Paper Template for COMP90049 Report

Anonymous

1 Introduction

With large user group and high extremely high frequency of use, social network has become an inseparable part of people's daily life and twitter is a typical example. User location is a kind of important information in social network, which can be applied to improve user experience, detect event and build earthquake reporting system (Sakaki et al., 2013). Nevertheless, according to Sloan et al. (2013), less than 1% of users are willing to disclose their location probably due to privacy issue. Therefore, it is of great significance to analyze and predict the user's location from the data level.

The main purpose of this research is to infer the user's location relying on the content of the tweets by the means of machine learning. To begin with, this paper will summary several related literatures concerning about machine learning. Secondly, it will provide a brief introduction to the dataset used in the research. Later, the methods applied in the project including Support Vector Machine and Random Forest will be explained. Then, it will present the results and analysis corresponding to that. Finally, a conclusion of the whole article will be drawn.

2 Related Literature

Since the 21st century, data scientists have done numerous researches on predicting user's location.

In the early studies, the prediction is based on friendship connection of the user. Backstrom et al. (2010) claim that the relationship between geography and society is becoming more precise, which make it possible to build a statistic model for description about their interaction. Rout et al. regard the task as a classification problem, and SVM as well as several features reflecting characteristics of user network are adopted in their research. Kong et al. (2014) build a location estimation system, SPOT, which is based on three various algorithm, friend-based, social closeness-based, and energy and local social coefficient-based.

In recent years, instead of user network, scientists are more willing to analyze from the perspective of text content. For example, a geolocation prediction algorithm is put forward by Chi et al. (2016) and it applies multinomial Naïve Bayes as a classifier, using text-based features which are extracted from tweets. Jayasinghe et al. (2016) solve this problem by using time zone text classifiers.

In general, due to these significant works, the accuracy rate of positioning user is increasing year by year.

3 Dataset

The data set used in the research comes from multitudinous tweets collected from twitter, and it will be divided into three parts, training set, development set and test set. The data is pre-processed by the method of Mutual Information and the best 200 terms (454 attributes) are selected. Below is a brief description of each set.

a. train-best200.arff: Function is used to fit the model. It contains 96586 instances.
b. dev-best200.arff: It is used to evaluate the model in order to find the model with the best performance. It contains 34029 instances.
c. test-best200.arff: Labels will not be given in this set and the results from using this data will be uploaded to Kaggle and given a final score. It contains 32978 instances.

Since R is adopted for modeling in this experiment, the above files will be transcoded into CSV format through weka.

4 Methodology

4.1 Data Pre-processing

Obviously, tweets with the same user ID have the same location. Therefore, it is feasible to integrate the data of the individual user for modeling. The author implements this process with two commands in R, "group_by" and "summarize", both of which are derive from the package "dplyr". After processed, the instances

of training data are decreased to 2396. This also greatly increases the time complexity of the algorithm.

4.2 Classification Method

This research will use two machine learning methods, Support Vector Machine (Hereinafter referred to as the SVM) and Random Forest, which come from packages "e1071" and "randomForest" in R, respectively.

4.2.1 Support Vector Machine

SVM is a kind of generalized linear classifier following the rules of supervised learning, of which the decision boundary is the maximum-margin hyperplane. Due to the principle of structural minimization, it can avoid over-learning problem and have strong generalization ability.

4.2.1 Random Forest

Random Forest belongs to supervised learning algorithm, which is an integrated learning algorithm based on decision tree, achieved by bagging method. It is very simple to implement with low computational overhead, but it performs amazingly in classification and regression

4.3 Zero-R

Zero-R is the simplest classifier. In a nutshell, it just chooses the class with the highest occurrence rate according to the historical data as the result of the unknown sample. In other words, for any unknown sample, the classification result is the same. Although this classifier has no predictive function, it can be used as a comparison classifier for other classifiers, that is, act as the baseline of performance. In the experiment, the author will measure the result of Zero-R classifier through weka and compared it with other results (SVM and RandomForest).

5 Evaluation

This research will evaluate the model from two aspects. One is evaluating directly with the development set, and the other is to evaluate based one the predictor score of test set given by Kaggle.

5.1 Evaluation Metrics

Precision, Recall and Accuracy will be employed as evaluative criteria in the research. Their formula is as follow:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

True Positive (TP): The prediction is a positive example, and the actual is a positive example;

False Positive (FP): The prediction is a positive example, and the reality is a negative example;

True Negative (TN): The prediction is a negative example, and the reality is a negative example;

False Negative (FN): The prediction is a negative example, and the actual is a positive example.

5.2 Results and Analysis

5.2.1 Result

Three different classifiers are utilized to model the training data, and the obtained models are tested with the development set. In addition, the predictors of SVM and Random Forest will be submitted to Kaggle to gain scores. The results were shown in table 1-3.

	New York	California	Georgia
Precision	0.644	/	/
Recall	1.000	/	/
Accuracy: 0.644			

Table 1- Zero-R classifier

	New York	California	Georgia
Precision	0.820	0.787	0.494
Recall	0.853	0.549	0.590
Accuracy: 0.751			
Kaggle score:0.736			

Table 2- SVM classifier

	New York	California	Georgia
Precision	0.743	0.748	0.803
Recall	0.968	0.383	0.309
Accuracy: 0.748			
Kaggle Score: 0.751			

Table 3- Random Forest classifier

5.2.2 Analysis and Comparison

According to table 1, it can be found that when all the positions of unknown samples are predicted as "New York", the accuracy of the algorithm reached 64.4%. This is a relatively high number, indicating that tweets released in New York accounts for a large part of the training data.

By comparing table 1 with table 2 and table 3, it can be discovered that both SVM classifier and random forest classifier have excellent performance in the overall accuracy rate, exceeding the base value by more than 10%.

However, from the observations in tables 2 and 3, several defects can still be found. For example, in table 2, although the precision of New York and California reaches a relatively high level (around 80%), Georgia precision is rather low, only 49.4%. This possibly due to the small number of Georgia samples leading to the overall underfitting of the model. It is well known that SVM performs well in machine learning problems with small sample size, however, according to this experiment, when processing large samples, the operation time is long, and the results are not accurate enough.

As can be seen from table 3, the prediction precision of all three sites performs well. The reason why the overall accuracy rate is not high enough is that the recall rate of California and Georgia is relatively low, only slightly more than 30%. In contrast, the recall rate of New York is unusually high reaching 96.8%. The reason for this consequence is that overfitting may occur to New York. Random Forest is a classifier with multiple decision trees, which performs well in many classification problems. However, as can be seen in this experiment, when dealing with classification problems with high dimensions, the random forest algorithm tends to overfit due to the high noise.

6 Conclusion

Overall, both SVM and Random Forest have good performance in this study, but they are not satisfactory enough.

In machine learning problems, the selection of features often determines the upper limit of the result. Therefore, in order to get the better result, feature engineering can be improved. For example, Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) can used for dimension reduction.

Additionally, it can also be optimized from the perspective of the classifier. For instance, some integrated machine learning algorithms can be adopted including bagging, boosting and stacking.

Reference

Sakaki, T., Okazaki, M., and Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans. Knowled. Data Eng. 25, 919–931. doi: 10.1109/TKDE.2012.29

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., et al. (2013). Knowing the tweeters: deriving sociologically relevant demographics from twitter. Soc. Res. Online 18, 1–11. doi: 10.5153/sro.3001

Backstrom, L., Sun, E., and Marlow, C. (2010). "Find me if you can: improving geographical prediction with social and spatial proximity," in Proceedings of the 19th International Conference on World Wide Web (New York, NY: ACM), 61–70. doi: 10.1145/1772690.1772698

Rout, D., Bontcheva, K., Preoţiuc-Pietro, D., and Cohn, T. (2013). "Where's@ wally? a classification approach to geolocating users based on their social ties," in Proceedings of the 24th ACM Conference on Hypertext and Social Media (New York, NY: ACM), 11–20. doi: 10.1145/2481492.2481494

Kong, L., Liu, Z., and Huang, Y. (2014). Spot: locating social media users based on social network context. Proc. VLDB Endowm. 7, 1681–1684. doi: 10.14778/2733004.2733060

Chi, L., Lim, K. H., Alam, N., and Butler, C.

J. (2016). “Geolocation prediction in twitter using location indicative words and textual features,” in *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)* (Osaka), 227–234

Jayasinghe, G., Jin, B., Mchugh, J., Robinson, B., and Wan, S. (2016). “Csiro data61 at the wnut geo shared task,” in Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT) (Osaka), 218–226.