Model comparison

Table 1. Accuracy and time used records

Model/Feature (accuracy/time)	TF-IDF	Bigram	Unigram	1-to-2 grams	1-to-3 grams
500 features					
SVC(linear)	0.845/109.7s	0.727/283.54s	0.837/5063.7s	0.837/298.4s	0.837/296.5s
SVC(rbf)	0.633/238s	0.643/340.9s	0.834/162.46s	0.836/187.3s	0.836/176.6s
SVC(sigmoid)	0.5/221.9s	0.616/314.63s	0.828/177.8s	0.830/184.3s	0.830/180.8s
LinearRegression	0.834/5.673s	0.717/17.52s	0.817/8.04s	0.820/24.59s	0.819/41.78s
RandomForestClassifier	0.783/6.328s	0.691/18.1s	0.783/9.05s	0.778/25.38s	0.783/40.05s
VotingClassifier	0.806/363.05s				
1000 features					
SVC(linear)	0.852/187.23s	0.764/288.3s	0.865/349.3s	0.863/379.05s	0.86/384.6s
LinearRegression	0.843/7.11s	0.742/19.2s	0.839/12.11s	0.841/26.78s	0.84/49.0s
VotingClassifier	0.815/671s	0.751/769s	0.854/634s		
SVC(linear) /1200 Features			0.867/962s		
SVC(linear) /1500 Features			0.866/32184s		
5000 features					
SVC(linear)	0.876/1004.4s		0.84/908.3s	0.846/969.5s	
LinearRegression	0.826/63.3s		0.774/64.78s	0.772/86.7s	
VotingClassifier	0.806/6644s		0.850/2794s		
10000 features					
VotingClassifier	0.82/18674.1s		0.85/7136.4s		
LinearRegression	0.66/579s				

SVC

SVM provides SVC with linear, RBF and polynomial kernel. SVC models can get higher accuracy as more features. Table 1 shows SVC with a "linear" kernel always gets the highest accuracy.

Linear regression

Linear regression and logistic regression can be used to solve classification problems, much more straightforward than Linear Classification in optimization (Beader.me,2020). Linear regression is chosen to investigate whether the regression model can replace Linear Classification. Table 1 shows that the accuracy of linear regression is almost equal to that of linear classification with 1000 features while decreasing sharply with more features. Surprisingly, this model spends the least time.

Random forest classifier

Random forest classifier is an ensemble algorithm that combines a set of decision trees and aggregates votes from different trees to decide the final class of the test object. Table 1 shows that random forest classifier has surprisingly high efficiency as an ensemble algorithm, even more, efficient than linear classification. Unluckily, this model results in poor accuracy.

Voting classifier

Voting Classifier combines multiple different models into a single model. Because Linear Regression does not belong to classification, Combinations of Random forest classifier and different kinds of SVC models are carried out. As table 1 shows, this model performance well by unigrams as features. Obviously, it has low efficiency.

Conclusion

In conclusion, the accuracy of SVC (linear) is the highest and can be improved at the cost of efficiency as the increase of feature dimension. For the other models, the effect is optimal when the feature dimension is 1000~1500. Linear regression has high efficiency, which can be used in classification problems instead of Linear classification after selecting features.

Reference

Beader.me. (2020). Linear Models for Classification | Machine learning notes. [online] Available at: https://beader.me/mlnotebook/section3/linear_models_for_classification.html [Accessed 7 Jan. 2020].