A Study of Tree Based Machine Learning Techniques for Restaurant Reviews

Shina[a], Shikha Sharma[b], Anshu Singla[c]
Chitkara University Institute of Engineering and Technology,
Chitkara University, India
shina.pahuja@chitkara.edu.in[a], shikha.sharma@chitkara.edu.in[b], asheesingla@gmail.com[c]

Abstract—**Online reviews have become the easiest way to share one's experience with others, be it regarding a product brought or any kind of service availed. As reviews are the eminent factors that enhances the best services to the customers. Zomato has become the most popular application in India for ordering food online or checking about the reviews of a restaurant. Our research includes classifying restaurants into several classes based on their service parameters. Popular machine learning algorithms like Decision Tree and Random Forest were applied over a dataset of over 8500 records. The results have proved that the Decision Tree Classifier is more effective with 63.5% of accuracy than Random Forest whose accuracy is merely 56%.**

Keywords—online reviews, zomato, decision tree, random forest, machine learning

I. Introduction

Opinion is the key aspect in decision making process. Everyday people share their experiences, thoughts corresponding to various services and products through the World Wide Web. As the web is a huge repository, it contains various experiences of different customers based on their personal encounters. Internet is an essential part now a days and people can easily access the different opinions of different products and services. Online reviews help the user to choose the best option as per his requirement depending upon the ratings given by various customers. However, these opinions can be mined and interpreted on various factors. These factors would help us to improvise analysis tasks using the several mining techniques. Feedback is very important requirement for the service provider in order to improve and provide the best services to the customer. These helps in reducing traffic on road, save time and also help the individuals.

Data mining techniques contributes optimistic solutions [1]. It extricates the meaningful data in the form of patterns and relationships from the Dataset. Several Machine Learning algorithms are applied over the dataset for retrieving results. Generally data mining and machine learning are melded with each other. The point that makes Data Mining and Machine Learning different from each other includes the type of dataset on which the technique is applied along with its interpretations. The technique of Machine learning helps in to prognosticate upcoming incidents which are based on combinations of patterns, whereas data mining proceeds as a informant of information from machine learning to withdraw data. Though Data mining uses the techniques of recognizing patterns using classification and analysis yet Machine Learning advances this concept and uses similar algorithms

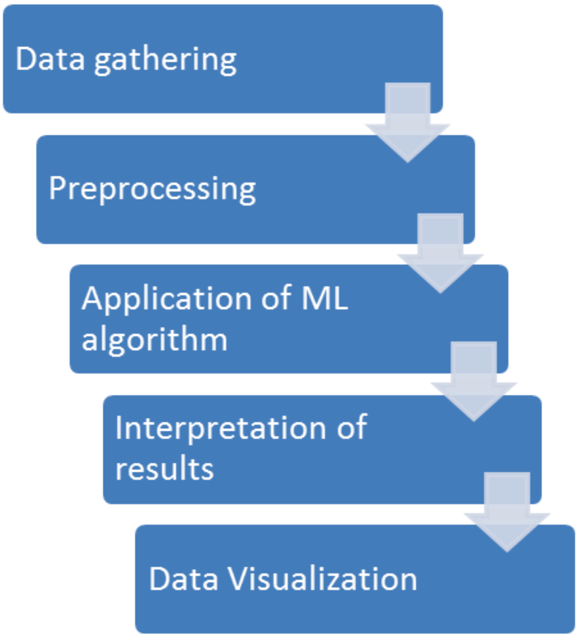


Fig. 1. Process of machine learning in data mining

According to the Fig. 1, the process followed to interpret the data obtained from the online reviews. Its initial phase carries the gathering of data and pre-processing which includes collection of data and removal of outliers. The subsequent phase includes applying Machine Learning algorithm on the new dataset generated from the initial phase. The algorithms used for classifying the overall class from the review parameters are Random Forest Algorithm and Decision Tree. Based on the outcomes generated by the algorithms applied the results are interpreted and visualized for better understanding and decision making. We here check the algorithm that performs better on our dataset. It compares the accuracy of the machine learning algorithms by comparing the already mentioned class of the restaurant in the dataset.

II. Related Work

The literature done so far includes the application of the mining techniques on various types of data sets such as review systems, feedback systems, business applications etc. Ngai, E. *et al.* [3] performed the literature review of the data mining techniques in duration of year 2000 to 2006 casing 24 journals and suggested classification technique to that can be applied on the dataset of gathered articles . Nine hundred articles were distinguished and reviewed for their straight appropriateness in data mining techniques to CRM. Eighty-seven articles were subsequently selected, classified and reviewed . Total 87 selected papers was segregated on four CRM dimensions (Customer Retention, Customer Identification, Customer Development and Customer Attraction) and total seven data mining roles(Regression, Classification, Clustering, Association, Forecasting, Visualization and Sequence Discovery). Further papers were categorized into nine sub-categories of CRM elements.

As per Cui, H. *et al.* [4] had computed fragments for negative and positive extrinsic expressions and its robustness can be crucial in single- or multi- document summarization, document ranking, data mining applications. This paper glanced at a clarified category of the problem: classifying online product reviews into positive and negative classes. Here, we discuss a sequence of experiments with various machine learning algorithms in order to evaluate trade-offs, using relatively 100K product opinions from the web.

Quang He. *et al.* [5] used the regression model for assessing the online reviews that creates a relationship between customer online review and business performance of hotels.

In [6], review reasons were found out on the basis of liking and disliking of product. Label sentences were applied on epinions.com and complaints.com. Feature selection computed 71% F-Score in reason identification task and in reason classification task it was computed to be 61% F-Score.

Machine Learning techniques [7] has been in Cantonese review classification. After data processing, Naïve Bayes classification found efficient than Support Vector Machine (SVM) in terms of comparisons on the dataset. As per the survey of 2000 American Internet Users, it has been found that people were keen to requital 20% more for an excellent rating rather than good rating.

Li *et al.* [8] filtered the comment section on basis of feature selection after analyzing the opinions given by the customers. They constructed a model for affinity propagation to obtain the desired results clustering algorithm were applied to extract the final hot opinions. On comparing the original comments with the resultant accurate mining could be done.

Zomato [9] is one of the big service providers that books the food order online. The main objective of the authors is to classify the restaurants that tie up with the Zomato based on different service parameters/feedbacks provided by the customers.This may help both (i) the restaurants to improve the services. (ii) the zomato application to prioritize the restaurants to serve the customers.

III. Data Collection

The dataset has been extracted from UCI Machine Learning Repository [10] which is a compendium of databases, domain theories, and data generators which is frequently used by the community of machine learning for pragmatic scrutiny of algorithms used for machine learning. We have used data for restaurant reviews of www.zomato.com. Zomato is one of the leading search applications for restaurants. It provides the menu, average cost for two, allows booking of table, offers delivery, price range and aggregate rating. The dataset contains reviews of restaurant collected from the customers. For our research, the data consists of rating from over 8500 restaurants from all over India.

IV. EXPERIMENT

The experiment has been conducted over the two most commonly used machine-learning algorithms namely: Decision Tree and Random Forest Algorithm [11].

A. DECISION TREE

Decision Tree algorithm is one of the most widespread used classification technique due to its human-friendly approach and close correspondence to the real time trees. The decisions are taken based on the traversal of tree depending upon the various parameters. The first node of the tree is known as the “root” node. The tree moves downwards, walking through the “internal” nodes to the “leaf” node. The leaf nodes are the ones which do not have any downward successor. The decision of placement of nodes is taken on the basis of some mathematical calculations of “Information Gain”. Information gain is based on the Entropy that bequeaths the capability of consistency of an attribute. The higher the information gain values, the more chances of an attribute of it becoming the root node. The data is split further and subsequently the information gain is calculated for each attribute. The decision tree uses the supervised approach to train the model. One of major flaw in this methodology is its inconsistency which leads change to change in whole structure of tree on occurrence of a small change in the data. Each node of the tree depicts an attribute and the branches are available options to choose from as shown in Fig. 2.

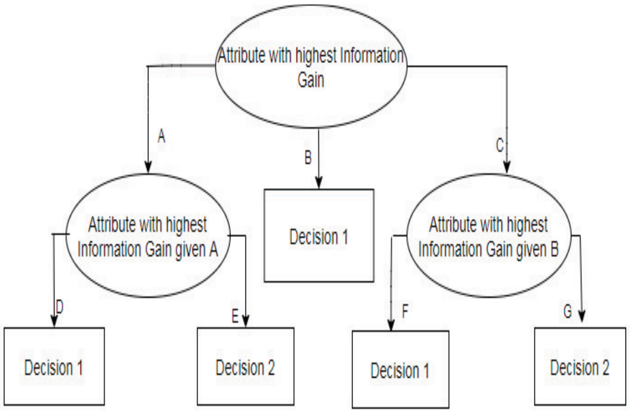

Attribute with highest Information Gain
A
B
C
Attribute with highest Information Gain given A
Decision 1
Attribute with highest Information Gain given B
D
E
F
G
Decision 1
Decision 2
Decision 1
Decision 2

Fig. 2. Decision Tree Structure

B. RANDOM FOREST

Random Forest Algorithm is a type of decision algorithm that can be used for intensive calculations. This method is though slow, but it is quite effective algorithm which can be used for estimating variable values. In Machine Learning, it comes under supervised learning algorithm. It can be used for both types of classification as well as regression techniques. As the name suggests, its task is to create forest containing number of decision trees. More the number of trees in the forest indicate higher accuracy of the classification. The result of the prediction is directly dependent on the randomness of the trees. Trees are extended very deep in order to learn more about the irregular patterns which can be used to handle outlier detection, missing values as well as categorical values. Here the splitting of node is based on searching the most important feature. It is based on random forest regression. This algorithm will first create pseudo code and then it will predict the values based on it by creating random classifier. In this node are selected based on available 'm' features and choosing 'k' features among them using this to find the root node and further used for splitting nodes. It selects some attributes and performs the splitting of data to fetch the required feature. It can even handle unsupervised, variable and unlabeled data. It averages the high bias and high variance and gives balance between them.

In this algorithm, basic idea is to merge the weak learners to form strong learners to utilize them efficiently and accurately. It can be to avoid over fit problem of data, for handling missing values used to identify the important features of training dataset. With the available features it can be used to train the dataset as shown in Fig 3.

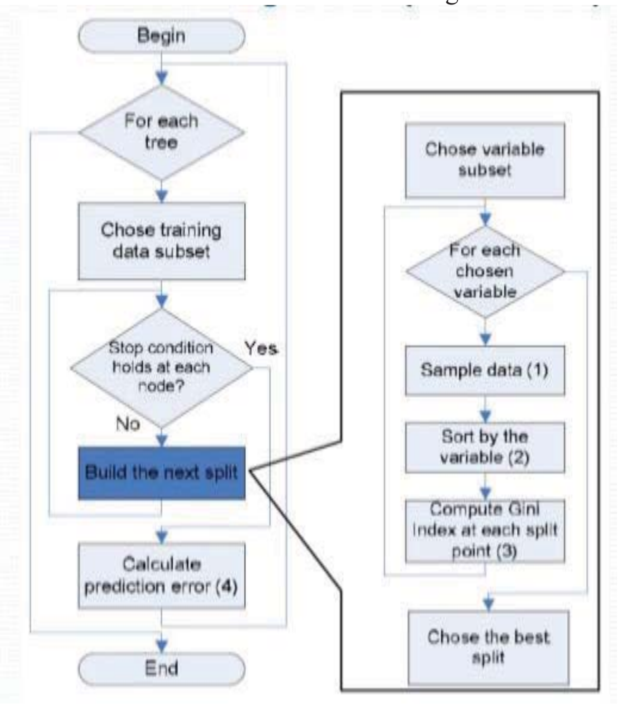


Fig. 3. Random Forest Algorithm

C. CROSS VALIDATION

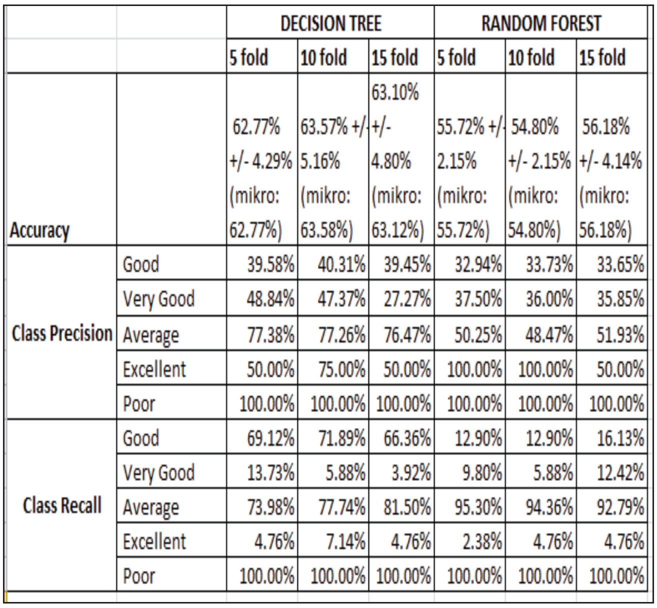
The approach of cross validation is a mathematical methodology that is used to split the dataset into several partitions and train and test the data by applying several iterations over it. The working of the method is by training the machine learning algorithm with a new data every time and tests it with the remaining dataset.

It is popularly referred as k-Cross Validation where in the value of k holds utmost importance. The data is divided into k partitions and then multiple iterations occur. The primary usage of the cross-validation methodology is in the field of machine learning that is used to approximate the competence of a model when applied over unrecognized dataset. The technique follows a simple rule to predict the efficiency of the model by applying it over limited data which was not used to train the model. The method gained popularity due to its non-biased output regarding the model's efficiency and it is also uncomplicated to follow. Hence this methodology is preferred over the classic training and testing of data.

D. RESULTS

Using the above discussed algorithms the accuracy received through 5 fold, 10 fold and 15 fold cross validation is obtained as shown in TABLE I below. This simply shows that the efficiency of the decision tree is higher than that of the random forest for prediction of the class of restaurant.

TABLE I. RESULTS OF CLASSIFIERS



		DECISION TREE			RANDOM FOREST		
		5 fold	10 fold	15 fold	5 fold	10 fold	15 fold
Accuracy		62.77% +/- 4.29% (mikro: 62.77%)	63.57% +/- 5.16% (mikro: 63.58%)	63.10% +/- 4.80% (mikro: 63.12%)	55.72% +/- 2.15% (mikro: 55.72%)	54.80% +/- 2.15% (mikro: 54.80%)	56.18% +/- 4.14% (mikro: 56.18%)
Class Precision	Good	39.58%	40.31%	39.45%	32.94%	33.73%	33.65%
	Very Good	48.84%	47.37%	27.27%	37.50%	36.00%	35.85%
	Average	77.38%	77.26%	76.47%	50.25%	48.47%	51.93%
	Excellent	50.00%	75.00%	50.00%	100.00%	100.00%	50.00%
	Poor	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Class Recall	Good	69.12%	71.89%	66.36%	12.90%	12.90%	16.13%
	Very Good	13.73%	5.88%	3.92%	9.80%	5.88%	12.42%
	Average	73.98%	77.74%	81.50%	95.30%	94.36%	92.79%
	Excellent	4.76%	7.14%	4.76%	2.38%	4.76%	4.76%
	Poor	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

The accuracy of the decision tree algorithm on performing 5,10and 15 fold cross validation is 62.77%, 63.57% and 63.10% respectively whereas in case of the Random Forest Algorithm the accuracy is substantially lower than in case of Decision Tree. Since the random forest generates multiple decision trees as a part of execution hence it is more expensive as compared to the decision tree. Another drawback it suffers is over fitting due to creation of multiple decision trees. Since as per our dataset the number of decision trees is more than required. Hence decision tree outperforms the random forest.

REFERENCES

[1] I. K. C. U. Perera and H. A. Caldera, "Aspect based opinion mining on restaurant reviews," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, 2017, pp. 542-546. doi: 10.1109/CIAPP.2017.8167276

[2] https://www.import.io/post/data-mining-machine-learning-difference/

[3] Ngai, E. W., Xiu, L., &amp; Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), 2592-2602.

[4] Cui, H., Mittal, V., &amp; Datar, M. (2006, July). Comparative experiments on sentiment classification for online product reviews. In AAAI (Vol. 6, pp. 1265-1270).

[5] Cao, Q. V., Burkhart, H. E., & Max, T. A. (1980). Evaluation of two methods for cubic-volume prediction of loblolly pine to any merchantable limit. *Forest Science*, *26*(1), 71-80.

[6] Soo-Min Kim and Eduard Hovy. 2006. Automatic identification of pro and con reasons in online reviews. In Proceedings of the COLING/ACL on Main conference poster sessions (COLING-ACL '06). Association for Computational Linguistics, Stroudsburg, PA, USA, 483-490.

[7] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li, Sentiment classification of Internet restaurant reviews written in Cantonese, Expert Systems with Applications, Volume 38, Issue 6, 2011, Pages 7674-7682,ISSN 0957-4174, doi.org/10.1016/j.eswa.2010.12.147.

[8] H. Li, Q. Peng and X. Guan, "Sentence level opinion mining of hotel comments," 2016 IEEE International Conference on Information and Automation (ICIA), Ningbo, 2016, pp. 2065-2070. doi: 10.1109/ICInfA.2016.7832160

[9] https://www.zomato.com

[10] https://archive.ics.uci.edu/ml/index.php

[11] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood, Random Forests and Decision Trees, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012