithms of Table II a classical test, which exams all the pairwise differences among these, is presented. After having computed the p-values, a second stage of correction is executed. For the first phase, Friedman's Aligned Ranks test is used and continuing to the next phase Juliet P. Shaffer's correction method is applied, as it is suggested in the literature for two different values of alpha parameter: 0.01 and 0.05. Computation and the illustration of this test was made in the R platform through scmamp [25] and Rgraphiz [26] libraries. Figure 2 visualizes this statistical test. Both of the diagrams consist of nodes that represent the algorithms of the experiment. The computed value inside each box is the ranking of the respective algorithm. The one with the higher ranking – Rel-RASCO (C4.5) on both cases – is highlighted in different color from the rest. The requirement for drawing a line is that two nodes are linked in the null hypothesis of being equal cannot be rejected.

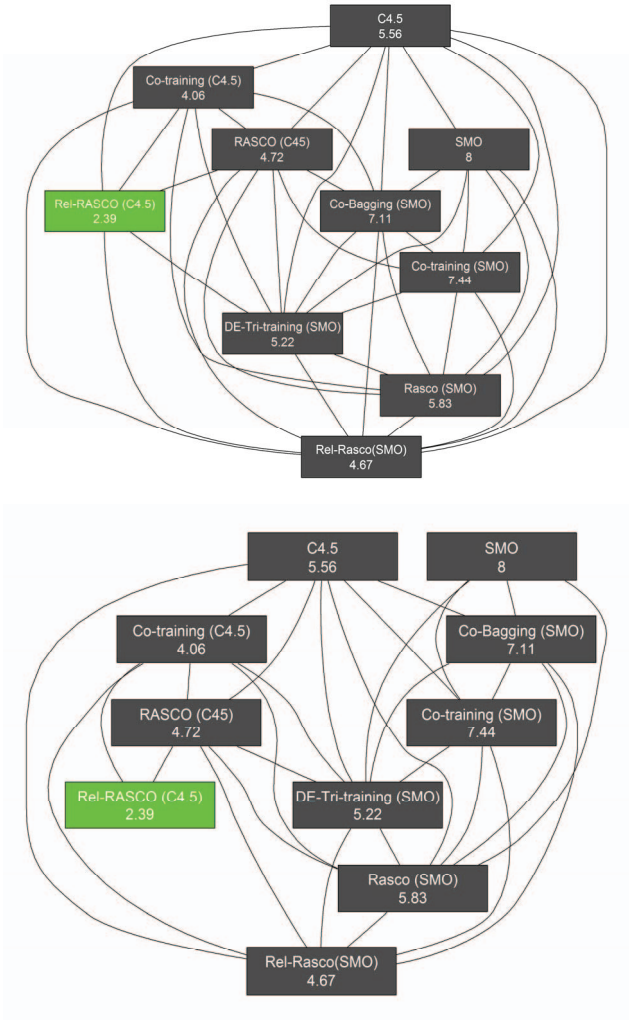

Fig. 2. Testing the null hypothesis that all the algorithms have the same accuracy for alpha =0.01 (upper) and alpha =0.05 (lower).

V. Conclusions

Machine Learning and Data Mining techniques can produce a number of rules that may improve the decisions or the predictions that have been made by human experts on several fields. As it concerns bankruptcy forecasting or other similar financial tasks, the consequences of wrong estimations may induce tremendous costs to the corresponding organizations.

The most related works have been oriented towards supervised algorithms or the construction of ensemble methods for increasing the classification accuracy rates or. However, usage of supervised algorithms require a large amount of data for performing well, which are not free of cost and time. Thus, semi-supervised schemes provide an efficient solution, since only a small amount of data are needed for achieving similar or even better learning ability and more robust classification behavior.

A number of SSL algorithms are being compared against their corresponding supervised under three different labeled ratios. The results prove the improvement of the most tested SSL methods. The best performance was achieved by Rel-RASCO scheme when it was combined with C4.5 as base learner, a well-known algorithm of decisions trees category. This algorithm builds a number of classifiers by splitting the initial feature set to smaller subsets. The difference with the RASCO scheme is that some restrictions are assumed during the choice of each feature. A limitation of this study was that only financial ratio variables were used by the learning models. There may be other key quantitative variables (i.e., stock data, market value, age) as well as qualitative variables (leadership, reputation, type of ownership, etc.) and there is rich literature in organization theory that reports the importance of these variables, too.

Some promising points for future work could be the feature selection for SSL algorithms [27] or the use of the datasets from different years as distinct views inside multi-view SSL algorithms. Furthermore, use of non-financial rates for constructing a separate view for Co-training scheme or exploitation of alternative weighting fucntions inside Rel-RASCO method, could provide significant improvement.

Finally, all the techniques employed in the problem of predicting bankruptcy can be straight forwardly used in other financial classification problems such as prediction of fraudulent financial statements [28].

References

[1] Edward I. Altman and E. Hotchkiss, *Corporate Financial Distress and Bankruptcy: Predict and Avoid Bankruptcy, Analyze and Invest in Distressed Debt*, 3rd Editio. Wley, 1993.

[2] F.-M. Tseng and Y.-C. Hu, "Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks," *Expert Syst. Appl.*, vol. 37, no. 3, pp. 1846–1853, Mar. 2010.

[3] V. García, A. I. Marqués, and J. S. Sánchez, "An insight into the experimental design for credit risk and corporate bankruptcy prediction systems," *J. Intell. Inf. Syst.*, vol. 44, no. 1, pp. 159–189, Sep. 2014.

[4] C.-L. Chuang, "Application of hybrid case-based reasoning for

enhanced performance in bankruptcy prediction," *Inf. Sci. (Ny).*, vol. 236, pp. 174–185, 2013.

[5] S. Balcaen and H. Ooghe, "35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems," *Br. Account. Rev.*, vol. 38, no. 1, pp. 63–93, Mar. 2006.

[6] P. Ravi Kumar and V. Ravi, "Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review," *Eur. J. Oper. Res.*, vol. 180, no. 1, pp. 1–28, Jul. 2007.

[7] S. Tian, Y. Yu, and M. Zhou, "Data Sample Selection Issues for Bankruptcy Prediction," *Risk, Hazards Cris. Public Policy*, vol. 6, no. 1, pp. 91–116, Mar. 2015.

[8] C. Serrano-Cinca and B. Gutiérrez-Nieto, "Partial least square discriminant analysis for bankruptcy prediction," *Decis. Support Syst.*, vol. 54, no. 3, pp. 1245–1255, 2013.

[9] 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.

[10] X. Zhu and A. B. Goldberg, *Introduction to Semi-Supervised Learning*, vol. 3, no. 1. Morgan & Claypool, 2009.

[11] 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.

[12] Z.-H. Zhou and M. Li, "Tri-training: exploiting unlabeled data using three classifiers," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1529–1541, Nov. 2005.

[13] F. Muhlenbach, S. Lallich, and D. A. Zighed, "Identifying and Handling Mislabelled Instances," *J. Intell. Inf. Syst.*, vol. 22, no. 1, pp. 89–109.

[14] 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.

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

[16] 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.

[17] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," *Proc. Ninth Int. Conf. Inf. Knowl. Manag. - CIKM '00*, pp. 86–93, 2000.

[18] 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.

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

[20] 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.

[21] J. Abellán and C. J. Mantas, "Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring," *Expert Syst. Appl.*, vol. 41, no. 8, pp. 3825–3830, 2014.

[22] C. F. Tsai, Y. F. Hsu, and D. C. Yen, "A comparative study of classifier ensembles for bankruptcy prediction," *Appl. Soft Comput. J.*, vol. 24, pp. 977–984, 2014.

[23] D. Deligianni and S. Kotsiantis, "Forecasting Corporate Bankruptcy with an Ensemble of Classifiers," in *SETN*, 2012, pp. 65–72.

[24] 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.

[25] B. Calvo and G. Santafé, "scmamp: Statistical Comparison of Multiple Algorithms in Multiple Problems." pp. 1–10, 2015.

[26] K. Kasper, J. Gentry, L. Long, R. Gentleman, S. Falcon, F. Hahne, and D. Sarkar, "Rgraphviz: Provides plotting capabilities for R graph ob-jects. R package version 2.15." 2016.

[27] W. Bo, J. Yan, and Y. Shuqiang, "Forward semi-supervised feature selection based on Relevant set correlation," *Proc. - Int. Conf. Comput. Sci. Softw. Eng. CSSE 2008*, vol. 4, no. 60703110, pp. 210–213, 2008.

[28] S. Chen, Y. J. Goo, and Z. Shen, "A Hybrid Approach of Stepwise Regression , Logistic Regression , Support Vector Machine , and Decision Tree for Forecasting Fraudulent Financial Statements," *Sci. World J.*, vol. 2014, p. 9, 2014.