Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 127 (2018) 511–520

Procedia Computer Science

www.elsevier.com/locate/procedia

The First International Conference on Intelligent Computing in Data Sciences

Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis

Yassine AL AMRANI[a,*], Mohamed LAZAAR[b], Kamal Eddine EL KADIRI[a]

[a]*Abdelmalek Essaâdi University, LIROSA Laboratory, Tetuan, Morocco*
[b]*Abdelmalek Essaâdi University, New Technology Trends Team, Tetuan, Morocco*

Abstract

Sentiment analysis becomes more popular in the research area. It allocates positive or negative polarity to an entity or items by using different natural language processing tools and also predicted high and low performance of various sentiment classifiers. Our work focuses on the Sentiment analysis resulting from the product reviews using original techniques of text's search. These reviews can be classified as having a positive or negative feeling based on certain aspects in relation to a query based on terms. In this paper, we proposed hybrid approach to identify product reviews offered by Amazon. The results show that the proposed system approach outperforms these individual classifiers in this amazon dataset.

Keywords: Sentiment analysis, classifiers, support vector machine, Random Forest, amazon, accuracy ;

1. Introduction

Classification is the process wherein a class label is assigned to unlabeled data vectors. It can be categorized into supervised and un-supervised classification which is also known as clustering. In supervised classification learning is done with the help of supervisor i.e. learning through example.

In this method, the set of possible class labels is known apriori to the end user. Supervised classification can be subdivided into non-parametric and parametric classification. Parametric classifier method is dependent on the probability distribution of each class. Non-parametric classifiers are used when the density function is unknown.

Recently, multiple platforms are developing very interesting either in terms of volume of data or according to the number of users around the world, they offer users all the possibilities to express their opinions and to exchange

* Corresponding author. Tel.: +212-658-200-678;
E-mail address: alamraniyassine@gmail.com

1877-0509
10.1016/j.procs.2018.01.150

their ideas with the others [1].

The sentiment analysis found in the form of comments, reviews and feedback and provides necessary information for various purposes. These opinions or sentiments can be divided into two categories: positive and negative; or also categories of different rating points (e.g. 3 stars, 4 stars and 5 stars). The polarity of sentiments like “good” and “bad” also identify the sentiments either positive or negative [2].

Sentiment analysis is the part of the text mining that attempts to define the opinions, feelings and attitudes present in a text or a set of text. It is particularly used in marketing to analyse for example the comments of the Net surfers or the comparatives and tests of the bloggers. It requires much more understanding of the language than text analysis and subject classification. Indeed, if the simplest algorithms consider only the statistics of frequency of occurrence of the words, it is usually insufficient to define the dominant opinion in a document. It is the process of determining the contextual polarity of the text, that is, whether a text is positive or negative. The use of this analysis helps researchers and decision-makers better understand opinions and client satisfaction using sentiment classification techniques in order to automatically collect different perspectives on from various platforms. There has been a large amount of research in the area of sentiment classification. Traditionally most of it has focused on classifying larger pieces of text, like reviews (B. Pang, L. Lee, and S. Vaithyanathan., 2002). In this paper, a comparison of two popular classifiers was performed to classify product reviews either positive or negative (Support Vector Machine (SVM), Random Forest) and our approach Random Forest Support Vector Machine (RFSVM).

This paper presents an approach to determine how sentiments can be classified using Support Vector Machine, Random Forest and RFSVM based Hybrid Approach. The paper provides the comparison with other existing technique, shows that the use of hybrid approach can improve the efficiency of sentiment analysis. The proposed hybrid approach gives better result as compare to the existing technique. The rest of the paper is described as follows: Section 2 describe sentiment analysis system. Section 3 introduces applied algorithms in this field. Section 4 discusses proposed methods. Section 5 explain the results and analysis obtained. Section 6 presents the conclusion and future work for the proposed work.

2. Sentiment Analysis System

Sentiment Analysis is the process of finding the opinion of user about some topic or the text in consideration [3], it determines whether a piece of writing is positive or negative.

The various challenges in sentiment analysis is one that the public don’t always express sentiments in same way means some express in the form of ratings and some in the form of comments and second involving sentences that don’t express any sentiment. The sentiment analysis process is shown in figure 1. The text preparation step performs required text pre-processing and cleaning on the dataset which including removal of stop words. Sentiment identification step determines the sentiment of people expressed in the text and analyzes it. Finally, sentiment classification is conducted to get the results [4].

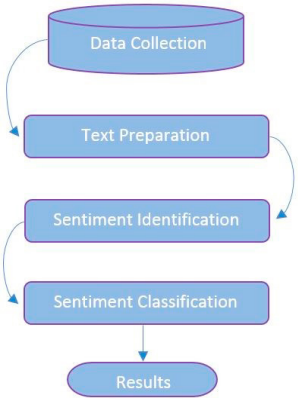


Fig. 1. Sentiment analysis model

Recently, sentiment analysis has attracted an increasing interest. It is a hard challenge for language technologies, and achieving good results is much more difficult than some people think. The task of automatically classifying a text written in a natural language into a positive or negative feeling, opinion or subjectivity (Pang and Lee, 2008), is sometimes so complicated that even different human annotators disagree on the classification to be assigned to a given text [5]. Personal interpretation by an individual is different from others, and this is also affected by cultural factors and each person's experience. And the shorter the text, and the worse written, the more difficult the task becomes, as in the case of messages on social networks [6].

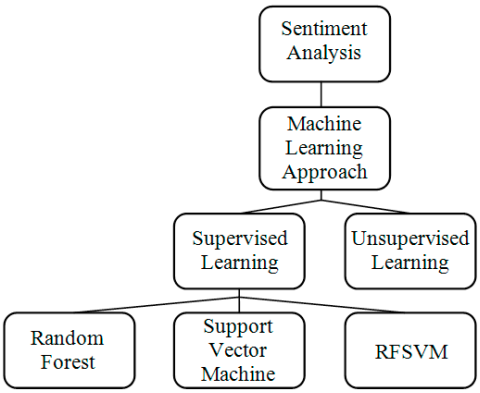


Fig. 2. Classification of sentiment analysis

In feature extraction, a sentence or document is broken into words to build up the feature matrix. In the matrix, each sentence or document is a row and each word form a feature as a column, and the value is the frequency count of the word in the sentence or document. Feature matrix is then passed to each classifier and their performance is evaluated [7]. In this work, we have studied the classification of sentiment using two popular algorithms, namely Support Vector Machine, and Random Forest.

Text classification play an important role in many applications, it assigns one or more classes to a document according to their content. Classes are selected from a previously established taxonomy (a hierarchy of categories or classes).

The text classification supports a variety of text classification scenarios like:

- Binary classification like simple sentiment analysis (positive, negative)
- Multiple class classification like selecting one category among several alternatives.

Most partitioning algorithms do not take raw text as input but numeric vectors. For this it is necessary to find a representative transformation that converts the text to digital vectors. A family of this transformation is called Bag-of-Words (BOW).

3. Applied Algorithms

In this work, we will apply tree supervised learning algorithms

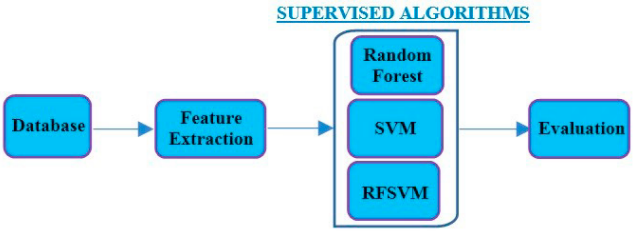


Fig. 3. Control flow of the system

3.1. Random forest

Random forest, which were formally proposed in 2001 by Leo Breiman and Adèle Cutler, are part of the automatic learning techniques. This algorithm combines the concepts of random subspaces and "bagging". The decision tree forest algorithm trains on multiple decision trees driven on slightly different subsets of data.

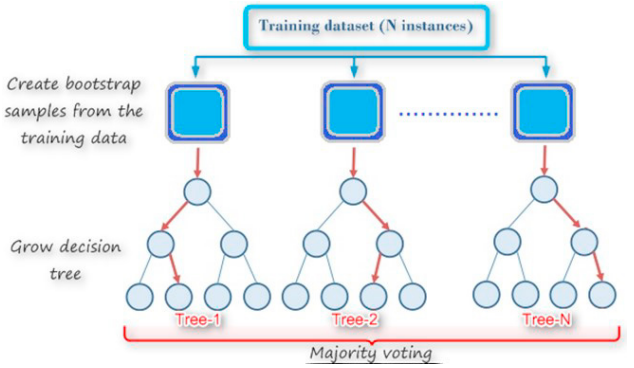


Fig. 4. Pictorial representation of random forest

The random forest is part of the family set methods that take the decision tree as an individual predictor, they are based on the methods of Bagging, Randomizing Outputs and Random Subspace excusing boosting. [8]

The random forest algorithm is one of the best among classification algorithms - able to classify large amounts of data with accuracy. It is an ensemble learning method for classification and regression that constructs a number of decision trees at training time and delivers the class that is the mode of the classes output by individual trees.

Random Forest Algorithm:

- For b = 1 to B Make
  - Draw a bootstrap sample Z* of size N from the training data.
  - Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
    - Select m variables at random from the p variables.
    - Pick the best variable/split-point among the m.
    - Split the node into two daughter nodes.
- Output the ensemble of trees $\{T_b{}_1^B\}$.

To make a prediction at a new point x:

$$\text{Regression: } \hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random forest tree.

$$\text{Then } \hat{C}_{rf}^{B}(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B.$$

In random forest classification method, many classifiers are generated from smaller subsets of the input data and later their individual results are aggregated based on a voting mechanism to generate the desired output of the input data set. This ensemble learning strategy has recently become very popular. Before RF, Boosting and Bagging were the only two ensemble learning methods used. RF has been extensively applied in various areas including modern drug discovery, network intrusion detection, land cover analysis, credit rating analysis, remote sensing and gene microarrays data analysis etc... [9] [10]

There are two ways to evaluate the error rate. One is to split the dataset into training part and test part. We can employ the training part to build the forest, and then use the test part to calculate the error rate. Another way is to

use the Out of Bag (OOB) error estimate. Because random forests algorithm calculates the OOB error during the training phase, we do not need to split the training data.

3.2. Support vector machine

The Support Vector Machine method is a statistical classification approach which is based on the maximization of the margin between the instances and the separation hyper-plane. It was considered the best text classification method (Xia, Rui, Chengqing Zong, and Shoushan Li, 2001). It is a non-probabilistic binary linear classifier, that has the ability to linearly separate the classes by a large margin, it become one of the most powerful classifier capable of handling infinite dimensional feature vectors.

Support Vector Machine is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. [11]

In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.

This method was used for the first time in the text classification by (Thosten Joachims, 1998). It has proved successful in classifying opinion documents, including style (Diederich et al., 2000). The principle for support vector machine algorithm is to solve classification and regression problems. It has been applied to many fields (bioinformatics, information retrieval, computer vision, finance, etc.).

SVMs were developed in the 1990s based on the theoretical considerations of Vladimir Vapnik on the development of a statistical theory of learning: Vapnik-Chervonenkis theory. SVMs were quickly adopted for their ability to work with large data, the small number of hyper parameters, their theoretical guarantees, and their good results in practice.

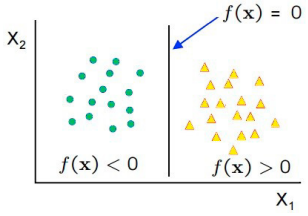
To make our discussion of SVMs easier we will be considering a linear classifier for a binary classification problem with labels y and features x. We'll use $y \in \{-1, 1\}$ to denote the class labels and parameters w, b:

$$\mathbf{f(x) = w^T x + b}$$

w: normal to the line.

b: bias.

SVM is represented by a separating hyper plane f (x) that geometrically bisects the data space into two diverse regions thus resulting in classification of the input data space into two categories [12]:



The function f(x) denotes the hyperplane that separates the two regions and facilitates in classification of the data set. The two regions geometrically created by the hyperplane correspond to the two categories of data under two class labels. A data point "a" belongs to either of the region depending on the value of f(a). If f(a) > 0 it belongs to one region and if f(a) < 0 it belongs to another region.

Assume that the input data consists of n data vectors where each data vector is represented by x i ∈ Rn, where i=1, 2, …, n. Let the class label that needs to be assigned to the data vectors to implement supervised classification be denoted by yi, which is +1 for one category of data vectors and -1 for the other category of data vectors. The data set can be geometrically separated by a hyperplane. Since the hyperplane is represented by a line it can also be mathematically represented by [13]:

$$w^T x_i + b \geq +1$$
$$w^T x_i + b \leq -1$$

The hyperplane can also be represented mathematically by [14]:

$$f(x) = sgn(w^T x + b)$$

where sgn() is known as a sign function, which is mathematically represented by the following equation [15]:

$$sgn(x) = \begin{cases} 1\ if\ x > 0 \\ 0\ if\ x = 0 \\ -1\ if\ x < 0 \end{cases}$$

The distance D of a data point x from the hyperplane is represented mathematically by the equation:

$$D = \frac{|w^T x + b|}{|w|}$$

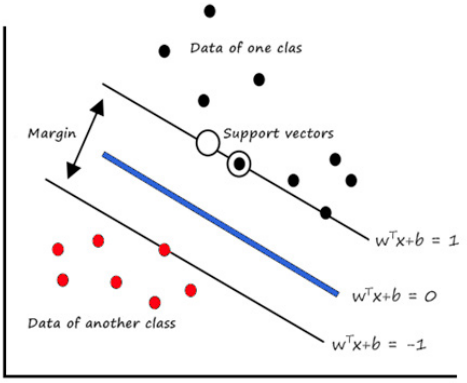


Fig. 5. Hyper plane of support vector machines

Then the margin is given by:

$$\frac{w}{\|w\|} \cdot (x_+ - x_-) = \frac{w^T(x_+ - x_-)}{\|w\|}$$
$$= \frac{w^T\left(\left(\frac{+1-b}{w^T}\right) - \left(\frac{-1-b}{w^T}\right)\right)}{\|w\|} = \frac{2}{\|w\|}$$

There are many such hyperplanes which can split the data into two regions. But SVM ensures that it selects the hyperplane that is at a maximum distance from the nearest data points in the two regions. There are only few hyperplanes that shall satisfy this criterion. By ensuring this condition SVM provides accurate classification results. [16]

The steps followed while using SVM in classifying data are mentioned in the below algorithm [17]:

Input: I: Input data
Output: V: Support vectors set
Begin
Step 1: Divide the given data set into two set of data items having different class labels assigned to them
Step 2: Add them to support vector set V
Step 3: Loop the divided n data items
Step 4: If a data item is not assigned any of the class labels then add it to set V
Step 5: Break if insufficient data items are found
Step 6: end loop
Step 7: Train using the derived SVM classifier model and test so as to validate over the unlabeled data items.
End

4. Proposed method

The proposed system takes a hybrid approach to solve the problem of classification and possesses the capabilities

of Random Forest and SVM at the same time. Firstly, random forest is an ensemble learning method that construct a number of decision trees at randomly selected features and predict the class of a test instance by voting of the individual trees. Support Vector Machine revolves around the notion of a margin — either side of a hyperplane that separates two classes. Maximizing the margin and with this way creating the largest possible distance between the separating hyperplane and the instances on either side of it has been proven to reduce an upper bound on the expected generalization error. RF was not sensitive to input parameters, thus, we just used the default parameters for each classifier. The trained classifiers return scores between 0 and 1, these scores are then transformed to a binary state indicating 'negative' or 'positive'. For each combination, the existence of element is considered positive (P) or negative (N). The notation of TP indicates True Positives: number of examples predicted positive that are actually positive, FP indicates False Positives: number of examples predicted positive that are actually negative, TN indicates True Negatives: number of examples predicted negative that are actually negative and FN indicates False Negatives: number of examples predicted negative that are actually positive.

The classification metrics considered for the sentiment analysis are Accuracy, Precision, Recall and F-Measure and these parameters are evaluated based on the calculated positivity and negativity of reviews by the proposed hybrid approach. The performance evaluation of classifiers is made according to the following formulas:

Report of the true positives. It corresponds to:

$$TP\ Rate = \frac{TP}{TP+FN}$$

It is thus the report between the number of positive instances classified well and the total number of elements which should be classified well.

Report of the false positive one. He corresponds, symmetrically in the previous definition:

$$FP\ Rate = \frac{FP}{FP+TN}$$

The datum of the rates TP Rate and FP Rate allows to reconstruct the matrix of confusion for a given class.

Precision is the report between the number of the true positive and the sum of the true positives and the false positive. A value of 1 expresses the fact that all the positive classified examples were really:

$$Precision = \frac{TP}{TP+FP}$$

Recall is the percentage of correct items that are selected. recall of 1 means that all the positive examples were found.

$$Recall = \frac{TP}{TP+FN}$$

Accuracy is a common measure for the classification performance and it's proportional of correctly classified instances to the total number of instances, whereas the error rate uses incorrectly classified rather than correctly.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

This quantity allows to group in a single number the performances of the classifier (for a given class) as regards Recall and the Precision:

$$F\text{-}Measure = 2 * \frac{Recall*Precision}{Recall+Precision}$$

5. Results and discussions

The training and test data we used for this work were taken from the "Amazon" which contains 1000 instances divided into positive (500) and negative (500). In this article, Cross Validation method with fold value equal to 10 has been used for training and testing phases

We will use some techniques that automatically extracts this data into positive or negative sentiments. By using the sentiment analysis, the customer can know the feedback about the product before making a purchase. Sentiment

analysis is a type of natural language processing for tracking the mood of the public about a particular product.

5.1. Using random forest

The following table shows the result obtained using the Random Forest algorithm:

Table 1. Cross validation results for random forest

	Positive	Negative	Total
Positive	402	98	500
Negative	92	408	500
Total	494	506	1000

From this table, we see that 810 reviews are correctly classified among 1000, and 190 reviews are misclassified.

5.2. Using support vector machine

The following table shows the result obtained using Support Vector Machine algorithm:

Table 2. Cross validation results for support vector machine

	Positive	Negative	Total
Positive	409	91	500
Negative	85	415	500
Total	494	506	1000

From this table, we see that 824 reviews are correctly classified among 1000, and 176 reviews are misclassified.

5.3. Using random forest support vector machine

The following table shows the result obtained using our approach RFSVM:

Table 3. Cross validation results for random forest support vector machine

	Positive	Negative	Total
Positive	416	84	500
Negative	82	418	500
Total	498	502	1000

From this table, we see that 834 reviews are correctly classified among 1000, and 166 reviews are misclassified. Table 4 represent results in terms of accuracy of proposed approach and the other algorithms used in this paper.

Table 4. Number of classified instances

	C.C.I	I.C.I	A (%)	TTBM (s)
RF	810	190	81.0	7
SVM	824	176	82.4	3.93
RFSVM	834	166	83.4	3.70

- C.C.I.: Correctly Classified Instances; - I.C.I.: Incorrectly Classified Instances;
- A.: Accuracy; - TTBM: Time Taken to Build Model.

From this table, it is represented that the accuracy computed in the case of proposed method (RFSVM) is better as compared to random forest and support vector machine. Improving the algorithm in different ways could improve the results further.

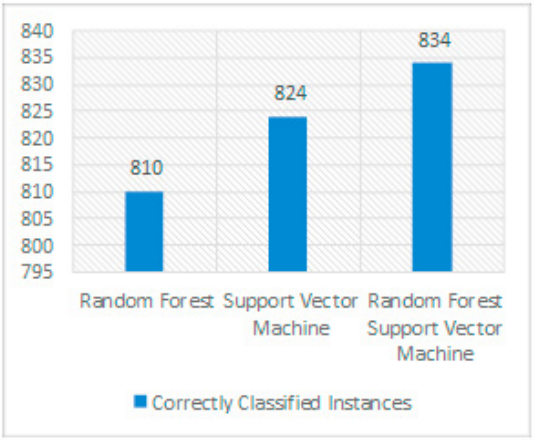


Fig. 6. Number of correctly classified instances

From figure 6 it is evident that Random Forest Support Vector Machine (RFSVM) shows the best performance as compare to other studied algorithms.

Table 5. Number of classified instances

	RF		SVM		RFSVM	
	Positive	Negative	Positive	Negative	Positive	Negative
TP Rate	80.4	81.6	81.8	83	83.2	83.6
FP Rate	18.4	19.6	17	18.2	16.4	16.8
Precision	81.4	80.6	82.8	82	83.5	83.3
Recall	80.4	81.6	81.8	83	83.2	83.6
F-Measure	80.9	81.1	82.3	82.5	83.4	83.4
ROC Area	88.4	88.4	91.4	91.4	91.7	91.4

For RFSVM, we found that the precision, recall and F-measure were 83.4 %, 83.4 % and 83.4 %, respectively. The F-measure of RFSVM was higher than that of others algorithms, which meant RFSVM fitted better than these classifiers. The hybrid approach combines the advantage of both the Random Forest and Support Vector Machine. It is inheriting more accuracy using supervised machine learning approaches and providing good stability against the other algorithms.

The paper considered the combination of supervised classification algorithms to product review data and also predicted the positive and negative reviews by people. The hybrid approach which contains the combination of Random Forest and Support Vector Machine produced better results on the basis of Accuracy, Precision, Recall and F- Measure. Random Forest approach improved the performance in the case of small reviews and Support Vector Machine improved the performance just in case of large reviews are working as a single hybrid approach.

6. Conclusion and perspectives

Sentiment analysis is essential for anyone who is going to make a decision. It is helpful in different field for calculating, identifying and expressing sentiment. Although the work has yielded interesting results, we plan to make some changes in future work to improve performance and achieve better results. In this paper, we have

compared Support Vector Machine, Random Forest and Random Forest Support Vector Machine algorithms (RFSVM) which are very suitable for generating rules in classification technique. From the experimental results, it is concluded that Random Forest Support Vector Machine algorithm seems better than the other algorithms for product reviews dataset offered by Amazon. The reason for better results in the case of hybrid classification methodology used in this paper is since it makes use of the advantages of each of the individual traditional RF, SVM classifications methods.

References

1. V. Umadevi, "Sentiment analysis using weka," *IJETT International Journal of Engineering Trends and Technology*, vo1. 8 (4), pp. 181-183, Dec. 2014.

2. R. Prabowo, and M. Thelwall, "Sentiment analysis: A combined approach," *Journal of Informetrics*, pp. 143-157.

3. Liu, Bing. "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies* vol. 5, no. 1, pp. 1-167, 2012.

4. V. S. Jagtap, and K. Pawar, "Analysis of different approaches to Sentence-Level Sentiment Classification," *International Journal of Scientific Engineering and Technology*, vol. 2, pp. 164-170, Apr. 2013.

5. Xia, Rui, Chengqing Zong, and Shoushan Li. "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences* vol. 181, issue 6, pp. 1138–1152, 15 Mar. 2011.

6. Melville, Prem, Wojciech Gryc, and Richard D. Lawrence. "Sentiment analysis of blogs by combining lexical knowledge with text classification," *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1275-1284, 2009.

7. Aditi Mahajan, Anita Ganpati. "Performance evaluation of rule based classification algorithms," *IJARCET International Journal of Advanced Research in Computer Engineering & Technology*, vol. 3, issue 10, pp 3546-3550, Oct. 2014.

8. R. Genuer, "Forêts aléatoires : aspect théoriques, sélection de variables et applications," *Thèse de Doctorat Mathématiques*, Université de Paris-Sud XI, 2010.

9. V.F. Rodríguez-Galiano, F.Abarca-Hernández, B. Ghimire, M. Chica-Olmo, P.M .Akinson, C. Jeganathan, "Incorporating Spatial Variability Measures in Land-cover Classification using Random Forest, " *Procedia Environmental Sciences*, vol. 3, pp. 44-49, 2011.

10. Reda M. Elbasiony, Elsayed A.Sallam, Tarek E. Eltobely, Mahmoud M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted k-means, " *A in Shams Engineering Journal*, Available online 7 Mar. 2013.

11. Witten, Ian H., and Eibe Frank. "Data mining: practical machine learning tools and techniques," *Morgan Kaufmann*, pp. 1-560, Jun. 2005 by Elsevier.

12. Y. Al-Amrani, M. Lazaar, and K. E. Elkadiri, "Sentiment Analysis using supervised classification algorithms," *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications (BDCA'17)*. ACM, New York, NY, USA, Article 61, 8 pages. DOI: https://doi.org/10.1145/3090354.3090417.

13. Chun-Xia Zhang, Jiang-She Zhang, Gai-Ying Zhang, "An efficient modified boosting method for solving classification problems," *Journal of Computational and Applied Mathematics*, vol. 214, issue 2, 1 May 2008, pp. 381-392.

14. Xinjun Peng, Yifei Wang, Dong Xu, "Structural twin parametric-margin support vector machine for binary classification, Knowledge-Based Systems," vol. 49, Sept. 2013, pp. 63-72.

15. J. T. Lalis, "A New Multiclass Classification Method for Objects with Geometric Attributes Using Simple Linear Regression," *IAENG International Journal of Computer Science*, vol. 43, no. 2, pp.198–203, 2016.

16. Hsun-Jung Cho, Ming-TeTseng, "A support vector machine approach to CMOS-based radar signal processing for vehicle classification and speed estimation," *Mathematical and Computer Modelling*, vol. 58, issues 1–2, Jul. 2013, pp. 438- 448.

17. S.N. Jeyanthi, "Efficient Classification Algorithms using SVMs for Large Datasets," A Project Report Submitted in partial fulfillment of the requirements for the Degree of Master of Technology in Computational Science, *Supercomputer Education and Research Center*, IISC, BANGALORE, INDIA, Jun. 2007.