Object: Classify companies into investment type and non-investment type according to their business descriptions.

Every company has an official business description [1] in the NECIPS system. By classifying every part in the description, we can generate class distribution for it, score the degree of investment, and use it to judge whether the company is investment type with other info including company name (including past names) and type.

In order to improve accuracy, we delete the phrases in business descriptions which have non-essential meaning. The phrases are generated in step1 and step2, which bases on the business descriptions of an official list of investment type companies provided by AMAC (Assets Management Association of China) and a part of the unknown sample.[2]

1. Use AMAC sample to generate a deleted list and a remaining list (500 phrases respectively) as the benchmark to judge whether the phrase has non-essential meaning (deleted phrases) or not (remaining phrases).[3]

1.1 Classify phrases roughly to deleted part and remaining part according to their location in business descriptions.[4] Deleted part contains the sentences that appear in brackets, in parentheses, and before colons.[5] All other sentences are in remaining part.

Results: deleted phrases list, sample size is 5,268 after dropping duplicate; remaining phrase list sample size is 14,738 after dropping duplicate.

1.2 To ensure accuracy, we manually inspect the two lists respectively, select 200 high frequency phrases, and use TFIDF to select 300 more phrases which have the highest similarity with the 200 phrases.

2. Use machine learning algorithm (binary classification) to build up a full list of

[1] Companies can change their business descriptions. We also collect their historical business descriptions.

[2] As we have an increasing sample size in the process of collecting companies invested by investment companies, we use a subsample of business descriptions in AMAC sample (sample size: 69,526) and unknown sample (sample size: 2,501,407) to build up the list of phrases having non-essential meaning (step1 and step2).

[3] The final lists of deleted and remaining phrases are on the website (yue-fei.com)

[4] Further inspection is needed because the descriptions are haphazard occasionally. For example, sometimes the sentences having non-essential meaning are not in brackets or parentheses.

[5] The sentences are usually restrictive statements (for example, 不得以公开方式募集资金开展投资活动 ("should not do investment activities using open funds raised from the public") , additional notes (for example, 法律、法规另有规定除外 ("except for specific requirements by the laws and regulations")) and phrases indicating item categories (for example, 一般经营项目 ("general business items")). These phrases can appear in both the business descriptions of investment type and non-investment type of companies and therefore lead to a high textual similarity without helping to determine whether the company is a investment type.

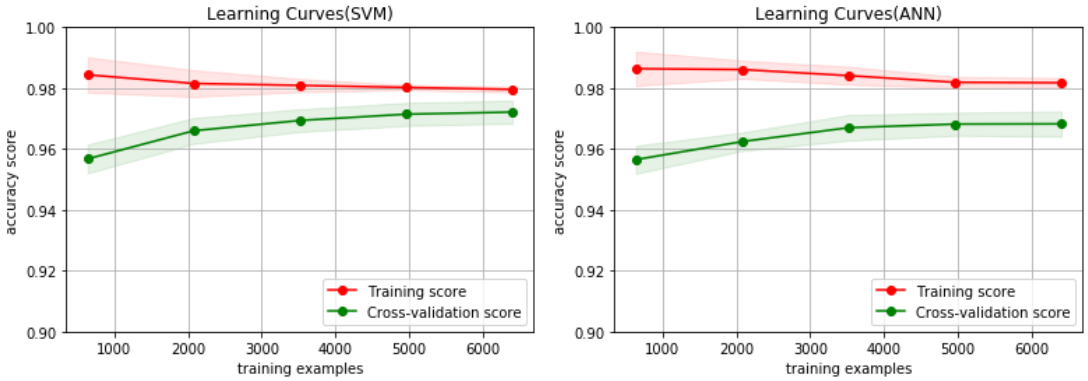
phrases have non-essential meaning.

2.1 Cut business descriptions into phrases at punctuation marks in AMAC sample and the unknown sample, and delete the phrases which have the opposite meaning of those in the remaining list in step 1.[6]

Results: AMAC sample (phrases), sample size is 18,890 after dropping duplicate; unknown sample (phrases), sample size is 2,131,229 after dropping duplicate.

2.2 For phrases in AMAC sample (phrases) and unknown sample (phrases), use TFIDF to calculate similarities of each phrase with the phrases in deleted list and remaining list build up in step 1. The 1000 TFIDF-similarities are the input variables for ML.

2.4 Randomly select 10,000 phrases and label manually for model training. In two candidate machine learning algorithms, Support Vector Machine (SVM) and Artificial Neural Network (ANN), SVM is selected.



Result: 134,386 phrases are classified to deleted part.

3. generate class distribution for business descriptions.[7]

3.1 To reduce the sample size for practical purpose, we only select the business descriptions including 投资 ("investment"), 管理 ("management") or 私募基金 ("private offering fund") as candidates, and others are regarded as non-investment companies.

3.2 Delete the deleted phrases generated in step 2, and the phrases which have the opposite meaning with those in the remaining list in step 1 (the same as 2.1). Cut

[6] For example, "should not XX", "XX is not allowed", "except XX", etc. where XX is in remaining list in step 1. These phrases have high TFIDF-similarity with the phrases in remaining list, but their meanings are totally contrary to those in remaining list. Therefore, we pre-exclude them to avoid misallocation in the following steps.

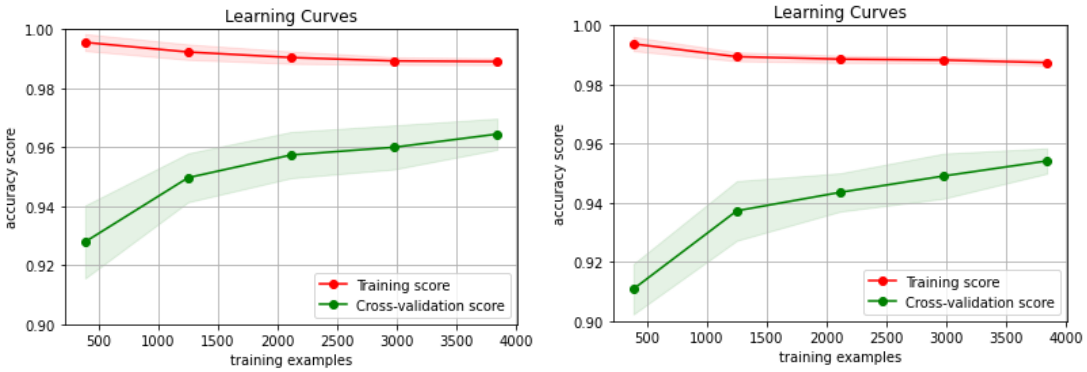
[7] Because a business description normally contains multiple business items, it is not easy to judge whether it is investment type as a whole.

business descriptions into business description categories[8] at semicolons and full stops.[9]

3.3 Using a high-frequency phrases list classified manually[10], two methods are applied for classification: (1) If an item can be totally covered by some classified phrases, then its class distribution is assembled by the classes of those phrases. (2) All items have 0 TFIDF-similarity with classified phrases except class 0 phrases are class 0.[11]

3.4 All items fail to be classified in 3.3 are cut off to phrases at punctuation marks and use the same methods as 3.3 for classification. In addition, some special investment type phrases are discerned by regex expressions.[12]

3.5 The remaining phrases are classified by two SVMs. The first one classify the phrases to class 0 and not class 0, and the second one classify those not class 0 to detailed classes.



3.6 Assemble every part of a business scope item to obtain class distribution.[13]

4. classify companies into investment type and non-investment type.

4.1 In China, companies with type “有限合伙” or “有限责任合伙” are definitely investment companies. They are found by company name or company type.

---

[8] Business description category: a list including all business items for a company.

[9] If a business description category only contains one item, then cut it at commas.

[10] The list contains over 600 phrases in high frequency items from AMAC sample. Investment type is class 1 and 2, management type is class 3, not-sure type is class 4, and industry type is class 0. We assign not-sure type because some phrases appear in both investment type and non-investment type. For example, 企业策划 (“enterprise planning”), 财务咨询 (“finance consulting”), etc.

[11] In practice, we only need to check whether the item has the same words in classified phrases (except class 0).

[12] Some companies list the industries they invest. These investment type phrases are hard to find by other methods because they contains many words expressing industry, For example, 在电子、通信领域进行投资 (“invest in electronics and communications"), 对石油制品、石化项目的投资 (“investment in petroleum products and petrochemical projects”), etc.

[13] Some phrases are only assigned 1 class number, but their meanings are mixed. For example, some companies invest companies and securities at the same time. We revise them manually and get revised class distribution.

4.2 Three criteria are designed to find investment companies: (1) Score class distribution by three different methods. The companies whose scores satisfying three threshold values simultaneously are investment type.[14] (2) It is hard to judge some companies by business descriptions[15], so the second criterion finds investment companies whose names contains some specific phrases.[16] (3) Because companies always put their main business in the front of their business descriptions, the third criterion concerns whether the first three items of business descriptions has enough investment-related phrases.

4.3 In all the investment companies selected by the three criteria, peculiar companies are picked out by finding keywords in company names.[17]

[14] Threshold values are determined by setting different values, manually judge the company within the scope and calculate accuracy.

[15] Some business descriptions don't say investment activities specifically, but they don't say industrial activities either.

[16] The names of companies with high investment scores are collected. Because normally, the front part of a company name is self-named, we delete it by locating investment-related keywords and only remain the part shows type and nature. The latter part is used to find more investment companies.

[17] The main types of the peculiar companies are: energy resource, infrastructure and real estate, tourism, finance. Companies are also excluded If their names show their in some specific industry.