COMP90049 Geolocation of Tweets with Machine Learning

1 Introduction

Machine learning is a new developed computer science technique which allows computer to learn the pattern behind the data. Industries are increasingly relying on machine learning to make predictions and decisions. Fraud detection and cancer prognosis and prediction are example applications of machine learning. There are a large amount of studies focus on the classification of users in social media and further predict their behaviour. Determining the location of a user is part of user classification which is also an important component of marketing and location based recommendation system (Bao et al., 2012). Nevertheless, users are reluctant to allow public access to their personal information from the aspect of privacy protection (Sloan et al., 2013). Inference technique is regarded as a dominated way to gain such information. To investigate the relationship of users' location and posts, *Twitter* are used as data source to classify users' geolocations. Feature engineering and several classification methods will be discussed in this report. The objective of this study is to identify the best approach to predict the geolocation.

2 Literature Review

The geo-spatial locum of user are proven that can be indicated by their posts in social media (Han et al., 2012). Language are considered as geographically biased (Pennacchiotti and Popescu, 2011). A user who mentioned NYC subway is more likely to be from New York and these lexical priors contribute to the prediction of geo-location. Compare with one single post, the post history enhances the correctness of prediction (Huang and Carley, 2017). In practical, the weakness of collecting posting history is the time cost. *Twitter* data can only be collected through *Twitter* API and the API speed is limited.

Chi et al. (2016) suggested that multinomial Naive Bayes Classifier is widely used for text classification tasks. A similar study claimed that the precision of multinomial Naive Bayes Classifier achieves 52.8% on city level and 91% on country level (Huang and Carley, 2017). Other popular supervised classification includes K-Nearest Neighborhood (KNN), Logistic Classification, Random Forest etc. The advantages and weakness of these classifiers will be investigated later in this study.

3 Method

Feature Selection is a vital process in machine learning. The purpose of feature selection is to optimize a given objective by searching the optimal set of features. It influences on the learning speed and the amount of features effects the performance of prediction (Post et al., 2016). In this experiment, user Id, tweet Id, post and the location label are collected but only posts are considered as feature. All posts are used to train learning model and further label the users as New York, California, and Georgia.

Due to the fact that post history improves the precision, posts for each user are grouped respectively. It assumes that the label of a user is constant. A count vectorizer is used to vectorize word with minimum 1-gram and maximum 5-gram. To further select the features, chi-square test and variance threshold are utilised to filter out dependent and irrelevant features. Additionally, the amount of features is investigated to see the impact on the performance of classification.

Top 2000 features are selected to train the model by using multinomial Naive Bayes, k-Nearest Neighbour (K-NN), Random Forest, and Logistic Classification. The parameters of classifiers are varied to optimize the models. An existing machine learning library is used to ap-

Figure 1: Precision of Prediction With and Without Post History

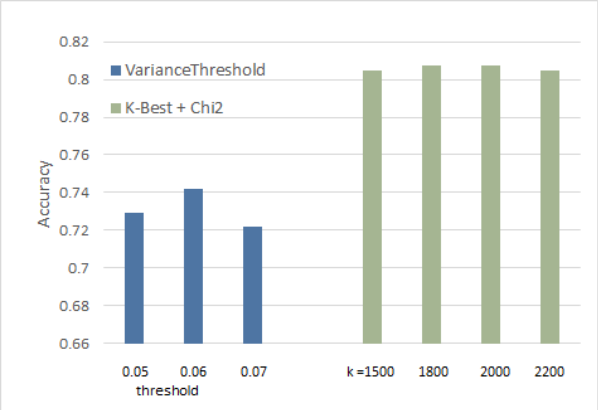

Figure 2: Feature Selection: Chi Square test, SVM

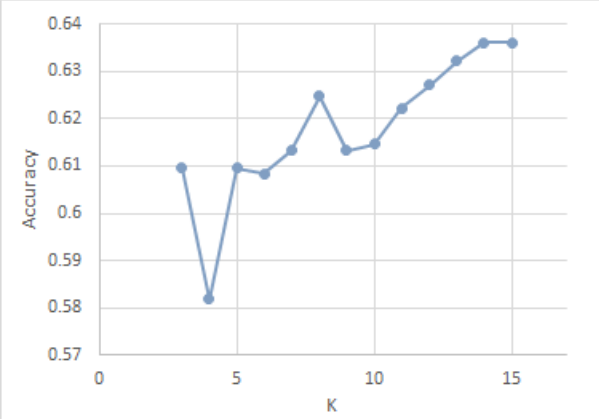

Figure 3: Classifier: K-Nearest Neighbour

ply the mentioned algorithms[1].

The outcomes of these classifier are assessed by accuracy. For the most ideal classifier, a confusion matrix of the labels are summarised. The precision and recall of them are also calculated individually to visualize the characteristic of dataset.

4 Result

4.1 Feature Engineering

Post history of a user is verified to be more powerful in classifying the user than a single post. In Figure 1, post history gives an increment of precision from 60% to 70-80% depend on the learning method.

Two feature selection methodologies with logistic classification are compared to assess their performance. Chi Square test deletes those dependent features and variance threshold is a low variance filter. As shown in Figure 2, Chi Square test has better performance which improves 10% precision than variance threshold. Both 1800 and 2000 best features offer a optimal data set. This fact suggests that the top 2000 features filtered by chi square test should be used for the further modeling.

4.2 K-nearest Neighbourhood Optimization

K-NN classifies the data based on defining the closet training data set in feature space. K-NN is desired for small scale data as it requires storing and searching (Wang et al., 2011). By merging the tweets of same user, the redundancy of data is alleviated and the scale of data is minimized. This process enhances the accuracy of KNN classification from 57% to 62%. As shown in Figure 3, the precision of model increase when the number of neighbours increases. The trend line behaves like an exponential function with a outlier point at k=4. As k increase, the run time takes longer without big improvement in accuracy. A large k might leads to over-fit of the model. Thus, a optimised K-NN classifier with precision of 63.6% with k=14 is took for future comparison.

4.3 Random Forest Optimization

Random forest classification is an ensemble of decision trees which avoids over-fit. A random forest model optimization involves a variety of parameters. Reif et al. (2012) claimed that the number of trees (called n-estimators), max tree depth, and maximum leaf nodes are the most influential parameters. As varying these parameters, the accuracy of classifier is jittered as shown in Figure 4 and Figure 5. As the outcomes do not converge, optimization becomes complex. Another characteristic of random for-

[1]https://github.com/scikit-learn/scikit-learn

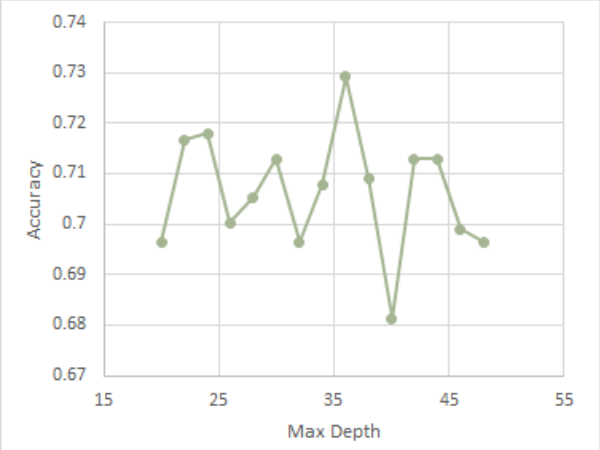

Figure 4: Random Forest Classifier with max depth

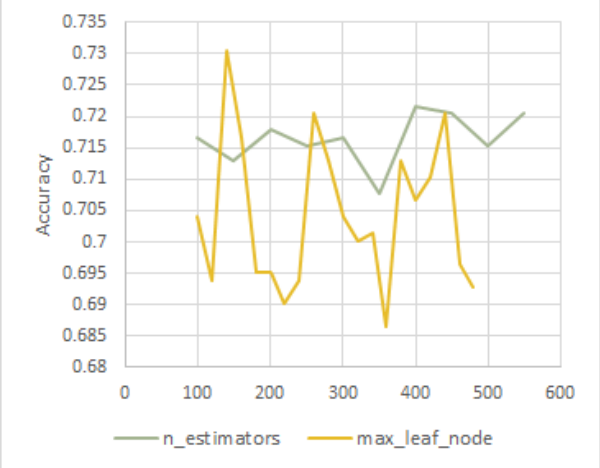

Figure 5: Random Forest Classifier with n-estimator and max leaf nodes

est classification is the accuracy change for each prediction. The range of its precision is 70% to 74% which is 10% above the K-NN classifier.

4.4 Four Classifiers Comparison

Other than random forest and K-NN, multinomial Naive Bayes and Logistic classification are assessed. A default random forest classifier with n-estimator = 10 and 14-NN classifiers are used. The best predictor is based on Logistic classification as shown in Figure 6. The accuracy of logistic classifier is 80.73%. Figure 6 also indicates a good fit of multi-nomial Naive Bayes and the dataset.

5 Evaluation

5.1 Classifier

By viewing Figure 6, K-NN classifier has a worst performance compared with other three classifiers. This might be caused by the nonuniform distributed amount of train dataset of labels and the noise data (Wang et al., 2011). This problem could be solved by weighting the dataset. In another way, random forest classification is a better learning method when dealing with unbalanced data (Kobyliński and Przepiórkowski, 2008).

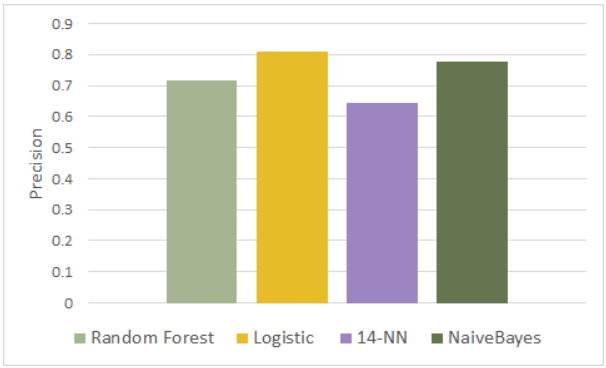

Figure 6: Classifiers Comparison

Random Forest classifier has less improvement than logistic regression after feature reduction. It indicates that random forest classifier has a better capability to process large amount of features and the noise has less effect on Random Forest classifier.

Multinomial Naive Bayes classifier obtains a more precise result than random forest classifier. That is because Chi Square test strengthen Naive Bayes classifier. However, a multinomial model has not only assumption on independence of words but also independence of multiple occurrences of same words lewis1998naive. The additional hypothesis cannot be implemented by Chi Square test.

Logistic Classification is the best approach over other three methods. It also take advantages of independence of features. According to Ho et al. (1994), to derive a more effective logistic method, the correlation is the issue to be solved.

5.2 Dataset

The visualization of the performance of logistic classifier is expressed in a confusion matrix, precision and recall as shown in Table 1 and Figure 7. From Figure 7, Georgia has a much higher precision than recall. This fact tells that the user who is not from Georgia may be easily be predicted as from Georgia. The classification of user from Georgia might involves many useless features which results in sparse prediction. The

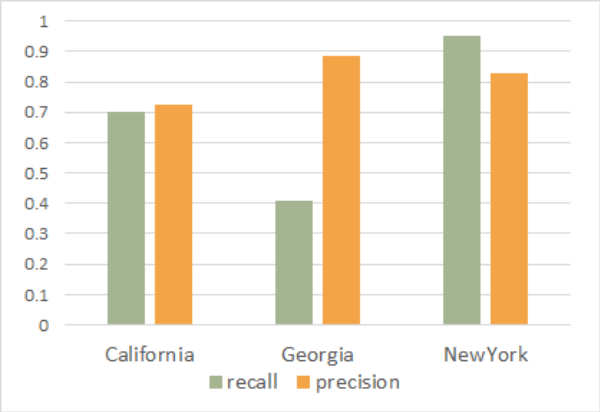

Figure 7: Evaluation of city classifications

recall of New York classification is higher than precision, which means the classifier misses out some user who is actually from New York. Low location-bias language can be one of the reasons that leads to this problem. Further, the number of users who is from New York is larger than other two cities in training dataset. Hence, the model of New York user classification is better than other two. This can be visualized by the higher recall and precision of New York user classification compared with the others.

City	California	Georgia	NewYork
California	118	3	47
Georgia	25	52	52
NewYork	20	4	472

Table 1: Confusion Matrix

6 Conclusions

The feature engineering has a big impact on the predictions. The Chi Square test is proven to be conducive to Naive Bayes and Logistic classifier. For further research, other feature selection strategy such as Mutual Information can be applied to evaluate the impact of feature selection.

Logistic classification is a best approach to estimate the geolocation of users. Naive Bayes also offers good results but it is more suitable for a model with less number of attributes. Random Forest classifier is costly to evaluate its parameters and the performance of K-NN classifier is not ideal. For future improvement, there are other classification methods available such as linear Support Vector Machine and neural network.

References

J. Bao, Y. Zheng, and M. F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th international conference on advances in geographic information systems*, pages 199–208. ACM, 2012.

L. Chi, K. H. Lim, N. Alam, and C. J. Butler. Geolocation prediction in twitter using location indicative words and textual features. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 227–234, 2016.

B. Han, P. Cook, and T. Baldwin. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062, 2012.

T. K. Ho, J. J. Hull, and S. N. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):66–75, 1994.

B. Huang and K. M. Carley. On predicting geolocation of tweets using convolutional neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 281–291. Springer, 2017.

Ł. Kobyliński and A. Przepiórkowski. Definition extraction with balanced random forests. In *International Conference on Natural Language Processing*, pages 237–247. Springer, 2008.

M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

M. J. Post, P. van der Putten, and J. N. van Rijn. Does feature selection improve classification? a large scale experiment in openml. In *International Symposium on Intelligent Data Analysis*, pages 158–170. Springer, 2016.

M. Reif, F. Shafait, and A. Dengel. Meta-learning for evolutionary parameter optimization of classifiers. *Machine learning*, 87(3): 357–380, 2012.

L. Sloan, J. Morgan, W. Housley, M. Williams, A. Edwards, P. Burnap, and O. Rana. Knowing the tweeters: Deriving sociologically rele-

vant demographics from twitter. *Sociological research online*, 18(3):1–11, 2013.

X. Wang, H. Zhao, and B.-l. Lu. Enhanced k-nearest neighbour algorithm for large-scale hierarchical multi-label classification. In *Proc Joint ECML/PKDD PASCAL Workshop on Large-Scale Hierarchical Classification*, 2011.