Proposal 349

1. What task? Why we are interested in this topic.

Nowadays social-related advertising is a common interest. When we go through some web pages especially the shopping websites, those websites can always know what items we bought most frequently and thus they can advertise specific items correspondingly, which brings up great profits to the company. A bunch of machine learning methods are used to improve advertising based on browser cache and user profiles.

We decide to design an advertising algorithm based on dynamic user profiles. We will calculate a similar group based on a seed group (the seed package) provided by advertiser and build our customer look-alike model. Look-alike models are used to classify larger audiences into smaller seed groups to create reach for advertisers. The audience classified into a specific seed group will reflect the benchmark characteristics of the seed. In summary, look-alike modeling can be used to reach new prospects that look like a marketer's best customers.

2. How to get our data?

We get our data from an online advertising algorithm competition. We have 4.5GB training data and 400MB data for testing. This dataset contains millions of seed group people, features of candidate group people, and features of seed crowd with corresponding advertising features. All the dataset has been divided into train set and test set. The training set categories users to users belonging to the seed package in the crowd and users who do not belong to the seed package (i.e. positive and negative samples). The test set will test whether our algorithm can accurately mark whether the user in the test set belongs to the corresponding seed package.

3. Which features/attributes will we use for the task?

Our user profile features include age, gender, marital status, education, consumption ability, interest, mobile carriers, operating system, house and so on. The advertisement features include ID of the advertisement, type of product, target of the advertisement, category of the advertisement. From our observation of this dataset, we find that most of the sample data features are sparse. We will find a way to deal with the sparsity.

4. What will our initial approach be?

We want to convert the dataset into one-hot encoding which is the standard way for categorical variables and the dimension of the whole dataset needs to be reduced because of the large size of the dataset.

Our baseline method would be decision trees and clustering because look-alike algorithm is a decision making problem. Therefore we think that the problem would be well suited for Decision Tree Classification.

Random Forests Classifier can also be applied to our model, because random forests is an ensemble learning method which is composed by a multitude of decisions trees. Instead of using all the features given in the training dataset, random forests utilize bootstrap aggregating to reduce the variance of the data and select significant features which should be more reliable than training a single decision tree.

Besides we will integrate our model with a couple of other classifiers as well, such as multilayer neural network, AdaBoosting. After researching on the task we found a recent released algorithm called light GBM which may work well on our task. Light GBM is an implementation of gradient boosting decision tree with two novel techniques: gradient-based one-side sampling and exclusive feature bundling. It is a modified decision tree classifier with built-in gradient boosting. It has great compatibility with large datasets and is histogram based, i.e., it buckets continuous feature values into discrete bins which fasten the training procedure. Therefore we assume it would be suited for our look-alike classification task.