Data Mining: Analyzing and Predicting the Quality of Product online

COMP3503 Project

Professor: Dr. Silver

Lei Xie

100123800

Jodrey School of Computer Science

Acadia University

1. Abstract

The Web has become a very common place for customers to leave comments and reviews. For example, Amazons and Best Buy have provided space for customers who have already bought the product items or have experience with online shopping to comment in each product page. This paper is focusing on online customer reviews of products. After gathering the reviews of competing products, comparisons were made and these comparisons will be a great source for potential customers and product manufacturer(Liu, Hu, & Cheng, 2005).

2. Introduction

The Internet has already changed the way people shop, and the way they make decisions when they purchasing products. There are so many related product information and reviews online. By dealing with such big data, data analysis methods and tools should be used. The goal of data analysis is to discover useful information, suggestions and to help make decisions, by processing of inspecting, cleaning, transforming, and modelling data. The tools used is called WEKA which is a open source software, and can do data testing, data virtualizing, and result predicting. The success criteria for the correctly rate should be around 95%.

The definition of data mining(DM) is that it is a process of discovering patterns in data. The process must be semiautomatic. “The patterns discovered must be meaningful in that they lead to some advantage, usually an economic one. The data is invariably present in substantial quantities. ”(Witten, Frank, & Hall, 2011)

There are two type of DM, which are Knowledge Discovery in Databases (KDD) process, and Cross Industry Standard Process for Data Mining(CRISP-DM), and I am choosing CRISP-DM for this project.

3. Method and Approach

a. Project Plan

CRISP-DM data mining model was used to solve this problem. CRISP-DM is a hierarchical process model including six steps, which are business understanding, data understanding, data preparation, modelling, evaluation, and deployment (CRISP-DM 1.0 Step-by-step data mining guide, Chapman, & Clinton, n.d.).

Business understanding

The original data was collected from epinions.com, which was a general consumer review website. The Epinions helps visitors decide on purchase by giving them reviews of items. So the business perspective requirements were, using machine learning tools to predict the product's quality, and helping customers make decisions easily.

Data understanding

The initial data set was quite straight forward. In MixedProsCons_train.txt file, each line contained target value "Cons" or "Pros" with a sentence or word inside. For the MixedProsCons_test_nolabel.txt file, instead of "Cons" and "Pros", the target value was "Labs".

In train data set, "Cons" and "Pros" were the classification values, which will be the predicted result for the test set.

Data preparation

The format of each data set was like “<Pros>SOME_WORDS</Pros>” or “<Cons>SOME_WORDS</Cons>”, which was not easy to convert into formatted XLSX or CSV file for future usage. The clean data format is “SOME_WORDS, Pros” or “SOME_WORDS, Cons”.

In order to read in by the machine learning tool WEKA, the string should be in quotes as well.

Modelling

There were several techniques that can be used for documentation classification, which were J48 , NaiveBayes, and NaiveBayesMultinomial classifier.

J48 classifier:

It was one of the decision tree modelling, and very common to use.

NaiveBayes classifier:

It implemented the probabilistic Naive Bayes classifier.

NaiveBayesMultinomial classifier:

It implemented the multinomial probabilistic Naive Bayes classifier(Witten, Frank, & Hall, 2011).

Evaluation

It is important to have high quality from a data analysis perspective. Before deploying the final test set in the model, we have to evaluate the training set and make sure the model we use is the better one and can help us achieve the business goal.

The value of Correctly Classified Instances will be the important part for evaluation. We would like our model to produce highest correct rate.

Deployment

In order to deploy the data mining result into the business, the final model should be evaluated before applying to the test data set. Finally, after the model was chosen, it deployed to the test data set, then got the predictions.

b. Solution

Firstly, a C program was written for data preparation, which made the data set easy to convert to CSV file, so that it can be read into the machine learning tool WEKA.

For the modelling, we choose J48 classifier and NaiveBayesMultinomial classifier for the final experimental design.

J48 classifier is one of the decision tree modelling, and it has two main parameters which are ConfidenceFactor value(0 ~ 1) and minNumObj value.

NaiveBayesMultinomial classifier implements the multinomial probabilistic Naive Bayes classifier, which have two parameters which are IDFTransform and TFTransform, and can be set as true or false.

c. Experimental design

In order to experiment these two models and find the best one, some parameters was adjusted to test the training data set.

J48

The ConfidenceFactor value and minNumObj value were set as the experimental parameters.Firstly, the minNumObj was set as default “2”, and changed the ConfidenceFactor value from 0.05 to 1.0 as eleven test sets, then recorded each results

in TABLE 1. Secondly, the ConfidenceFactor value was set of 0.6 for all the three test cases, which the minNumObj value is 1, 2 and 3, and recorded the data in TABLE 2.

NaiveBayesMultinomial

The IDFTransform and the TFTransform were set as the experimental parameters. For the first test, both IDFTransform and TFTransform were set as False, and for the second test, both of the parameters were set as True, and recorded the data in the TABLE 3

4. Results

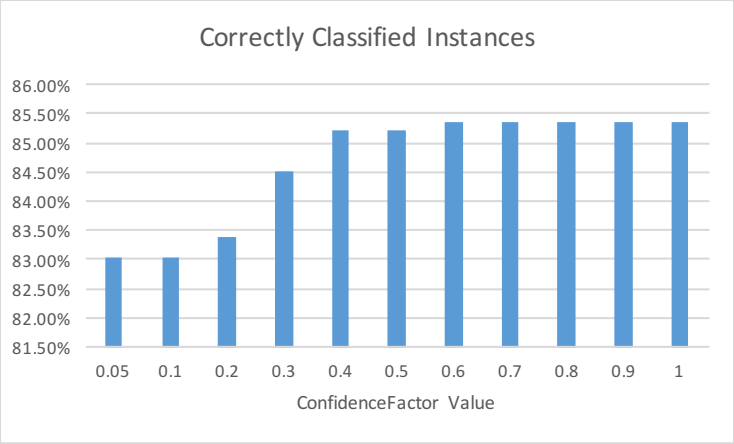
a. J48

The results using FilteredClassifier, and choose J48 as classifier, StringToWordVector as filter.

---

1. ConfidenceFactor Value

ConfidenceFactor Value	Correctly Classified Instances
0.05	83.05%
0.1	83.05%
0.2	83.40%
0.3	84.50%
0.4	85.20%
0.5	85.20%
0.6	85.35%
0.7	85.35%
0.8	85.35%
0.9	85.35%
1.0	85.35%

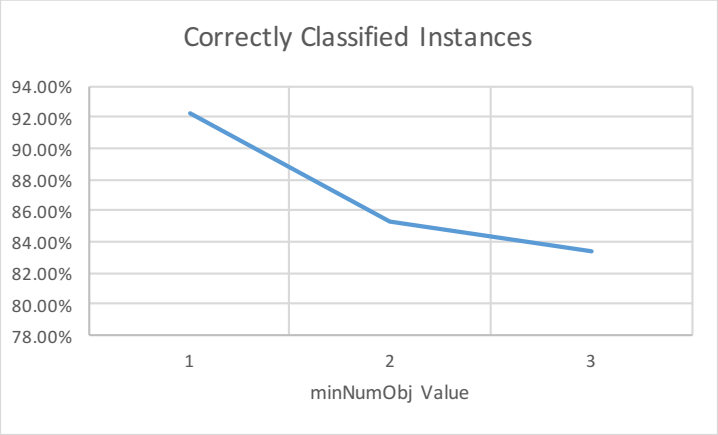


The table and chart both showed that the correctly classified instances increased from 83.05% and peaked at 85.35% in confidence factor value equal to 0.6, and remained the same after that.

---

2. minNumObj Value

minNumObj Value	Correctly Classified Instances
1	92.20%
2	85.35%
3	83.84%



The table and chart both show that the correctly classified instances had a decrease when the minNumObj value increased, and the maximum correctly classified instances was 92.20% when minNumObj value equal to 1.

3. The whole test set for training data

Here is the table that recorded the whole results for training data.

	A	B	C
1	ConfidenceFactor Value	minNumObj Value	Correctly Classified Instances
2	0.05	1	86.30%
3		2	83.05%
4		3	81.80%
5	0.1	1	86.30%
6		2	83.05%
7		3	81.80%
8	0.2	1	88.75%
9		2	83.40%
10		3	82.15%
11	0.3	1	89.65%
12		2	84.50%
13		3	82.50%
14	0.4	1	89.95%
15		2	85.20%
16		3	82.75%
17	0.5	1	90.30%
18		2	85.20%
19		3	83.20%
20	0.6	1	92.20%
21		2	85.35%
22		3	83.40%
23	0.7	1	92.20%
24		2	85.35%
25		3	83.40%
26	0.8	1	92.20%
27		2	85.35%
28		3	83.40%
29	0.9	1	92.20%
30		2	85.35%
31		3	83.40%
32	1	1	92.20%
33		2	85.35%
34		3	83.40%

The best correctly classified instances was 92.20%(flagged in red) in the whole test set, which the ConfidenceFactor value was larger or equal to 0.6, and minNumObj value is equal to 1.

b. NaiveBayesMultinomial

Here are two test sets for NaiveBayesMultinomial model. The test one is using default values, and for the second test, set both IDFTransform and TFTransform as true in StringToWordVector.

NaiveBayesMultinomial model	Correctly Classified Instances
TEST 1	95.4%
TEST 2	97.3%

The test one had 95.4% correct rate, and test two had 1.9% higher than the previous one.

5. Discussion

According to the result we got from the experiment.

In J48 model, the correctly classified instances peaked at 92.20%, when setting the parameter ConfidenceFactor value between 0.6 and 1.0, and set the minNumObj value as 1.

In NavieBayesMultinomial model, it had higher correctly classified instances as 97.3%, when setting both IDFTransform and TFTransform as true.

The best result of NavieBayesMultinomial model, had 5.1% more correct rate than the best one in J48 model which also achieved the success criteria.

6. Conclusion

Firstly, this paper stated the problem that how to predict the quality of product when given the reviews online. Secondly, it described the CRISP-DM methods we are using to approach the problem. Thirdly, it designed a experiment to test the training data set, to find out which model has the highest correct rate. Finally, according to the result,The NavieBayesMultinomial classifier was chosen as the final best model.

7. Bibliography

What is the CRISP-DM methodology? (n.d.). Retrieved November 29, 2015 from http://www.sv-europe.com/crisp-dm-methodology/

Chapman, P, & Clinton, J. (n.d.). CRISP-DM 1.0 *Step-by-step data mining guide*. Retrieved November 29, 2015, from http://www-staff.it.uts.edu.au/~paulk/teaching/dmkdd/ass2/readings/methodology/CRISPWP-0800.pdf

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. Elsevier.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In Proceedings of the 14th International Conference on World Wide Web (pp. 342–351). New York, NY, USA: ACM. http://doi.org/10.1145/1060745.1060797