Research on the Knowledge Map in the Prediction of Bond Default

Hui Hu[1], Anxu Bu[2*], Le Yang[3], Chi Ma[4]

[1]e-mail: huhui@hzu.edu.cn

[2*] Corresponding author: buanxu@ustl.edu.cn

[3]e-mail: yangle@ustl.edu.cn

[4]e-mail: machi@hzu.edu.cn

[1]School of Computer Science and Engineering, Huizhou University, Huizhou 516007, China

[2]School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China

[3]School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China

[4]School of Computer Science and Engineering, Huizhou University, Huizhou 516007, China

Abstract—When solving the problem of bond default, because there is a lot of relational and categorical data in the bond data, reasonable use of these data to predict bond defaults is of great significance. Based on the construction of bonding knowledge map, in this paper, knowledge representation learning is used to vectorize the knowledge of picture, and the extracted vector is input into the deepfm model as a feature to predict whether the bond will default. Compared with the general traditional bond default prediction method, in this paper, knowledge map is introduced as the feature embedding of bond default prediction model. The experimental results show that it has higher prediction accuracy.

Keywords-component; default prediction; DeepFM; knowledge map; knowledge representation learning

1 INTRODUCTION

In recent years, with the spread of credit risk in the debenture market, events of defaults on corporate debenture occur continually in our country. Hence, forecasting bonds default has important practical significance default by using objective data and computer technology.

Researchers have done more study on debenture default forecasting and measure of credit risk. Ohlson[1] applied the logic function to compute the breach probability of debenture. Altman[2] proposed the Z-score trustworthiness scoring model using multiple discriminant method to analysis of the likelihood of bank failure or default. J. P. Morgan[3] introduced a credit

ICIDC 2022, June 17-19, Qingdao, People's Republic of China
Copyright © 2022 EAI
DOI 10.4108/eai.17-6-2022.2322865

measurement model to assess credit risk. The KMV model was proposed by KMV, the KMV is a conventional model of financial primarily used to forecast domestic defaulted bonds.

In the wake of the evolution of deep learning, many excellent models are gradually applied to credit risk prediction. The support vector machine was used to forecast enterprise credit rating by Young Chan Lee[4]. Dutta et al.[5] used CNN to predict bond credit rating, and the validity of CNN is proved.

In addition, in the wake of the fast-evolving of knowledge map, scientists have recognized that epistemology map can complement features and be fed into deep learning models to improve their performance. The CKE model was proposed by Zhang et al.[6] , it embeds a structured epistemology map into the network by the Bayesian improved TransR model. The text knowledge characteristic and picture knowledge characteristic are fused to obtain the vectorized representation of the film. This expression is fed into a synergistic ensemble studying frame, and personalized recommendation is carried out by integrating these knowledge.Through experiments, it can be found that the performance of the model can be effectively improved by using the thin-film vectorized expression .

However, owing to bond information contains many types of characteristics. For example, information, version, and a few recessive shareholding nexus, Alone rely on the deep learning model can not better reflect the sophisticated relation of the debenture market. Hence, in the light of the features of bond information, and the bond epistemology map is constructed. So by using this epistemology to express learning models, to learn semanteme and structural information about epistemology maps, as the transcendental knowledge of bond default, it supplements the input of the model in order to enhance its performance.

In this study, the model used to forecast debenture default is DeepFM[7]. This paper chiefly studies how to build the bond knowledge map, vectorize the knowledge in the debenture epistemology map by using epistemology express learning, and input the extracted vector into the prediction model as features to improve the prediction accuracy.

2 RESEARCH ON KNOWLEDGE MAP

Google proposed the concept of knowledge map[8] in 2012. In essence it is similar to the semantic network[9]. It links different kinds of information together to obtain a relational network. In the light of the knowledge reach, the epistemology map is generally divided by generic epistemology map and field epistemology map. The generic epistemology map is oriented to all fields, such as Zhishi.me, ConceptNet, BabelNet, etc. The generic epistemology map focuses more on data breadth, so it emphasizes entities more. The domain specific epistemology map has many entity attributes and industry significance, it requires accurate data, close to the industry, and strict and rich data patterns. Field epistemology map, also called industry epistemology map or vertical epistemology epistemology, is an industry epistemology base consists ospecialized data for a specific field. Facing the bond field, this paper collects data from the bond announcements on the Internet and constructs a knowledge map.

At the moment, the build of epistemology map is primarily carried out on the collected and processed data sets, and then extract information to build field epistemology map. When we want to construct the epistemology map of a new subdivision field, we need to face a problem:

the acquisition of field data. In the face of this situation, the construction of field epistemology map must start from the acquisition of domain knowledge. First, we have gain enough adequate field knowledge, and then carry out data screening, cleaning and format conversion to meet the input of the model. The extracted structured information can not be diametrically leading-in the map database. Only afterwards the fusion of knowledge, knowledge processing, elimination of redundancy and errors in information, and consolidation of information can the warehouse be updated.

In the wake of knowledge map, knowledge map has been extensive attention and deeply studied by academics and industry domestic and foreign, and its application scenarios are more and more. For example, semantic search[10], smart question-and-answer[11], recommendation system[12]. This paper chiefly talk over the construction of field epistemology map. The build methods of epistemology map chiefly contain of bottom-up and top-down[13]. The bottom-up build method pick up entities, attributes and relationships from datasets, they are added to the data layer of the epistemology map, then summarizes and organizes these elements of knowledge, gradually abstracts them into concepts, and finally forms the pattern layer. The top-down method is the exact opposite. The bottom-up method takes the network public data set as the data source, extracts the ontology structure from the public data source, and merges the entities, relationships and attributes extracted from the domain data into the epistemology map.

3 KNOWLEDGE VECTOR REPRESENTATION OF THE DEBENTURE KNOWLEDGE MAP

When building the knowledge map, the structured data in the Wind database is used as the data source, and use the top-down method to build the debenture epistemology map.

The entire procedure of epistemology vector expression based on a debenture epistemology map as shown in Figure 1, it is divided by the following steps.

- Build a debenture epistemology map in view of existing data.
- Learn the built epistemology map through the epistemology expression learning model to obtain the the relation matrix and entity matrix.
- In order to obtain the needed bond knowledge representation, the entity matrix needs to be corresponding to the entity.

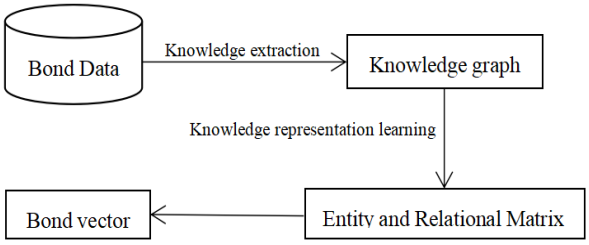

Figure 1. Knowledge vector representation based on a debenture knowledge map.

3.1 Construction Of The Debenture Knowledge Map

In the map, nodes represent entities, and relationships between entities are expressed as edges connecting nodes. The relation is directed, and the end outcome is a directed picture. Table 1 shows the specific entity and relationship information.

TABLE 1. ENTITY, RELATION AND ATTRIBUTE STATISTICS

Type	Name	Quantity
Entities	Debenture, location (debenture listing place, issuer's province/city), company, person, industry, bond type, shareholder type	24865
Relationship	Issue relationship, listing location relationship, industry relationship, provincial-urban relationship, bond type relationship, legal person relationship, chairman relationship, general manager relationship, shareholder relationship, major shareholder type relationship, actual controller relationship	10
Attributes	Bond attributes: securities code, coupon rate, total issuance, issuance time, maturity time, bond rating and other attributes Corporate attributes: date of establishment, registered capital, whether listed companies, total number of employees, major products and business attributes	/

3.2 Knowledge Vector Representation Of The Debenture Knowledge Map

It is difficult for computers to directly use the knowledge map represented by symbols. Nevertheless, the entities and relation in the epistemology map can be embedded into the vector space by knowledge representation learning, represented in the form of vectors, and import them as features into the models. The semanteme information and architecture of the primitive map are retained by the training vector.

First,The bond knowledge map needs to be preprocessed to produce the data form required for the epistemology expression model. Then we numbered all entities; every entity has a only id, and every relation number has a only ID. Then, in the light of the marked entity ID and relation id, map every pair of triples (h, t, r) with ID to obtain triples, tail entity ID and relation ID. After this above processing, Table 2 show the epistemology expression of learning file generated from the epistemology expression.

TABLE 2. KNOWLEDGE REPRESENTAION LEARNING DOCUMENTS

file name	Content	Quantity
entity2id.txt	entity-id pair	24865
relation2id.txt	relation-id pair	10
triple2id.txt	Triple represented by id	71803

The main thought of the knowledge expression model is to insert the entities and relation of knowledge map into the m-dimensional room, and learning the low-dimensional dense vector for each entity. The includes the similitude between entities and information about the network architecture of the map. Epistemology expression learning cut down the high-dimensionality and heterogeneousness of the epistemology map, and abate the other calculative overhead caused by the introduction of epistemology map. Meanwhile, it is also convenient to input continuous low-dimensional vectors into models, so as to the symbol knowledge of knowledge map can be better utilized, and the performance of the model will be further improved. The TransE model is a commonly used knowledge representation pattern, it was put forward by Borders in 2013 [14],which is a translation-based epistemology expression learning pattern. The TransH model[15] was put forward by Wang et al. in 2015. This relation is mapped to the hyperplane by TransH, thereby balancing the complicacy and expressive ability of the model. The TransR model[16] was proposed by Lin et al. The TransE and TransH embed complicacy and entities relations into the identical vector space, anyway they are diverse objects in essence, and it may not be well expressed in the identical vector room. Meanwhile, an entity may have various semantical properties, and these semantic properties corresponding to different relationships. Although the relationships will be mapped to the hyperplane by TransH, it still cannot break the restrictions of relations and entities in the identical room.

Thus, for the preprocessed epistemology map, the TransR is used as the training entities vector of the knowledge expression model[16]. In this model, entities and relations are embedded into two distinct rooms, and the entities in the entity room are mapped to the related room by the entity-relation projected matrix M_r. For triples (h, t, r), head vector h and tail vector t are mapped by projection matrix M_r to gain the expectant tail vector t_r and head vector h_r, the relationship between vector and matrix is shown in Formula 1.

$$\begin{aligned} h_r &= hM_r \\ t_r &= tM_r \end{aligned} \tag{1}$$

The relational vectors r will connect h_r and t_r. Entities that were approach to one another before in physical room will be far apart in a particular relevant room, as shown in Figure 2.

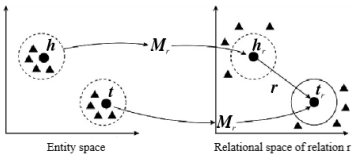

Figure 2. TransR model.

Formula 2 is the corresponding scoring function.

$$f_r(h,t) = \left\| h_r + r - t_r \right\|_2^2 \tag{2}$$

The loss function is shown in Formula 3.

$$\mathcal{L}=\sum_{(h,r,t)\in S}\sum_{(h',r,t')\in S'_{(h,r,t)}}[\gamma+f_r(h,t)-f_r(h',t')]_+ \tag{3}$$

Here, γ is a edge arguments, $[x]_+$ is a hinge loss function, and $S'_{(h,r,t)}$ serves as a built mistake tuple.

Negative samples will be produced when training the model, and then select the Berne negative sampling way of the TransH model to produce negative samples, because this sampling way will be more rational than else methods. This sampling way put forward in TransE is called the Unif sampling way, and this sampling way put forward in TransH is called the Bern sampling way.

This Bern sampling way makes use of different chance to take the place of the tail and head entities for many-to-one, and many-to-many relationships triplets. For many-to-one relationships, a greater chance displaces the tail node, and for one-to-many relationships, a greater chance displaces the head nodes. This sampling method would be more rational, thereby choosing the approach to produce negative samples

Afterwards the model is trained, and generate the following three matrices: the relation matrix, the projection matrix and the entity matrix. Figure 3 shows the bond vector of the entity matrix.

2667,14 Zhong Ye SCP004,-0.038404092,0.01705473,-0.015120124,-0.0013837904,-0.039
9385,15 Ping AN CD185,0.017166318,-0.037501357,0.015627874,-0.042882107,0.077600
5463,14 Gan Gong Tou CP001,0.012970981,-0.030862646,0.029915318,-0.02410593,0.087
13728,16 Pang Da Qi Mao CP001,0.0010844995,-0.0053431815,-0.0039749043,-0.014842
3912,14 Tian Fu CP001,-0.0024695555,0.023539018,0.00796744,0.007286392,-0.0146530

Figure 3. Bond vector representation

As shown in Figure 3, the first column represents the ID of bond, the second column represents the name of bond, and the third column represents the vector expression of bond.

4 EXPERIMENTS AND RESULTS ANALYSIS

4.1 Data Analysis

This paper selects the debenture of the inter-bank market and take the exchange market as the test data. Start date January 1, 2010 and end date September 1, 2018, Excluding government bond data. Finally, 17624 mature keys were obtained as experimental objects; and the default bond is also tagged. Then mark the default sample as positive sample 1, and sign the rest debenture as 0. In total 118 defaulted debenture were flagged.

For samples in the training set at each training time, the positive samples are copied through the up-sampling approach, thereby the final ratio of the positive samples to the negative samples set is about 1:15 in the training. Using the precision-recall curve (PRC) as the

evaluation index, the prediction results of the classifier are evaluated.

4.2 Experimental Results

By building the debenture map and training the debenture epistemology expression, we design an majorizing deep learning model according to the epistemology map. The epistemology expression of debenture is taken as part of the model input, and train the model to forecas debenture defaults. Afterwards multiple sets of comparative experiments, we gain the best arguments of the model. Table 3 shows the concrete arguments.

TABLE 3. OPTIMIZED DEEPFM MODEL PARAMETERS

Parameter	Value
DNN Hidden Size	3
DNN Activation	ReLU
Learning rate	0.001
Batch size	128
Epoch	10
Optimizer	Adam
Loss	binary_crossentropy

If the distribution of positive and negative samples is particularly asymmetrical, namely, when the quantity of negative samples exceeds the quantity of positive samples by a large margin, PRC can gauge classifier more effectually. The samples of this paper are uneven, thereby we ultimately select PRC as the appraisat criteria to assess the performance of model through contrasting the region the PRC curve. Due to PRC is very susceptible for imbalanced data, can assess thoroughly classification of the classifier. The calculation of recall ratio and precision is shown in Formula 3 and Formula 4, the abscissa of PRC indicate the recall ratio and the ordinate indicate the precision.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

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

TP indicates the quantity of specimens forecast to be 1 and effectively 1. FP indicates the quantity of specimens forecast to be 1 and effectively 0. FN effectively the quantity of specimens forecast to be 0 and effectively 1.

This paper makes use of the DeepFM model that incorporates the semantical message of the epistemology map to test and train, and ultimate sort the scores from high to low. Select the first 100 bonds for every test and conduct the experiment for 5 times. The test results are shown in Table 4, that is, the ranking of practical default debenture among the top 100

possible default debenture. It can be seen from the table that among the top 100 types of debenture estimate to default, 7 types of debenture have erious breach of contract

TABLE 4. RANKING OF DEBENTURES PRACTICALLY DEFAULT IN THE TOP 100 POSSIBLE DEFAULT DEBENTURES

Bond name	Rank
15 Le Shi 01	2
13 Bo Yuan MTN001	5
15 Chuan Mei Tan PPN001	17
13 Dan Dong Gang MTN1	21
17 Hu Hua Xin SCP002	63
15 Dan Dong Gang PPN001	77
15Dan Dong Gang PPN002	91

Table 5 compares the quantity of breach debentures in the top 100 possible breach debentures in the five groups of experiments.

TABLE 5. QUANTITIES OF PRACTICAL DEFAULTY DEBENTURES IN THE TOP 100 DEBENTURES

Bond name	Occurrence number
15Chuan Mei TAN PPN001	4
13 Bo Yuan MTN001	5
13 Dan Dong Gang MTN1	3
17 Hu Hua Xin SCP005	2
17 Hu Hua Xin SCP004	5
15 Dan Dong Gang PPN001	5
17 Hu Hua Xin SCP003	3
17 Hu Hua Xin SCP002	2

By calculating the average of the five groups of experimental results, among the top 100 bonds that may default, the practical number of debenture that got defaulted was 7.6.

In order to prove the validity of the put forward way, the forecast outcomes of DeepFM algorithm are compared with LR algorithms and Xgboost algorithms general used in default forecast. The top 100 debenture, the top 150 debenture and the top 200 debenture that probably default are chosen to compare the forecast outcomes.Table 6 is the experimental outcomes, and Figure 4 is the comparison figures.

TABLE 6. CONTRASTS BETWEEN DEEPFM MODEL AND CONVENTIONAL MEANS

Mode	100	150	200
LR	4.2	6.1	11.3
Xgboost	7.3	9.2	11.4
DeepFM-KG	7.7	9.8	11.9

As can be seen from the Figure 4, among the top 100, top 150, and top 200 debenture that probably default, the accuracy of DeepFM model is the highest. By and large, this forecast outcome of the DeepFM is analogous to that of Xgboost. This cause is that for deep learning models, the quantity of specimens is too little. Thus, as an integrated learning model, Xgboost can learn characteristics better with a small quantity of specimens. This is what people believe when the specimens are relatively large, the practical forecast performance of the DeepFM will be obviously mapped. Among the three models, the prediction performance of the LR model is the worst and relatively is rely on characteristics engineering.

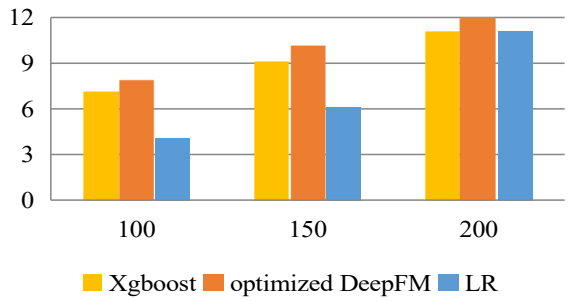

Figure 4. Practical quantity of breach debentures in the top 100, 150 and 200 debentures

5 CONCLUSION AND FUTURE STUDIES

In the light of the features of debenture data, this paper uses the epistemology expression learning model, embeds the discrete symbolic epistemology expression into the vector room to gain the epistemology expression of debentures, embeds the semantic message of the epistemology expression into the DeepFM model to predict bond default. The high-order features are automatically learned through the deep learning model, and the high-order cross features are extended by taking the combined knowledge map as a priori knowledge.

The experimental outcomes show that the DeepFM algorithm integrating the semantical information of knowledge map greatly improves the prediction accuracy. At the same time, it proves the feasibility and effectiveness of knowledge map fusion in-depth learning.

In the following study, more epistemology message will be added in the debenture epistemology map. At the moment, the data we gain is not enough, it is possible to add more epistemology to help researchers gain more overrall debenture semantic information, it is more favorable to knowledge expression learning.

Acknowledgments. This research was partially supported by a grant from the Foundation of Guangdong Education Committee 2021ZDJS082, and partially by the Subject of Hunan Social Science Achievement Evaluation Committee under Grant No. XSP22YBZ054.

References

[1] Ohlson J A. Financial Ratios and the Probabilistic Prediction of Bankruptcy[J]. Journal of Accounting Research, 1980, 18(1): 109-131.

[2] Altman E I. Finacial Ratio, Discriminant Analysis and the Prediction of Corporate Bankruptcy[J]. Journal of Finance, 1968, 23(4): 589-609.

[3] Morgan J. Creditmetrics-Technical Document[J]. JP Morgan, New York, 1997(1): 102-127.

[4] Lee Y C. Application of Support Vector Machines to Corporate Credit Rating Prediction[J]. Expert Systems with Applications, 2007, 33(1): 67-74.

[5] Dutta S, Shekhar S. Bond Rating: a Nonconservative Application of CNNs[C]// IEEE International Conference on CNNs, 1988:443-450.

[6] Zhang F, Yuan N J, Lian D, et al. Collaborative Knowledge Base Embedding for Recommender Systems[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 353-362.

[7] Guo H, Tang R, Ye Y, et al. DeepFM: A Factorization-Machine based CNN for CTR Prediction[J]. ArXiv Preprint ArXiv:170304247, 2017.

[8] Amit S. Introducing the knowledge graph[R]. America: Official Blog of Google, 2012.

[9] Sekine S. NYU: Description of the Japanese NE System Used For MET-2[C]. Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998.

[10] Ruan Guangce, Xia Lei. Research on Knowledge Association of Retrieval Results Based on Word Co-occurrence Relationship [J]. Journal of the China Society for Scientific and Technical Information,2017,36(12):1247-1254.

[11] Ma Feicheng. Knowledge Organization and Provision in Digital Environment [J]. Journal of Zhengzhou University(Philosophy and Social Sciences Edition),2005(04):5-7+14.

[12] Pan J Z, Vetere G, Gomez-Perez J M, et al. Exploiting Linked Data and Knowledge Graphs in Large Organisations[M]. Springer International Publishing, 2017.

[13] Bollacker K, Evans C, Paritosh P, et al. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge[C]. Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2008:1247-1250.

[14] Bordes A, Usunier N, Garcia-Duran A, et al. Translating Embeddings for Modeling Multi-relational Data[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013: 2787-2795.

[15] Wang Z, Zhang J, Feng J, et al. Knowledge Graph Embedding by Translating on Hyperplanes[C]// Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014: 1112-1119.

[16] Lin Y, Liu Z, Sun M, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion[C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015: 2181-2187.