Classification of Product Rating Using Data Mining Techniques

Pinku Deb Nath, Sowvik Kanti Das, Fabiha Nazmi Islam, Kifayat Tahmid, Raufir Ahmed Shanto and Rashedur M. Rahman

Abstract Data mining is the procedure to find patterns and necessary details from huge amount of data collected from various sources for a period of time. The target of our research is to classify the rating of individual products in an online shopping website based on price, discount, number of items left, sellers, count of likes and seller followers. The online shopping website from where we collected the data for prediction is kaymu.com.bd which is an online store in Bangladesh. The product rating that we are going to predict gives the correct rating of each product that not only depends on a single user's rating but the overall rating considering views of every other user. This helps user to decide what product to buy and how good it actually is.

Keywords Data mining · Product rating · Online shopping · Machine learning

P.D. Nath · S.K. Das · F.N. Islam · K. Tahmid · R.A. Shanto · R.M. Rahman (✉)
Department of Electrical and Computer Engineering, North South University,
Plot – 15, Block – B, Bashundhara, Dhaka 1229, Bangladesh
e-mail: rashedur.rahman@northsouth.edu

P.D. Nath
e-mail: pinku.nath@northsouth.edu

S.K. Das
e-mail: sowvik.das@northsouth.edu

F.N. Islam
e-mail: fabiha.islam@northsouth.edu

K. Tahmid
e-mail: Kifayat.navid@northsouth.edu

R.A. Shanto
e-mail: Raufir.ahmed@northsouth.edu

D. Król et al. (eds.), *Advanced Topics in Intelligent Information and Database Systems*, Studies in Computational Intelligence 710,
DOI 10.1007/978-3-319-56660-3_3

1 Introduction

Shopping is a common interest of people. At present the Internet provided us the facilities of online shopping that has made our life easier. Now anyone can shop from home, saving all the effort to go to showrooms and buy things. But since people do not need to go shopping themselves how would they know the things they are buying are of good quality and whether they serve well. Our research comes up with assistance to customers providing the reliable product ratings out of 5, when no E-Commerce center in Bangladesh has done a thorough analysis of correct rating of products. The research has basically been done on kaymu.com.bd using web crawler to extract and create dataset necessary to find the product rating. In most online shopping websites, we see the rating of product, but this is often not satisfactory since if only one customer gives the product a rating of 5, it remains as 5 which might be different from other people who might have not rated the product but given bad reviews. Therefore, the given rating of product is somewhere vague and might not be correct. Our research performs an extensive analysis using price, discount, number of items left, count of likes and seller followers. The data mining methods and other techniques that are used here in our research are: Import.io for data collection, J48, Naïve Bayes, Multilayer perceptron in Weka.

2 Related Works

Many e-commerce related researches have been already done to improve the marketing strategy and profit margin of many organizations. In [1] the authors discussed various techniques of clustering, association rule mining and other data mining procedures applied on user internet usage to understand the behavior and motive of the customers. The authors extracted patterns based on the purchase and browsing information of the past customers to predict the behavior of the future customers. Similar work has been shown on a research by Patels and Chauhan [2] where they have used web mining to find plausible patterns in the usage of the customers to improve the user interface of the websites and provide customized services. Pattern discovery techniques were also used here. The authors analyzed on perspective of business shopping organizations but we predicted to help customers to get the right product.

Relatively new methods like Neural Networks have been suggested for web mining in the research paper by Crone and Soopramanien [3] to estimate the probabilities of class labels. In this research, several variables related to user information, product and past internet shopping data are used to classify consumers into “online shoppers”, “browsers” and “non-internet shoppers” using the data collected from surveys in UK. The authors in [4] predicted the success of future films with selected/information provided on IMDB before release. They faced few

problems initially with the format of source data which they converted to suitable formats using data mining tools and developed java application to process and extract data to overcome this problem. The prediction techniques are not as same as ours. We provide an easier way to collect data from an e-commerce website like kaymu.com.bd and process it but outcome is similar to find out the rating. In addition to finding patterns in the data, it is also important to visualize the patterns and effectiveness of the generated model. There are many visualization tools such as scatter plots, grand tour, data cube etc. [5]. The types of data required for the application of web mining in e-commerce are customer information, commodity information and server information [6]. Customer information refers to the personal data of the customers, commodity information refers to the product features such as price, amount left etc. and server information refers to the cookies, logs generated by a user session [6]. There are three sub-sections of web data mining and they are web content-mining, web structure-mining and web usage-mining [6]. Web content-mining is the process finding patterns, models or knowledge from the contents of a web page, web structure-mining is the process of recognizing the underlying correlations among the web pages and other online objects and web usage-mining is the process of mining browsing patterns from the usage information of the customers [6]. In our paper, we have explored the domain of web content-mining using import.io. Collection of web content information is primarily done using a crawler. There are some bespoke software tools and services to crawl webpages and collect data such as Kimono Labs, import.io and Crawlbot [7].

In addition, there are many scientific approaches for collecting data such as Mining Data Records in Web Pages (MDR) algorithm proposed in [8] which discovers sets in web pages with some common features by comparing the attributes of the child node in the tree generated by following hyperlinks etc. [7]. Since there are many e-commerce sites in many regions of the world and selling products of high diversity, the data analysis should be independent from both various languages and specific product domains [7]. The quick discovery of patterns and reacting to those patterns are of paramount importance to the online retailers. For example, in one study it was found that if a product stays online for a long time then the probability that it will be sold becomes very low [9].

3 Dataset and Tools

The online shopping website that is used in our research for data extraction to classify the product rating is kaymu.com.bd and import.io to scrap the html pages. We have used “WEKA” for implementing different algorithms’ of pruning and preprocessing the dataset and applying various classification techniques.

4 Methodology

Our approach is divided into three categories: “Data Collection”, “Preprocessing” and “Classification”.

4.1 Data Collection

Here we have used import.io. Import.io is a web application where there are various modules to collect data in various fashions from the webpages. We use the crawler module to collect data. Crawler was used to crawl the entire website, specifically the product pages within the website.

From the product page it is decided that to collect 10 information given in the Fig. 1 (relationship is shown in Table 1). We consider them important for the classification of rating. For doing this the attribute names are set manually from the crawler interface of import.oi. There are two ways to train the crawler so that it can retrieve data from all the product pages of the website. One way is to click the item from the webpage which can be accessed from the crawler interface. Another way is

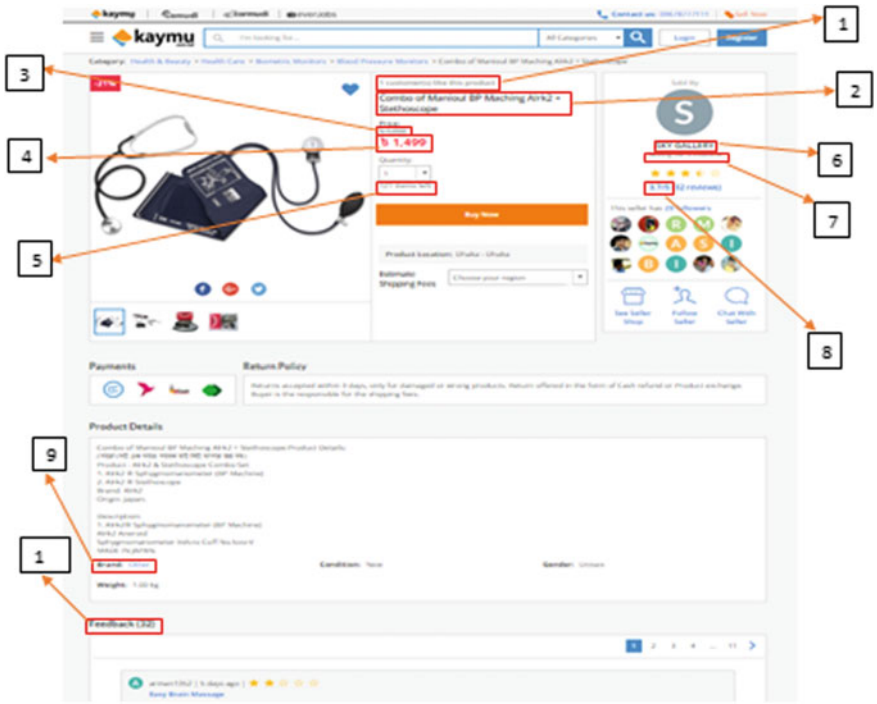

Fig. 1 Kaymu.com.bd product page

Table 1 Attribute relation and description

Attribute number from Fig. 1	Attribute name	Attribute type	Attribute description
1	product_like	Numerical	Number of likes of the product given by customers
2	product_name	Nominal	Name of the product
3	price_before	Numerical	Price of the product without discount
6	seller	Nominal	Name of the seller/company who is selling the product
7	seller_selling_for	Nominal	The number of years, month or days the seller is selling products
8	rating	Numerical	Rating out of five (average of all the ratings given by the customers). In this example 3.7/5 we took the numerical value 3.7
9	brand	Nominal	The name of the brand of the product
10	product_feedback	Numerical	Number of people who have given a comment on this product

to manually override the Xpath (reads through the source code of the webpage). The attributes where Xpath was used are given below:

rating: .//*[@itemprop="rating"]/@content

reviews: .//*[@itemprop="votes"]/@content
product_feedback: .//*[@class="row-small-12 mtl pbn"][contains(.,"Feedback")]
product_likes: .//*[@class="s-bold gray-medium wishlist likes"]/@data-wishlist-count
products_left: .//*[@class="quantity gray-medium s-bold"]/@data-qty

It is even possible to collect more attributes but it seems that the 10 attributes are enough to classify the rating. Once all the attributes are finished training from one product page we have to test it on four more product pages to see if we have made our training right. After finishing the training phase, we start our crawler. The starting page of the crawler: http://www.kaymu.com.bd/. Data extracted form pages with template (product page): http://www.kaymu.com.bd/{words-num}.html$. The collected data was downloaded from import.io as csv (comma separated format).

4.2 Data Preprocessing

The collected data which is in csv format is converted to arff format which is weka's default format. First we detect the outliers and extreme values using weka filters and

then we discretize some of the numerical attribute for better classification results. The discretization of the attribute seems to worsen the performance of the classification so we have decided later to omit the discretization. Among the attribute it is observed that the seller_follower has 8%, product_feedback has 8%, product_likes has 51%, and rating has 8% missing values. Two of the nominal attributes like Sellers and Brand are removed as they seem to impact the results of the classification in a negative way. We calculate the discount from the price_before and price_now attribute using excel and give name the new attribute discount because removing the price_before attribute gives better classification results. The attribute seller_selling_for has a number of varying nominal values like "1 days", "2 days", "3 days" etc. As a result, it takes a good amount of time to build our classification model so we decide to merge these attributes into three categories:

New: less than 1 month
Oldnew: less than 1 year and greater than 1 month
Old: more than 1 year

The new merged seller_selling_for attribute has not made a significant impact on the classification results but it seems to reduce the time to build classification models. The final distribution of the tenfold cross validation training data is given in Fig. 2.

4.3 Classification

Generally, there are two types of classifiers and they are discriminative and generative classifiers. Discriminative classifiers build a function from an input set to class label and generative classifiers build a model of a joint probability and predict the class label of an input instance using Bayes rules [10]. In Weka, there are many ways to configure the classifiers which affect the model constructed by the classifiers. For example, we can choose the minimum number of records to consider at each node of the tree. This will affect the size of the tree, for example, the decision tree with higher minimum number of records per node will have less nodes compared to the decision tree with fewer records per node. If we set the "prune tree" option to be true, then the generated tree will be even smaller. However, greater accuracy of the decision tree is achieved when the minimum number of records per node is low and the tree is unpruned. We have used tenfold cross validation which lowers the possibility of over fitting the training set into the model.

We have used oneR classifier to determine the most significant attribute. OneR classifies based on one of the attributes that gives the minimum number of error. In our case product_feedback gave the minimum number of errors. The model generated by oneR is 91% accurate with 5245/5704 instances correctly classified on training set.

We can also determine the baseline accuracy of the dataset using the zeroR classifier. In the zeroR classifier, the class label which occurs frequently is determined and then the classifier labels all the records with the frequently occurring class label. In our dataset, the baseline accuracy was 83%. Hence, if the accuracy of any model is greater than 83% then we can say that the model is useful. A model is considered good if the Area under the curve (AUC) of the Receiver Operating Curve (ROC) is above 0.5. If the AUC value close to 0.5 then the model randomly assigns class labels to the records. If the AUC value is less than 0.5 then the model is more likely to give wrong labels to records.

We experimented with three classifiers to observe which classifier created the most accurate model for our dataset. The classifiers used are J48 Decision Tree, Naïve Bayes and Multilayer Perception Neural Network classifiers.

4.3.1 J48 Decision Tree

The J48 Decision Tree is based on the updated version of the C4.5 Decision Tree algorithm. The J48 Decision Tree is a discriminative classifier. We have used tenfold cross validation for building the model. We have experimented and saw that when the minimum number of records per node is low and the tree is unpruned then the accuracy is high. Since we have used cross validation, the chances of over fitting was low. Hence, we have used minimum 2 records per node and left the tree unpruned.

The AUC value of the J48 model is above 0.98 for all the class labels except for the class label 2, which is satisfactory. The correctly classified instances are 98.1241% and incorrectly classified instances are 1.8759% which is also appreciable. Figure 2 shows the ROC plot of the J48 classifier.

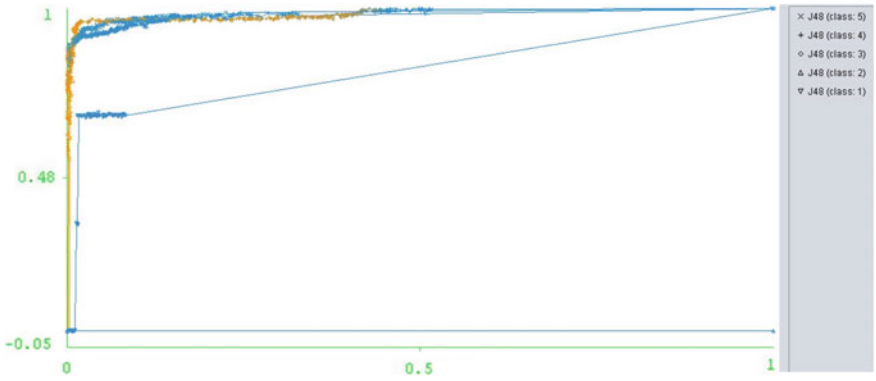

Fig. 2 J48 ROC Plot

4.3.2 Naive Bayes

Naïve Bayes is a simple probabilistic classifier which uses Bayes theorem to predict the probabilities of class labels. In Naïve Bayes, the attributes are considered independent of each other and they contribute to the class label of a record independently [11]. Maybe there is some inter-dependence in the attributes of the data-set for which reason Naïve Bayes performed poorly as a classifier.

4.3.3 Multilayer Perception

Multilayer Perception is an implementation of Neural Network which has two non-linear activation functions each of which maps weighted inputs to the output of each neuron. The activation functions are:

$$y(v_i) = \tanh(v_i) \quad \text{and} \quad y(v_i) = (1 + e^{-v_i})^{-1}$$

Here, the first function is the hyperbolic tangent and its value ranges from −1 to 1, y(i) is the i th neuron output and v(i) is the weighted sum of input synapses and the second function is the logistic function.

5 Evaluation

From the application of the three classifiers, we have observed that the model generated by the J48 Decision Tree algorithm has performed the best in classifying the rating for a product with a satisfactory AUC in ROC curve for the class labels 3, 4 and 5. We have also observed that the Naïve Bayes created a bad model for the classification of the rating for a product. The Neural Network based Multilayer Perception has created a model that has a satisfactory performance in the classification of the ratings. The model generated by the Multilayer Perception has taken around 10 s to construct the model while the models for the J48 and Naïve Bayes are generated within 3 s. The performance of the classifiers can be observed using the bar charts in Figs. 3, 4 and 5 which represent the Average Accuracy, Average F-Measure and Average AUC of the J48 Decision Tree, Naïve Bayes and

Fig. 3 Performance statistics chart of J48 decision tree classifier

Fig. 4 Performance statistics chart of multilayer perception classifier

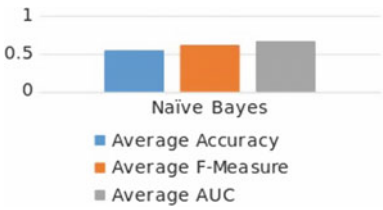

Fig. 5 Performance statistics chart of Naïve Bayes classifier

Multilayer Perception classifiers respectively. The unique characteristic of our data is that most of the products have rating 4 while some of the records have rating 3 and 5. Only three instances have rating 1 and none of the records have a rating of 2. Hence, generated models have performed well in classifying products that should receive rating 4, 5 or 3 because of the abundance of such products for training and testing datasets. We faced many problems and discrepancies while collecting data such as missing values etc. In addition, there were products that received many likes but few people rated that product. There were also products that had many good reviews but their rating did not reflect the positive sentiments of the customers. There might also be the situations where a product received a rating from a customer but the customer did not buy the product. The presence of such issues can reduce the effectiveness of the models generated by the data mining tools.

6 Conclusion and Future Work

The online shopping website from where we collected the data for prediction is kaymu.com.bd which is an online store in Bangladesh. The product rating that we are going to predict gives the correct rating of each product that not only depends on a single user's rating but the overall rating considering views of every other user. This helps user to decide what product to buy and how good it actually is. Possible future works include collaborating with the e-commerce owners and working on customer usage data that can be collected from the server end of the e-commerce websites. We can also suggest e-commerce website to ask their customers to fill a survey some days after they have bought a product because these survey data will more accurately reflect the sentiments of the customers.

References

1. Rastegari, H., Sap, M.N.M.: Data mining and e-commerce: methods, applications, and challenges. Jurnal Teknologi Maklumat 116–128 (2008)
2. Patel et al.: Web mining in e-commerce: pattern discovery, issues, and application. Int. J. P2P Netw. Trends Technol. **1**(3), 40–45 (2011)
3. Crone, S.F., Soopramanien, D.: Predicting customer online shopping adoption—an evaluation of data mining and market modelling approaches. In: Data Mining Conference (DMIN), pp. 20–23 (2005)
4. Saraee, M., et al.: A data mining approach to analysis and prediction of movie ratings. In: Data Mining V, pp 344–352. WII Press, UK (2004)
5. Zhang, F.: The application of visualization technology on e-commerce data mining. In: Second International Symposium on Intelligent Information Technology Application (2008)
6. Zhao, W., Lin, H.: WEB data mining applications in e-commerce. In: 9th International Conference on Computer Science and Education, Canada, pp. 557–559 (2014)
7. Horch, A., et al.: Mining e-commerce data from e-shop websites. IEEE Trustcom (2015)
8. Liu, B., et al.: Mining data records in web pages. In: SIGKDD USA, 2003, pp 601–606
9. Dlamini, M.G., et al.: Extracting interesting patterns from e-commerce databases to ensure customer loyalty. In: Proceedings of 2015 IEEE 12th International Conf. Networking, Sensing and Control, Taiwan, pp. 382–387
10. Andrew, Ng.Y., Michael, J.L.: On discriminative vs. generative classifiers: a comparison of logistic regression and Naïve Bayes. http://ai.stanford.edu/~ang/papers/ (2002)