Decision Tree, Naïve Bayes and Support Vector Machine Applying on Social Media Usage in NYC / Comparative Analysis

Ahmed Burhan Mohammed
Collage of Arts , Kirkuk University , Kirkuk , Iraq

Abstract

Data mining and classification are most research idea that used in many topics by researchers. This study presents the comparison of three algorithms for classifications such as (Decision Tree, Naïve Bayes and Support Vector Machine), applying for social media usage dataset by NYC, to get the best result of the classification algorithm that can classify the instances according to the platforms. The final result of this research refer to the Support Vector Machine returned the best result among these techniques.

Keywords: Social Network, Weka, Decision Tree, Support Vector Machine, Naïve Bayes, Machine Learning, Classification, Recall and Precision

Introduction

Data Mining (DM) is the upcoming research area to solve various problems, its applications are used in different areas such as social media, marketing, banking, health care, insurance and medicine. There are different research fields such as web mining, text mining, image mining, sequence mining, etc.

The frequently applied data mining technique is the classification. That is, a method of data analysis which obtains models to describe and separate data classes and concepts [1]. Classification builds model by using training data and uses model to test data to estimate the accuracy of the classification. The algorithm analysis is the input and it generates a prediction.

The social media sites have become part of our daily life and have become wide spread. A brief history of social network: the first social network launched in 1997 and in 1999 the platforms launched to allow people to answer messages and publish photos and videos and invite others to join them as friends. the first and largest site for social networking a Facebook launched in 2004. In 2006 tweets appeared on the site Twitter. Today many private social sites and networking applications deployed on mobile devices make it easier to use and upload data.

Literature Review

Milan Kumari, Sunila Godara [1]: comparing the output result of four classification algorithms of data mining (RIPPER, SVM, Decision Tree, and ANN). Next, according to the sensitivity, specificity and accuracy of the result of applying classification methods in Cardiovascular Disease, dataset SVM model turned out to be best classifier algorithm for cardiovascular disease prediction.

Rohit Arora, Suman [2]: introduce to apply two classification algorithms MLP and J48 on five different datasets with less than 1000 instances. Then, according to TPR, FPR, Precision, Recall, F-measure and ROC Area. Finally, found that Multilayer Perceptron is the better algorithm in most of the cases after computing the result of each dataset.

S. Vijayarani, M. Muthulakshmi [3]: present the comparison of two classification algorithms, Bayes and Laz. Which are applied to dataset of files stored in hard disk. Finally, the lazy classifier's IBK classification technique has yielded better result than other techniques for this dataset.

Tina. R. Patil, S. S. Sherekar [4]: in this paper two classification methods applied on bank dataset which are Naïve Bayes and Decision Tree. Depending on the sensitivity and specificity of the output results of these algorithms. assuming that Naïve Bayes returned best classifier result.

Swasti Singhal, Monika Jena [5]: introduce a WEKA as data mining tool, relating classification and clustering algorithms present in this paper. And how can get the output results in analyzing, applying and virtualize these techniques.

Bhakti Ratnaparkhi, K. Rajeswari, Paritam H. Patil and Suvarna Thube[6]: present three types of data mining tools that can be used for classification, clustering and association rule mining. Then, comparing the results of these tools according to the accuracy note that Weka gets 83%.

R. Nivedha, N. Saram [7]: propose the classification of CART algorithm applied on social media messages dataset of twitter. Next, depending on Precision, Error Rate and Accuracy. Comparing the result with Naïve Bayes algorithm, to sign the CART as best algorithm return high classifier result.

Bahadopreza Ofoghl, Meghan Mann, Karin Verspoor [8]: present two types of classification methods. First, Lexicon-based classification and Machine Learning based classification which are applied on the emotion of the tweets in London. In addition, this paper shows how we can classify the emotion according to different types. Last, found the ML-based classifier achieved.

Material and Methods

A. Dataset

This dataset is taken from http://www.data.gov [9], NYC Social Media Usage dataset we are using in this paper contains Twitter and Facebook and etc… statistics from various NYC agencies and organizations [10]. It contains total 5758 instances of social media pages [11]. It includes 5 attributes and class information as listed below:

Agency *{name of agency}*

Platform *{social media like Facebook, Twitter, YouTube, linked ln etc.}*

URL *{web pages}*
Date Sampled *{date of collection using social media sites}*
Likes/Followers/Visits/Downloads *{numeric of using site for these actions}*

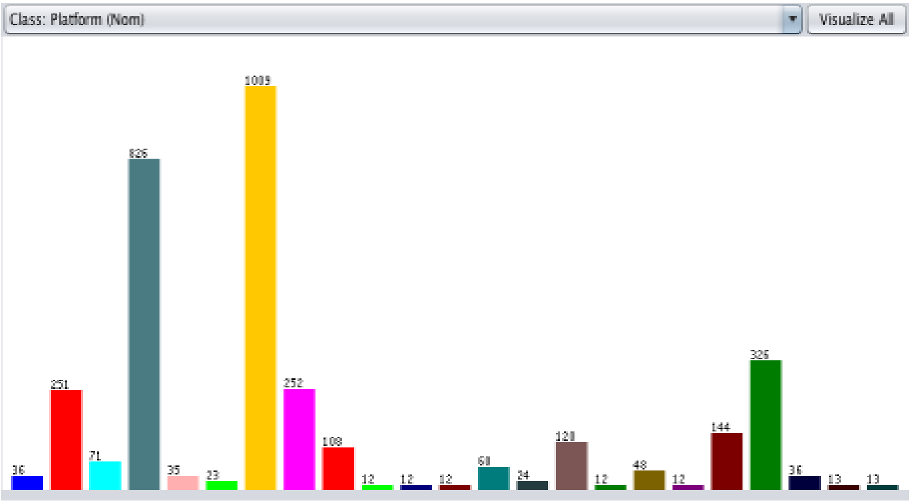

Figure 1: Dataset with platform

B. Decision Tree (DT)
Decision Tree Algorithm, a popular machine learning of the classification techniques, is based upon J.R. Quilan C4.5 [4]. Decision tree creates a binary tree. This technique recursively separates observation in branches to construct a tree for the purpose of improving the prediction accuracy [2]. All data examined will be of the categorical type [1].

C. Naïve Bayes (NB)
Naïve Bayes is a simple probabilistic classifier that evaluates a set of probabilities by calculating the frequency and arrangements of value in a given dataset [3]. NB is used for determining the probability of another element that has already occurred using Bayesian theorem [12].

D. Support Vector Machine (SVM)
Support Vector Machine is a classification technique that seeks to find a hyperplane that partitions the data by their class label and at the same time avoid over-filtering the data [13]. The learning of the hyperplane in linear SVM is done by transforming the problem by using linear algebra [14]. And it makes a binary classification based on separating hyperplane on a remapped instance space.

E. Recall and Precision
The decision made by the classier can be signified in a structure known as a confusion matrix or contingency table [15]. The confusion matrix has four categories:

TN / True Negative: predict negative instance as negative.

TP / True Positive: predict positive instance as positive.

FN / False Negative: predict positive instance as negative.

FP / False Positive: predict negative instance as positive.

Recall or Sensitivity: recall (also known as sensitivity) is the fraction of related cases that are recovered [16]. While precision (called PPV) is the fraction of recovered cases that are related. Both precision and recall are based on a considerate and measure of weight [15].

Classification is the most famous method used in the analysis. It arranges very large data in a world of huge data, which can be difficult to take the decision to analyze this data, so the classification and reliance on several applications of mathematical algorithms and the process make it easier to perform operations of classification of data into small groups making the right decisions of this data analysis more accurate.

That classification process depends on two main steps, the first, is using the training data is created by using the model rules of classification or decision tree, or mathematical formulas. Each object from the data set before the sorting must belong to a class known in advance The second step uses the model to predict the classification of new data in designated test data set or others to know the data. Accuracy is calculated by comparing and analyzing the results of the application of classification using the form on others with knowledge of the data before the data classification and has been reached for classification accuracy, while the accuracy rate is the percentage of samples in the test group, which have been classified correctly by the model. It should be noted that the test data must be independent of the training data in order to be acceptable precision and adopting the model to classify new data and knowledge of others unnamed.

Algorithm: *Steps of applying classification algorithms*

1. Obtain the data from NYC social media usage dataset

2. Convert the dataset format to be available for work on it in WEKA Platform

3. Extract 40% of dataset for training
4. Examine the classification algorithms
5. Build the model DT
6. Examine the model on test dataset 60% Of overall Dataset
7. Get final result of classification
8. Calculate the results of TPR, FPR, Recall and Precision

Experimental Works and Results

We have preformed techniques of classification with three algorithms; Naïve Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM). The dataset used in this paper is NYC Social Media Usage. Now we will explain how to prepare the data to be available for our work design, split the dataset in two parts. First, 40% from the dataset is for training extract. Last, 60% from the dataset is for test.

Additionally, the output of applying the algorithms depends on, using training dataset to create a model, and then applying the model on test dataset to get the output result of the classifications. And the comparison of the results is considered by evaluating the TPR, FPR, Recall and Precision, for each algorithm.

A. Comparisons of correctly and incorrectly classifying instances

The classification algorithms for the data were applied to the test data set to reach the best algorithm among the three algorithms applied for this research to users of social networks according to Table 1 and Fig 2, which are categorized according to the platform of social networking sites such as Facebook, Twitter, Instagram etc. First, when an NB algorithm was applied, 2320 cases were classified correctly and 344 cases were not categorized correctly. Second, when applying the DT algorithm, it was found that the algorithm was correctly assigned to 3349 cases but failed to classify 106 cases correctly. Last, when applying the SVM algorithm, it was found that it correctly classified 3395 and 60 incorrectly cases from the total of 3455 cases being worked on in test data. The SVM algorithm yields the best results of the classification with a few cases that are not properly categorized.

Since the dependence on algebraic equations in the SVM algorithm had the best results from the expectations on which the NB algorithm is based. Also, the SVM is better than the Decision Tree algorithm that depends on the branch.

Table 1: Output Results of Correctly and Incorrectly Instances Classified

	Train NB	Train DT	Train SVM	Test NB	Test DT	Test SVM
No. of correctly instances	1963	2231	2249	2320	3349	3395
Percentage of correctly classified	85.08%	96.76%	97.48%	67.14%	96.93%	98.26%
No. of incorrectly instances	344	76	58	1135	106	60
Percentage of incorrectly classified	14.91%	3.29%	2.51%	32.85%	3.06%	1.73%

According to the results obtained from the applied of algorithms to the social networking users data, as showed in Table 1 we find the highest percentage recorded by the algorithm of SVM is 98.28%. which was able to classify as many as possible data onto the platforms of the affiliated. While the lowest percentage was recorded by the NB algorithm is 67.14% where it showed a significant weakness for the classification of social networking usage data.

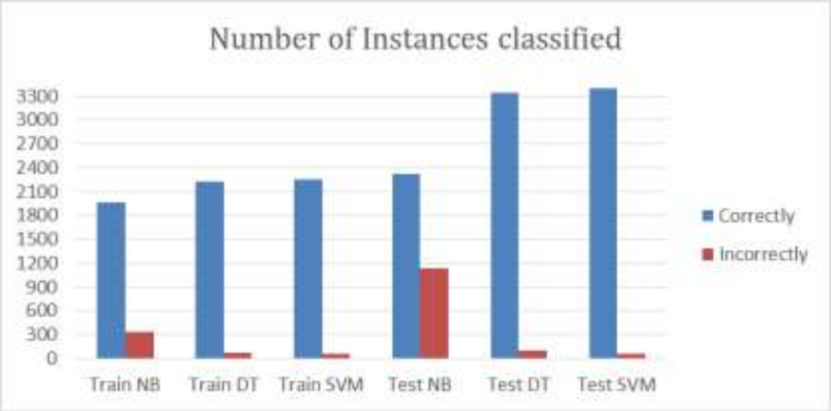

Figure 2: Number of Correctly and Incorrectly Instances Classified

B. Comparisons of TPR and FPR

Depending on the true positive rate values and the false positive rate obtained from the application of the algorithms on the test data set, which is supposed to be the best value for the ratio of the true positive rate 1, and that the best value for the false positivity rate is zero. According to Table 2, it is found that the ratios of the DT algorithm are (TPR = 0.969, FPR = 0.011), the ratios of the NB algorithm are (TPR = 0.671, FPR = 0.013), finally the ratios of the SVM algorithm are (TPR = 0.983, FPR = 0.004). In view of the values mentioned above, we note that SVM algorithm returned the highest ratio close to 1 and the lowest ratio close to 0. While we find a weak NB algorithm to reach ideal ratios or to be close to idealism based on a true positive rate and a false positive rate.

Table 2: Output Results of TPR and FPR

	TP Rate	FP Rate
Train NB	0.851	0.027
Train DT	0.967	0.012
Train SVM	0.975	0.007
Test NB	0.671	0.013
Test DT	0.969	0.011
Test SVM	0.983	0.004

For the purposes of understanding the results more clearly, the diagram shown in Fig 3. shows us the results obtained from applying for a true positive rate equation and false positive rate, which shows that the highest value obtained is the value of the SVM algorithm and that the lowest value obtained are for the NB.

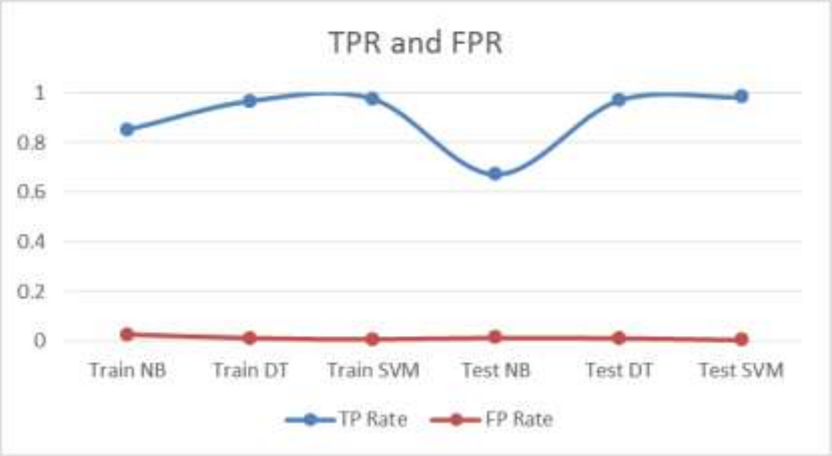

Figure 3: TPR and FPR Output Result

C. Comparisons of Precision and Recall

According to the results obtained by applying the Precision and Recall equation of the three algorithms on the test data set, where the Recall = 1 when the false negative = 0, meaning that 100% of positive integer was detected. And that Precision = 1 when false positive = 0, where no false results. Comparing these defaults with the values we obtained, we find that the SVM algorithm has the highest value of (Precision = 0.986, Recall = 0.983). While the NB algorithm got the lowest value of Precision and less than expected to Recall (Precision = 0.842, Recall = 0.671).

Table 3: Output Results of Precision and Recall

	Precision	Recall
Train NB	0.851	0.851
Train DT	0.971	0.967
Train SVM	0.977	0.975
Test NB	0.842	0.671
Test DT	0.973	0.969
Test SVM	0.986	0.983

Furthermore, for understanding the results more clearly, the diagram shown in Fig 4. shows us the results obtained from applying the algorithms, the results of Precision and Recall shows that the maximum value obtained is the value of the SVM algorithm and that the minimum value obtained are for the NB.

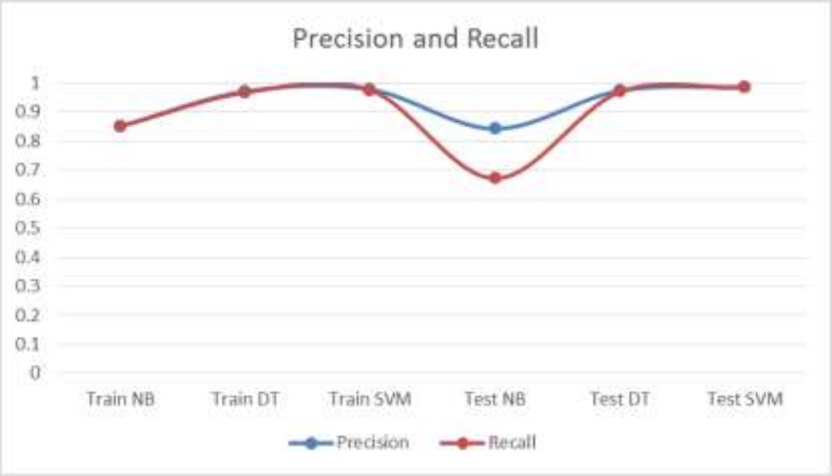

Figure 4: Precision and Recall Output Result

Conclusion

Data mining techniques are commonly used in many fields, especially the classification of data used to access new data and decisions derived from the application of algorithms to large amounts of data. In this paper, three classification algorithms are applied to data onto social media usage, namely Naïve Bayes, Decision Tree and Support Vector Machine.

According to the results obtained from applying for these algorithms, three types of comparisons were made to the values obtained, which worked to classify these data according to their platforms. In the first part of the comparison, it was found that the number of cases that were properly classified is as follows (NB 2320, DT 3349, SVM 3395) indicating that the best result recorded for SVM algorithm and the worst result recorded for the NB algorithm. The algorithms that failed to classify the cases were as follows (NB 1135, DT 106, SVM 60) case of each algorithm. It is clear that the NB algorithm has not been able to classify more than 1,000 cases. In the second part of the comparison of the results of these algorithms were based on the values of the true positive rate and false positive rate. The result of TPR which appeared as follows (NB 0.671, DT 0.969 and SVM 0.983) and for the FPR as follows (NB 0.013, DT 0.0101 and SVM 0.004). Furthermore, this indicates that the best rating results were for the SVM algorithm which reached the highest percentage of the classification. In addition, the results were compared based on the values of the precision and recall, which used for the evaluation of the final results of the algorithms applied to the data. The output result of precision was as follows: Algorithms (NB 0.842, DT 0.973 and SVM 0.986), and the output result of recall was as follows (NB 0.671, DT 0.969 and SVM 0.983).

Finally, this study demonstrated the weakness for the NB algorithm as a result of having obtained the minimum percentage of correct classification cases of the test dataset. Additionally, as shown in Fig 5. that the machine vector support algorithm was the best model used to classify social media data cases.

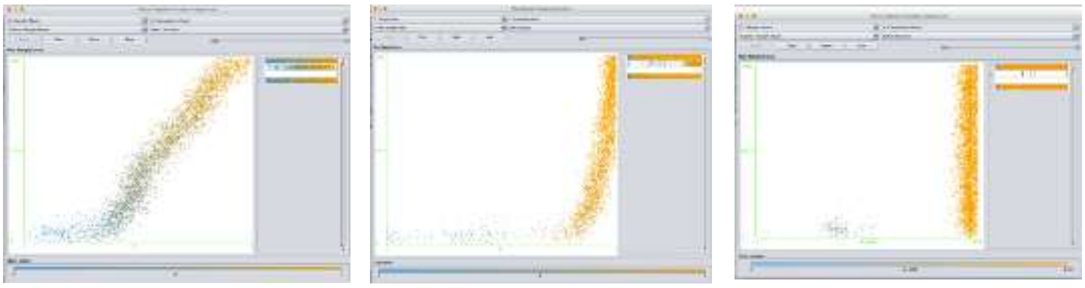

Figure 5: The Output Result Virtualization

In future we intend to improve other type of data mining classification methods and compare them with the results of this study.

References

1. **Milan Kumari, Sunila Godara,** "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", International Journal of Computer Science and Technology, Vol. 2, Issue 2, June 2011.

2. **Rohit Arora Suman,** "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA", International Journal of Computer Applications (0975 – 8887), Vol. 54, No.13, September 2012.

3. **S. Vijayarani, M. Muthulakshmi,** "Comparative Analysis of Bayes and Lazy Classification Algorithms ", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013.

4. **Tina R. Patil, Mrs. S. S. Sherekar,** "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications Vol. 6, No.2, Apr 2013 ISSN: 0974-1011.

5. **Swasti Singhal, Monika Jena,** "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Vol. 2, Issue-6, May 2013.

6. **Pritam H. Patil, Suvarna Thube, Bhakti Ratnaparkhi, K.Rajeswari,** "Analysis of Different Data Mining Tools using Classification, Clustering and Association Rule Mining", International Journal of Computer Applications (0975 – 8887), Vol. 93 , No.8, May 2014

7. **R. Nivedha, N. Sairam,** "A Machine Learning based Classification for Social Media Messages", Indian Journal of Science and Technology, Vol. 8(16), July 2015, ISSN: 0974-6846.

8. **Bahadorreza Ofoghi, Meghan Mann, Karin Verspoor,** "Towards Early Discovery of Salient Health Threats: A Social Media Emotion Classification Technique", Pacific Symposium on Biocomputing 2016.

9. https://catalog.data.gov/dataset/nyc-social-media-usage-555a2, "Social Media Usages Dataset.

10. file:///Users/sewerae/Desktop/social%20media/Introduction%20to%20machine%20learning:%20Classification%20of%20news%20with%20the%20help%20of%20the%20working%20environment%20We.webarchive, "Introduction to Machine Learning:

Classification of NEWS with the Help of the Working Environment WEKA".

11. **Chris Barrows, Eileen Reynolds,** "New York University Social Media Style Guide", Last Edit: Chris Barrows NYU Social Media Team, (Summer 2014).

12. **Deepa S. Deulkar, R. R. Deshmukh,** "Data Mining Classification", Imperial Journal of Interdisciplinary Research (IJIR), Vol-2, Issue-4, 2016, ISSN: 2454-1362.

13. **Jason Brownlee,** "Support Vector Machines for Machine Learning", Machine Learning Algorithms, April 20, 2016.

14. **Ranjini Srinivas,** "Managing Large Data Sets Using Support Vector Machines", A THESIS Presented to the Faculty of the Graduate College, University of Nebraska, master degree 2010.

15. **David M. W. Powers,** "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation", School of Informatics and Engineering, Flinders University Adelaide, Australia, Technical Report SIE-07-001, December 2007.

16. https://en.wikipedia.org/wiki/Precision_and_recall , Recall and Precision.

17. **Han J., Kamber M., Pei J.,** "*Data Mining Concepts and Techniques*", Elsevier, Massachusetts, no. 3, pp. 443-490, 2012.

18. **Mandeep Kaur, Pravneet Kaur,** "A Review on Automatic News Classification using the Probabilistic Classification Algorithms", International Journal of Science and Research (IJSR) ISSN: 2319-7064, August (2015).

19. **Amrita Naika, Lilavati Samantb,** "Correlation review of classification algorithm using data mining tool: WEKA Rapidminer, Tanagra, Orange and Knime", International Conference on Computational Modeling and Security (CMS 2016), Procedia Computer Science 85, 2016.

تحليل ومقارنة لتطبيقات تقنيات التصنيف على مستخدمي وسائل الاعلام الاجتماعية

احمد برهان محمد

كلية الآداب ، جامعة كركوك ، كركوك ، العراق

ahmedlogic79@yahoo.com

الملخص

ان استخراج البيانات وتصنيفها من أكثر الافكار التي تستخدم في العديد من الموضوعات من قبل الباحثين. في هذا البحث نقدم ثلاثة خوارزميات تصنيف البيانات وهي (شجرة القرارات، السذاجة بايز ودعم الة المتجهات) على مستخدمي وسائل الاعلام الاجتماعي في مدينة نيويورك للحصول على أفضل نتيجة من هذه الخوارزميات والتي يمكنها تصنيف الحالات وفقا للمنصات التابعة اليها. النتيجة النهائية لهذا البحث يشير الى ان خوارزمية دعم الة المتجهات اعادت لنا أفضل نتيجة للتصنيف من بين هذه التقنيات.