Physica A 492 (2018) 389–394

Contents lists available at ScienceDirect

Physica A

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

A hybrid network-based method for the detection of disease-related genes

Ying Cui [a,c,d], Meng Cai [b,c,*], Yang Dai [e], H. Eugene Stanley [c]

[a] School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China
[b] School of Economics and Management, Xidian University, Xi'an 710071, China
[c] Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA
[d] Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an 710071, China
[e] School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China

HIGHLIGHTS

- Comparatively analyzed the topological properties between disease-related genes and non-disease genes in protein–protein interaction network.
- Disease-related genes were found have distinct network topological properties with non-disease genes.
- An improved forest-based model was applied as classifier.
- The proposed hybrid networked based disease gene detection method was proven to perform better than previous similar studies in accuracy.

ARTICLE INFO

Article history:
Received 2 June 2017
Received in revised form 1 October 2017
Available online 31 October 2017

Keywords:
Disease gene detection
Topological properties
PPI network
Random forest

ABSTRACT

Detecting disease-related genes is crucial in disease diagnosis and drug design. The accepted view is that neighbors of a disease-causing gene in a molecular network tend to cause the same or similar diseases, and network-based methods have been recently developed to identify novel hereditary disease-genes in available biomedical networks. Despite the steady increase in the discovery of disease-associated genes, there is still a large fraction of disease genes that remains under the tip of the iceberg. In this paper we exploit the topological properties of the protein–protein interaction (PPI) network to detect disease-related genes. We compute, analyze, and compare the topological properties of disease genes with non-disease genes in PPI networks. We also design an improved random forest classifier based on these network topological features, and a cross-validation test confirms that our method performs better than previous similar studies.

1. Introduction

Detecting disease-related genes, an important yet challenging task in human genetics, enables us to upgrade disease diagnostic tools and drug design. A disease is not caused by an aberration in a single gene but by perturbations in entire cellular systems, especially molecular networks [1]. The effort to identify disease genes in molecular networks has led to the development of "network medicine" [1,2], which recapitulates the molecular complexity of human disease and offers network-based computational methods to identify how molecular complexity manipulates human disease.

* Corresponding author at: School of Economics and Management, Xidian University, Xi'an 710071, China.
E-mail addresses: ycui@xidian.edu.cn (Y. Cui), mcai@xidian.edu.cn (M. Cai).

https://doi.org/10.1016/j.physa.2017.10.026
0378-4371/

Research dedicated to systematically capturing the properties of disease-associated genes in molecular networks has found that genes related to the same or similar diseases tend to cluster and interact with each other [3–5]. This finding has promoted the development of network-based approaches to identifying and prioritizing candidate disease genes by using such biological network data as protein–protein interaction (PPI) networks [6–8], disease phenotype networks [9–13], regulatory networks [14–16] and co-expression networks [17–19].

Although there has been a substantial effort to detect disease genes, the current results are far from satisfactory [20]. Most efforts use sophisticated integrated data sources [9–19] that are time consuming and that demand huge computer resources. Our response is to propose a novel network-based approach to detecting disease-related genes using only PPI network data.

A PPI network consists of physical interactions between proteins and is a powerful data source when detecting disease genes [21]. The strong connection between proteins and human disease confirm that proteins that physically interact with each other share a common function [3,22]. Thus an aberration in one protein tends to replicate similar disease phenotypes.

In this paper we propose a hybrid network-based method for disease-related gene detection. We analyze the topology of a PPI network and find that disease genes have properties that enable us to distinguish them from non-disease genes. We use these topological features and an improved random forest model ("CForest") to detect disease genes. The simulation results indicate that our method achieves 85.01% accuracy, outperforming previous similar studies [6,8]. By deciphering the topical properties of PPI networks, we have significantly improved the disease gene detection accuracy and expanded our understanding of complex genotype–phenotype relationships.

2. Materials and methods

2.1. Data sources

To construct a PPI network, we derive 37336 protein–protein interactions in 9213 proteins from the Human Protein Reference Database (HPRD, release 9) [23], one of the most reliable databases for PPI data [24]. We obtain the list of disease-associated genes and non-disease genes supplied by the Online Mendelian Inheritance in Man (OMIM) [25]. Only genes that the molecular basis for their related disorders is known or a mutation has been found in this gene are considered as disease gene samples. We find 2512 genes that have at least one related disease phenotype and one PPI in the PPI network. We also find 5534 genes with interactions in HPRD but no disease association record in OMIM, and we classify them non-disease genes.

2.2. Network topological properties analysis

Disease genes have distinct network topological properties that distinguish them from non-disease genes [26]. We construct a PPI network and analyze how the topological properties of disease genes in the network differ from those of non-disease genes. We use seven network topological measurements, (i) degree, (ii) average nearest-neighbor degree, (iii) authority centrality, (iv) betweenness centrality, (v) closeness centrality, (vi) eigenvector centrality, and (vii) PageRank to capture the topological properties of the PPI network. The topological features are listed in Table 1, which also introduces their function, description, and reference.

These features indicate the topological importance of a gene in a PPI network from different perspectives. For example, the node degree indicates the number of edges connecting to this node, and the more neighbors a node has, the more influential it is. The K-nearest neighbor is the average degree of a node's nearest neighbors, which complements node degree because if the degree of a node's nearest neighbor changes it changes the node's topological importance. Betweenness and closeness are path-based centrality measurements. Betweenness measures the control of nodes flowing along the shortest path in the network, and closeness measures node importance using the average path length of information propagation in the network. The remaining three are eigenvector-based measurements that assign relative scores to all nodes in the network. Here connections with high-scoring nodes contribute more to the score of the node than an equal number of connections with low-scoring nodes.

We compute all of the topological properties in R 3.2.3, and compare the mean value and the median of these features in (i) disease genes, (ii) non-disease genes, and (iii) all genes to topologically discriminate disease genes from non-disease genes. Fig. 1 shows the analytic results.

Fig. 1 shows that disease genes score higher in degree, authority centrality, betweenness centrality, eigenvector centrality, and PageRank, but lower in average nearest-neighbor degree and closeness centrality. These network properties enable us to distinguish disease genes from non-disease genes.

2.3. Classifier

For this work we use CForest, an improved random forest (RF) classifier [34,35]. Unlike the standard RF classifier that uses a majority voting rule with unfair splitting criterion, CForest uses an unbiased base classification trees in a conditional inference framework. In CForest framework, Strobl et al. proposed a conditional permutation importance scheme [35] that partitions the entire feature space using a fitted forest model. This determines the influence of a variable and computes

Table 1
Topological properties of PPI network.

Features	Function	Description	Reference
Degree	k_i	The number of edges connected with node i.	[27]
K—Nearest Neighbor	$\frac{\sum_{i=1}^{k} k_i}{k}$	The average nearest neighbor degree of node i with degree k.	[28]
Authority Centrality	$t(A) * Ax = \lambda x$	The principal eigenvector of $t(A) * A$	[29]
Betweenness Centrality	$g(i) = \sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$	σ_{jk} is the total number of shortest paths from node j to node k, $\sigma_{jk}(i)$ is the number of that paths going through node i.	[30]
Closeness Centrality	$C(i) = \frac{N}{\sum_{i \neq j} d(i,j)}$	N is the network size, $d(i, j)$ refers to the shortest path between node i and j.	[31]
Eigenvector Centrality	$Ax = \lambda x$	λ is the eigenvalue of adjacency matrix A.	[32]
PageRank	$PR(i) = (1-d) + d\sum_{j \in N(i)} \frac{PR(j)}{k_i}$	$N(i)$ stands for the neighbors of node i, d is damping factor.	[33]

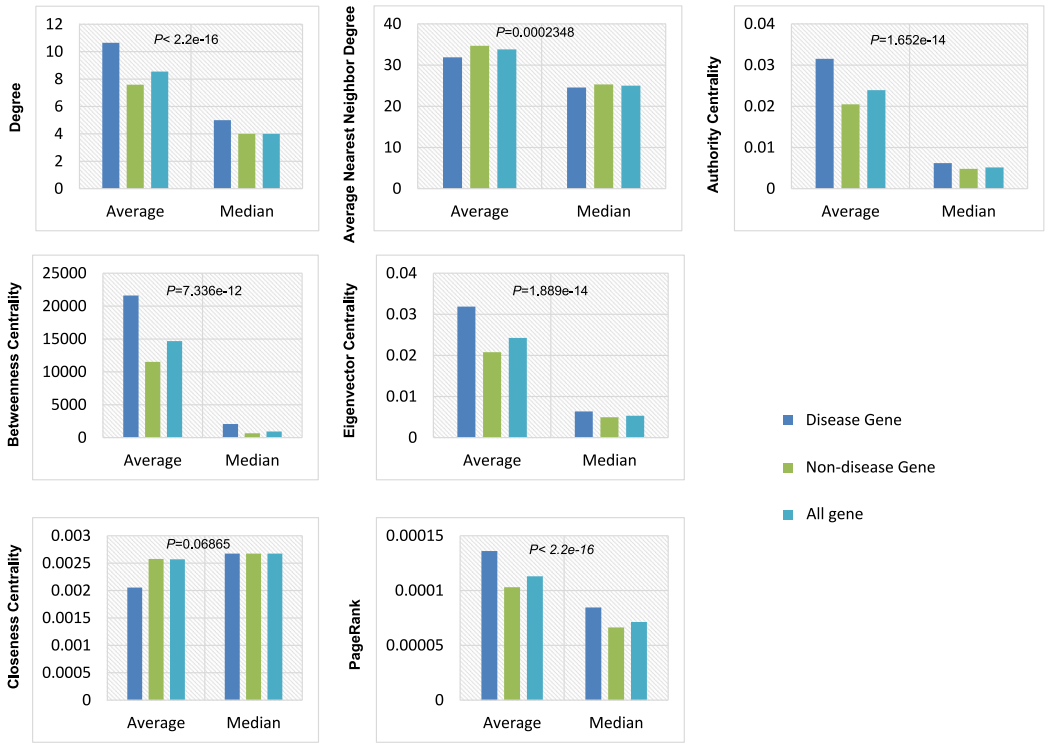

Fig. 1. Comparatively analysis on topological properties between different gene sets.

its permutation importance irrespective of correlated covariate type, i.e., CForest provides an unbiased variable importance measurement based on a conditional permutation scheme for evaluating feature importance.

In the CForest importance measurement, the importance of a predictor variable X_j is evaluated by the difference between the prediction accuracy before and after the permutation. Let $\bar{B}^{(t)}$ represent the out-of-bag (oob) sample for a tree t, with $t \in \{1, \cdots\cdots, ntree\}$. Here $ntree$ is the number of trees in the forest. The oob-prediction accuracy for tree t before the permutation is

$$\frac{\sum_i \bar{B}^{(t)} I\left(y_i = \hat{y}_i^{(t)}\right)}{|\bar{B}^{(t)}|}, \tag{1}$$

where $\hat{y}_i^{(t)} = f^{(t)}(x_i)$ is the predicted class for observation i before permutation.

After permuting the value of X_j, the new accuracy is,

$$\frac{\sum_i \bar{B}^{(t)} I\left(y_i = \hat{y}_{i\pi_j|Z}^{(t)}\right)}{|\bar{B}^{(t)}|}, \tag{2}$$

Table 2
Classification performance of different classifiers.

Classifier	Precision (%)	Recall (%)	Accuracy (%)
K-Nearest Neighbors	65.72	66.01	65.93
Support Vector Machine	69.82	71.57	70.98
Random Forest	70.01	71.76	71.25
CForest	84.83	85.27	85.01

where Z refers to the remaining predictor variables $Z = X_1, \cdots\cdots, X_{j-1}, X_{j+1}, \ldots$. Then the variable importance of X_j in tree t can be expressed

$$VI^{(t)}\left(X_j\right) = \frac{\sum_i \bar{B}^{(t)} I\left(y_i = \hat{y}_i^{(t)}\right)}{|\bar{B}^{(t)}|} - \frac{\sum_i \bar{B}^{(t)} I\left(y_i = \hat{y}_{i\pi_j|Z}^{(t)}\right)}{|\bar{B}^{(t)}|}. \tag{3}$$

Finally the importance score of each variable X_j for the forest is calculated as an average over all trees,

$$VI\left(X_j\right) = \frac{\sum_{t=1}^{ntree} VI^{(t)}\left(X_j\right)}{ntree}. \tag{4}$$

2.4. Performance assessment

We use precision, recall (sensitivity), and accuracy, which are commonly-used evaluation indexes in machine learning methods [36,37], to assess the performance of our approach,

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP}. \tag{7}$$

We calculate these performance measurements using the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) at a score cut-off that distinguishes predicted from non-predicted. Positives are disease genes and negatives are non-disease genes.

3. Results

We combine all seven network topological properties to create an input classifier vector. We then apply CForest as the classifier on the combined feature sets in the R 3.2.3 system by using a 10-fold cross-validation. We consider all 2512 disease genes to be positive samples, and we randomly select 2512 non-disease genes to be negative samples. All 2512 gene pairs are randomly and equally divided into ten sets. Each set is used as a testing set, and the remaining nine are used for training. We use a total of ten testing sets, and each training set is nine times the size of its corresponding testing set. In CForest model, two parameters play important role — *ntree* and *mtry*. The *ntree* parameter is the number of trees in the forest, and the *mtry* parameter is the number of variables available for splitting at each tree node. A grid-search with 10-fold cross-validation are used to determine the best *ntree* and *mtry* in our classification model. We find that the CF classifier performs best when $ntree = 500$ and $mtry = 3$. Our method proves to be effective in disease gene detection with 84.83% precision, 85.27% recall, and 85.01% accuracy.

To highlight the good performance of CForest in discrimination of disease and non-disease genes according network topological features, we attempted to compare the classification results of CForest with other classical classifier including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF) in this work. In the conduction of each classifier model, parameters were optimized by using grid-search with 10-fold cross-validation. The classification performance of different classifiers are displayed in Table 2. It is obvious that CForest classifier perform better than other classifiers.

As described in Section 2.3, because CForest provides an unbiased measure of variable importance, we can evaluate the classification importance of a feature. Fig. 2 shows the seven features ranked according their importance score generated by CForest classifier. The rank of a features indicates their importance in discriminating disease genes from non-disease genes.

To demonstrate the superiority of our method, we compare its precision, recall, and accuracy to those of other PPI network-based methods [6,8]. Fig. 3 shows the results of this comparison, which indicate that our method is significantly better than the methods proposed in Refs. [6] and [8].

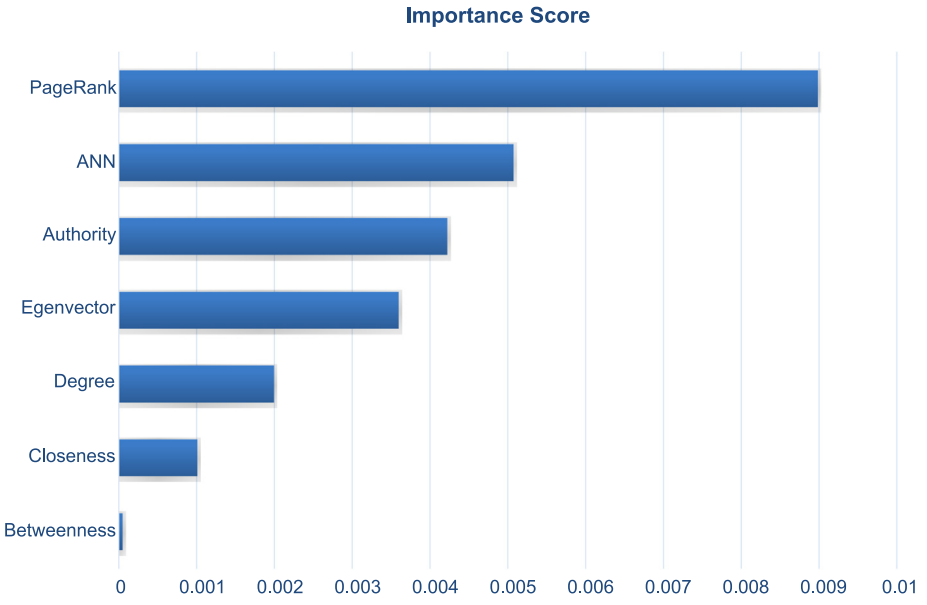

Fig. 2. Importance rank of topological properties.

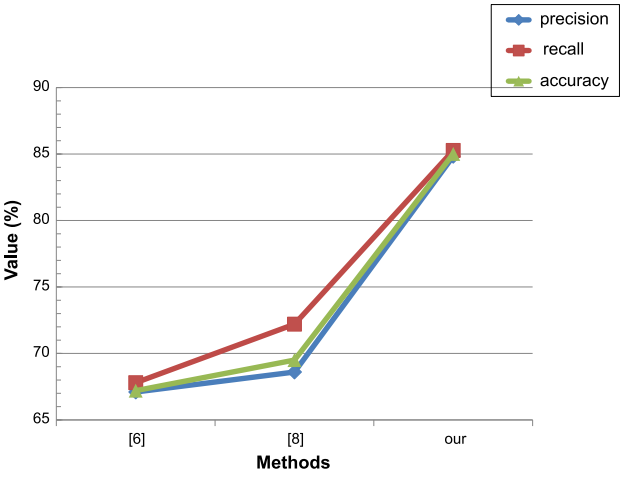

Fig. 3. Comparative performance of different methods.

4. Conclusion

We have used the topological properties of a PPI network to detect disease-related genes. We first compute the topological properties, i.e., degree, average nearest-neighbor degree, authority centrality, betweenness centrality, closeness centrality, eigenvector centrality, and PageRank to evaluate the topological importance of a gene in the PPI network from various perspectives. We then compare these topological properties and find that they clearly distinguish disease-related genes from non-disease genes. Disease genes score higher in degree, authority centrality, betweenness centrality, eigenvector centrality, and PageRank than non-disease genes, but lower in average nearest-neighbor degree and closeness centrality.

We then combine these topological properties to serve as classifier input. We use an improved random forest classifier model (CForest) based on unbiased base classification trees in a conditional inference framework to distinguish disease genes from non-disease genes. In addition, we rank the importance of topological properties using the variable importance measure supplied by CForest, which expands our understanding of sophisticated genotype–phenotype associations. We also compare the precision, recall, and accuracy performance of our method with those of other PPI network-based methods. We find that the results produced by our method are a significant improvement over those of other methods.

Acknowledgments

This work is financially supported by grants from Natural Science Foundation of Shaanxi Province of China (No. 2016JQ6072), National Natural Science Foundation of China (No. 71501153), Xi'an Science and Technology Program (No. XDKD001), and China Scholarship Council (201506965039, 201606965057).

References

[1] A.L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: A network-based approach to human disease, Nat. Rev. Genet. 12 (2011) 56–68.
[2] W. Wang, M. Tang, H.E. Stanley, et al., Unification of theoretical approaches for epidemic spreading on complex networks, Rep. Progr. Phys. 80 (2017) 036603.
[3] M. Oti, H.G. Brunner, The modular nature of genetic diseases, Clin. Genet. 71 (2007) 1–11.
[4] K.I. Goh, M.E. Cusick, D. Valle, et al., The human disease network, Proc. Nat. Acad. Sci. USA 104 (2007) 8685–8690.
[5] I. Feldman, A. Rzhetsky, D. Vitkup, Network properties of genes harboring inherited disease mutations, Proc. Nat. Acad. Sci. USA 105 (2008) 4323–4328.
[6] J. Xu, Y. Li, Discovering disease-genes by topological features in human protein-protein interaction network, Bioinformatics 22 (2006) 2800–2805.
[7] J. Chen, B.J. Aronow, A.G. Jegga, Disease candidate gene identification and prioritization using protein interaction networks, BMC Bioinformatics 10 (2009) 73.
[8] S. Wu, F. Shao, R. Sun, et al., Analysis of human genes with protein–protein interaction network for detecting disease genes, Physica A 398 (2014) 217–228.
[9] J. Li, X. Lin, Y. Teng, et al., A comprehensive evaluation of disease phenotype networks for gene prioritization, PLoS One 11 (2016) e0159457.
[10] Y. Li, J.C. Patra, Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network, Bioinformatics 26 (2010) 1219–1224.
[11] U.M. Singh-Blom, N. Natarajan, A. Tewari, et al., Prediction and validation of gene-disease associations using methods inspired by social network analyses, PLoS ONE 8 (2013) 58977.
[12] Y. Chen, T. Jiang, R. Jiang, Uncover disease genes by maximizing information flow in the phenome-interactome network, Bioinformatics 27 (2011) 167–176.
[13] O. Vanunu, O. Magger, E. Ruppin, et al., Associating genes and protein complexes with disease via network propagation, PLoS Comput. Biol. 6 (2010) 1000641.
[14] S. Zickenrott, V.E. Angarica, B.B. Upadhyaya, et al., Prediction of disease-gene-drug relationships following a differential network analysis, Cell Death Dis. 7 (2016) e2040.
[15] J.C. Chen, M.J. Alvarez, F. Talos, et al., Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks, Cell 159 (2014) 402–414.
[16] E. Khurana, Y. Fu, J. Chen, et al., Interpretation of genomic variants using a unified biological network approach, PLoS Comput. Biol. 9 (2013) e1002886.
[17] S. van Dam, U. Võsa, A. van der Graaf, et al., Gene co-expression analysis for functional classification and gene-disease predictions, Brief. Bioinform. (2017) bbw139.
[18] M. Ernst, Y. Du, G. Warsow, et al., FocusHeuristics–expression-data-driven network optimization and disease gene prediction, Sci. Rep. 7 (2017) 42638.
[19] P. Maji, E. Shah, S. Paul, RelSim: An integrated method to identify disease genes using gene expression profiles and PPIN based similarity measure, Inform. Sci. 384 (2017) 110–125.
[20] C.J. Mungall, N.L. Washington, J. Nguyen-Xuan, et al., Use of model organism and disease databases to support matchmaking for human disease gene discovery, Hum. Mutat. 36 (2015) 979–984.
[21] S. Navlakha, C. Kingsford, The power of protein interaction networks for associating genes with diseases, Bioinformatics 26 (2010) 1057–1063.
[22] H.G. Brunner, M.A. van Driel, From syndrome families to functional genomics, Nat. Rev. Genet. 5 (2004) 545–551.
[23] T.K. Prasad, R. Goel, K. Kandasamy, et al., Human protein reference database2009 update, Nucl. Acid. Res. 37 (2009) D767-D772.
[24] S. Mathivanan, B. Periaswamy, T.K.B. Gandhi, et al., An evaluation of human protein-protein interaction data in the public domain, BMC Bioinformatics 7 (2006) S19.
[25] V.A. McKusick, Mendelian inheritance in man and its online version, OMIM, Amer. J. Hum. Genet. 80 (2007) 588–604.
[26] X. Wang, N. Gulbahce, H. Yu, Network-based methods for human disease gene prediction, Brief. Funct. Genomics Proteomi. 10 (2011) 280–293.
[27] R. Diestel, Graph Theory, Springer, Heidelberg, 2000.
[28] A. Barrat, M. Barthelemy, R. Pastor-Satorras, et al., The architecture of complex weighted networks, Proc. Natl. Acad. Sci. USA 101 (2004) 3747–3752.
[29] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (1999) 604–632.
[30] U. Brandes, A faster algorithm for betweenness centrality, J. Math. Sociol. 25 (2001) 163–177.
[31] L.C. Freeman, S.P. Borgatti, D.R. White, Centrality in valued graphs: A measure of betweenness based on network flow, Social Networks 13 (1991) 141–154.
[32] P. Bonacich, Power and centrality: A family of measures, Amer. J. Sociol. 92 (1987) 1170–1182.
[33] S. Brin, L. Page, Reprint of: The anatomy of a large-scale hypertextual web search engine, Comput. Netw. 56 (2012) 3825–3833.
[34] T. Hothorn, K. Hornik, A. Zeileis, Unbiased recursive partitioning: A conditional inference framework, J. Comput. Graph. Stat. 15 (2006) 651–674.
[35] C. Strobl, A.L. Boulesteix, T. Kneib, et al., Conditional variable importance for random forests, BMC Bioinformatics 9 (2008) 307–317.
[36] Y. Zhu, C. Xie, B. Sun, et al., Predicting Chinas SME credit risk in supply chain financing by logistic regression, artificial neural network and hybrid models, Sustainability 8 (2016) 433.
[37] Y. Zhu, C. Xie, G. Wang, et al., Predicting Chinas SME credit risk in supply chain finance based on machine learning methods, Entropy 18 (2016) 195.