Journal of Engineering and Applied Sciences 13 (3): 552-556, 2018
ISSN: 1816-949X
© Medwell Journals, 2018

Boosting Decision Trees for Prediction of Market Trends

[1]Sharmishta Desai and [2]S.T. Patil
[1]Department of Computer Engineering, Pune University, Pune, India
[2]Vellore Institute of Technology (VIT), Pune, India

Abstract: Usage of social sites like Facebook, Twitter is increasing rapidly. People are using these sites for getting feedback about any product or service. Social data is the best data which business analyst can use for getting analysis results. From social data the investor can predict in which products the users are more interested in or what changes the users want in service. The social data analysis will definitely increase the sale or profit gained by the investor. The use of machine learning algorithms for analysing market data will add more knowledge into the knowledge of investor. In this study, we have proposed a method for analysing market data collected from social sites. We have shown the behaviour of different machine learning algorithms against market data. It is found that AdaBoost algorithm with one level decision trees perform best against market data. We have used AdaBoost to boost the performance of decision trees and proved with results. Different phases of social media data mining are also explained in detail.

Key words: Decision trees, random decision trees, AdaBoost, social media data, phases, business analyst

INTRODUCTION

Investors are always worried about their products sale and profit. Every time it is not possible to get feedback about the product or service from individual customer. As usage of social sites is growing rapidly, this data can be used for getting feedback about the product from bulk users.

Prediction and classification of social network data is new area of research. It will help to gather business intelligence information from social media data (Tang *et al*., 2015). This study will help to enhance the knowledge of machine learning domain. It will add values to existing machine learning algorithms which will work efficiently for social network data (Global Pulse, 2012).

The predictive analysis of social network data will help business analytics to understand market trends, understand customer behaviour and take feedback on different products and services (Cho *et al*., 2011). As usage of social network websites like Facebook, Twitter is growing rapidly, this data will help different analyst to do analysis (Bawa, 2011).

Literature review: In literature many researchers have tried to exploit the machine learning algorithms for different structured and unstructured data. There is lot of work available on social media data analysis. Bichen has used neural network to analyse behaviours of customers using social media data set. Desai and Patil (2014) has explained a way to find link between to users social media like Twitter or Facebook using machine learning algorithm. Chanchal *et al*. (2013), Bakshi (2012) and Global Pulse (2012) have explained big data architecture, challenges, etc. Characteristics of social activities and patterns of communication in Twitter are studied by Naaman *et al*. (2010). Davidov *et al*. (2010) have used hash tags and other sentiment labels for sentiment analysis. An effective and efficient followee recommender system built by Hannon *et al*. (2010). Methods to recommend influential users proposed by Kwak *et al*. (2010). Twitter use within and across organizations and geographic markets comparison is proposed by Burton and Soboleva (2011). Kim and Tran (2011), explained how to maximize the outcomes of SMM through Word-of-Mouth (WOM) marketing by identifying the core group of users. Liyang *et al*. (2014) recommendation system based on ranking Support Vector Machine (SVM) is explained. Researcher has shown SVM performs better than neural network.

MATERIALS AND METHODS

As social media data is unstructured, it contains large amount of noisy data (Tang and Liu, 2011). So, for applying machine learning techniques accurately, our research proposal is using efficient pre-processing and cleaning technique. In our proposed method, we have divided whole classification process into different following parts (Fig. 1):

- Data pre-processing and cleaning
- Feature selection
- Thresholding

Corresponding Author: Sharmishta Desai, Department of Computer Engineering, Pune University, Pune, India

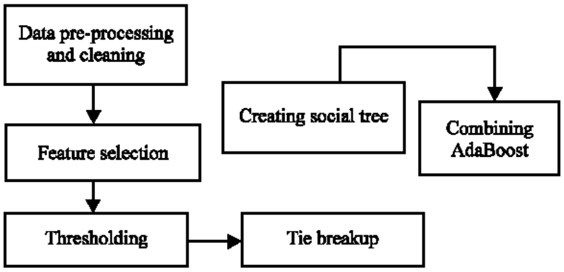

Fig. 1: Proposed architecture for social media data classification

- Creating social tree
- Tie breakup
- Combining AdaBoost with one level decision trees

Data pre-processing and cleaning: When data is collected from social media, there is a high possibility of noisy or missing data due to loss of connection. Noisy data cause model size explosion or over fitting of the model.

In social media, there are some users active for whole duration while some users are active for small duration. This information does not give us meaningful information for developing a model. Sometimes there are some unobserved links available. All this data is used for data cleaning, so that, it would not waste our training time.

Biased data also create a problem in data classification. It is imbalanced class distribution. To do proper data pre-processing and cleaning, following algorithm is proposed:

- Input: -x1, x2, ……… (Non ending stream)
- Output: -x1, x2, ……… (Processed stream)

Algorithm 1:

```
For all attributes Xi∈{X1, X2, ……………}
For all Xi∈{' '}
Calculate:
    $\dot{X} = \frac{\sum_{i=1}^{n} Xi}{n}$
Calculate:
    $S2 = \frac{\sum_{i=1}^{n} Xi\text{-}Xavg}{n\text{-}1}$
If S2>90% then remove (Xi)
For all Xi∈{leaf nodes}
Fnb: arg max r = {n'i, j1, n'i, j, 2, ….., n'i, j, r}
n'i, j, r = P(X|cf). P(cf)/P(X)
```

Feature selection: Proper selection and construction of features is a critical task. It affects the result of machine learning algorithm execution. Features are evaluated based on their information gain. The feature or attribute having larger information gain considered as a node in decision tree. If several features are representing same information then some other features are combined or some are deleted from this. Every feature is ranked according to its information gain. Features are selected in a subset. The number of attributes in one subset is defined by thresholding:

$$H(X) = \sum_{i=0}^{n} p(x_i)\log(p(x_i)) \quad (1)$$

Info_Gain(X) = H(X1)-H(X2) where H(X1)-Entropy before split and H(X2)-Entropy after split. The thresholding method is explained.

Thresholding: Thresholding is required to limit features into subset. Thresholding will set the number of features in a data set which are sufficient for finding information gain on that set. For finding threshold there is no thumb rule. Based on trial and error in learning phase threshold value can be calculated.

Tie breakup: There are some attributes for which there is no difference in information gain in two attributes. Such attributes are called as tie attributes. When such attributes occur then one out of two is selected.

Creating social tree: Convert the social media data into social graphs. Pre-process raw data and social graphs so that they become suitable for applying ML algorithms. Read social media data into leaf node. Calculate information gain on attributes using following Eq. 2:

$$H(X) = \sum_{i=0}^{n} p(x_i)\log(p(x_i)) \quad (2)$$

Info_Gain(X) = H(X1)-H(X2) where H(X1)-Entropy before split and H(X2)-Entropy after split. Select the attribute with highest gain using Hoeffding Bound (HB) criteria. H(Xa)-H(Xb)>HB then select attribute Xa or select Xb. HB is calculated using following Eq. 3:

$$HB = \sqrt{\frac{R^2 \ln\left(\frac{1}{\delta}\right)}{2n}} \quad (3)$$

Split the node and add two leaf nodes. Xa→left = for all attributes Xi, H (Xi)<H (Xa). Repeat above steps till $X \neq \theta$.

Combining AdaBoost with one level decision trees: AdaBoost is an algorithms used for boosting weak classifiers. It can be combined with any other classifier to improve its performance. One level decision tree is a

Table 1: Dataset for one product (P)

User ID	Price	Quality	Performance	Delivery	Buying preference
1	Costly	Good	Good	In time	Yes
2	Cheaper	Average	Good	Delayed	No
3	Affordable	Average	Average	Delayed	No
4	Cheaper	Good	Average	In time	Yes
5	Costly	Average	Average	Delayed	No
6	Costly	Average	Good	Delayed	No

Table 2: Dataset is used for tree model formation

User ID	Price	Quality	Performance	Delivery	Buying preference
1	1	3	3	3	Yes
2	3	2	3	1	No
3	2	2	2	1	No
4	3	3	2	3	Yes
5	1	2	2	1	No
6	1	2	3	1	No

decision tree with one internal node which is immediately connected with leaf nodes. These trees are also called as one level trees or decision stump trees. These trees make prediction based on value of single input feature. Based on values of input features variations can be done. For example, if input feature is continuous then based on threshold, stump can be formed. Value less than threshold will be one leave and greater than threshold will be another leaf. AdaBoost is based on weighted sum of output of other classifiers. It is represented as given:

$$F(x)=\sum_{t=1}^{T}f_t(x) \tag{4}$$

Social network datasets: There are many social network datasets are available like Brightkite, Gowalla, Twitter, etc. We have used Twitter dataset for our experimental research because it is the one of the commonly used social site. Its information is given.

Twitter: Twitter is a social news website. It can be viewed as a hybrid of email, instant messaging and SMS messaging all rolled into one neat and simple package. It's a new and easy way to discover the latest news related to subjects you care about.

In the study, we have used market dataset that is comments given by users about the product. We have written one python script to extract one week data from Twitter. Then, this data filtered according our requirement like we want only specific product related comments, time data as well as location data. The example of data set for one Product P is given in Table 1. These are comments of users about product P1.

This text dataset is converted into numeric by applying weight to each text comment. This dataset is used for tree model formation (Table 2).

RESULTS AND DISCUSSION

Experimental results are generated using Weka 3.6 on windows platform. The market dataset extracted from Twitter is used for experimentation. Number of iterations are 10 and threshold value set is 100. For market data set decision tree, NaivesBayes and AdaBoost with one level decision tree algorithms are evaluated. Following steps are followed to extract Twitter data.

Create a script having a unique id or Twitter account to extract data. Use Twitter API to extract the tweets, the extracted tweets are the tweets of the current day. We can also narrow our search down to extract tweets related to a particular domain.

Once, the data has been extracted, it is organized based on required functions in a database. Social parameters tell us about the closeness between two users, i.e., user affinity and the influence of one user over other, i.e., user influence. Social parameters are instantiated by creating a graph of social network and observing the interactions between nodes.

Location parameters are based on the current location of the user which is narrowed down by time zones and the places where the user checks-in. Time parameters are just the regency of a mention.

Machine learning algorithm is executed on above created dataset. Once, the classification is achieved its trained offline as well as online. The complete classification would be able to suggest mentions to a publisher for targeted advertising. The result of different machine learning algorithms on market data set is shown in Table 3.

In Fig. 2, we can observe AdaBoost with decision trees classify data more accurately. AdaBoost is combined with one level decision trees, so, the performance of decision trees has been improved.

Table 3: Comparison of different machine learning algorithms performance

Algorithm names	Correctly classified	Incorrectly classified	F-measure	ROC	Time (sec)
Decision tree C4.5	2948	1679	0.496	0.499	0.03
Decision tree CART	2948	1679	0.496	0.499	0.49
Random forest	2948	1679	0.496	0.500	0.45
AdaBoost with one level decision trees	3464	1163	0.742	0.795	0.33
NaiveBayes	2948	1679	0.499	0.496	0.02
SVM	2948	1679	0.496	0.500	0.19

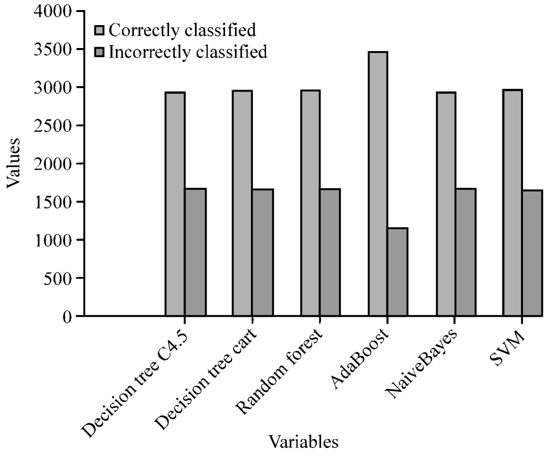

Fig. 2: Comparison of attribute classification

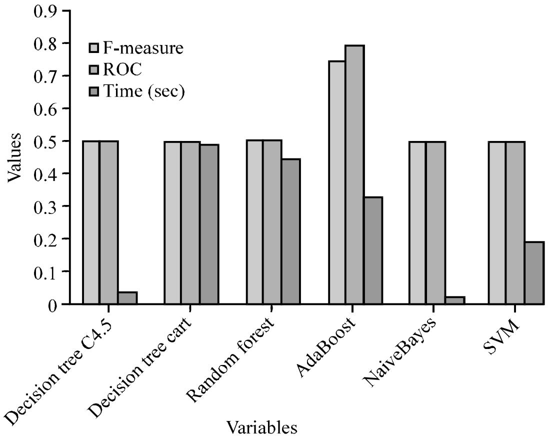

Fig. 3: Comparison of performance of ML algorithms

In Fig. 3, we can see ROC and F-measure for AdaBoost is greater than other algorithms. More is the ROC and F-measure, more accurate is the algorithm. Time required for AdaBoost is less in comparison with other decision tree algorithm but it is more in comparison with NaiveBayes and SVM (Support Vector Machines).

CONCLUSION

As data is growing rapidly day by day due to wide usage of social media, this social data will help business analyst as well as researchers to get the feedback about any service or product. Machine learning algorithms are very much useful for doing this analysis. In this study, we have collected data from Twitter by executing python script. Then, we have filtered it according to our requirement. We have executed different Machine Learning (ML) algorithms like Naive Bayes, Support Vector Machine (SVM) onto it. It is found that decision tree algorithm's performance can be boosted by combining it with AdaBoost. Accuracy of AdaBoost with decision trees is more than other ML algorithms. Also, this algorithm takes less execution time as compared to other decision tree algorithms.

ACKNOWLEDGEMENTS

We would like to thank MIT College of Engineering, Pune, India for providing infrastructure to do this research. Also, we would like to thank Savitribai Phule Pune University, Pune, India for sponsoring our research topic under BCUD research grant scheme.

REFERENCES

Bakshi, K., 2012. Considerations for big data: Architecture and approach. Proceedings of the IEEE Conference on Aerospace Conference, March 3-10, 2012, IEEE, Herndon, Virginia, ISBN:978-1-4577-0556-4, pp: 1-7.

Bawa, C.A., 2011. Sensing the urban: Using location-based social network data in urban analysis. Proceedings of the 1st Workshop on Pervasive Urban Applications, June 12-15, 2011, University of San Francisco, San Francisco, California, pp:1-7.

Burton, S. and A. Soboleva, 2011. Interactive or reactive? Marketing with Twitter. J. Consum. Marketing, 28: 491-499.

Chanchal, Y., W. Shuliang and K. Manoj, 2013. Algorithm and approaches to handle large data-a survey. Int. J. Comput. Sci. Netw., Vol. 2,

Cho, E., S.A. Myers and J. Leskovec, 2011. Friendship and mobility: User movement in location-based social networks. Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining, August 21-24, 2011, ACM, San Diego, California, ISBN:978-1-4503-0813-7, pp: 1082-1090.

Davidov, D., O. Tsur and A. Rappoport, 2010. Enhanced sentiment learning using Twitter hashtags and smileys. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, August 23-27, 2010, ACM, Beijing, China, pp: 241-2491.

Desai, S. and S.T. Patil, 2014. Differential evolution algorithm with support vector machine to classify objects efficiently. Int. J. Adv. Res. Comput. Sci Manage. Stud., 2: 71-74.

Global Pulse, 2012. Big data for development: Challenges and opportunities. White Paper, Global Pulse, New York, May 2012.

Hannon, J., M. Bennett and B. Smyth, 2010. Recommending Twitter users to follow using content and collaborative filtering approaches. Proceedings of the 4th ACM Conference on Recommender Systems, September 26-30, 2010, Barcelona, Spain, pp: 199-206.

Kim, Y.S. and V.L. Tran, 2011. Selecting core target users for online social networking marketing with target marketing: A preliminary report. Proceedings of the 17th Americas Conference on Information Systems (AMCIS), August 4-7, 2011, University of Detroit Mercy, Detroit, Michigan, pp: 1-9.

Kwak, H., C. Lee, H. Park and S. Moon, 2010. What is Twitter, a social network or a news media? Proceedings of the 19th International Conference on World Wide Web, April 26-30, 2010, Raleigh, NC., USA., pp: 591-600.

Naaman, M., J. Boase and C.H. Lai, 2010. Is it really about me? Message content in social awareness streams. Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, February 06-10, 2010, ACM, Savannah, Georgia, ISBN: 978-1-60558-795-0, pp: 189-192.

Tang, L. and H. Liu, 2011. Leveraging social media networks for classification. Data Min. Knowl. Discovery, 23: 447-478.

Tang, L., Z. Ni, H. Xiong and H. Zhu, 2015. Locating targets through mention in Twitter. World Wide Web, 18: 1019-1049.