VOL. 15, NO. 6, MARCH 2020 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

AN EFFICIENT FEATURE SELECTION SYSTEM FOR AUTOMOTIVE SENTIMENT CLASSIFICATION IN HADOOP FRAMEWORK USING NAÏVE BAYES CLASSIFIER

K. Vimal Kumar Stephen, Faiza Rashid Ammar Al-Harthy and Mohammed Tariq Shaikh
Department of Information Technology, Ibra College of Technology, Ibra, Oman
E-Mail: vimal@ict.edu.om

ABSTRACT

Hadoop is a MapReduce framework with an open source implementation which is basically employed in scholastic and engineering for big data analysis. The MapReduce framework is usually employed to assess huge datasets like tweets collections, online documents or large scale graphs. Sentiment classification is the significant part in text mining to categorize documents based on their opinion or sentiment. In sentiment classification, documents can be signified in the feature vector form utilized in machine learning algorithms to carry out classification. The feature selection process with the feature vectors is essential. In this work, Term Frequency (TF) based feature extraction method is employed. A feature selection method called Information Gain (IG) and Particle Swarm Optimization (PSO). Binary PSO (BPSO) is the binary version of PSO and can be useful to feature selection domain. The presented feature selection methods object to remove noisy, unrelated, or inappropriate features that may worsen the performance of classification. Naive Bayes (NB) classifier helps to develop the classification presentation. Investigational consequences prove that the PSO based feature selection process attains greater NB classification performance than IG based feature selection.

Keywords: big data, sentiment classification, hadoop framework, mapreduce, feature selection, information gain (IG), particle swarm optimization (PSO) and naïve bayes (NB) classifier.

INTRODUCTION

Big data consigns to a dimension of datasets beyond the capability of characteristic database software implements to confine, amass, handle, and examine. Major big data features are high volume, diversity and velocity. Volume stands for huge sizes that is difficult to be practiced with conventional database and single computing machine. Velocity denotes that data is continually generated in a faster rate and range relates to various patterns like texts, images and videos [1]. Sentiment analysis is termed as opinion mining denotes to an information extracting method on subjectivity from a textual input. To attain this, it unites methods from Natural Language Processing (NLP) and text analysis. Sentiment mining abilities permit finding whether the text input is either objective or subjective whereas in polarity mining text is categorized as positive or negative.

To execute a classification of sentiment analysis for detecting the sentiment in the messages/text, [2] a pattern is framed using given training and test datasets. Another way in supervised learning concept is to manually identify the sentiment in the messages and train the algorithm with the training dataset. Considering huge data quantities to be categorized and the input are text, Naïve Bayes (NB) method is widely used because of its fast training method even with huge training data volumes. A Mapreduce based NB is effective in executing sentiment analysis for big data.

Apache Hadoop is an open-source MapReduce framework implementation, which is based on parallel programming method for distributed processing, proposed by Google. It permits the distributed dataset processing in petabytes crosswise hundreds or thousands of computer products linked to a system. The Hadoop Distributed File System (HDFS) is the Hadoop storage space section. It is intended to amass huge dataset consistently on clusters, and to flow such data to user applications at higher throughput. HDFS accumulates file system metadata and application data alone. To assure consistency, accessibility and performance of the system, it amasses three autonomous data block (replication) prints [3].

MapReduce is a novel distributed programming concept extensively employed to function parallel applications for huge scale datasets processing. MapReduce employs key/value pair data type in Map and Reduce functions and it is instigated by the map and reduce primitives in Lisp and several other functional languages. The Map function needs the user to deal the input key pairs value and creates intermediate key and value pairs group. Reduce function deals with the intermediate key pairs and the value set related to the intermediate key value. Reduce function joining such values, to obtain a smaller value set [4].

The Map algorithm comprises three stages: (i). Hadoop and MapReduce framework create a map task for each Input Split, and is produced by the job Input Format. Every <Key, value>relates to a map task. (ii) Execute Map task, course the input <key, value> to model a new<key, value> and (iii). Mapper's output is arranged towards every allotted Reducer. Reducer algorithm has the following 3steps: first, MapReduce will allot associated block for each Reducer (Shuffle). After that, the reducer input is combined based on the key (sorting step) and at last step is Secondary Sort. The Hadoop cores are Hadoop MapReduce [5] and HDFS.

Feature selection try to decrease a dataset by eliminating inappropriate or superfluous features. This method searches to attain a least attributes set, so the

consequences of data mining are functional over the decreased dataset are as near as probable to the consequences attained with all attributes. This diminution assists the extracted pattern considerate and increases the posterior learning stages speed. Feature selection techniques can be classified as: (i) Wrapper approach: The selection condition is fitness function fraction. (ii) Filtering approach: The selection depends on data associated measures, like crowding. (iii) Embedded approach: The optimal feature subset is created in the classifier construction [6].

Raw data is obtained as the input in classification. This raw data is categorized as apt to a definite class on the basis of needed parameter set. To make decision within the uncertainty conditions, the Bayesian theory is employed as the framework that implies a probabilistic concept to inference. It has been examined by Bayes that by finding the previous frequency events, the forthcoming event probability can be calculated. The Bayesian system is deceptively not difficult. As well, the forecast that it creates is on the basis of data from real life cases- The huge the data, the better it functions. As well, such systems are auto correcting that defines the results change with the change in data [7].

Classical and syntactic are the two concepts that are implemented by the classification system. It is contemplated that a probabilistic system produces patterns and Statistical is based on arithmetical pattern characterizations. The structural inter-relationships of features created based on structural pattern recognition. An algorithm range from simple Bayesian classifiers to potent neural networks is adapted in pattern recognition. Here, NB classifier is employed is a functional Bayesian learning system based on Bayesian theorem. This is well acclimatized for greater dimensional inputs. NB frequently executes greater over the more sophisticated classification systems, in spite of its ease. In definite domains, the theorem performance is approximately on par with neural network and decision tree learning.

Feature selection involves some classification pattern features, classification accuracy, the time required for learning classification functions, the sample quantity required for learning and costs regarding the structures. Feature selection is an optimization method to decrease a huge original feature to a relatively smaller features subset that are important to increase the classification accuracy rapidly and efficiently [8]. PSO is extensively employed to resolve optimization setbacks and a feature selection issue. In engineering PSO, there are many methods to execute optimizing with raised weight attribute for all features employed, choosing the aspect and feature selection. Here, presents the IG and PSO based feature selection techniques in sentiment classification for Hadoop framework. The rest part of the paper is systematized as follows. Section two discusses related works in literature. Section three describes different techniques employed in this investigation. Section four discusses outcomes and Section five concludes the work.

RELATED WORKS

Parveen & Pandey [9] focused on sentiment extraction from a well-known micro blogging website, Twitter in which the views and opinion are posted. It has made sentiment analysis on tweets that aid to offer a few predictions on business intelligence. It employs Hadoop Framework for handling movie dataset accessible on the twitter website as reviews, feedback, and comments. Sentiment analysis effects on twitter data will be revealed as diverse sections handling positive, negative and neutral sentiments.

Huge feature set created in the sentiment analysis of movie reviews slows the process and is less sensitive. To deal a huge number of features and enhance sentiment classification, an information-based feature selection and classification was suggested by Pratiwi [10]. The suggested technique decreases over 90% redundant features whereas the presented classification system attained 96% accuracy.

Satish & Kavya [11] presented a proficient method to categorize the big data from e-mail based on Firefly Algorithm (FA) heuristic and NB classifier. Suggested method had two phases, Map reduce framework for training and testing. The conventional FA is adapted and the optimized feature space is applied on behalf of the optimal consequences. When the optimal feature space was determined via FA, NB classifier is applied. These two methods are efficiently shared in Map-Reduce framework. The experimental consequences are verified by assessment measures like computation time, accuracy, specificity and sensitivity.

Cao *et al.*, [12] presented a parallel intend and understanding technique for a PSO-optimized Back Propagation (BP) neural network on the basis of MapReduce on the Hadoop. The PSO was employed towards optimization of BP neural network's initial weights and thresholds and develop the classification algorithms accuracy. The MapReduce parallel programming pattern was exploited in BP algorithm to attain parallel processing. The parallel PSO-BP neural network algorithm classification accuracy is about 92%, and the system effectiveness is about 0.85, that proposes clear benefits as processing big data.

Ahmad *et al.*, [13] proposed system architecture by selecting features through Artificial Bee Colony (ABC). Also, a Kalman filter had been employed in Hadoop ecosystem which is employed to remove noise. Also, usual MapReduce with ABC had been employed to improve efficiency of processing. Furthermore, an absolute four-tier architecture was presented effectively combined the data, remove redundant information, and assess the data using Hadoop-based ABC algorithm. The methodology is accessed through Hadoop and MapReduce with the ABC algorithm. ABC algorithm had been employed to choose features, while, MapReduce had been maintained through a parallel algorithm to perform on a large dataset.

METHODOLOGY

In this section, the various techniques: feature extraction by TF method, IG based feature selection; PSO based feature selection and Naïve Bayes classifier is explained.

A. Term Frequency (TF) based feature extraction

The TF scrutinizes the single term frequency on the features set, whereas dataset has regularity on the basis of set of definite term feature [14]. It calculates the term frequency with its function, TF(t), specifying t as the number of terms. TF (t, oi) is percentage of datasets fit in to a class o_i. And |c| signifies the number of available classes in equation (1).

$$TF(t) = \sum_{i=1}^{|b|} TF(t, o_i) = \sum_{i=1}^{|b|} R(t \mid o_i) \quad (1)$$

B. Information Gain (IG) based feature selection

For every feature, IG is measured through the training data. It is an arithmetic property used to measure goodness of feature parts of the training instances depending on its class [15]. Based on measure entropy IG is measured as in equation (2):

$$Entropy(S) = -\sum_{i} p(i \mid S) \log_2 p(i \mid S) \quad (2)$$

Here S signifies all aspects, p(i|S) specifies its portion to class i. The class can be in the form of either positive, negative, or neutral. The changes occur in entropy when a training instances are divided into smaller subsets. Based on this, IG specifies the expected reduction. The IG of a feature t comparative to a group of aspects S, is represented in equation (3):

$$IG(S,t) = Entropy(S) - \sum_{v \in Values(t)} \frac{|S_v|}{|S_t|} Entropy(S_v) \quad (3)$$

Here, Values (t) specifies the set of all promising values for feature t and it can be either positive, negative, or neutral. Svrepresents the subset of S with aspects of class related to feature t. St represents all aspects set fit in to feature t. |.|signifies the cardinality set.

C. Particle Swarm Optimization (PSO) based feature selection

PSO motivated by social behaviour in a bird flock. In this algorithm, every particle flies in the search space with a velocity which is altered based on its past experience and the behaviour of other particles in the neighborhood [16]. Individual particle takes its objective function value as in (4):

$$v_{id}^{t} = w \times v_{id}^{t-1} + c_1 \times r_1 (p_{id}^{t} - x_{id}^{t}) + c_2 \times r_2 (p_{gd}^{t} - x_{id}^{t}) \quad (4)$$

Where i symbolizes the ith particle and d is the dimension of the solution space, c_1 and c_2 are constants representing the cognition and social learning factor respectively, r_1 and r_2 are random numbers between 0 to 1, p_{id}^{t} and p_{gd}^{t} signifies the position with the best fitness for the ith particle and the best position in the neighborhood, v_{id}^{t} and v_{id}^{t-1} are the velocities at time t and time t - 1, and x_{id}^{t} is the position of ithparticle at time t. Individual particle then update to new position as in equation (5):

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t}, \; d = 1, 2, \ldots, D, \quad (5)$$

Binary PSO is employed to feature selection problem. BPSO in which a particle shifts in a state space limited to 0 and1 on every element, depending on the variations in probabilities is represented in equation (6 and 7):

$$x_{id} = \begin{cases} 1, & rand() < S(v_{i,d}) \\ 0, & otherwise \end{cases} \quad (6)$$

$$S(v) = \frac{1}{1 + e^{-v}} \quad (7)$$

The function S(v) is a sigmoid limiting transformation and rand( ) is a random number between 0 to 1.

D. Naïve Bayes classifier

Owing to its simplicity, the NB classifier is a population algorithm. It also is computationally efficient and its performance is verified. It is a supervised learning system and statistical method for classification. This is particularly functional in high input dimensionality. An essential probabilistic model is presumed by this system. This permits for attaining the model's uncertainty in a scrupulous method through determining the probabilities outcomes [17]. It is a text classification allocates class $c^* = argmaxcP(c \mid d)$, to a given document d. The NB classifier uses Bayes' rule equation (8):

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} \quad (8)$$

Where, P(d) helps to choose c*. To evaluate P (d/c), assumed d's class as in (9).

$$P_{NB}(c \mid d) = \frac{P(c)\left(\prod_{i=1}^{m} P(f_i \mid c)^{n_i(d)}\right)}{P(d)} \quad (9)$$

Where, the quantity of features is denoted by m. The feature vector is denoted by f_i .The working of naïve bayes classifier is given as below [18]:

Assume D as the training dataset where the n-dimensional vector, $X = (x_1, x_2, ..., x_n)$ represents each tuple.

Say the number of classes are m given by C_1, C_2, C_3...., C_m. If it wants to classify an unknown tuple X, the Naive Bayesian classifier allocates an indefinite tuple X to the class C_i if and only if $P(C_i \mid X) > P(C_j \mid X)$. For $1 \le j \le m$, and $I \ne j$, above posterior probabilities are designed by Bayes Theorem.

Uses of NB classification:

This theorem is used for classifying text documents. It is a probabilistic learning scheme.

It can filter spam, which is its biggest advantage. Using a NB classifier, it can detect a spam e-mail and can differentiate illicit spam mail from genuine ones.

Hybrid recommender systems that employ the NB classifier and collaborative filtering recommender systems employ machine learning and data mining systems for filtering unfamiliar information. They can forecast if a user can be given with a resource.

RESULTS AND DISCUSSIONS

For evaluating the different methods, the subset of Amazon automotive product review dataset is employed (15000 positive, 20000 negative and 15000 neutral).The IG based feature selection, PSO based feature selection and NB methods are employed. The classification accuracy, average recall, average precision and average f measure is represented in Tables 1 to 4 and Figures 1 to 4.

Table-1. Classification accuracy for PSO based feature selection.

	IG based feature selection	PSO based feature selection
Naïve Bayes	78.3	88.5

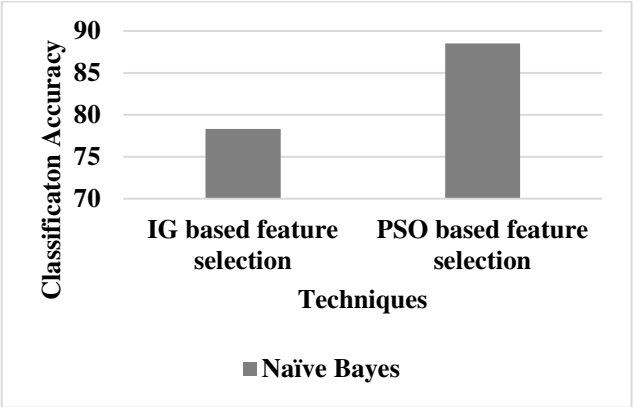

Figure-1. Classification accuracy for PSO based feature selection.

From the Figure-1, it is seen that the PSO based feature selection consists of higher classification accuracy by 12.23% on behalf of Naive Bayes when compared with IG based feature selection.

Table-2. Average recall for PSO based feature selection.

	IG based feature selection	PSO based feature selection
Naïve Bayes	0.7861	0.888333

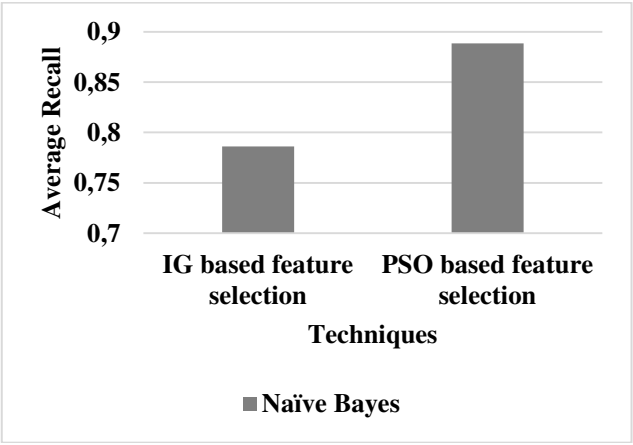

Figure-2. Average recall for PSO based feature selection.

From the Figure-2, it is seen that the PSO based feature selection consists of higher average recall by 12.21% on behalf of Naive Bayes when compared with IG based feature selection.

Table-3. Average precision for PSO based feature selection.

	IG based feature selection	PSO based feature selection
Naïve Bayes	0.780933	0.884933

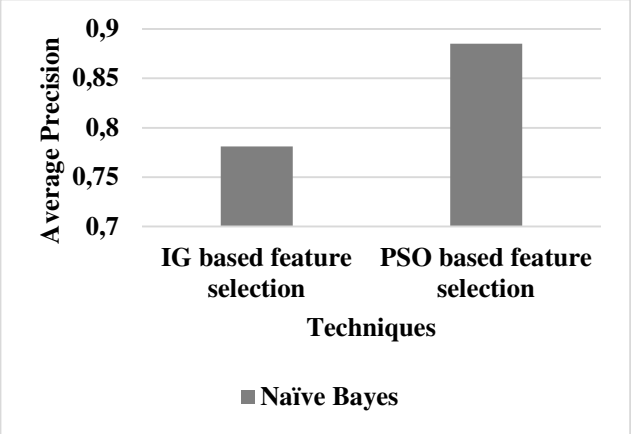

Figure-3. Average precision for PSO based feature selection.

From the Figure-3, it is seen that the PSO based feature selection consists of higher average precision by 12.48% on behalf of Naive Bayes when compared with IG based feature selection.

Table-4. Average F measure for PSO based feature selection.

	IG based feature selection	PSO based feature selection
Naïve Bayes	0.7817	0.884467

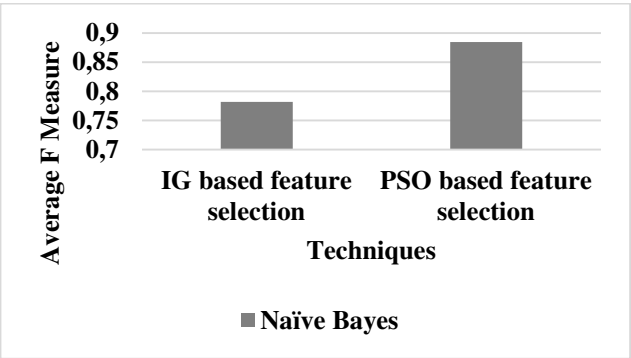

Figure-4. Average F measure for PSO based feature selection.

From the Figure-4, it is seen that the PSO based feature selection consists of sophisticated average f measure by 12.33% on behalf of Naive Bayes when compared with IG based feature selection.

CONCLUSIONS

Data analysis is the greatest challenging problem in the world specifically in huge data volume. Sentiment analysis is the method of employing text analytics to extract different opinions of data sources. Feature selection is the significant features manipulate the classification accuracy rate. A proficient and strong feature selection means can remove noisy, inappropriate and superfluous data. Here, proposed an easy and absolute system for sentiment mining on huge datasets with TF feature extraction method, IG and PSO based feature selection process with the Hadoop framework.TF of any given term is described since the number of times a provided term has appear in a specific given text or document. It normally how regularly a term happens in a document. IG is the significant well known feature filtering approaches. By IG, it can calculate a score for each individual feature that symbolizes better than the feature partitions the features among the different sentiment classes. PSO is a computational concept based on combined behaviour inspired through the social behavior of bird flocking or fish schooling. The presented PSO-based feature selection algorithm is used to investigate the feature space for the optimal feature subset in which features are chosen on the basis of definite discrimination condition. The naïve bayes classifier develops the classification performance by eliminating the unrelated features. Results demonstrate that the proposed method has higher classification accuracy by 12.23% for Naive Bayes when compared with IG based feature selection.

REFERENCES

[1] Sehgal D. & Agarwal A. K. 2018. Real-time Sentiment Analysis of Big Data Applications Using Twitter Data with Hadoop Framework. In Soft Computing: Theories and Applications (pp. 765-772). Springer, Singapore.

[2] Skuza M. & Romanowski A. 2015, September. Sentiment analysis of Twitter data within big data distributed environment for stock prediction. In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on (pp. 1349-1354). IEEE.

[3] Ayma V. A., Ferreira R. S., Happ P., Oliveira D., Feitosa R., Costa G. ...& Gamba P. 2015. Classification algorithms for big data analysis, a Map Reduce approach. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences. 40(3): 17-21.

[4] Pakize S. R. & Gandomi A. 2014. Comparative study of classification algorithms based on MapReduce model. International Journal of Innovative Research in Advanced Engineering (IJIRAE). 1(7): 251-254.

[5] Liu B., Blasch E., Chen Y., Shen D. & Chen G. 2013, October. Scalable sentiment classification for big data analysis using Naive Bayes classifier. In Big Data, 2013 IEEE International Conference on (pp. 99-104). IEEE.

[6] Peralta D., del Río S., Ramírez-Gallego S., Triguero I., Benitez J. M. & Herrera F. 2015. Evolutionary feature selection for big data classification: A mapreduce approach. Mathematical Problems in Engineering.

[7] Islam M. J., Wu Q. J., Ahmadi M. & Sid-Ahmed M. A. 2007, November. Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. In Convergence Information Technology, 2007. International Conference on (pp. 1541-1546). IEEE.

[8] Wahyudi M. & Kristiyanti D. A. 2016. Sentiment Analysis Of Smartphone Product Review Using Support Vector Machine Algorithm-Based Particle Swarm Optimization. Journal of Theoretical & Applied Information Technology. 91(1): 189-201.

[9] Parveen H. & Pandey S. 2016, July. Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. In Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on (pp. 416-419). IEEE.

[10] Pratiwi A. I. 2018. On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis. Applied Computational Intelligence and Soft Computing.

[11] Satish K. R. & Kavya N. P. 2014, November. Big data processing with harnessing hadoop-MapReduce for optimizing analytical workloads. In Contemporary Computing and Informatics (IC3I), 2014 International Conference on (pp. 49-54). IEEE.

[12] Cao J., Cui H., Shi H. & Jiao L. 2016. Big data: A parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce. PloS one. 11(6): e0157551.

[13] Ahmad A., Khan M., Paul A., Din S., Rathore M. M., Jeon G. & Choi G. S. 2018. Toward modeling and optimization of features selection in Big Data based social Internet of Things. Future Generation Computer Systems. 82, 715-726.

[14] Joseph S., Mugauri C. & Sumathy S. 2017, November. Sentiment analysis of feature ranking methods for classification accuracy. In IOP Conference Series: Materials Science and Engineering (Vol. 263, No. 4, p. 042011). IOP Publishing.

[15] Schouten K., Frasincar F. & Dekker R. 2016, June. An information gain-driven feature study for aspect-based sentiment analysis. In International Conference on Applications of Natural Language to Information Systems (pp. 48-59). Springer, Cham.

[16] Liu Y., Wang G., Chen H., Dong H., Zhu X. & Wang S. 2011. An improved particle swarm optimization for feature selection. Journal of Bionic Engineering. 8(2): 191-200.

[17] Sumathi T., Karthik S. & Marikkannan M. 2014. Artificial Bee Colony Optimization For Feature Selection In Opinion Mining. Journal of Theoretical & Applied Information Technology. 66(1): 368-379.

[18] Dinu L. P. & Iuga I. 2012, March. The Naive Bayes classifier in opinion mining: in search of the best feature set. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 556-567). Springer, Berlin, Heidelberg.