Solution To Kaggle Yelp Contest

Wei Yuan

June 2, 2016

1 Introduction

In this competition, we are given photos that belong to a business and asked to predict the business attributes. This task can be treated as a multi-label learning problem, i.e. each instance can have multiple labels. I choose Multi-Label KNN, a variant of KNN, to solve the problem. Compared with other popular algorithms, ML-KNN can be easily implemented and can naturally handle the multi-label cases, even though the cost of prediction is relatively high.

2 Solution

Due to time limitation, we simplify the task a little bit. Instead of directly estimating labels for each business, we choose to estimate labels for each photo. After that, all the photos belonging to the same business vote its label list.

2.1 Data Preparation

We use the toolkit [1] to extract features for each photo in the dataset. Each photo is represented by a vector of 64 features, including 44-d color correlogram, 14-d color texture moment, and 6-d RGB color moment [1].

2.2 ML-KNN

ML-KNN [3] is an extension for KNN. For each unseen instance, its K nearest neighbors in the training set are firstly identified. After that, based on statistical information gained from the label sets of these neighboring instances, i.e. the number of neighboring instances belonging to each possible class, maximum a posteriori (MAP) principle is utilized to determine the label set for the unseen instance. Please see the details in [3].

2.3 Voting

For each business in the test set, we aggregate its photos' label lists. Then we compute the average vote for each possible label and set a threshold to filter it.

[1] https://github.com/li-xirong/features

Table 1: Offline Evaluation Results

Model	F1 Score
ML-KNN	0.678
DBPNN	0.646
BCC	0.641
RAkEL	0.664
BPNN	0.628
BR	0.649

Table 2: Online Evaluation Results

Submission	Public Score	Private Score
u1234x1234	0.83177	N/A
rec	0.61235	0.60612
rec20	0.61412	0.60973

3 Experiments and Results

The whole project is developed under Scikit-Learn [2].

3.1 Offline Evaluation

Here, we conduct 5-fold cross validation on the training set. Besides ML-KNN, I also tried several popular algorithms provided in Meka[2]. For ML-KNN, we choose K=10, s=1.0 as Laplace smoothing. The voting threshold is set to 0.44. For all other models, I simply adopt the default settings in Meka.

According to Table 1, it is clear that ML-KNN gives the best results.

3.2 Online Evaluation

Table 2 shows results of top-one and two of my submissions. I have set k to different values, i.e. 10 and 20. Due to the large gap top-one, I didn't try more parameter settings.

3.3 Analysis

Multi-label learning and Image Processing are new to me. Therefore, I chose an algorithm which can be quickly implemented. But this leaves plenty room for improvement.

First, I underestimated the problem at the beginning. The whole solution is divided into 2 steps, i.e. photo label estimation and business label voting. These two steps are treated independently. Therefore, I didn't make full usage of the data. For example, the photoid2bizid information in training data is not used at all. An alternative of ML-KNN is that we can apply it directly to each restaurant. In fact, some researches in multi-instance multi-label learning [4] fit better in this task.

[2]http://meka.sourceforge.net/

Second, it is a pity that I didn't have time to do data analysis this time. It is possible to find some relations between labels, photos and restaurants, as well as the statistics, which could be very useful.

Third, the feature extraction step could affect the final results a lot. It would be interesting to see how different image feature extraction algorithms behave.

4 Conclusion

Overall, it is a nice experience to finish this competition. Exploring something new and learning is of great fun. Thanks for the opportunity.

References

[1] Xirong Li, Cees GM Snoek, and Marcel Worring. Learning social tag relevance by neighbor voting. *Multimedia, IEEE Transactions on*, 11(7):1310–1322, 2009.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[4] Zhi-Hua Zhou, Min-Ling Zhang, Sheng-Jun Huang, and Yu-Feng Li. Multi-instance multi-label learning. *arXiv preprint arXiv:0808.3231*, 2008.