Automatically Selecting Best Features During Semantic Similarity Measure

Lasang Jimba Tamang
University of Memphis, Memphis, Tennessee
ljtamang@memphis.edu

Abstract. Feature selection during semantic similarity measure is one of the most important task. The performance of model depends on both the selection of features and appropriate number of features. In this project, we select best Fi number of features for i <= total number of features using chi square feature selection, build model, and evaluate the model, and finally the feature Fi that yields best performance is selected as best feature to be used.

1 Introduction

Semantic similarity measure uses different classification and regression machine learning algorithm. The performance of these algorithm depends largely on the feature set we use. However, choosing the feature set that best contribute to model performance is tedious job. Most of the time, we manually select different features to be used or use hit and trail method which are inaccurate very often. Even the small difference in the result due to failure in choosing best feature set can make the system behind from the competitor during compaction. In this scenario, our system attempts to use the approach for automatically selecting best feature set that can lead us to best semantic measure.

The overall idea is to iteratively build different model for all possible number of features available choosing best n features, build model on them and evaluate the model and finally report the feature as best feature which yielded best correlation coefficient.

Rest of the paper is organized as approach we took for our system construction, design we did for the system, implementation detail, result and future work that can be performed.

2 Approach

Out of many approach for the selection of best features, we are interested in in this project with chi-squared feature selection approach. In our approach, we incrementally select different number of best features starting from one to total number of features available using chi-squared test, build gradient boosting regression model for each of

those selected features, evaluate the performance using 10-fold cross validation, and finally choose the features set which yields best model accuracy during the evaluation process and report it as best features selection.

2.1 Chi-Squared Feature Selection

Chi Square Test is used in statistics to test the independence of two events. Given dataset about two events **A** and **B**, we can get the observed count **O** and the expected count **E**. Chi Square Score measures how much the expected counts **E** and observed Count **O** derivate from each other.

In feature selection, the two events are occurrence of the feature and occurrence of the class. In other words, we want to test whether the occurrence of a specific feature and the occurrence of a specific class are independent. If the two events are dependent, we can use the occurrence of the feature to predict the occurrence of the class. We aim to select the features, of which the occurrence is highly dependent on the occurrence of the class. When the two events are independent, the observed count is close to the expected count, thus a small chi square score. So a high value of χ^2 indicates that the hypothesis of independence is incorrect. In other words, the higher value of the χ^2score, the more likelihood the feature is correlated with the class, thus it should be selected for model training.

The χ^2 score can be computed as

$$\chi 2 = \sum_{i=1}^{n} \frac{(a_n - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i}$$

where,

χ^2 = Pearson's cumulative test statistic, which asymptotically approach a χ^2 distribution.

O_i = the number of observations of type *i*.

N = total number of observations

$E_i = Np_i$ the expected (theoretical) frequency of type *i*, asserted by the null hypothesis that the fraction of type *i* in the population is

n = the number of cells in the table.

We find the χ^2 score between each of our feature and the gold score, and the selection of best n feature are the features with are in the range of first best n scores.

2.2 Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Like other boosting methods, gradient boosting combines weak "learners" into a single strong learner, in an iterative fashion. It is easiest to explain in the least-squares regression setting, where the goal is to "teach" a model F to predict values in the form $\hat{y}$ = F(x), by minimizing the mean squared error $(\hat{y} - y)^2$ to the true values y (averaged over some training set).

We use gradient boosting to build model our regression model for text similarity.

2.3 Cross validation

Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset). The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like overfitting, give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem), etc.
One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

One of the main reasons for using cross-validation instead of using the conventional validation (e.g. partitioning the data set into two sets of 70% for training and 30% for test) is that there is not enough data available to partition it into separate training and test sets without losing significant modelling or testing capability. In these cases, a fair way to properly estimate model prediction performance is to use cross-validation as a powerful general technique.

In summary, cross-validation combines (averages) measures of fit (prediction error) to derive a more accurate estimate of model prediction performance. We use 10-fold cross validation method in which we divide our data set into 10 folds.

3 Design

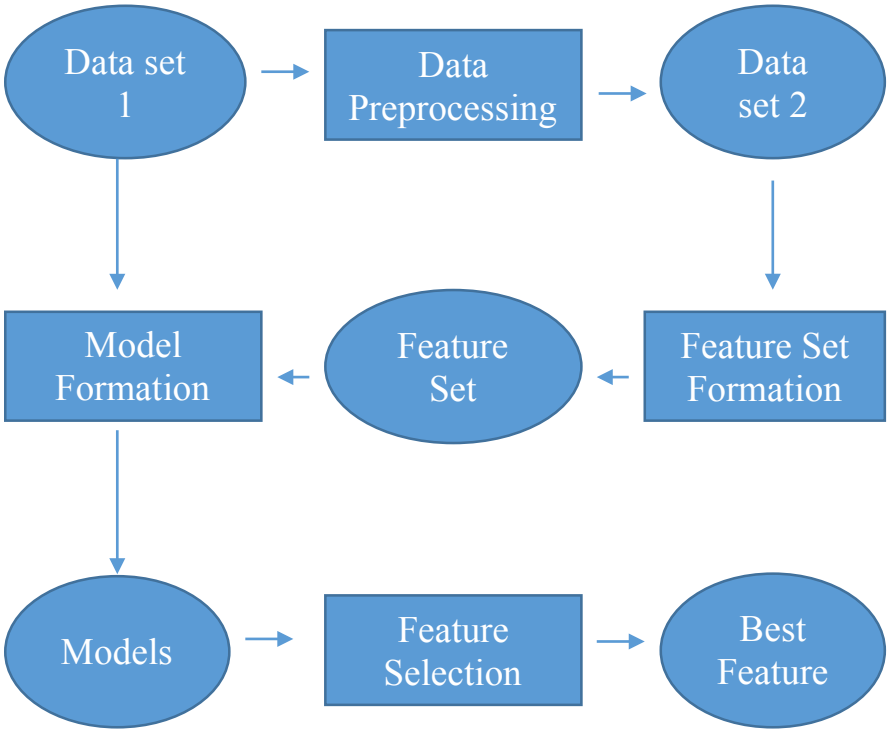

Figure 1: Architecture of Automatic Feature Selection

The architecture of automatic feature selection is as shown in the figure 1. In general, the data are preprocessed and different regression model are formed using the feature set given by chi-squared test, and the model is evaluated using 10-fold cross validation. Finally, the feature set that yielded best model with best accuracy is reported as best feature by our system.

The major building blocks are described in the following sections.

3.1 Data Preprocessing

Data preprocessing is used to obtain data set 2 to be used for chi-squared test for feature set formation. Feature set formation uses chi-squared test which only used non-negative values and discrete gold values. So, in this preprocessing step, we convert all non-

negative values to 0 and all continuous gold values to discrete values, which we call as data set 2.

3.2 Feature Set Formation

For n number of total feature, we can have 1, 2, … n possible ways of choosing best features. For this, we use chi-squared test score method using data set 2. While we have to find best n feature, we choose first n feature with best n scores. For example, if the score of features A, B, C with gold score G is 1, 2 and 3 and we have to find best 2 feature, we choose A and B, and A if we have to find best 1 feature only. In this way, we have total n feature set F_i where i<=n.

3.3 Model Formation

We take n (e.g. 50) different features set from step 4.2 during the model formation and for each feature set F_i where i<=n, we form gradient boosting regression model RM_i. We use data set 1 for training the models.

3.4 Feature Selection

Each RMi model is evaluated using 10-fold cross validation. The feature set F_i is selected as best feature that yield best average accuracy for RMI.

4 Implementation

For the implementation of the system, we took as data set generated Deep Tutor Lab, Institute of Intelligent System at University of Memphis, during SemEval-2017 International workshop on semantic evaluation, as data set 1. The data set consists of 4514 row with 51 columns. The first 50 column consists of score for different kind of features and the last column consists of similarity score between two sentence. We also call the last column as gold value and we call this dataset 1. The data set was preprocessed and data set 2 was formed. We generated 50 different feature set (F_1, F_2, F_{50}). We then build 50 different regression model (RM_1, RM_2, RM_{50}) using gradient boosting repressor in sckit-learn library using the feature set. We then evaluated accuracy of each of this model using 10-fold cross validation. The accuracy score used was Pearson correlation coefficient between predicted value and actual gold score value. Finally, we selected the feature set as best feature set which gave us the model with best accuracy.

5 Result

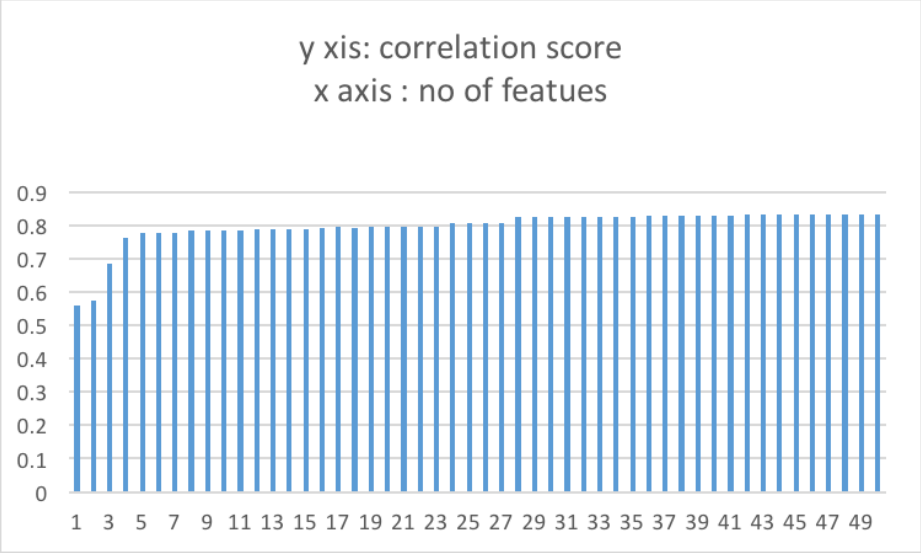

Figure 2 : Correlation score vs number of features used

The graph above shows the Pearson correlation score VS no. of selected features used. The score starts from 0.55 for single feature used, increases until 28 and then remains more or less constant then onward around 0.83 correlation score. The best score is achieved when number of features used is 45 and the best score is 0.833. The corresponding features used to build this model is then reported as best feature set.

6 Future Work

The above implementation is one approach in effort to automatically select best number of features. We have just used chi-squared method to select best n features and used only regression model. Besides, we have not put effort on parameter tuning during model implementation.

In the future we may extend this work by approaches other than chi-squared test for feature set formation, and different other model formation other than gradient boosting, and take the average accuracy of all of them and decide selection of feature based on that. Other than this, we can also extend this work to include parameter tuning during model formation process.

References

1. http://www.learn4master.com/machine-learning/chi-square-test-for-feature-selection
2. https://en.wikipedia.org/wiki/Gradient_boosting
3. https://en.wikipedia.org/wiki/Cross-validation_(statistics)