Default Prediction Using Network Based Features

Poenaru-Olaru, Lorena; Redi, Judith; Hovanesyan, Artur; Wang, Huijuan

DOI
10.1007/978-3-030-93409-5_60

Publication date
2022

Document Version
Final published version

Published in
Complex Networks and Their Applications X

Citation (APA)
Poenaru-Olaru, L., Redi, J., Hovanesyan, A., & Wang, H. (2022). Default Prediction Using Network Based Features. In R. M. Benito, C. Cherifi, H. Cherifi, E. Moro, L. M. Rocha, & M. Sales-Pardo (Eds.), *Complex Networks and Their Applications X : Proceedings of the 10th International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2021, Volume 1* (1 ed., pp. 732-743). (Studies in Computational Intelligence; Vol. 1015). Springer. https://doi.org/10.1007/978-3-030-93409-5_60

Default Prediction Using Network Based Features

Lorena Poenaru-Olaru[1(✉)], Judith Redi[2], Arthur Hovanesyan[3], and Huijuan Wang[4]

[1] Distributed Systems, Delft University of Technology, Delft, The Netherlands
L.Poenaru-Olaru@tudelft.nl
[2] Data Science, Miro, Amsterdam, The Netherlands
[3] Data Science, Exact, Delft, The Netherlands
[4] Multimedia Computing, Delft University of Technology, Delft, The Netherlands

Abstract. Small and medium enterprises (SME) are crucial for economy and have a higher exposure rate to default than large corporates. In this work, we address the problem of predicting the default of an SME. Default prediction models typically only consider the previous financial situation of each analysed company. Thus, they do not take into account the interactions between companies, which could be insightful as SMEs live in a supply chain ecosystem in which they constantly do business with each other. Thereby, we present a novel method to improve traditional default prediction models by incorporating information about the insolvency situation of customers and suppliers of a given SME, using a graph-based representation of SME supply chains. We analyze its performance and illustrate how this proposed solution outperforms the traditional default prediction approaches.

Keywords: Default prediction · Transactional network · Network features · Network-based models · Network centrality

1 Introduction

Small and medium enterprises play a key role in economy. In the Dutch economy for example, they not only generate 61.8% of the overall value of the country but also maintain a high employment rate[1] (64.2% of the total employment). Furthermore, according to Chong et al. [1], they act as important suppliers for large corporates, ensuring in this way the country's product exports and, thereby, the economical growth. Despite being considered the backbone of the economy, SMEs suffer from a higher exposure rate to default than large corporates. The primary cause of this fact is their tremendous vulnerability to economic change [2]. Predicting beforehand that an SME will default in the future could be beneficial in preventing this event, as certain measures could be taken earlier.

[1] Small Business Act for Europe (SBA) Fact Sheet - Netherlands.

R. M. Benito et al. (Eds.): COMPLEX NETWORKS 2021, SCI 1072, pp. 732–743, 2022.
https://doi.org/10.1007/978-3-030-93409-5_60

A plethora of learning models were proposed in literature when it comes to default prediction and improving their accuracy is a direction that plenty of authors are focusing on. The default prediction problem is often referred to as *credit scoring prediction* in literature. For instance, Sang and Nam et al. [3] are investigating the effect of parallel random forest on credit scoring prediction models, while Dayu et al. [4] are looking into extreme learning machines classifiers. However, most of these approaches are only relying on the financial situation of each SME and improving the prediction technique instead of looking into other features. In this paper, we are referring to this type of models as the traditional default prediction models, which consider SMEs as isolated entities instead of treating them as part of a supply chain ecosystem. Recently, Misheva et al. [5] constructed a synthetic network based on the similarities between the financial features of SMEs. It has been shown that traditional default prediction models could be improved by the addition of graph features, such as node degree and closeness centrality.

Beyond financial features of SMEs, we aim to explore whether the consideration of interconnections between SMEs in the supply chain ecosystem could further improve the default prediction.

In this paper, we firstly construct a real-world transactional network composed of around 228.000 Dutch SMEs. It is a temporal (time evolving), undirected and unweighted network, which is measured annually. Two nodes are connected by a link in a year if they have monetary transaction in that year. We analyze the transactional network to identify evidence that the default of an SME is also related to the position of that SME in the transactional network. Furthermore, we propose a novel yearly default prediction model which incorporates both financial and network-based features.

The traditional model that only contains financial features is considered as our baseline. In the hybrid model that we proposed, we systematically consider diverse nodal network features including both network centrality features, beyond the node degree and closeness considered in [5] and graph embedding features. Hence, nodal properties derived from the network topology and the embedding space have been taken into account. We further perform an in-depth analysis of existing methods of handling the issue of training on highly imbalanced classes, since the phenomena of high-class imbalance in default prediction may significantly affect the performance of machine learning models.

This paper is structured as follows: In Sect. 2 we present the intuition behind our idea, as well as some network analysis to motivate why the network features could be relevant in default prediction. Section 3 introduces our prediction method, while Sect. 4 depicts the performance analysis of the proposed method as well as the interpretation of the obtained results. Section 5 contains our conclusions and proposals for future work.

2 Temporal Transactional Network

The transactional network is an abstraction of SMEs interactions in which nodes represent SMEs and an edge between nodes in a given year indicates that the

SMEs are in a business relationship (customer and supplier). The network is evolving over time in the sense that new SMEs are joining the network and edges may appear or disappear over time, in case business relationships are created or broken, respectively. This type of networks is referred to as temporal networks, and they have been proven to be successful in studying epidemic spreading, for instance [6]. The temporal transaction network can be regarded as 9 static network snapshots, measured yearly from 2011 until 2019. In each year, one node can be in either of the two states, *defaulted* or *non-defaulted*. The defaulted state means that the SME has serious financial issues reported in a specific year or it is bankrupt, while the non-defaulted state is assigned to financially healthy SMEs. We will also refer to the transactional network as a business graph.

Intuitively, the interactions between SMEs could be relevant for their financial situation. Assume that the SME supplier A has two customers SME B and C in a given period. The default of either B or C may result in their inability to pay their debts to the supplier A and, therefore, in A*'s* financial stability degradation. However, if SME A would collaborate with far more SMEs than two, the default of one of its counterparts would not be as tremendous as in the first case. Therefore, the interactions between SMEs could be relevant indicators of their financial situation. This could be further supported via the following basic network analysis.

Defaulted Sub-network Extraction. This technique was previously used by Yoshiyuki [7] to understand the phenomena of bankruptcy and check whether it could be modelled as an epidemic spreading process. We used the last snapshot of the network in 2019 including both defaulted and non-defaulted nodes, from which we further extracted the sub-network which contains only defaulted SMEs (nodes) and the links between them. We focused solely on the connected components of this sub-network, in which every defaulted node could reach any other defaulted node via a path. The existence of connected components supports the possibility that defaulted nodes could contribute to default of their neighbours.

We found 3 such connected components of 7, 8 and 24 nodes, respectively. We further checked the year when the default of each node started. As an example, the component of 7 nodes is shown in Fig. 1. These defaulted nodes are close to each other in the network such that they form a connected component. Moreover, these nodes started to default in a similar time.

Distance Between Defaulted Nodes. We further employ a statistical method to evaluate whether the defaulted nodes are close to each other in the network. In this sense, we examine whether the distance between defaulted nodes is smaller than the ones of non-defaulted nodes. We compare the average shortest path between defaulted nodes $E_D[H]$ with the one between non-defaulted nodes $E_{ND}[H]$ on the transactional network snapshot 2019. Given the large number n of nodes and the complexity of computing the shortest path between two nodes, we made use of sampling and statistical tests to obtain the approximate $E_{ND}[H]$. We initially randomly chose 735 non-defaulted nodes (the same as the number defaulted ones) and we calculated the average shortest path between these nodes on the transactional network in 2019. We repeated this procedure 20 times and,

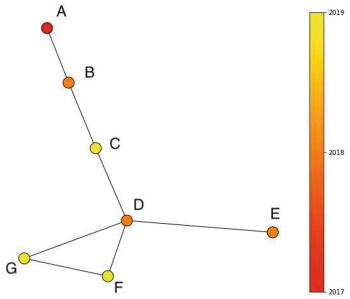

Fig. 1. A representation of 7 interconnected defaulted SMEs extracted from out transactional network with respect to the year of their default.

in the end, we took the mean of the average shortest path over the 20 iterations as $E_{ND}[H]$. Indeed, the average shortest path between defaulted nodes is smaller $E_D[H] = 3.26 < E_{ND}[H] = 3.49$. We further employ a paired difference statistical test, the Wilcoxon signed-rank test, and validate that defaulted nodes are closer to each other in the network. Hence, the position of SME in the network is relevant for its default.

3 Prediction Method

The objective is to predict whether an SME will default or not in a given year $t + 1$ based on its financial and network characteristics in the previous year t. When designing the model we pair the features calculated at the end of year t with labels of year *t+1*. In the following two subsections we will motivate our choices in terms of both financial features and network features.

3.1 Financial Features

One important step of model construction was determining appropriate features. In default prediction literature, *financial features*, which indicate the financial situation of an SME at a particular moment have been widely studied.

For the financial baseline model we consider the following financial coefficients as features: *Cash*, *Current Assets*, *EBIT*, *EBITDA*, *Equity Book Value*, *Interest Expenses*, *Retained Earnings*, *Revenue*, *Short Term Debt*, *Total Assets*, *Total Liabilities* and *Working Capital*.

Given that we worked with a real-world data-set, we encountered the situation of having missing data in terms of financial coefficients. The missing of financial coefficients can be due to the inactivity of a company. Thus, we considered that financial coefficient for which the data is missing are 0. This explains also why we have not considered ratios of financial coefficients as features, as in other works. [8]

3.2 Network Features

In order to incorporate graph information, we extracted some *network features* from the nodes that we, thereafter, combined with the financial ones to understand whether they improve the accuracy of traditional models. Through our analysis we observed that network features taken alone are not informative enough to predict default, thus we opted for the combination. We consider the following representative nodal network properties [9], also called centrality metrics:

Node Degree, which is the number of links that incident to the node.
Clustering Coefficient, which is the probability that the neighbours of a node are connected. It measures the probability that two collaborators of an SME also collaborate.
Eigenvector Centrality, which is principal eigenvector component corresponding to the node. The principal eigenvector is the eigenvector corresponding to the largest eigenvalue of the adjacency matrix of the network. A node tends to have a large eigenvector centrality if it is connected to many well connected nodes.

Centrality metrics like *closeness* and *betweenness* [10] will not be considered because of their high computational complexity, actually associated with the shortest path computation. Li et al. have investigated the correlation between the network centrality metrics via both theoretical analysis and experiments in real-world networks [11]. They found that metrics with a high computational complexity like the betweenness tend to be correlated with metrics with a low computational complexity in diverse types of networks. This supports that information of closeness and betweenness could be captured by the three centrality metrics that we consider and the graph embedding features that we will introduce.

Another type of network derived features are the *graph embeddings* also known as *network embeddings*. Network embedding aims to represent a network by assigning coordinates to nodes in a low-dimensional vector space[6,12]. The embedding vectors of the nodes will also be considered as network features. We use node2vec, a random walk based network embedding to derive the embedding vectors of the nodes [13]. Specifically, the following configuration was considered: p = 1, q = 1, number of walks = 10, walk length = 80. We experimented with different dimensions of the embedding vectors and chose the best performing model. Thereby, the optimal dimension of graph embedding in our case was 4.

We further created multiple *hybrid models* which contains both financial and network based features. In this sense, we extended the financial coefficients model with the network features in order to understand whether the classification accuracy could be improved by incorporating network information.

3.3 High Class Imbalance Problem

An ubiquitous issue that rises when predicting default is the problem of having the 2 classes, defaulted and non-defaulted, extremely imbalanced. This is the

result of having annually a significantly higher number of non-defaulted SMEs than defaulted SMEs. This usually has tremendous effects on the classifier's performance to distinguish between the two classes. The reason for this is the fact that the classifier does not see enough samples of the minority class to be able to further extrapolate. For instance, if the training set is composed of data from 2011 until 2018, then the percentages between defaulted and non-defaulted samples would be 0.06% and 99.94%, respectively.

In order to overcome this problem, we employed 2 *data-driven methods*, *undersampling* and *oversampling*, and one *algorithm driven method weighting*, presented in Leevy et al.'s survey [14].

We, thereby, undersampled the majority class to lower the number of non-defaulted SMEs and oversampled the minority class to increase the number of defaulted SMEs in the training set. The undersampling was done such that the 2.5% default rate of SMEs in the Netherlands was preserved. Thus, every year we should have around 2.5% defaulted SMEs and around 97.5% non-defaulted SMEs in the training set. We applied stratification to perform undersampling, which ensures the inclusion of SMEs from different categories. Our chosen categories were sector and company size. In terms of oversampling, we used SMOTE [15], which creates synthetic samples of the minority class by interpolating between similar samples. The weighting method assigns a higher weight to the minority class to penalise the model when having the tendency of classifying everything as non-defaulted in order to preserve accuracy.

Besides the previously mentioned techniques, we also considered combinations between them, such as *undersampling + SMOTE* and *undersampling + weighting the minority class* as they are commonly used in class imbalance literature.

3.4 Classifiers

As for the machine learning algorithms, we consider diverse *tree based classification models* as they have been proved efficient in prediction problems where the classes are strongly imbalanced, such as anomaly detection, default prediction and fraud detection [16,17]. Specifically, we employed multiple tree based classifiers, ranging from *simple* classifiers, such as Decision Tree and Random Forest, to *boosting* algorithms, such as AdaBoost, XGBoost and LightGBM.

In Sect. 4, we will firstly evaluate all classifiers and class imbalance methods using the baseline model where only financial features are taken into account. The best combination of the classifier and class imbalance method will be identified and used further to compare our hybrid model that incorporates both network and financial features with the baseline.

4 Performance Analysis

In this section, we will design the experiments to evaluate our methods.

4.1 Experimental Setup

First, we give a comprehensive picture of our experimental setup. Our data-set records the financial and network information of Dutch SMEs from 2011 until 2019. We split into the training set (samples from 2011 to 2018), which was used in order to learn the behaviour of defaulted and non-defaulted SMEs, and testing set (2019), which was employed to evaluate the model's performance. This choice is motivated by our objective to evaluate the model's performance close to current time and the fact that most of the defaulted samples are reported in 2019.

In Sect. 4.4, different choices of the training and test sets will be considered to explore the robustness of the model against fluctuation of the economy.

4.2 Preliminary Selection of Classifier and Class Imbalance Method

We evaluated the performance of diverse combinations of the aforementioned classifiers and class imbalance methods using the baseline model where only financial features are considered. We measured the performance in terms of *Area Under the ROC Curve (ROC AUC)* score, where the ROC Curve can be obtained by plotting the *True Positive Rate (TPR)* against the *False Positive Rate (FPR)*. The metric is suitable for class imbalance prediction problem since it shows how well the classifier is distinguishing between the 2 classes. Its output rages from 0 to 1, where 1 corresponds to perfect prediction. [18].

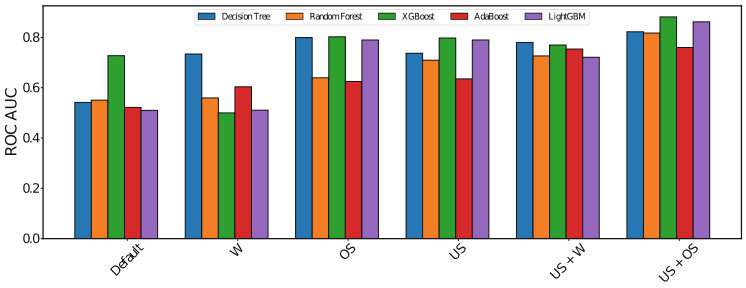

Fig. 2. The predition quality ROC AUC when using different classifiers and the following 6 class imblance methods: **default** in which we do not apply any method to handle class imbalance, **W** in which we assign a higher weight to the minority class, **OS** in which we oversample the minority class using SMOTE, **US** in which we undersample the majority class and the 2 combinations **US + W** and **US + OS**.

Figure 2 shows the results of all the employed classifiers and class imbalance methods in terms of ROC AUC. We observe that the class imbalance method composed of an undersampling technique combined with an oversampling technique (SMOTE) outperforms the other 4 for each classifier. This could be possibly explained by the fact that some non-defaulted samples could be redundant.

Thus, removing the redundancy increases the probability that the classifier sees more relevant samples and its ability to distinguish between the defaulted and non-defaulted increases. Another important observation is that the XGBoost classifier achieves the highest result. Thereby, in our further experiments, we are only considering this particular classifier XGBoost and the undersampling + SMOTE method to overcome the problem of having a high-class imbalance.

4.3 Comparison Between Models

Furthermore, we evaluate whether the hybrid models that include both financial and network features outperform the baseline model that incorporates financial features alone. Besides ROC AUC, we also considered the TPR. Since misclassifying a defaulted SME could result in high financial losses, we need the TPR to understand the percentage of the correctly identified defaulted SMEs.

Different combinations of graph features will be considered in the hybrid model. We use the following abbreviations to denote features:

- **Baseline financial features - B**;
- **Eigenvector Centrality - EC**;
- **Clustering Coefficient - CC**;
- **Node Degree - ND**
- **Graph Embedding - GE**

We began with adding each graph feature to the baseline features and observing whether it improves the ROC AUC or the TPR. Table 1 shows that only the hybrid model that consider the eigenvector centrality outperformed the baseline. Furthermore, we can also observe that by adding the network information eigenvector centrality into the baseline, the model was able to detect with 1% more defaulted SMEs than the initial one. Although this improvement does not seem significant, the model B + EC is more suitable than the baseline, in the sense that correctly classifying as many defaulted SMEs as possible could possibly prevent the loss.

Table 1. Predictions quality of the baseline and of the hybrid model that includes baseline financial features and one graph feature. *The highest ROC AUC and TPR are highlighted.*

Model	ROC AUC	TPR
B	0.881	0.815
B + EC	**0.884**	**0.827**
B + CC	0.872	0.794
B + ND	0.878	0.809
B + GE	0.875	0.803

Furthermore, we evaluate the hybrid model that includes not only EC but also other graph features. The objective is to understand whether considering more network features could further improve the performance. All possible combinations with other network features beyond EC are evaluated in Table 2 .

We find that the combination between the network features leads to an even higher improvement in the model's accuracy. By adding the node degree (ND) or graph embedding (GE) to the B + EC model, the ROC AUC improves further by around 1% and its TPR increases further by around 2%. This improvement is statistically significant according to McNemar's statistical test. The addition of network features to the traditional financial models could indeed improve the default prediction.

Table 2. Prediction quality of the baseline compared with the complex hybrid models that consider the financial feature and the eigenvector centrality and/or other network features. *The highest ROC AUC and TPR are highlighted.*

Model	ROC AUC	TPR
B	0.881	0.815
B + EC	0.884	0.827
B + EC + CC	0.863	0.782
B + EC + ND	**0.894**	**0.848**
B + EC + GE	0.892	0.842
B + EC + CC + ND	0.878	0.815
B + EC + CC + GE	0.864	0.782
B + EC + ND + GE	0.874	0.806
B + EC + ND + GE + CC	0.861	0.776

4.4 Robustness of the Optimal Model

Within this subsection we explore how robust the best hybrid model (B + EC + ND) is. In the previous experiments, we have trained the model on data from 2011 until 2018 and tested on 2019. A robust model is supposed to be able to perform well when tested on different years, thus robust against fluctuation of the economy. To do so, we include the samples of 2019 into the training set and extract the samples from the other years, one by one, in order to use them as testing sets. We depict our results in Table 3.

From Table 3 we can observe the fact that the lowest obtained TPR was 0.8, which means that our model succeeded in correctly determining more than 80% defaulted SMEs each year. Hence, the B + EC + ND model is relatively robust to the changes in the economy, thus reliable when deployed into production. The tests since 2014 are more representative in the sense that more defaults occur since 2014 than before 2014. Thus, there is a higher probability that a model misclassifies a higher number of defaulted samples compared to a lower one.

Table 3. Evaluation of the B+EC+ND model when tested in each possibly year.

Year	ROC AUC	TPR
2011	0.998	1
2012	0.997	1
2013	0.997	1
2014	0.914	0.833
2015	0.898	0.8
2016	0.928	0.862
2017	0.932	0.870
2018	0.906	0.819
2019	0.894	0.848

4.5 Interpretation of Results

From our previous experiments, we observed that there were only 3 hybrid models that were able to achieve better performance than the baseline (B), namely B+EC, B+EC+ND and B+EC+GE. In the following part of this subsection, we are focusing on analyzing the meaning of the chosen network features within the business aspects.

The B+EC model is composed of financial features and one particular network feature, the eigenvector centrality. In complex networks theory, the eigenvector centrality shows whether one particular node has many highly connected neighbors. In our case, we observed that an SME with a high eigenvector centrality has a lower likelihood of default. In other words, a company that is surrounded by many highly connected neighbours is less likely to default. This explains why the consideration of the eigenvector centrality could improve the prediction.

The B+EC+ND model adds two types of network features to the traditional financial features, namely the node degree and the eigenvector centrality. The node degree shows how many connections does a particular node have. Our findings reveal that the further augmentation of the role of degree beyond the eigenvector centrality could improve the performance of the model.

The B+EC+GE model includes financial features, eigenvector centrality and the embedding vector of an SME. The embedding vector of a node is supposed to capture the information of a network that is different from but possibly correlated with centrality metrics like degree and eigenvector. Nodes play a similar role, e.g. being the hub of a local community could possibly have a similar embedding although they are not close to each other in the network topology. Hence, graph embeddings could possibly carry valuable information regarding the company status. We observed the hybrid model B+GE that includes the graph embedding features and financial features performs even worse than the baseline. However, the model B+EC+GE that combines the eigenvector centrality, embedding features and financial features performs better than both the baseline and the B+EC model. In summary, adding network features does not

necessary improve the traditional default prediction model. By adding the appropriate network features, e.g. the embedding combined with the eigenvector of a node, the prediction be evidently improved.

5 Conclusions and Future Work

In this paper, we have developed the method to improve financial feature based default prediction models by incorporating network-based features extracted from a real-world transactional network composed of Dutch SMEs. This method entails the construction of the transactional network and the systematic inclusion of diverse network features including centrality metrics in the network topology domain and embedding vectors of nodes in the embedding space, beyond the choice of the classifier and method to overcome the class imbalance problem. We observed and demonstrated that our hybrid model performs better than the baseline financial model, especially in terms of identifying as many defaults as possible when the network features have been appropriately chose. The combination of node degree and the eigenvector centrality enhances the traditional default prediction model the most. Moreover, through our evaluation over years, we demonstrated that the hybrid model, which achieved the highest performance, is robust to economical changes. Additionally, we provided an interpretation of the network features in a business context in order to explain why they improve the baseline.

In terms of future work, we believe our hybrid model including its design and performance analysis can be further explored in the following directions. As a start, we have considered the undirected and unweighted transactional network. The volume and the direction (the customer-supplier relationship) of the monetary transactions between SMEs can be relevant for default prediction. Hence, the weighted and directed transactional network can be further investigated. To illustrate our method, we have selected the classifier and class imbalance method that performed the best in the baseline model to further evaluate the hybrid model. Other combinations of the classifier and class imbalance method, further fine-tuned hyperparameters in the network embedding could be used to evaluate the hybrid model. Regarding of the choice of the network features, more combinations of network features could be considered especially those with a low computational complexity.

Disclaimer. The information made available by Exact for this research is provided for use of this research only and under strict confidentiality. We would, therefore, like to thank Exact for providing us with resources to pursue this project.

References

1. Chong, S., et al.: The role of small- and medium-sized enterprises in the Dutch economy: an analysis using an extended supply and use table. J. Econ. Struct. **8**, 12 (2019)

2. Asgary, A., Özdemir, A., Özyürek, H.: Small and medium enterprises and global risks: evidence from manufacturing SMEs in Turkey. Int. J. Disaster Risk Sci. **11**, 59–73 (2020). https://doi.org/10.1007/s13753-020-00247-0
3. Ha, S., Nam, N., Nhan, N.: A novel credit scoring prediction model based on feature selection approach and parallel random forest. Indian J. Sci. Technol. **9**, 05 (2016)
4. Xu, D., Xuyao, Z., Hu, J., Chen, J.: A novel ensemble credit scoring model based on extreme learning machine and generalized fuzzy soft sets. Math. Probl. Eng. **2020**, 1–12 (2020)
5. Misheva, B.H., Giudici, P., Pediroda, V.: Network-based models to improve credit scoring accuracy. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 623–630 (2018)
6. Zhan, X.-X., Li, Z., Masuda, N., Holme, P., Wang, H.: Susceptible-infected-spreading-based network embedding in static and temporal networks. EPJ Data Sci. **9**, 30 (2020)
7. Yoshiyuki, A.: Bankruptcy propagation on a customer-supplier network: an empirical analysis in Japan (2018)
8. Altman, E.I., Sabato, G.: Modeling credit risk for SMEs: evidence from the us market (2007)
9. Rodrigues, F.A.: Network centrality: an introduction. arXiv: Physics and Society, pp. 177–196 (2019)
10. Wang, H., Hernandez, J.M., Van Mieghem, P.: Betweenness centrality in a weighted network. Phys. Rev. E **77**, 046105 (2008)
11. Li, C., Wang, H., Haan, W., Stam, C., Mieghem, V.: The correlation of metrics in complex networks with applications in functional brain networks. J. Stat. Mech: Theor. Exp. **2011**, 11 (2011)
12. Cui, P., Wang, X., Pei, J., Zhu, W.: A survey on network embedding. IEEE Trans. Knowl. Data Eng. **31**, 833–852 (2019)
13. Grover, A., Leskovec, J.: node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2016, pp. 855–864. ACM Press (2016)
14. Leevy, J.L., Khoshgoftaar, T.M., Bauder, R.A., Seliya, N.: A survey on addressing high-class imbalance in big data. J. Big Data **5**, 1–30 (2018)
15. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
16. Brennan, P.J.: A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection (2012)
17. Maurya, C.K., Toshniwal, D., Venkoparao, G.V.: Online anomaly detection via class-imbalance learning. In: 2015 Eighth International Conference on Contemporary Computing (IC3), pp. 30–35 (2015)
18. Bekkar, M., Djema, H., Alitouche, T.: Evaluation measures for models assessment over imbalanced data sets. J. Inf. Eng. Appl. **3**, 27–38 (2013)