ISSN: 2277-9655
Impact Factor: 4.116

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

ANALYSIS OF TWITTER DATA WITH MACHINE LEARNING TECHNIQUES

Mudit Rastogi*, Ankur Singh Bist
* Department of Computer Science and Engineering Krishna Institute of Engineering and Technology Ghaziabad, India
Department of Computer Science and Engineering Krishna Institute of Engineering and Technology Ghaziabad, India

DOI: 10.5281/zenodo.57978

ABSTRACT

Classification of data is an important aspect of getting vigorous knowledge and help to analyze and perform any further action. This paper deals with how different Machine-Learning Techniques classify on features of time-windows of Twitter, a micro-blogging social media and to determine whether or not these times-windows are followed by Buzz events. In particular, we compare different machine learning techniques like Naïve Bayes and SVM, to find the accuracy of classification with or without applying dimensional reduction in the number of attributes with the help of PCA algorithms.

KEYWORDS: Support Vector Machine; Accuracy; PCA

INTRODUCTION

Using data provided by UC Irvine Machine Learning Repository, our goal is to apply machine-learning techniques to successfully classify on features of time-windows of Twitter, a micro-blogging social media and to determine whether or not these times-windows are followed by buzz events. In this paper we classify the data set using different approaches and compare the accuracy of different machine learning techniques.

One of the challenges is facing in present time is the number of features for evaluating classifier. Naïve Bayes would be one way to make classification in order to get baseline accuracy of classifier. Apart from this, SVM is another way that would be use for mapping our features to a higher dimensional space. Our objective of this paper was use of Naïve Bayes as an introductory measure followed by SVM to achieve better results. Finally, we applied PCA and find the optimal classifier doing dimensional reduction.

DATA SET

The data we used for our project was provided by the UC Irvine Machine Learning Repository website under the topic Buzz Prediction in Social Media. We were provided with 140707 record samples for our training set and testing set, and their associated label of whether the event is Buzzing or not in that social network site. For each record, we were given 77 attributes (features) of real data types. This dataset was published using the UCI guidelines. Hence, examples were stored using a standard comma separated value (CSV) format. We used less number of records like in hundreds or in thousand to train or test our classifiers.

DATA ANALYSIS

In order to prepare data for training in our Naïve Bayes classifier and SVM algorithms, we generated scientific multidimensional array from the .csv file and used the float data type. For our SVM, also same technique was used to save data in array form. There were 77 attributes that contain real type values entries with no missing values. We were provided two label classes represented by 0 and 1. Here, 0 represented Non-Buzzed Event and 1 represented Buzzed Event.

ISSN: 2277-9655
Impact Factor: 4.116

APPROACH/METHOD

Gaussian Naïve Bayes classification was used as an introductory result to see what is achievable. More sophisticated and advance techniques like SVM in addition to Principal Component Analysis (PCA) were used to see if improvements could be made in the classification test. We experimented with using different feature sets of each method. We also worked with diiferent set of training and testing data.

NAÏVE BAYES METHOD

Naive Bayes methods are supervised learning algorithms based on popular known Bayes' theorem. It is one of the most basic statistical tools in term of classification of data before predicting or applying any condition. For any class variable say, y (here in our data set there are two class, first one is 0 representing NonBuzzed Event and second one is 1 representing Buzzed Event) and a number of many feature vectors say, from $x_{\_1}$ through $x_{\_n}$,(in our dataset there are about 77 features) Bayes' theorem states the following mathematical relationship.

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Using the naive independence assumption for all the features that

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

For all i, this relationship is further simplified to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Since we know that $P(x_1 \ldots x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

Gaussian NB implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

where the parameters σ_y and μ_y are estimated using maximum likelihood which is used as method of estimating the parameters of a statistical model for the given data.

SVM

The Support Vector Machine (SVM) was first proposed by Vapnik is one of the important supervised algorithm that has been best in offering optimal marginal classification. Recent studies and result of the experiment has shown that SVM is highly effective in terms of classification accuracy than the other data classification algorithms. SVM comes into existence to separate the large chunk of the available data with a gap. These gaps separate the data points belonging to a different class. The data points which lie on this gap are the Support Vector Points. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. Support Vector Algorithms work on different parameters which affects the result and the optimal time to achieve it.

We have experimented with different parameter for result in better accuracy. These parameters include different kernel functions, the standard deviation of the Gaussian kernel and the number of training examples.

Mathematical discussion of support vector algorithms is provided taking n features.
Let us assume different data points as:

$\{(x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)\ldots\ldots\ldots,(x_n, y_n)\}$.

And there are two classes for $y_n = 1$ or -1. These data points can be visualized as by means of the separating hyper plane, which can be mathematically represented as

$$w.\ x + b = 0 \qquad (1)$$

Where b is scalar (similar to a bias feature in Regression analysis) and w is n-dimensional Vector.

Factor b restricts solution by avoiding the hyper plane pass through origin all the time. We are focused to get high margin classification and there exists two classes y_n= -1 or 1. So, hyper plane which is parallel for both class share same features and scalar factor b,which are mathematically described as

$$w.\ x + b = 1$$
$$w.\ x + b = -1$$

If the training data are linearly separable, we can select these hyper planes so that there are no points between them and then try to maximize their distance. By geometry, we find the distance between the hyper planes is $2 / |w|$. So we want to minimize $|w|$. To excite data points, we need to ensure that for all I either

$$w.x_i - b \geq 1 \text{ or } w.\ x_i - b \leq -1$$

This can be written as

$$y_i\ (w.\ x_i - b) \geq 1\ ,\ 1 \leq i \leq n \text{ ------(2)}$$

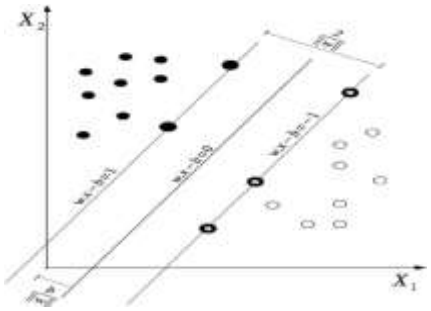

Figure.1 Maximum margin hyper planes for SVM trained with samples from two classes

As discussed earlier Data Points that reside along the hyper planes or decision boundary are called Support Vectors (SVs). A separating hyper plane with the largest margin defined by $M = 2 / |w|$ that is specifies support vectors means training data points closest to it.

$$y_j\ [w^T.\ x_j + b] = 1\ ,\ i = 1 \qquad (3)$$

We will be working for different kernel (parameter) which will influence our testing result and its accuracy.
These kernels are:

- Linear kernel: $K(x_i, x_j) = x_i^T \cdot x_j$.
- Polynomial kernel: $K(x_i, x_j) = (\gamma\ x_i^T\ x_j + r)^d\ ,\ \gamma > 0$
- Radial Basis Function(RBF) kernel: $K(x_i, x_j) = \exp(-\gamma\ \|x_i - x_j\|^2),\ \gamma > 0$

ISSN: 2277-9655
Impact Factor: 4.116

Here, γ, r and d are kernel parameters. In these popular kernel functions,

Today SVM has been employed in a wide range of real world problems used in various engineering application like image recognition. The performance of SVM is very sensitive to how the cost parameter and kernel parameters are set. As a result, the user normally needs to conduct extensive cross validation in order to figure out the optimal parameter setting.

PCA

PCA was invented in 1901 by Karl Pearson as an analogue of the principal axis theorem in mechanics; but later it was developed and named by Harold Hotelling. Our data set have lots of features or large dimension of features. High dimensional data can pose problems for machine learning as predictive models based on such data run the risk of over fitting. Such large number of features may reduce the possibility of getting higher accurate results from our testing data sets. Also, many of the features may be redundant or highly correlated to each other, which can also lead to a low accuracy. So in order to work for higher accuracy, we need to consider the important features that only affect the region of the classifier for various classes. PCA is a classical statistical method of transforming features of dataset into a new set of non-correlated features called Principal Components (PCs). The number of principal components is less than or equal to the number of original variables. PCA can be used to reduce the dimensionality of a data set, while still retaining as much of the variability of the dataset as possible.

EXPERIMENT

The classification experiments were conducted on Buzz in social media. Data Set could be taken from https://archive.ics.uci.edu/ml/datasets/Buzz+in+social+media+ . We used Python language and its tools for experimentation. We employed both methods on a different data set with and without dimensional reduction. First, using Naïve Bayes for getting baseline accuracy followed by SVM with different kernel linear, polynomial, and RBF. The result of different machine learning techniques has been studied and the result is framed on a table with pictorial representations of Machine learning Techniques w.r.t Accuracy at given algorithms.

We get the results after applying the Machine Learning Algorithm to train the classifier using all the 77 attributes which are the 77 dimensions of the multidimensional array of data sets.

RESULT

Table I

Comparison of accuracy of different kernel of SVM without applying dimensional reduction.

S No.	Kernel	Training data	Testing data	Accuracy
1	Linear	1000	1000	0.927
2	RBF	1000	1000	0.958
3	Polynomial(3)	1000	1000	0.92
4	Polynomial(4)	1000	1000	0.923

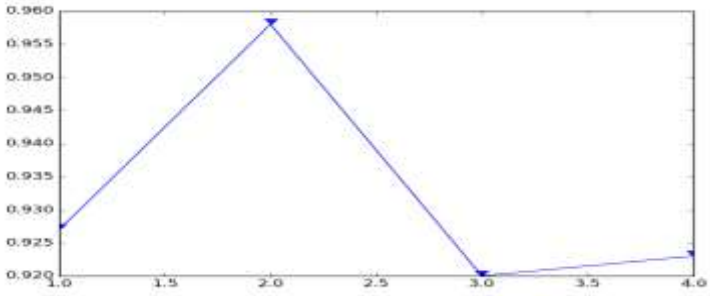

Figure: Comparing accuracy for different kernels

ISSN: 2277-9655
Impact Factor: 4.116

X coordinate represents a machine learning technique, serial number
Y coordinate represents the accuracy of algorithms out of 1.

1: Linear 2: RBF 3: Polynomial (degree 3) 4: Polynomial (4)

Table Ii
Comparison of accuracy of machine learning classifiers without dimensional reduction

S No.	Technique	Training data	Testing data	Accuracy
1	Naïve Bayes	1000	1000	0.569
2	Linear SVM	1000	1000	0.927
3	RBF SVM	1000	1000	0.958
4	Polynomial(4) SVM	1000	1000	0.923

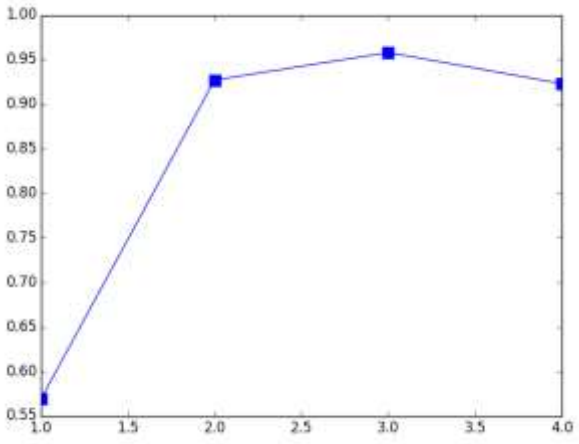

Figure: Comparing accuracy for different machine learning algorithms classifiers

X coordinate represents a machine learning technique, serial number
Y coordinate represents the accuracy of algorithms out of 1.

1: Naive Bayes 2: Linear 3:RBF 4:Polynomial(degree 4)

Now we performed dimensionality reduction in the attributes in order to check whether the accuracy of the classifier increases or not. We reduced the attributes from 77 to 3 attributes and hence removed major redundant features.

Table Iii
Comparison of accuracy of machine learning after applying dimensional reduction.

SNo.	Technique	Training data	Testing data	Accuracy
1	Naïve Bayes	1000	1000	0.883
2	Linear SVM	1000	1000	0.958
3	RBF SVM	1000	1000	0.958

ISSN: 2277-9655
Impact Factor: 4.116

4	Polynomial(degree 3) SVM	1000	1000	0.923

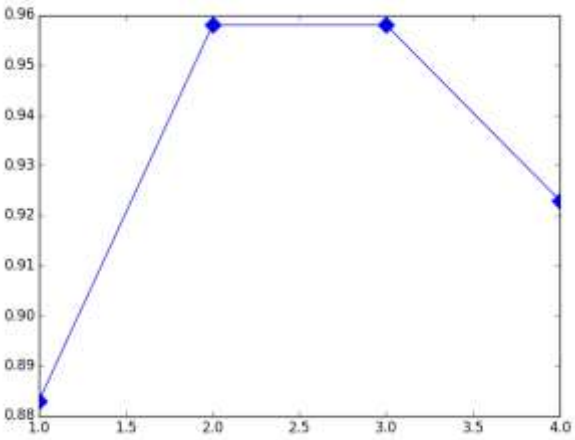

Figure: Comparing accuracy for different machine learning algorithm classifiers after applying PCA

X coordinate represents a machine learning technique, serial number
Y coordinate represents the accuracy of algorithms out of 1.

1: Naive Bayes 2: Linear 3:RBF 4:Polynomial(degree 4)

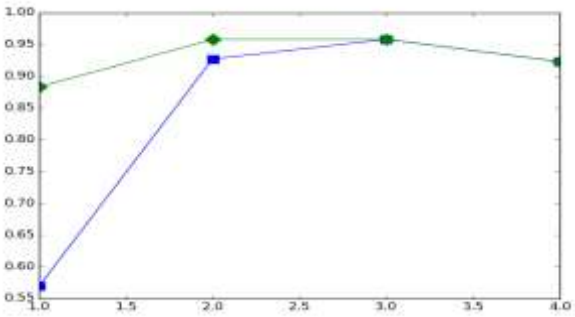

Figure: Comparing accuracy for different machine learning algorithm classifiers before and after applying PCA

X coordinate represents a machine learning technique, serial number
Y coordinate represents the accuracy of algorithms out of 1.

1: Naive Bayes 2: Linear 3: RBF 4: Polynomial (degree 4)

Green Dots represent the accuracy after PCA
Blue Dots represent the accuracy before PCA

INSIGHT KNOWLEDGE

- Among the three methods, Naïve Bayes performed the worst and SVM performed the best which differs by 38.9% (approx..). All the methods have roughly the same performance on our data set, excluding the Naïve Bayes. This is probably because there was one feature that was NOT strongly associated WITH BUZZ EVENT.

- The Naïve Bayes model assumes that all features are independent. which shows that, assuming that features are independent is not necessarily a bad assumption for our problem.
- Table II offers a summary of the achievable accuracy, using Naïve Bayes and SVM. In SVM, we could see that RBF kernel (non-linear) outperforms the Linear SVM and gives better result but we cannot question the optimal result on the basis of large dataset.
- Value of Table III says that after applying dimensional reduction in our datasets we saw increase in accuracy of the classifier. This gives knowledge that, even though we were given many features of in our data, we found that most of the features were not useful in classification.
- The result of the above practical work emphasis that Naïve Bayes classifier is not able to give higher accuracy, but showed vast improvement after the features are transformed through PCA algorithms.

FUTURE WORK

We tested these algorithms using training data and testing data with maximum 1000 records. The experiment can produce more precise or more accurate result when more testing and training data are used and for that system with higher configuration or performing experiments in different clusters will be useful.

REFERENCES

[1] Pandey, Neha, B. K. Singh, and Ankur Singh Bist. "A novel feature learning for image classification using wrapper approach in GA." Signal Processing and Integrated Networks (SPIN), 2015 2nd International Conference on. IEEE, 2015.
[2] Andrew Ng. CS229 Lecture Notes. Stanford University, 2012
[3] Andrew Ng Naïve Bayes Lecture Notes, Stanford University 2012.
[4] Cortes, Corinna; and Vapnik, Vladimir N.; "Support-Vector Networks", Machine Learning, 20, 1995.
[5] Andrew Ng Principal Component Analysis Lecture Notes, Stanford University 2012.
[6] Sharma, Rudranshu, Ankur Singh Bist, and Vikas Kumar. "INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY EXTREME LEARNING MACHINE."
[7] Bist, Ankur Singh. "INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY DETECTION OF COMPUTER VIRUSES USING WELM_ FPSO_FBFO."
[8] Kumar, Amit, Vikas Chauhan, and Ankur Singh Bist. "Role of Artificial Neural Network in Welding Technology: A Survey." International Journal of Computer Applications 67.1 (2013).
[9] Bist, Ankur Singh, and Neha Pandey. "INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY FEATURE SUBSET SELECTION: A REVIEW."
[10] Bist, Ankur Singh. "INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY VIRUS GENERATION KITS: A SURVEY."
[11] Bist, Ankur Singh. "Fuzzy Logic for Computer Virus Detection." IJESRT, ISSN: 2277-9655.
[12] Mishra, Pushplata, and Ankur Singh Bist. "INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY RECENT TRENDS IN FACE DETECTION."
[13] Mehta, Pankaj, et al. "Image Classification using NPR and Comparison of Classification Results."
[14] Das, Purushottam, Shambhu Prasad Sah, and Ankur Singh Bist. "INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY RECENT TRENDS IN XOR PROBLEM."
[15] Kumar, Vikas, Ankur Singh Bist, and Rudranshu Sharma. "INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY OUTLIER DETECTION USING INNER AND OUTER RADIUS BASED METHOD."