Selection of Effective Features Using a Neural Network

Ken Ito (1110008)

School of Information Science,
Japan Advanced Institute of Science and Technology

August 7, 2014

Keywords: Feature Selection, Text Categorization, Neural Network, Support Vector Machine.

Machine learning is often used for natural language processing. It automatically learns several patterns for classification of data that is hidden in a sample data, and classifies an unknown data with learnt patterns. One of the serious problems in machine learning is overfitting. It is the phenomenon that the trained model is too fit to the classification of the training data and often fails to classify unknown data. Overfitting is often caused when too many features are used for training. The feature selection is a technique to tackle this problem. It is a method to find effective features from a given feature set and reduce the number of the features by removing ineffective features.

In this study, a method using a neural network (NN) for feature selection is proposed. It evaluates effectiveness of each feature by considering weights of edges in NN, then obtains a subset of only effective features. In this study, text classification is chosen a task for feature selection, since the number of features used in text classification is very large. It is supposed that the number of text categories is two for simplicity. After feature selection, Support Vector Machine is trained with the selected feature set for binary text classification.

The proposed method consists of the following 6 steps. In step 1, the features are extracted from texts. In step 2, a neural network consisting of three layers is learned from the training data. In this thesis, the feature

set is partitioned into small subsets, and then NNs using one of subsets are trained. Because we cannot train a single NN with all features due to the lack of computational resources. In step 3, the scores of features are calculated based on the learned NN. In step 4, the top T highly ranked features are selected. In step 5, SVM is trained with the selected feature set. Finally, in step 6, a category of an unknown text is classified by SVM.

Three kinds of features for text classification are used in this study. The first feature is uni-gram which is a single word that appears in a document. It is widely used in various natural language processing. The second one is bi-gram that is a sequence of two adjacent words in a document. Co-occurrence word is the third feature, which is a combination of words that appear in the same document. Unlike bi-gram, co-occurrence words are not necessary to be adjacent.

We propose three models to calculate the score of the feature from Neural Network. We call them as $score_A$, $score_B$ and $score_C$. $Score_A$ is calculated only from weights of edges between a node corresponding to the i-th feature in the input layer and nodes in the middle layer. $Score_B$ is calculated from weights of edges between nodes in the middle and output layers in addition to nodes between input and middle layers. It is defined as an average sum of weights of edges on paths from the node of the i-th feature to nodes in the output layer. $Score_C$ is also calculated from weights of edges between nodes in the input and middle layers, and middle and output layers, but the latter is different. Between two edges from a single node in the middle layer to two nodes in the output layer, only higher weight is accumulated to $score_C$.

Experiments to evaluate the proposed method are carried out. Eight feature sets are compared in the experiments. Three are feature sets of uni-gram, uni-gram and bigram, uni-gram and co-occurrence word that frequently appear in the training data. There are used as baseline. Another three are uni-gram, uni-gram and bi-gram, uni-gram and co-occurrence word that are selected in terms of both frequency and scores of the proposed method. The seventh feature set is uni-gram that is selected only by the scores of the proposed method. The last one is 'uni-gram and highly ranked co-occurrence word', where co-occurrence word features are produced as follows: first the highly ranked uni-grams (words) by the scores of the

proposed method are obtained, then pairs of one word in the list and another word in a document are used as co-occurrence word features.

Reuters Text Collection, an English text collection for evaluation of text classification, is used in the experiments. Ten TOPIC categories in Reuters Text Collection are used as the target categories. The performance of text classification is measured by accuracy, precision, recall and F-measure. Note that F-measure is the most important indicator.

First, to examine effectiveness of each type of the feature, three baselines are compared. Uni-gram achieved the best F-measure 0.838, followed by co-occurrence word 0.742 and bi-gram 0.734. The reason may be that the models using bi-gram and co-occurrence word are overfitted, since the number of features is more than 100,000. Next, three scores ($score_A$, $score_B$ and $score_C$) of the proposed method are compared. $Score_B$ was best when the top 15,000 uni-gram and bi-gram features are used, but differences among three scores were not so great. Next, the proposed method is compared to the baseline using the same type of the features where frequent features are simply chosen. When uni-gram is used as the feature set, the F-measure of the proposed method using 5,000 features highly ranked by $score_B$ was 0.045 lower than the baseline (0.884). When uni-gram and bi-gram are used, it was also 0.05 lower than the baseline. When uni-gram and co-occurrence word are used, the F-measure of the system with the 20,000 features highly ranked by $score_C$ was 0.092 higher than the baseline. However, it was comparable to the baseline with frequent uni-gram features. Next, the proposed feature set 'uni-gram and highly ranked co-occurrence word ' is evaluated. Unfortunately, the F-measure was much lower than the baseline with the frequent uni-gram. The highest F-measure was 0.791, where the top 25 words are used to construct co-occurrence feature, but it was 0.047 lower than the baseline. Finally, the validity of the scores of the proposed method is verified. In this experiment, the change of the F-measure is observed when the features are removed in the descending order of the frequency or $score_A$, $score_B$, $score_C$. If the proposed scores precisely represent effectiveness of features, the F-measure may decrease monotonously. When the features are removed in the order of frequency, the F-measure raises or drops many times. However, the features are discarded in the order of $score_A$, $score_B$ or $score_C$, we can observe that the

F-measure decreases almost monotonously. Therefore, we can conclude that our proposed scores are appropriate for features selection.

The proposed method requires further improvement in future. Instead of training multiple NNs for the partitioned subsets of features, a single NN should be trained to evaluate effectiveness of all features simultaneously. Furthermore, we can find that the proposed method using uni-gram and co-occurrence word outperforms the baseline for several target categories. If we can judge individual categories if our method works well and apply our method for those categories, the overall performance will be improved.