RESEARCH ARTICLE OPEN ACCESS

Data mining of restaurant review using WEKA

Gayathri.T[1]
Assistant Professor, Department of Computer science, New Horizon college of Engineering, Bangalore

Abstract:

Many customers visit a restaurant based on food critics and reviews on websites such as Zomato.com. Restaurant strive at the initial stages of opening but their demand deteriorates after the initial hype. Further business for these restaurants are largely based on their reviews. What can the restaurant do to make their ratings better? Food taste is an obvious trigger to improve the ratings of a restaurant, but there are other factors that improve the ratings of a restaurant. Such as inclusion of cuisine, option for home delivery, presence of table etc. This paper aims a creating a prediction model for the reviews and analyze the trigger event that would improve the ratings.

***Keywords*- Restaurant review, Zomato, Multilayer Perceptron, Naïve Bayes, J48, Data mining**

I. INTRODUCTION

Machine learning sheds light on various domains unexplored by human analyses. It provides a view point that are not visible in general. The prediction and classification models which took scientist decades to created are achieved in days. Data when available in huge amount can be studied through machine learning algorithms to arrive at meaningful information.

Data miningused in predicting disease diagnosis, weather, customer expectations; learning the data to create automation, purchase pattern etc.There are four steps in the process of Data mining.Data collection, Data pre-processing, machine learning and Data visualisation.

Data collection is a predominant and difficult step in data mining. The data that is collected should be relevant and should cover all the spaces of the domain. The concentration on one sample space would lead to bias in the prediction or classification of result. After data collection comes data pre-processing. When data is collected not all information are relevant in the machine learning. For example, the age and date of birth can be two attributes in an employee data, this information is dependent on one another. The presence of such redundant information would lead to decreased accuracy. Next step is machine learning, Machine learning is the process where the system learns the data. From the learned knowledge it predicts, associates, classifies and clusters the data. For this purpose, various algorithms are used.The information gained from machine learning is better understood through visualization.

Restaurant is a domain which are traversed by small as well as big players. Data mining provides a way for both to improve their business with minimum effort. Restaurant business rely on the taste of food, the variety of cuisine that is provided in the restaurant, ambience, availability of home delivery, online booking, price etc. When any of the factor is improved or included it is possible to increase customer attention and thus increase productivity in business.

Zomato is a webpage and a mobile application which provides information about restaurants, reviews of restaurants and allows online ordering from the restaurants. The data is extracted using Zomato API [1] by Shruthi Metha. This dataset was downloaded from Kaggle an online repository for dataset [2].

II. LITERATURE SURVEY

There are a couple of research papers published based on restaurant reviews and hotel reviews. Following is a survey of such papers, [3] is a paper which reviews the Thai restaurants around the world. It attempts to find classify the restaurant

based on the reviews. The model proposed in this paper is, extraction of review from social networking site using text processing, artificial neural network is used to classify the dataset as positive and negative. mRMR feature selection technique is used for selecting the features of data set.

[4] paper analyses the fast food franchise data to help franchise reap benefit. Time series data from store as well as corporate is used with ARIMA model understand data. Outlier detection is used to identify sales opportunities and risk.

In [5] Yelp restaurant review dataset is used to model a system to improve restaurants. Here Latent Dirichlet Allocation (LDA) algorithm is used to find subtopics from the review. The ratings for the hidden topic allowed to understand the reason for rating.

In paper [6] the reviews are scraped from ww.tripadvisor.com using web crawler. The reviews are distinguished into positive and negative polarity using sentiwordnet and various machine learning algorithm are used to check their accuracy.

In most of these research papers reviews are extracted from one website and classification model is created. This paper is an attempt to create a trigger model to improve restaurants based on Zomato dataset.

III. DATA COLLECTION

Zomato data set is downloaded from Kaggle data repository. The dataset contains 22 attributes and 9552 instances. The attributes present in the dataset are: Restaurant Id, Restaurant Name, Country Code, City, Address, Locality, Locality Verbose, Longitude, Latitude, Cuisines, Average Cost for two, Currency, Has Table booking, Has Online delivery, Is delivering, Switch to order menu, Price range, Aggregate Rating, Rating color, Rating text, Votes.

IV. DATA PRE-PROCESSING

Data pre-processing can be data cleaning or data transformation. Dataset in Kaggle can be used for classification or association mining. When used in classification or prediction it is necessary to identify the features that would enable higher accuracy in classification. [7] suggests the use of data pre-processing to improve machine learning. Classification and clustering accuracy is predominantly dependent on the proper representation of data. Correlation based feature selection is used to reduce the number of features.

V. MACHINE LEARNING

Machine learning literally means, make the machine learn, machine learns by processing the data with various machine learning algorithm[7]. There is no fixed algorithm to provide high accuracy this is called No Free lunch theorem [8], however deep learning provides a better accuracy in most cases.

For any application it is important to apply few machine learning algorithms to find out the best suited model. Machine learning algorithms can be grouped under Bayes, Rule Based, Neural network and Decision tree.

A. Naïve Bayes

Naïve Bayes theorem is the best machine learning algorithm to use when the features are independent of one another[10]. Each instance is considered as a vector. The posterior probability of a class given a predictor is found with

P(h|d) = (P(d|h) * P(h)) / P(d)

P(d|h) - the posterior probability of class given a predictor

P(h) - Prior probability of a class

P(d) - Prior probability of a predictor

B. Decision Tree

Decision tree is arrived at by finding the optimum way to arrange the various nodes. There are two ways to identify the best partition of dataset at node, information gain or gain ratio. The decision tree model which uses information gain is ID3 and gain ratio is J48 [11]

C. Multilayer Perceptron

Multilayer perceptron contains large number of nodes called as neurons, joined together so that they for input layer hidden layer and output layer. The

instances are supplied though the input layer, bias and weight are added at the hidden layer and supplies the class in output layer [12].

VI. EXPERIMENTATION

The dataset acquired from Kaggle, first undergoes data preprocessing. From the information about dataset it was found that some attributes were redundant, restaurant id and restaurant represented the same information; Locality, locality verbose and latitude longitude represented the same information; rating color and rating text represented the same information. To avoid redundancy of attributes only one the representation was kept. Average cost for two is an attribute whose value is not standard. It depends on the currency attribute. Using the currency information, the average price is converted into standard US dollar format. Correlation based feature selection with ranker algorithm is done to reduce the number of dataset.

Machine learning algorithm such as J48, Naïve Bayes and Multilayer perceptron are prone to reap better results in most dataset. So Mutilayer perceptron, J48, naïve bayes classification is used learn the algorithm in WEKA is free online data mining tool published by Waikato University. The dataset is preprocessed, Feature selected, trained and tested using WEKA.The algorithm found to reap better result is J48.

TABLE II
ACCURACY RESULTS FOR CLASSIFICATION MODEL FOR ZOMATODATSET

Algorithm	Accuracy
J48	97.2%
Multilayer Perceptron	78.16%
Naïve Bayes	82.2%

To find the trigger to improve ratings, a sample record of poor rating is taken and modified to reduce the price range to one. This sample record is tested on J48 Zomato model. It was found that there was no change in rating. Whereas when the country code was changed there was change in rating

VII. CONCLUSION AND FUTURE WORK

Zomato dataset is used to a create classification model for restaurant rating. It was found that Multilayer perceptron work well with this dataset. In this paper an attempt is made to predict the trigger which would further enhance the rating of the review. This project can be further extended to create a tool to evaluate the trigger to improve the ratings.

ACKNOWLEDGMENT

I thank my college New Horizon college of engineering for providing support and tools for this research. I thank Head of Department, Dr.B.Rajalakshmi for her support and guidance.

REFERENCES

1. *https://developers.zomato.com/api#headline1*
2. https://www.kaggle.com/shrutimehta/zomato-restaurants-data
3. *Claypo, Niphat, and SaichonJaiyen. "Opinion mining for Thai restaurant reviews using neural networks and mRMR feature selection." Computer Science and Engineering Conference (ICSEC), 2014 International. IEEE, 2014.*
4. *Liu, Lon-Mu, et al. "Data mining on time series: an illustration using fast-food restaurant franchise data." Computational Statistics & Data Analysis 37.4 (2001): 455-476.*
5. *Huang, James, Stephanie Rogers, and EunkwangJoo. "Improving restaurants by extracting subtopics from yelp reviews." iConference 2014 (Social Media Expo) (2014).*
6. *V. B. Raut and D. D. Londhe, "Opinion Mining and Summarization of Hotel Reviews," 2014 International Conference on Computational Intelligence and Communication Networks, Bhopal, 2014, pp. 556-559. doi: 10.1109/CICN.2014.126*
7. *D. H. Deshmukh, T. Ghorpade, and P. Padiya, "Improving classification using preprocessing and machine learning algorithms on nslkdddataset," in Communication, Information & Computing Technology(ICCICT), 2015 International Conference on. IEEE, 2015, pp. 1–6*
8. *Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24.*
9. *Wolpert, David H., and William G. Macready. "No free lunch theorems for optimization." IEEE transactions on evolutionary computation 1.1 (1997): 67-82.*
10. *Lewis, David D. "Naive (Bayes) at forty: The independence assumption in information retrieval."*

European conference on machine learning. Springer, Berlin, Heidelberg, 1998.

11. *Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993*

12. *Goodman, Rodney M., and Zheng Zeng. "A learning algorithm for multi-layer perceptrons with hard-limiting threshold units." Neural Networks for Signal Processing [1994] IV. Proceedings of the 1994 IEEE Workshop. IEEE, 1994.*