An Industry Classification Model of Small and Medium-sized Enterprises based on TF-IDF Characteristics

Chen Jiahao, Zhang Jiayi

School of Economic and Trade management, Zhejiang University of Technology, Hangzhou 310023, Zhejiang

Keywords: text classification model; Bayesian model; VSM model; Xgboost; convolution neural network

Abstract: This paper selects the data of the national SME Information Disclosure System, uses the TensorFlow in Python to establish the corresponding learning framework, according to its business scope to carry on the corresponding classification. The Jieba participle in Python is first used to remove extraneous words from the business scope of the enterprise. Secondly, using the simple Bayesian text classification model, using Chi as the basis of feature selection, the multi-dimensional characteristics of each type of business scope are selected and re-weighed. After that, the VSM model is constructed for each business scope, which classifies it according to probability. Then, XG-boost is used to encode all the words one-hot, the tree-based model XG-boost is used to make decisions on the processing capacity of tabular data, and prune categories below the threshold. Then, the convolution neural network is used to encode the vocabulary, the lexical annotation is added to the participle, the Gensim training word vector is used, then the cosine similarity is used to calculate, and the classification results are finally obtained.

1. Introduction

National Industry Classification standards for the economic development of various industries, enterprise innovation, a variety of legal disputes have an important impact, in China's industry classification standards every 4-5 years will be adjusted, the current rapid development of Internet technology, a variety of new types of small and medium-sized enterprises continue to emerge, thus, It is very meaningful to make a new classification attempt for the SME industry. and the scientific industry Classification of enterprises is the basis for the study of enterprise risk and evaluation of enterprise development trend, but the current widely used industry classification GB content is relatively traditional, can not do an accurate industry division of new Industries, while the industry lacks an effective and accurate industry classification method, Therefore, it is of practical significance and theoretical significance to carry out the research on the industry classification model of small and medium-sized enterprises based on open data.

2. Basic setting

- It is assumed that the collection of data through network statistics has a certain representation
- The structure of the factors in the text does not change in a short period of time
- Assuming that the relevant data does not change dramatically for extreme reasons
- Assuming that the relevant resources and structures are at a reasonable average level

3. Mark

Table 1 mark

mark	describe
W	Neural network weight
b	Neural network bias Vector
P	Bayesian probability
$P(\varpi_i)$	Probability of Word vector being deleted
$f(\varpi_i)$	Frequency of Word vectors
X_i	Classify pattern
$t(i)$	Threshold value

4. Establishment and solution of models

4.1 Handle data

First of all. We preprocessed the data obtained using Python to eliminate unrelated characters. After elimination, we can make the data more suitable and efficient for the model. We can be found that the segmentation word has a better effect and retains the feature information to the greatest extent.

4.2 Simple Bayesian text Classification model

By using the simple Bayesian text classification model, we use Chi as the basis of feature selection to select multi-dimensional features and go heavy for each type of business scope. In this way, we can get about 1000 dimensions of the characteristics. With the characteristic, it is to construct the VSM model for each business scope, that is, to calculate the weight of each feature by using the TFIDF method to get the eigenvectors that represent the text, and the final classification results can be obtained by classifying it according to the probability [8]. Here are the feature words we've chosen.

Table 2 Partial feature word vectors

Word vector	The first	Second	Third	Fourth	Fifth	Sixth	……	n
Property	0.314128927	0.478855606	0.72843667	0.707720643	0.069274044	0.257035075	……	0.374274477
Finance	0.678035273	0.95431116	0.238027146	0.379611358	0.391845306	0.09380912	……	0.969999853
Environmental protection	0.676731656	0.711482018	0.437760564	0.491912321	0.816370635	0.22407557	……	0.376756456
Computer	0.787864535	0.861159692	0.377945529	0.752928633	0.499859928	0.203514481	……	0.77017168
Hospital	0.034724794	0.755343257	0.489490775	0.129076421	0.434595297	0.781515479	……	0.752642856
Agriculture	0.529228702	0.715174682	0.267972339	0.418549646	0.47299284	0.868378445	……	0.730944624

We use Bayesian probabilities for classification, the results are as follows, only the first 20 sets of results are taken here:

Table 3 Enterprise Classification results

Order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Classification	1	2	3	12	2	2	2	7	6	12	1	1	12	9	6	12	12	5	1	7

We can find through the analysis that in addition to the serial number 4 of the enterprise classification may have problems, other enterprise classification is more scientific, its accuracy in the sample reached 95%.In addition, the distribution below the 2 categories is more dense, thus indicating that the number of information technology enterprises is extremely large, with the first category accounting for 27.15%.

4.3 XGBoost text classification model

We used the Xgboost model to experiment with it to prove our effect. Given a series of classifiers, randomly remove the training set that obeys the random vector y,x distribution. Define the marginal function as:

$$m_g(X,Y) = a\nu_k I(h_k(x) = y) - \max_{j \neq y} a\nu_k I(h_k(x) = j) \tag{1}$$

The marginal function depicts the extent to which the votes of X under the correct classification y exceed the maximum average number of votes in other classifications. When the number of trees is large, it follows the law of large numbers, so the structure of the tree is: as the number of classification trees increases, it converges almost everywhere to

$$p_{X,Y}(p_\theta(h(X,\theta) = y) = y - \max_{j \neq Y} p_\theta(h(x,\theta) = j < 0) \tag{2}$$

Theta is a random variable corresponding to a single tree decision tree, and H (x,θ) is based on the output of X and Theta. In experiments on xgboost, the bagging method and stochastic feature selection are applied in parallel. The bag method of each new training set is obtained in the original training concentration through a random double sampling called step by step method. We use the following methods for data reading and operation.As a result, the following results are obtained.

Table 4 Partial feature word vectors

Word vector	The first	Second	Third	Fourth	Fifth	Sixth	……	n
Property	0.772370143	0.710867163	0.612608517	0.492673298	0.588842156	0.793924981	……	0.407384188
Finance	0.571463102	0.119581309	0.765989581	0.110794255	0.686101457	0.249600696	……	0.719340678
Environmental protection	0.472605106	0.929120312	0.217257632	0.836241397	0.02905922	0.787314736	……	0.40908253
Computer	0.637524153	0.663089624	0.728584035	0.946055082	0.30130209	0.270440229	……	0.268470882
Hospital	0.347440996	0.087352965	0.163204292	0.113976607	0.911691414	0.331244428	……	0.427946748
Agriculture	0.477884236	0.744724586	0.467455837	0.812844391	0.850432219	0.792682133	……	0.133154499

Table 5 Enterprise Categories

Order	Enterprise No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Classification	Category	1	2	3	7	4	4	2	7	6	12	1	1	12	9	6	12	12	5	1	7

The classification of categories 5th and 6th is wrong, so its correct rate is 90%, which also

confirms the correctness of the Bayesian method we have chosen from the opposite side.

4.4 Convolution neural Network text classification model

The typical structure of convolution neural network is to enter the original data directly into the input layer (UC1), the size of the original information determines the size of the input vector, the neurons extract its local characteristics. The feature mapping structure adopts the sigmoid function which affects the kernel of function as the activation function of convolution network, which makes the feature mapping have displacement invariance. As a result, the output of each unit of the middle layer is:

$$h_j = f(\sum_{i=0}^{N-1} V_{ij} x_i + \phi_j) \tag{3}$$

And the output of each unit of the output layer is：

$$y_k = f \sum_{j=0}^{L-1} W_u h_j + \theta_k \tag{4}$$

$$f(x) = \frac{1}{1+e^{-kx}} \tag{5}$$

Under the above conditions, the training process of the network is as follows:

1) Select the training group. 300 samples were randomly selected as training groups from the set.

2) Place the weights vij,wjk and threshold φj,θk into small random values close to 0, and initialize the precision control parameter epsilon and the learning rate alpha.

3) Take an input mode X and add it to the network, and given its target output vector d.

4) Calculate a middle-layer output vector h, and then use the formula (2.10) to calculate the actual output vector y of the network. 5) Compare the element yk in the output vector with the element DK in the target vector to calculate the m output error item [13]

Similarly, we used Gensim to get the word vector:

Table 6 Part feature word vector

Word vector	The first	Second	Third	Fourth	Fifth	Sixth	……	n
Property	0.387548491	0.527286248	0.307476984	0.114293522	0.200060678	0.82681353	……	0.310926562
Finance	0.287627028	0.067839344	0.017797899	0.8176298	0.047428042	0.102962436	……	0.146115982
Environmental protection	0.286554735	0.41659743	0.056298657	0.332801575	0.936386288	0.058813726	……	0.60457781
Computer	0.911616951	0.968187065	0.537687564	0.294214666	0.690861778	0.234108683	……	0.903495357
Hospital	0.362721126	0.663313926	0.347427444	0.797056767	0.415634572	0.716840403	……	0.029415253
Agriculture	0.126265527	0.439994989	0.720770787	0.62723263	0.991869121	0.758965642	……	0.25210902

Table 7 Enterprise Categories

Order	Enterprise No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Classification	Category	1	4	5	7	2	2	2	7	6	12	1	1	12	9	6	12	12	5	1	7

Through the analysis, we can find that the 2nd group and the 3rd group error, so its accuracy in the sample reached 90%, but also confirmed the scientific nature of the above method.

5. Analysis

Different machine learning models are suitable for different types of tasks, and deep neural networks can capture high-dimensional data such as images, voice and text well by modeling space-time positions. The tree-based model, on the other hand, handles tabular data [15] and has features that are not available in deep neural networks. Both types of models are important and are widely used in data science competitions and industry. We can use these models for more precise text classification. In some data processing, we have dealt with it as necessary to remove less trusted data points, which may have some impact. The rated index may have some impact on the outcome with some extreme special factors.

Acknowledgement

An Industry classification model of small and medium-sized enterprises based on TF-IDF characteristics

References

[1] Wangxiangxiang, Fang Vera, Chen Chongcheng. Study on the classification technology of cultural tourism text based on simple Bayesian [J/ol]. Journal of Fuzhou University (natural Science Edition): 1-6.
[2] Zhou Yuncheng, Xu Dongyu, Deng. Agricultural text classification method based on NB and Chi values [J/ol]. Jiangsu Agricultural Science: 1-5[2018-10-03].
[3] Xia Binghong, MA, Panri, Zhang handsome. Automatic classification method of coal mine safety hidden danger information [j/ol]. Industrial and mining Automation: 1-5[2018-10-03].
[4] Yin ya bo, Yang Wenzhong, Yang Huiting, Hushuying. KNN text classification algorithm based on search improvement [J]. Computer Engineering and Design, 2018 (09): 2923-2928.
[5] Cai Lizhong, Cai. Application of DBN in Chinese text classification [J]. Computer Engineering and Design, 2018 (09): 2974-2978+2991.
[6] Marcun, Guo Rui, Gauzen, Sun Yong. Essay Clustering algorithm for improving feature weights [J]. Computer system applications, 2018,27 (09): 210-214.
[7] Zxs, Meng Fan Rong, Zhou Yong, Liu Bing. Character convolution neural network short essay This classification algorithm [J/OL]. Computer Engineering and Applications: 1-11[2018-10-03].
[8] Song Ching Cheung, Chen Xiohong, Niu Qiang. Feature Selection method based on Chi improvement in text classification [J]. Microelectronics and Computers, 2018,35 (09): 74-78.
[9] Wang Hechen, Wang Yang. Classification model of very short text based on emotional tendency and SVM [J]. Scientific and Technical Bulletin, 2018,34 (08): 149-154.
[10] David. Logistic regression solves text classification problems [J]. World of Communications, 2018 (08): 266-267.
[11] Li Cenre, Wang Hao, Liu Xiaomin, Deng Sanhong. A comparative study of the text-to-quantization method of this classification for Weibo short essay [J]. Data analysis and Knowledge discovery, 2018,2 (08): 41-50.