ECE 681 Course Project Proposal

SHENGXIN QIAN

1. Problem Description

1.1. Project name

Classification on sentiment of Amazon product reviews

1.2. Project details

I plan to use Sentiment analysis to analyze customers' reviews. The result of sentiment analysis could represent true attitude behind feedbacks and help us quickly determine whether a product is our best choice or not without looking through all the reviews. I plan to use a classifier to determine the sentiment of reviews from Amazon. It is a 3-class classification which would label reviews with "positive", "negative" or "neutral".

1.3. The reason why I choose this topic

When we want to buy something on Amazon, some useful indicators which could help us to make a choice including star rating, sales number, price, customer reviews and Amazon's "best seller" tags. Intuitively, customers' feedback is the most useful and authentic indicator which including star rating and reviews. Even though star rating comes from customers themselves but that is not always consistent with people's reviews. Because people have a different understanding to the number of stars. So, when measuring customers' sentiment, the review is more accurate compared with star rating system. However, when the number of reviews is too large, we could only see some most useful reviews at the top. We have no time to look through all the reviews by ourselves. That is why an indicator which could indicate the portion of positive reviews could be very useful. Classification of reviews is a good solution because it could quickly get the percentages of different sentiments without the human being as long as it was well-trained.

1.4. Previous work on this problem

Sentiment analysis is a very popular topic in PPML area. In 1994, Daniel Jurafsky finished THE BERKELEY RESTAURANT PROJECT [1] which is a research on the sentiment of restaurant reviews and built the corresponding database. In 2005, Pang and Lee [2] collected many linguistic structures from movie reviews. Many PPML algorithms were used including SVM, NB and many other advanced ML algorithms [3]. The accuracy of sentiment classification has been more than 80% [4].

2. Data Description

2.1. The source of dataset

I plan to use Amazon product reviews as my dataset for training and testing. The most important reason is that the number of reviews is large and those reviews could offer me limitless training and testing dataset. At first, I plan to train my

classifier with 10000 reviews, 5000 of them are training data and the other are testing data. The number of datasets comes from papers of the previous study in this area [4]. I may adjust the number of dataset and portion of training data based on my own test. Also, many product reviews on Amazon are detailed and useful, which means it is highly possible to extract useful linguistic features from those reviews. More than that, I choose to use those data because Amazon offers APIs for developers to download products' reviews. Most of these reviews are well organized with a uniform format, which means I do not need to use the crawler to collect reviews by myself and consider about parsing problems.

There are a lot of Treebank datasets on the Internet which contains a lot of previous useful fully labeled parse trees. Those datasets are extracted from a large number of reviews and could be directly used with the corresponding algorithm as a part of the classifier. I would consider about using those labeled datasets to training my algorithm if I plan to use corresponding classifiers.

3. Initial Plan

3.1. Pre-processing

The feature selection is the most important part of the first stage. I plan to use bag of words representations [2] as the rule to extract sentiment features because it based on keywords frequency and location. All the keywords selection would base on related work in this area and words frequency analysis from my own dataset. Even though bag of words representations is out of data and the accuracy is not as high as RNTN [4], but it is the most intuitive way to estimate the sentiment of reviews.

3.2. Classifier selection

I plan to choose my classifier from BPNN, SVM, Decision Tree and Bayes [3]. All of them are sophisticated classifier but I don’t know how it works on my dataset and the features I choose. I tend to use SVM and BPNN to classify because both of them are my most interested algorithm. More than that, I plan to evaluate the performance of my classifier from these aspects: the running speed, the required number of datasets, the required training time and the accuracy of classification.

3.3. Statement

The problem I want to solve in this course project is similar to my another course project, but they focus on different aspects. My another course project is a shopping recommendation engine focusing on the realization of client and server application. The sentiment of reviews is just one of many arguments and the accuracy of sentiment analysis is not important. Other arguments include the star rating, price history, … I use Deeplearning4j to do the sentiment analysis in that project.

On the contrary, in ECE681, I only focus on the accuracy of sentiment analysis. I plan to write my own classifier without using open source library. I may use the reviews collected by the server application in another project.

4. Reference

[1] Jurafsky, Daniel, et al. "The Berkeley restaurant project." ICSLP. Vol. 94. 1994.
[2] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
[3] Hastie, Trevor, et al. "The elements of statistical learning: data mining, inference and prediction." The Mathematical Intelligencer 27.2 (2005): 83-85.
[4] Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." Proceedings of the conference on empirical methods in natural language processing (EMNLP). Vol. 1631. 2013.