Adaptive Personalisation
COMP40320

Review Classification Assignment

Name: Li Kang

Student Number: 14203445

29/APRIL/2015

Section 1. Description of Attributes

Attribute Name	Attribute Description
Rating	The rating from the review of every single product
SMOG	The Simple Measure Of Gobbledygood(SMOG) is a way of estimating the difficulty of writing
WordsPerSentence	The mean number the words for the sentence in the review
PercentComplexWords	The percent of the complex words in a review
SyllablesPerWords	The average number of syllables for every word
FleschReadingScore	The higher the score, the easier it is to understand the text
FogIndex	A simple formula for measuring readability
FleschKincaid	This score rates text on U.S. grade school level
NumberOfSentence	The number of sentences in the review
Width	The total number of the features in this review
Depth	The average number of words per sentence which includes a feature word
NumComplexWords	The number of complex words that occur in this review
NumberOfSentiment	The number of the sentiment number in this review
PopularityOfProduct	The number of the reviews about this product in the dataset
StdDevRating	The standard deviation of the ratings for a certain product from the whole reviews
MeanRatingForProduct	The mean rating for a certain product from the whole reviews
NumberOfWords	The number of words is this review
reviewHelpfulness	Class label (helpful or unhelpful)

Section 2. Analysis of Datasets

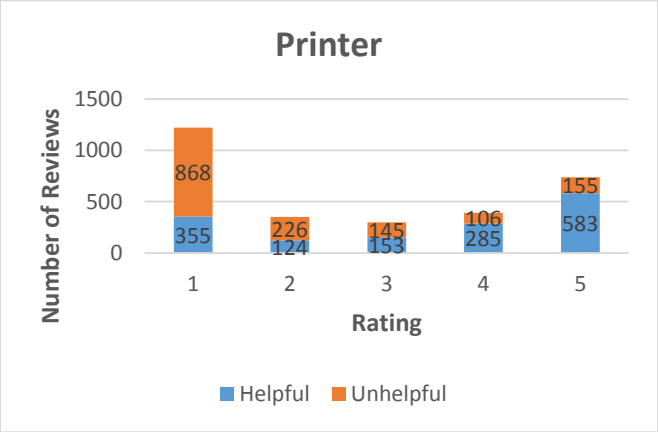

Fig2.1 Number of helpful and unhelpful reviews according the rating for printer

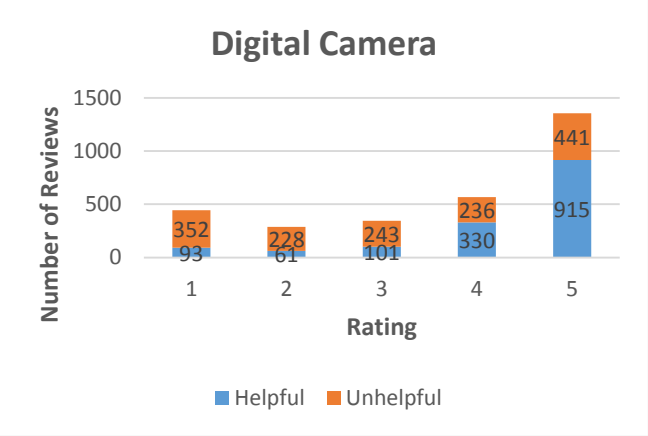

Fig2.2 Number of helpful and unhelpful reviews according the rating for digital camera

From the tables above, we can see that there is a huge difference between them. Generally speaking, the digital camera category has more helpful reviews in the five-star rating, which is nearly two times more than that of the printer category. While, the number about the helpful reviews for printer category in the one-star rating far weights than that of the digital camera, however the amount of the unhelpful reviews accounts most of them, about 868.

In fact, nearly 17% reviews is less than 4 star in the digital camera, but the percentage goes up to about 42% for the printer. Besides, the dataset received fewer reviews (less than 40%) from 2 to 4 star than 1-star or 5-star reviews in the two categories, and majority of the number of reviews in the first range above is fewer than 400. Moreover, the proportion about the helpful reviews from 2 to 4 star in the printer is more than the unhelpful reviews compared to the digital camera. Finally, the common feature for camera and printer is that the biggest part of the useful reviews both are in 5 star, and the number is 915 and 583 respectively.

In conclusion, the helpfulness increases as the rating grows, however the digital camera received more rating reviews in the 4 and 5 star.

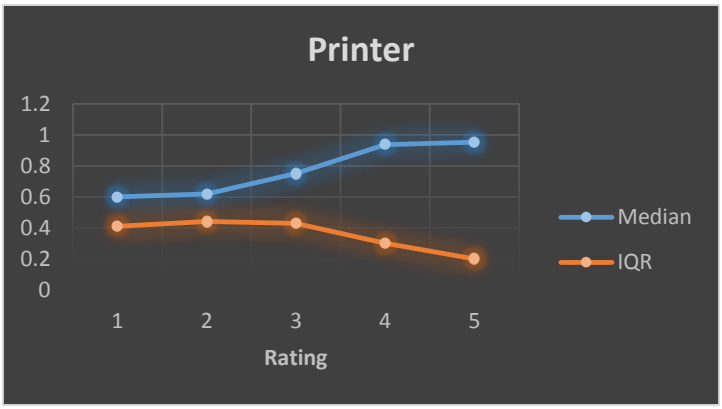

Fig2.3 Median and IQR of review helpfulness for printer

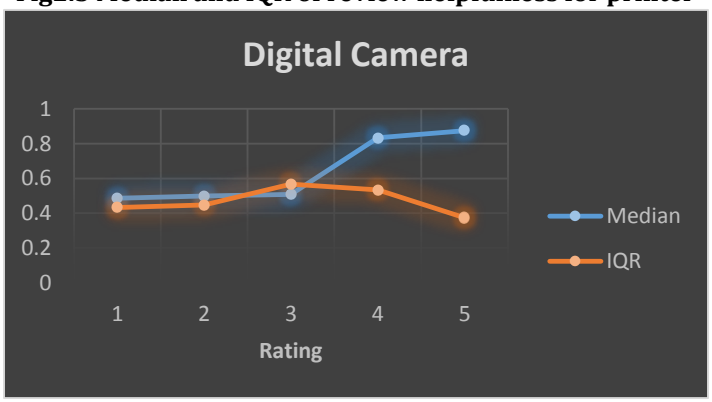

Fig2.4 Median and IQR of review helpfulness for printer

In this two line charts, the median for the printer and digital camera category increases gradually as the ratings grows, which means the higher the ratings, the more useful and meaningful advises the reviews will provide to the customers. Besides, the median hits the peak about 0.952 and 0.875 respectively for printer and camera in the rating 5 star from the lowest level (both are around 0.55) in the 1 star rating.

The IQR (interquartile range) is a measure of statistical dispersion being equal to the difference between the upper and lower quartiles, and it is defined as Q3-Q1; in other words, the IQR is the 1st quartile subtracted from the 3rd quartile. Therefore, a lower IQR value indicates the dataset more concentrates in a small range; a higher IQR value means that the dataset is diverging. In a word, this score is kind of divergence measures how much dataset are diverging. Also, what we can find is that the IQR is the lowest in the 5 star rating at about 0.2 and 0.4 respectively for printer and camera. The higher median score in 5 star means there are more useful reviews in this rating, and the lower IQR also indicates the most of 5 star reviews are more likely to be evaluated as helpful; so both of the two values indicate the higher rating reviews are more reliable. In addition, the trends of the either category are similar: increase from 1 star to 3 star, and reach the peak at 0.4 (printer) and 0.6 (camera), after that there is a dramatically decline. However, the IQR for printer is always lower than the camera.

Section 3. Analysis of Results

Classifier	TP Rate	FP Rate	ROC Area	Lable	Accuracy(overall)
J48					0.6997
	0.633	0.234	0.718	helpful	
	0.766	0.367	0.718	unhelpful	
IBK K = 1					0.6423
	0.681	0.396	0.642	helpful	
	0.604	0.319	0.642	unhelpful	
IBK K = 5					0.6963
	0.711	0.318	0.756	helpful	
	0.682	0.289	0.756	unhelpful	
IBK K = 10					0.701
	0.752	0.35	0.771	helpful	
	0.65	0.248	0.771	unhelpful	
RandomForest numTrees = 10					0.7053
	0.713	0.302	0.778	helpful	
	0.698	0.287	0.778	unhelpful	
RandomForest numTrees = 50					0.734
	0.685	0.217	0.805	helpful	
	0.783	0.315	0.805	unhelpful	
RandomForest numTrees = 100					0.737
	0.683	0.209	0.809	helpful	
	0.791	0.317	0.809	unhelpful	

Fig3.1 Results for the printer

In this section, using those two datasets to test the performance of three classifiers (J48, IBK and Random Forest), and later I will discuss the quality of the datasets as well as the attributes. Generally speaking, the accuracy is increasing as the number (K) of nearest neighbors grows in IBK and the number of the generated trees goes up in RandomForest. But the growth rate for accuracy sharply falls when the options have reach the threshold, so we should think about other methods to increase the performance. Moreover, the highest accuracy is at RandomForest when the number of generated tree is 100, and there is another thing that we should pay attention to: in this classifier, each time only considers 5 random features to generate the tree. Therefore, sometimes the high-dimensionality will bring in some noises to damage the accuracy. In the real world, we should extract features that are more compact and less noisy. Thus, I used ***wrappers*** in Weka to reduce the dimensions of the features, and use ***WrapperSubsetEval*** as the evaluator, ***BestFirst*** as the search method (***Backward***) to select features. Finally, there are 4/16 selected features:

```
Selected attributes: 1,11,14,15 : 4
                     rating
                     stdDevRating
                     numberOfSentiment
                     popularityOfPriduct
```

In the next step, **only** the selected 4 features and class label are used to test the same classifiers, and the following is part of the results:

Classifier	TP Rate	FP Rate	ROC Area	Lable	Accuracy(overall)
J48 Selected 4 Features					0.738
	0.643	0.167	0.78	helpful	
	0.833	0.357	0.78	unhelpful	
RandomForest numTrees = 100 Selected 4 Features					0.7073
	0.691	0.276	0.776	helpful	
	0.724	0.309	0.776	unhelpful	

Fig3.2 Results for the printer with selected features

It is obvious from figure 6 that the fewer but more compact features contribute the accuracy for J48 dramatically, from 69.96% to 73.8%, which is even higher than the previous highest accuracy (73.7%) in the figure 5. However, the accuracy for RandomForest decreases from 73.7% to 70.73%, which is mainly because that there are so limited feature to choose from when every time start to construct the trees, so there will have many duplicated trees. Thus, we always need to extract features from the feature set, and select suitable classifier to analysis data. Different classifier fits different data.

Another value we should discuss is ROC Area. Again, the value for the RandomForset classifier is the highest among them, which indicates the great performance of this classifier in this dataset. Increasing the number of trees will promote the ROC (from 0.778 to 0.809), which indicates the reliability for RandomForest in this situation. The reasons why RandomForset performs better, I think, is that:

- IBK will classify an object by a majority vote of its K neighbors, and the fewer neighbors (for example 1) is too subjective, but the more neighbors will introduce the vote from unrelated objects.
- J48 builds decision trees using the concept of information entropy, however, in this dataset there are some features tend to be the continuous or redundant attribute, which need to be handled better.

However, when every time start to generate the trees, the RandomForset classifier will avoid the shortcomings largely by choosing the random features, and it is robust to the presence of noisy features.

In addition, for the TP rate and FP rate, both for the helpful reviews are always higher than those of unhelpful ones.

Classifier	TP Rate	FP Rate	ROC Area	Lable	Accuracy(overall)
J48					0.7445
	0.752	0.263	0.753	helpful	
	0.737	0.248	0.753	unhelpful	
IBK K = 1					0.6923
	0.723	0.339	0.692	helpful	
	0.661	0.277	0.692	unhelpful	
IBK K = 5					0.7233
	0.755	0.309	0.792	helpful	
	0.691	0.245	0.792	unhelpful	
IBK K = 10					0.7263
	0.809	0.357	0.813	helpful	
	0.643	0.191	0.813	unhelpful	
RandomForest numTrees = 10					0.758
	0.798	0.282	0.836	helpful	
	0.718	0.202	0.836	unhelpful	
RandomForest numTrees = 50					0.7767
	0.78	0.227	0.85	helpful	
	0.773	0.22	0.85	unhelpful	
RandomForest numTrees = 100					0.7783
	0.779	0.223	0.854	helpful	
	0.777	0.221	0.854	unhelpful	

Fig3.3 Results for the digital camera

For the dataset for the digital camera, we can see that every ROC Area and Accuracy are higher than those of the printer in the tests, while, the trends for the different type of scores are similar as printer's when to change the options. To be concise, the accuracy is normally increased by 0.04 than that of the printer. In the J48, the TP and FP rates have raised for the helpful reviews than that of printer, and the rates for the unhelpful reviews are lower than that of printer; The TP rate in the IBK for camera is always higher than that of the printer, but the FP rate presents the opposite trend. Finally, the highest accuracy (0.737) in the Print from RandomForest is even lower the lowest accuracy (0.758) from the Camera classified by the same classifier.

In conclusion, the two dataset are analyzed by the same classifiers with the same configurations, but the accuracy and ROC Area are higher for the camera category than that of the printer category. Therefore, the reviews for the digital camera are more reliability to the customers.

Conclusion:

1. The growth rate for the accuracy will slow down when the configuration has reached a threshold.

2. Different classifier fits different dataset, and it is a common method to improve the performance by extracting a subset of the attributes with less noise.

3. Compared to the printer, the reviews for the digital camera are more reliable and will provide more meaningful advices to the customer.

Section 4. Analysis of The Attribute Selection

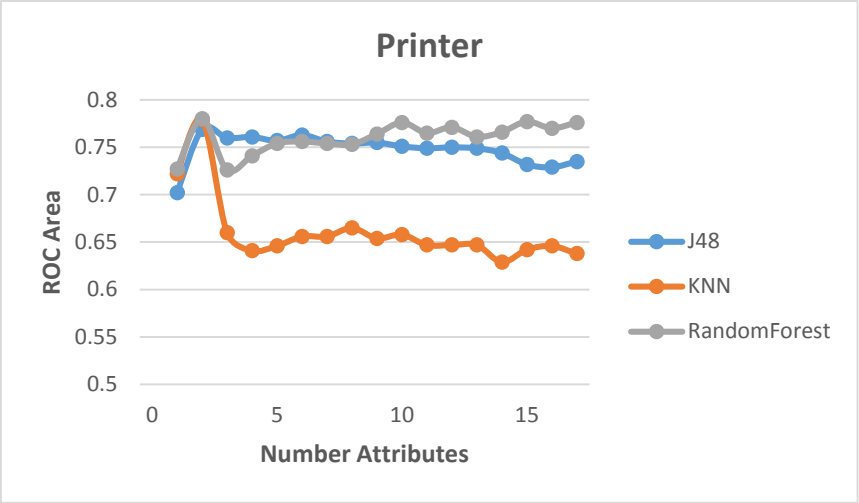

Fig4.1 ROC Area about Attribute Selection for Printer

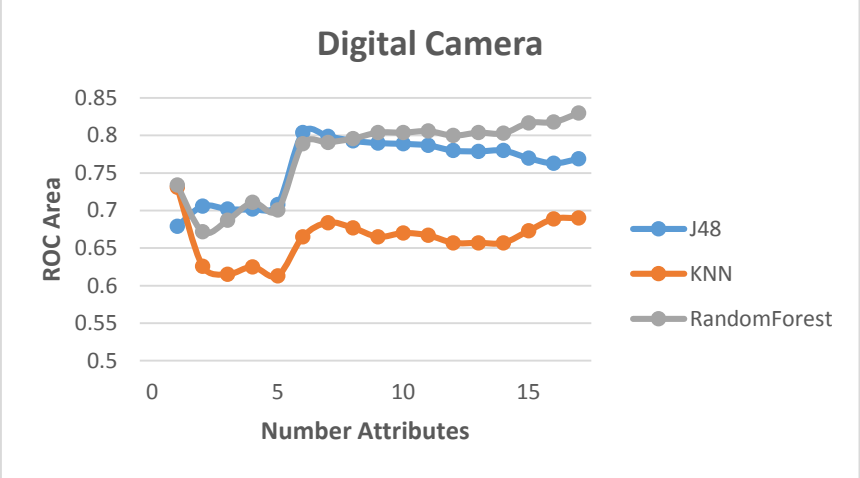

Fig4.2 ROC Area about Attribute Selection for Digital Camera

From the graphs, we can find that the trends for the lines about ROC Area in the two dataset are totally different. It is obvious that only two lines (J48 and RandomForest) overtakes 0.8 in the Digital Camera category from the two figures, which can be considered to be "Good" for the performance of those two classifiers. However, the IBK classifier always performs "Poor" (ROC Area is lower than 0.7) in either of the categories.

For the J48 and RandomForest, the performances of them in printer and camera are similar; both of them are fluctuating after a growth, but the peak points in the two categories are different, at 2 attributes in printer and at 6 attributes in camera respectively. Besides, for the IBK classifier in the printer, there is an increase from 1 attribute to the 2 atttibutes, and then the line fluctuates around 0.65 from the 3 attribute to the 17 attributes; the digital camera saw that the line for IBK increases

from the beginning to the 2 attributes, and it stays constant to the 5 attributes, then there is another fluctuation between 0.65 and 0.7 until the end.

However, the reasons why the same classifiers perform different for the different dataset are that:

- There are almost 200 words to define the features for the digital camera, while the available features for the printer are only 84 in the **Printer_Features.txt**. Therefore, the more features will provide more aspects for the customers to describe the camera in the reviews, which will offer more clues to distinguish the quality of the reviews. As a consequence, the performance of the classifiers in the camera will be better than those in the printer.
- 1 star review tends to be more subjective and shorter, and the number of the 1 star review in printer and camera are 1223 and 445 respectively. Besides, the majority of the 1 star review in the printer is Unhelpful, which will bring more negative influences into the quality of the printer dataset.

The next difference for those tow category is the top-ranked attributes:

```
Ranked attributes:
 0.14221    1 rating
 0.08789   15 numberOfSentiment
 0.08613    3 numberOfWords
 0.07746   14 numComplexWords
 0.07302    4 width
 0.07241    2 numberOfSentences
```

(a)

```
Ranked attributes:
 0.15252   15 numberOfSentiment
 0.14909    3 numberOfWords
 0.14605    4 width
 0.14545   14 numComplexWords
 0.12502    2 numberOfSentences
 0.12233    1 rating
```

(b)

Fig4.3 Top-6 attributes for the database about printer (a) and camera (b).

From the figure 4.3, we can find that the top-6 attributes ranked by the Information Gain are same, but the order is different: in the printer, the top-1 attribute is Rating, however the Rating attribute is ranked as the sixth in the digital camera. Moreover, the attributes of NumberOfSentiment and NumberOfWords enter into the top three ranks in the both of the datasets. Therefore, those mentioned attributes are good at distinguishing whether the review is Helpful or not.

In addition, compared to the IG of the top-ranked attributes in the printer (only one is higher than 0.1), those in the digital camera are all higher than 0.12. The top-ranked attributes in the camera are more reliable than those of the printer, since an attribute with high IG usually should be preferred to other attributes. Therefore, this is another reason why the classifiers perform better in the camera category.