Evaluation Grade Prediction Method with Limited Information from Experts

Xiao Wei*, Hongliang You, Jiajun Cheng, Qiang Gao, Shanhong Tang
Information Research Center of Military Science, Chinese Academy of Military Science, No.26, Fucheng Road, Beijing100042, China

*1441124101@qq.com

Abstract. With the development of artificial intelligence, using machine learning methods to build evaluation models has attracted more and more attention. However, training an evaluation model often needs a lot of labeling samples annotated by experts. It is very difficult to get enough labeling data with a limited group of experts. This paper proposes a method to learn the evaluation model with limited information from experts. This method has two stages. In the first stage, we build a large training set with ordinary people by comparing every two samples. After that, we train a Siamese Network with the paired comparison data set to get a score for each sample. In the second stage, we map the scores to evaluation grades with the help of experts. In the experiments, we use the UCI wine quality data set to evaluate our method. Experimental results demonstrate that we get a basically equivalent accuracy (0.5% decrease) with only 1.45% samples labeled by experts.

Keywords: Limited expert information; Siamese Network; Binary search; Evaluation grade

1. Introduction
Evaluation grade prediction is an important basis for decision-making. It is widely used in military[1], medical industry[2], society[3], environment[4], finance[5] and other aspects. Traditionally, we ask experts to give an evaluation score or grade to an object with their experience. Further, we find some indices to represent an object, assign a weight for each index with the help of experts, and calculate an evaluation score to evaluate the object. The two kinds of methods often get subjective results, which is greatly affected by the knowledge background and experience of experts. With the development of artificial intelligence, using machine learning methods to build evaluation models has attracted more and more attention. However, it is very difficult to obtain a large number of data labeled by experts due to the high requirements of experts' level in evaluation grade judgment. Therefore, it is of great significance to study how to learn a relatively accurate evaluation model with less information from experts.

This paper aims to reduce expertise dependence when training an evaluation model. The motivation is inspired by a well-known phenomenon that comparing two objects relatively is much easier than give each object an absolute evaluation grade, and the comparison results are easier to reach consensus than absolute grades. Fig.1 shows a different evaluation ability. Therefore, giving two samples with a certain gap, ordinary people who know some basic understanding of the evaluation problem can estimate which one is better. Motivated by this phenomenon, we put forward a two-stage method to reduce expertise dependence for evaluation grade prediction. In the first stage, we use a lot of labeling data
from ordinary people to get an evaluation score for each sample. We first build a large training set with ordinary people by asking them to compare every two objects. After that, we build a Siamese Network, which has two sub-networks with the same structure and parameters. Each sub-network inputs the feature of an object and outputs a score from 0 to 1. The output of the Siamese Network is the difference between the two outputs of the sub-networks. Training the Siamese Network with the labeling pairwise data, we get a model that can give a score from 0 to 1 for each sample. In the second stage, we invite experts to map the score to evaluation grades. In this way, we can train a good model with limited information from experts.

2. Related work
In recent years, grade prediction has been an area of active research in evaluation. The past researchers mainly studied along two lines of investigation. According to the idea of regression, a line of research is to use the algorithm to generate a fitting function to constantly approximate the evaluation grade results. The specific methods include artificial neural network[6-12], multiple linear regression model[13], gradient boosting[14]. Another line of research is to take the samples of each evaluation grade as a class according to the idea of classification and transform the grade prediction into sample classification for analysis. It mainly includes the Bayesian network[15-16], support vector machine[17-20], decision tree[21], and so on. All of the methods so far are based on complete expert information to predict the evaluation grades. However, in most practical scenarios, the evaluation grade judgment requires experts with a high level. Therefore, it is necessary to further study the evaluation grade prediction method with limited information from experts.

Siamese Network is a kind of "connected neural network", which is realized by sharing weights. The core idea is to find a mapping function, which can transform the input samples into a feature space. Each sample data corresponds to a feature vector. The difference between vectors is represented by some simple "distance measures". Finally, the similarity difference between input samples is fitted by this distance. In the training stage, the loss function value of a pair of samples from the same category is minimized, and the loss function value of a pair of samples from different categories is maximized[22]. The Siamese Network weakens the impact of sample labels, improves the model scalability, and can classify the untrained categories[22].

3. Evaluation grade prediction method with limited information from experts
This section mainly introduces the framework of the evaluation grade prediction model with limited information from experts. Fig.2 shows the model framework.
3.1. Comparison Module
The comparison module uses the decision-making judgment of ordinary people to compare the characteristics of input samples by pairwise[23] and label them with relatively objective relations. Suppose there are a total of \( n \) samples. The feature vector of samples is \( X_i (i = 1, 2, \ldots, n) \). Ordinary people pair the samples and divide the data pairs into two categories through decision-making judgment. Given a pair \( X_i, X_j \), they give a label \( Z_{ij} = 1 \) when they think \( X_i \) is obviously better than \( X_j \), and they a label \( Z_{ij} = 0 \) when they cannot distinguish which is better.

3.2. Scoring Module
The scoring module is mainly composed of Siamese Network, which consists of three layers—an input, hidden, and output layer.

Input layer: The input of this network is the feature of a pair \( X_i, X_j \).

Hidden layer: The hidden layer is a fully connected network with tanh activation function. We use a partial dropout layer to prevent overfitting.

Suppose the neuron input sample eigenvector \( X_i = [x_1, x_2, \ldots, x_m] \), weight is \( w_h \), bias is \( b_h \), and output \( Y_h \) is shown below.

\[
Y_h = \tanh(w_h X_i + b_h)
\]

Output layer: The sigmoid activation function is used to map the sample score to \([0, 1]\). Suppose that the input of the output layer is the output \( Y_h \) of the previously hidden layer, the weight is \( w_o \), the bias is \( b_o \), and the output \( Y_o \) is shown below.

\[
Y_o = \text{sigmoid}(w_o Y_h + b_o)
\]

The \( Y_o \) is the output of one side neural network. It means the score of the sample. The score of the sample \( i \) is \( f(X_i) \), and the score \( j \) of the sample is \( f(X_j) \). The prediction score difference is defined as

\[
Z'_{ij} = f(X_i) - f(X_j)
\]
Because the values of the new label are 0 and 1, which is equivalent to the binary classification problem, we compile the model using binary cross-entropy as the loss function. It is the loss function corresponding to the sigmoid.

\[
\text{Loss} = - \sum_{i,j=1}^{n} (Z'_{ij} \log Z_{ij} + (1 - Z'_{ij}) \log(1 - Z'_{ij}))
\] (4)

Loss is 0 only if \(Z'_{ij}\) is equal to \(Z_{ij}\), otherwise, loss is a positive number. Moreover, the greater the difference, the greater the loss. In the model, the order is the minimum loss of \(Z'_{ij}\) and \(Z_{ij}\).

3.3. Grading Module
The second stage is to find the score ladder and establish the mapping between the evaluation score and grade. The method used in this paper is a binary search.

Suppose the score of the stage 1 scoring model output is \(f(X_i), (i = 1, 2, 3 \ldots n)\), we arrange it into a column from small to large and record it as an ordered array \(F(X)\). \(f(X_i)\) means element in the array. First, experts mark the evaluation grade of starting point \(f(X_1)\), endpoint \(f(X_n)\), and middle point \(f(X_{n/2})\). If two grades are equal, then all samples between the two points are the same grade. For example, if there is \(R_1 = R_{n/2} = K\), then according to this method, it can be determined that all sample grades between No. 1 and No. \(n/2\) are K. If two grades are not equal to each other, \(f(X_{n/2})\) is regarded as the dividing point, the labeling process is recursive according to left ordered array and right ordered array respectively, it is to judge the grades of array \([f(X_1), f(X_2) \ldots f(X_{n/2})]\) and \([f(X_{n/2}), f(X_{n/2+1}) \ldots f(X_n)]\). Until there are two points of equal grade, all samples are labeled.

Then the time complexity of the algorithm is analyzed and deduced. There are \(n\) samples, \(m\) grades to be evaluated \((n > m)\). Then we need to find the score interval \(m - 1\) and establish the score ladder \(m\). If the evaluation process is multiple independent repeated experiments, the number of cycles is

\[(m - 1) \times K = (m-1)\log_2 n \] (5)

The time complexity of the sample evaluation algorithm is \(O((m-1)\log_2 n)\). if \(m\) is a constant, then the time complexity of the algorithm is \(O(\log_2 n)\). From the viewpoint of experts, the search process of the fractional interval is synchronous. If the process of bisection labeling is regarded as one round, the corresponding grade of all samples can be calibrated by using at most the marking \(\log_2 n\) rounds. At the same time, the upper and lower limits of expert labeled samples can be derived according to the quantitative relationship of \(n, m\). If \(n \gg m\) (far greater than), the maximum number of expert evaluation samples is \((m - 1)\log_2 n\). If \(n = m + 1\), the minimum number of expert evaluation samples is \(m + 1\).

4. Experiment
In the study, we use the UCI wine quality dataset. The data set contains 4898 samples, including 11 physical and chemical indicators, and 1 indicator of wine quality[18]. The quality grade is 3-9.

In the experiment, 80% of the data is used as the training set, 20% of the data is the test set. The accuracy of the test set is taken as the evaluation standard of the model. In order to verify the evaluation grade prediction method under the condition of limited information, the data set is constructed from the perspective of ordinary people. Ordinary people can not distinguish wine samples with a grade interval of 1, but can only distinguish wine samples with a grade interval greater than 1, that is, 5 is better than 3, 6 is better than 3, and so on. The correct rate was only 53.86%. The positive and nega-
tive samples are unbalanced, and the method of down-sampling is used to adjust the training data. Each sample is scored by training a neural network. And according to the order from small to large. Tab.1 shows the grade of samples labeled by experts.

| Rounds No. | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | ... | 12 | 12 | 12 |
|------------|---|---|---|---|---|---|---|---|---|     |    |    |    |
| Sample No. | 0 | 1958 | 3917 | 979 | 2937 | 489 | 1468 | 3427 | ... | 3058 | 3817 | 3916 |
| Grade      | 3 | 6 | 9 | 5 | 6 | 5 | 6 | 7 | ... | 6 | 7 | 8 |

**Tab.1** Sample Grades Labeled By Experts

So far, it is determined that No.0 sample is grade 3; all samples between No.1 and No. 110 are grade 4; all samples between No.111 and No.1348 are Grade 5; all samples between No.1349 and No.3058 are Grade 6; all samples between No.3059 and No.3817 are grade 7; all samples between No.3818 and No.3916 are grade 8; the only sample 3917 is determined as grade 9. To sum up, experts marked 12 rounds of samples, a total of 57 samples. At that time, through the derivation, we can get that experts need to label \(\log_{2}3918 = 12\) rounds. The number of samples ranged from 8 to 72. Tab.2 shows the comparison with previous research models.

| Model                   | Best Accuracy |
|-------------------------|---------------|
| Linear / Multiple Regression | 51.8%         |
| SVM                     | 65.0%         |
| Neural Network          | 52.9%         |
| Siamese Network         | 68.4%         |

**Tab.2** Comparison With Previous Research Models

Compared with the huge data set which needs expert annotation under the condition of complete information, the evaluation grade prediction method with limited expert information only needs to label 57 samples. Only 1.45% of the samples are labeled by experts, and 68.4% accuracy can be obtained by training on the training set. In the data set, only 53.86% of data are correct.

The structure of the model is clear and easy to construct. Depending on the current computer performance, it can quickly predict the evaluation grade of a large number of samples. By scoring, we can directly reflect the sample difference of each grade, so as to improve the prediction accuracy.

| Condition                        | Best Accuracy |
|----------------------------------|---------------|
| complete expert information      | 68.9%         |
| Limited expert information       | 68.4%         |

**Tab.3** Comparison Result

In order to explore the effectiveness of the evaluation grade prediction method under the condition of limited expert information, a Siamese Network with the same structure is adopted in the paper. Tab. 3 shows the comparison results. Under the condition of complete expert information, the prediction accuracy of the test set evaluation grade is 68.9%.

Tab.2 and Tab.3 show that the accuracy rate under the condition of limited expert information is only 0.5% lower than that under the condition of complete expert information, even higher than that of traditional evaluation grade prediction method based on complete expert information.

5. Conclusion
In this paper, an evaluation grade prediction method with limited information from experts is proposed. The key idea is to transform the relatively difficult grade judgment into a relatively easy judgment. People can judge the better one in two objects easily. On the same data set, we compare the accuracy of evaluation grade prediction under the condition of limited expert information and complete expert information. Although the expert labeling data is greatly reduced, the accuracy rate is not significantly reduced. Even more than part of the traditional evaluation grade prediction method is based on complete expert information. The experimental results show that the method achieves a good tradeoff between the cost and accuracy of expert labeling, reduces the number of expert judgment samples, and reduces the difficulty of data acquisition. The next step is to continue to improve the accuracy of evaluation and prediction and improve the structure of the neural network.

References
[1] HaoZ and Liang M 2019 Battle Fielddamage level assessment for submarine torpedo weapon system based on evidential reasoning algorithm J.Torpedo Technology 1 pp52-55
[2] Jia S, Mengjiao J, Zhe D 2019 High quality and superior effect of Dachuanxiong Formula based on quality constant grading evaluation technology of Chinese medicine pieces J. China Journal of Chinese Materia Medica 9
[3] Luo J, Jiangle L and Luo H 2019 Rural development level of villages in Wuhan City's new urban districts and its hierarchical structure J. Progress in Geography 38(9) pp1370-1381
[4] Bo W and Damei L 2019 The calculation details to determine the assessment grade of atmospheric environmental impact assessment J. Jiangxi Chemical Industry 3 pp8-11
[5] Wangbin H, Donghua Z, Xuefeng W 2005 Evaluation and prediction of personal credit risk in commercial banks J.Commercial Economy Studies, 9 pp53-54
[6] Hua X, Yuyin L, Minhua H and Haiyan C 2019 Prediction of airport departure delay level based on BP Neural Network J. Aeronautical Computing Technique 3
[7] Jianbo W, Fayang N, Jia Z and Longbiao P 2019 Construction of metropitbasedriskassessment DEA-AHP-BP Neural Network J. Architecture Technology 3
[8] Huijing W and Xiaohui H 2019 Research on the prediction of air quality index based on GA-BP Neural Network J. Journal of Anhui Normal University (Natural Science) 4 pp360-365
[9] Yue Q, Xu Q and Hongqing S 2019 Machine-learning-based model and simulation analysis of PM2.5 concentration prediction in Beijing J. Chinese Journal of Engineering 41(3)
[10] Guangyin J, Yan C, Liang Z, Hanbo T, Yanghe F and Jincai H 2020 Urban ride-hailing demand prediction with multiple spatio-temporal information fusion network J. Transportation Research Part C: Emerging Technologies 117 pp102665
[11] Qi W, Guangyin J, Xia Z, Yanghe F and Jincai H 2020 CSAN: A neural network benchmark model for crime forecasting in spatio-temporal scale J. Knowledge-Based Systems 189 pp105120
[12] Guangyin J, Zhexu X, Hengyu S, Yanghe F and Jincai H 2020 Deep multi-view spatiotemporal virtual graph neural network for significant citywide ride-hailing demand prediction arXiv preprint arXiv:2007.15189
[13] Xin C 2013 on the quality prediction model of wine J. Journal of Xi'an University of Arts & Science (Nat Sci Ed) 16(2) pp45-47
[14] Song H and Na L 2019 Evaluation of competitiveness of logistics enterprises in China J. Logistics Management 15
[15] Jing Y, Mei Z and Wei H 2019 The prediction model of network public opinion crisis level based on Dynamic Bayes Network J. Information Science 37(5) pp92-97
[16] Jiayu L, Zaiwei L and Yuelei H 2019 Study on early warming of temperature gradient of ballastlesstrack in summer in East China J. Railway Standard Design 4 pp40-46
[17] Tianyu Z, Yi L and Shilian S 2019 Prediction on coal spontaneous combustion risk for multi-coal based on SVM J. Journal of Hunan University of Science & Technology (Natural Science Edition) 2 pp11-17
[18] Enlai L 2016 Research on wine quality prediction based on support vector machine method
*J. Journal of Data Mining* 6 (1) pp42-53

[19] Yun Yand Meijing L 2017 Classification method based on AdaBoost-SVM for wine quality
*J. Computer Engineering and Design* 9

[20] Ruihong W, Yali W and Huanchong Z 2013 An evaluation model of wine quality based on Least Square Support Vector Machine
*J. Journal of Huaqiao University*(Natural Science) 1 pp36-41

[21] Juntao C, Zhenlan S and Yongjian Z 2019 Application of decision tree classification in evaluation of nursing workload in surgical ward
*J. Chinese Journal of Practical Nursing* 35(27) pp2148-2151

[22] Sainan Q 2019 Multi source image matching algorithm based on Siamese Network
*D. CASM*

[23] Xuan Z 2018 Research and implementation of ranking model based on deep learning
*D. CASM*