An Efficient Approach for Sentiment Analysis in a Big Data Environment

Khalid Ait Hadi, Rafik Lasri, Abdellatif El Abderrahmani

Abstract*: *Sentiment analysis is a very substantial area of research. Numerous studies have examined the subject in recent years. It has rapidly gained interest by reason of the unusual volume of opinionated data on the Internet. Significant research has been accomplished to study sentiment by resorting to diverse machine learning techniques. Nevertheless, the downward trend of the accuracy rates in these studies often impacts the approach's efficiency. With the aim of surmounting this obstacle, we introduce an efficient technique for sentiment mining in big data context. The data collected are cleaned using a preprocessing data mining technique before proceeding to the selection of the optimal features with the use of a versatile approach of greedy algorithms, called Carousel greedy, combined with a bio-inspired metaheuristic algorithm. The classification is subsequently performed by Cat Swarm Optimization Based Functional Link Artificial Neural Networks classifier and the performance of the approach is discussed through experimental results.

Index Terms*: *Big data, Bio-inspired intelligence, Carousel greedy algorithm, Opinion mining, Sentiment analysis.

I. INTRODUCTION

Sentiment analysis has become a very interesting area of research. It uses a combination of processes so as to spot, acquire and differentiate emotions, assessments, and opinions around people, topics, ideas, experiences, facts, events, and their properties. Sentiment analysis has been used to detect new trends, to analyze user intents and to gain knowledge. It could be used on various data sources. Discovering users' sentiments and opinions is very profitable in a lot of application fields such as customer satisfaction, political opinions, predictive analytics and a lot more [14]. Sentiment analysis (evoked also as opinion analysis, opinion mining or sentiment mining) is a Natural Language Processing problem. It is a multidisciplinary research discipline whose theoretical bases encompass computer science, linguistics, semantics and many more.

Sentiment analysis of textual substance from microblogging platforms, forums, electronic businesses, or social media in general, faces many difficulties and challenges owing to the volume, the velocity and the heterogeneousness of the obtainable data [7]. In addition, deciphering sentiments and opinions is a serious problem. Internet users are often inclined to use ambiguous and even cryptic styles and ways of expression. These challenges generate many fascinating research problems, such as determining the influence of comments on consumer's trends, discerning the opinion strewing, or determining the online reputation of a product or a brand. Sentiment analysis is a diversified process with several steps containing opinionated data collection, feature selection and sentiment classification. Applying efficient feature selection incarnates a crucial role in enhancing the performance of sentiment analysis in term of accuracy.

Taking into account the new constraints imposed by characteristics of big data content, essentially linked to volume, velocity and variety, mounting procedure for sentiment identification and classification becomes arduous. For that reason, feature selection constitutes a key step for sentiment analysis in big data context.

As a dimension reduction technique, feature selection has proved efficient in processing large dimensional data. It is functioning by excluding features that are superfluous or are not congruent. However, feature subset selection leads, in many cases, a NP-hard (nondeterministic polynomial-time hard) problem, which has an effect on the system's efficiency [10]. In this field, intelligent algorithms have been commonly used over the last years to deal with complex problems and enhance classification research. Principally, nature-inspired algorithms have been widely used in this branch of research. During the classification process, accuracy is particularly taken into account by the nature-inspired algorithms. Nevertheless, some of these algorithms meet two important problems: outdated memory and diversity loss. However, these difficulties can be got over with enhanced approaches like Carousel greedy algorithm with Cat Swarm Optimization based Functional Link Artificial Neural Networks (CSO-FLANN), which can increase the performance of the system in term of accuracy. In addition, the Carousel greedy algorithm can examine a wider spectrum of solutions without inducing a consequent increase of computational cost [5].

This work introduces an enhanced big data and machine learning method which can be carried out in several sentiment mining contexts. The proposed Carousel greedy algorithm combined with CSO-FLANN attains performing accuracy compared to the usual Particle Swarm Optimization algorithm.

The layout of this paper is as follows. Section II is devoted to describe related works.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Khalid Ait Hadi:, Laboratory of Sciences and Advanced Technologies, Department of Computer Sciences, Polydisciplinary Faculty of Larache, Abdelmalek Essaadi University, Morocco.
Rafik Lasri, Laboratory of Sciences and Advanced Technologies, Department of Computer Sciences, Polydisciplinary Faculty of Larache, Abdelmalek Essaadi University, Morocco.
Abdellatif El Abderrahmani, Laboratory of Sciences and Advanced Technologies, Department of Computer Sciences, Polydisciplinary Faculty of Larache, Abdelmalek Essaadi University, Morocco.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

In Section III, we present the Carousel greedy feature selection process with CSO-FLANN algorithm and the related sentiment word extraction and sentiment classification methods. Experimental results and conclusion are given in Section IV.

II. LITERATURE SURVEY

Sentiment classification is a particular operation whose purpose is to classify a text in relation to the polarities of the opinions it encloses. This notion has attracted wide interest by dint of its possible openings and applications. Overcoming the insufficiency of labeled data is the prevailing concern of sentiment classification and the underlying challenge is to formulate a classification task that uses both labeled and unlabeled data such that generalization of the classification model can be improved.

In recent years, various techniques have been proposed to achieve improved results when labeled data are deficient. In [15] a co-training algorithm that leans on k-nearest neighbor technique is proposed. A Support Vector Machines algorithm is introduced in [2] to deal with the scarcity of labeled data.

In [13] a dynamic label propagation based semi-supervised multi-label classification scheme is presented, and a regularized Kernel Spectral Clustering based multi-class semi-supervised learning algorithm is used in [8].

Furthermore, an algorithm for multi-class semi-supervised that even takes advantage of a circumscribed amount of labeled data is proposed in [12]. Another technique of the type multiclass semi-supervised learning with Markov random walks is introduced in [3] to approximate a distribution over the lacking labels.

Through this literature review, it is deducted that growing accuracy is the main challenge encountered during execution. Here, Carousel greedy feature selection combined to an efficient classifier CSO-FLANN are implemented to perform sentiment mining of voluminous amount of data obtained from Amazon's website.

III. METHODOLOGY

This section is devoted to analyze and comprehend large amounts of customers' opinions, where different companies are using the digital platforms for promoting their products and where customers' online reviews can influence the purchase decisions of a product. Here, sentiment analysis concerns data extracted from e-commerce platform Amazon. The classification will be conducted by searching first the class attribute and discarding noise with the use of a preprocessing technique. The class associated with each sentiment is thereafter determined based on new datasets. Sentiment classification consists of assigning a sentiment, from a set of possible values, to a given portion of text, and class (or topic) detection consists of assigning a class from a set of predefined classes to a given content.

A. Opinionated data collection and preprocessing

The data collected from the Amazon platform consist of information containing appreciations, feedback and criticisms about various products, which are important resources for exploring user's opinions on products. The acquired information includes various instances of incongruous and insufficient information that can impact the efficiency of the whole sentiment analysis methodology. Thus, noisy data (values which deviate from the expected) are removed from the dataset by use of a min-max normalization technique, where the data are scaled so as to project it in a small interval. This technique eliminates noise and helps avoid attributes with sizable ranges from overflowing on those with feebler ranges. It performs a linear transformation on the original data with the following equation:

$$S_n = \frac{S - S_{\min}}{S_{\max} - S_{\min}}$$

where S is the set of attributes, $S_{\min}$ and $S_{\max}$ are the minimum and maximum values of attributes and S_n is the novel normalized data which vary between *0* to *1*. The normalization comes to replace existing noise in the dataset, and small values of ratings are considered to be inchoate and are substituted.

B. Fast feature selection using Carousel greedy approach

Here, feature selection is processed using Carousel greedy (CG) algorithm [5]. This choice was made taking into consideration that CG is efficient in enriching the performance of solutions produced by the greedy algorithm. CG method, which manages to get over the usual weaknesses of greedy approaches, can manipulate attributes adequately [4].

Over the selection process, the noise-free datasets are analyzed using several constituents and the best attributes are selected with the use of the following optimization problem:

$$\min J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c} \mu_{ij}^2 d_{ij}^2$$

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^2}, 1 \le i \le n, 1 \le j \le c$$

$$d_{ij} = \left\| x_i - v_j \right\| and\, v_j = \frac{\sum_{i=1}^{n} \mu_{ij}^2 x_i}{\sum_{i=1}^{n} \mu_{ij}^2}, 1 \le j \le c$$

where $U = \left\lfloor \mu_{ij} \right\rfloor_{n \times c}$ is the fuzzy matrix of mean values, $V = \{v_1, v_2, ..., v_c\}$ represents the c-vector of the variance of attributes, x_i is the existing attribute in the dataset and $\|.\|$ is the Euclidean distance.

The aforementioned problem can be solved with the help of the Lagrange multipliers and the Fuzzy C-Means technique. After selecting feasible features, the Butterfly Optimization Algorithm (BOA) is then employed to engender the optimal feature associated to the solution. BOA is a nature inspired metaheuristic algorithm based on the food foraging strategy of butterflies [1]. In BOA, it is admitted that a butterfly will produce a fragrance with certain concentration that is related to its fitness.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

The fragrance will spread beyond location and other butterflies can feel it, and this is how butterflies can share their individual details with other butterflies and generate some common social intelligence. In this algorithm, fragrance is related to the stimulus according to this equation:

$$f_i = \sigma I^a$$

where f_i is the amplitude of fragrance discerned by *i*th butterfly, σ is the sensory modality, I is the stimulus intensity and a represents the level of absorption. During the movement, the butterfly performs a step in the direction of the most suitable butterfly g^* that can be formulated as:

$$X_i^{t+1} = X_i^t + \left(g^* - X_i^t\right) f_i$$

with X_i^t is the position of *i*th butterfly in iteration number *t*.

Eventually and throughout the algorithm, optimal features are extracted and provided to classifier to spot sentiments from dataset.

C. Classification process

Before proceeding with classification, opinionated words extraction proves necessary.

C.1 Sentiment extraction

The appraisal is linked to the content expressed along a sentence. At this stage, word extraction based on nouns and verbs modifiers and descriptors is taken into consideration. Content is obtained by use of dynamic conditional random fields, described in [11].

C.2 CSO-FLANN classifier

Once the sentiment word extraction is done, the related sentiments can be recognized using CSO-FLANN algorithm, which consists of input and output gates, with maintainability of word-related connections so as to examine exact sentiments. Over the classification process, a gradient function is used to drive sentiment features. The features are analyzed based on the use of an activation function of the type sigmoid function, defined by:

$$Sig(x) = M\left(1 + e^{-sx}\right)^{-1}$$

where *M* is the features' maximum value, and *s* the steepness value.

All over the process, the calculated value is analyzed in comparison with the one of the driven feature, in the ultimate target to extract the exact sentiment word. Thus, connection is updated using the Cat Swarm Optimization (CSO) [6], where the behavior of a cat inspires the resolution of the optimization problem and cat progressions are partitioned into seeking and tracing phases.

In such a way, optimization process can be performed as follows:

$$V_{k,d} = \beta V_{k,d} + a_c\left(x_{best,d} - x_{k,d}\right)$$

where $V_{k,d}$ is the velocity at the *k*th position, β is the inertia weight, a_c is the acceleration constant, $x_{k,d}$ and $x_{best,d}$ are respectively the position and the position of the best fitness value of the feline.

The above process is reiterated up to classify all the opinionated features. The FLANN architecture [9] provides an expansion function to enhance the input vector dimensionality. It is also appreciated for its faster convergence rate and lower computational effort.

The algorithm can then be schematized as follows:

ALGORITHM

1. Sentiment-related data acquisition.
2. Elimination of the noisy data by normalization as follows: $S_n = \frac{S - S_{min}}{S_{max} - S_{min}}$
3. Fast features selection using Carousel greedy algorithm and several functions such candidate set, selection function, feasibility function, objective function and solution function.
4. Selection of the best attributes by solving the following optimization problem:
$$\begin{cases} \min\ J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c} \mu_{ij}^2 d_{ij}^2 \\ \mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^2},\ 1 \le i \le n,\ 1 \le j \le c \\ d_{ij} = \|x_i - \nu_j\| \ and\ \nu_j = \frac{\sum_{i=1}^{n} \mu_{ij}^2 x_i}{\sum_{i=1}^{n} \mu_{ij}^2},\ 1 \le j \le c \end{cases}$$
5. Based on the feasible features, selection of the optimal features according to the food foraging strategy of butterflies, represented as: $X_i^{t+1} = X_i^t + (g^* - X_i^t) f_i$
6. Extraction of sentiment-related words and processing by the Cat Swarm Optimization based Functional Link Artificial Neural Networks approach.
7. Examination of the output of each feature using sigmoid function: $Sig(x) = M\left(1 + e^{-sx}\right)^{-1}$
8. Throughout the classification process, execute the necessary updates and perform also the following one: $V_{k,d} = \beta V_{k,d} + a_c(x_{best,d} - x_{k,d})$
9. Repeat until the sentiment classification is successfully fulfilled.

IV. EXPERIMENTAL RESULTS AND CONCLUSION

In this section, the data were handled using the Java integrated development framework Eclipse. The efficiency of the CSO-FLANN algorithm is evaluated comparing to the performances accomplished with Particle Swarm Optimization (PSO) algorithm, and using precision, recall and accuracy metrics in different datasets. The performance of the proposed algorithm is as follows:

Datasets	CSO-FLANN			PSO		
	Precision	Recall	Accuracy %	Precision	Recall	Accuracy %
10^5	0.95	0.86	94.66	0.85	0.73	87
10^6	0.93	0.79	96.01	0.77	0.65	91
10^7	0.89	0.73	97.37	0.85	0.59	94
10^8	0.85	0.66	98.25	0.59	0.57	96

Table 1: Performance of the CSO-FLANN algorithm with PSO

It is perceived from Table 1 that accuracy rates obtained from the CSO-FLANN algorithm are better compared to those issued from the PSO algorithm. It is concluded therefore that the implementation of a Carousel greedy algorithm with CSO-FLANN obtains better outcomes compared with alternate techniques.

Consequently, this work presents an improved technique that can be explored to probe several cases of sentiment analysis in big data context.

REFERENCES

1. S. Arora and S. Singh, An improved butterfly optimization algorithm with chaos, Journal of Intelligent & Fuzzy Systems, 32, 1079–1088, (2017).
2. A. Astorino and A. Fuduli, Support vector machine polyhedral separability in semi-supervised learning, J. Optim. Theory Appl., 164, 1039–1050, (2015).
3. A. Azran, The rendezvous algorithm: multiclass semi-supervised learning with Markov random walks, Proc. of the 24th International Conference on Machine learning, 49–56, (2007).
4. W. Bednorz, Advances in greedy algorithms, Published by In-Teh, (2008).
5. C. Cerrone, R. Cerulli and B. Golden, Carousel greedy: a generalized greedy algorithm with applications in optimization, Computers & Operations Research, 85, 97–112, (2017).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

6. S. C. Chu and P. W. Tsai, Computational intelligence based on the behavior of cats, International Journal of Innovative Computing, Information & Control, 3 (1), 163–173, (2007).
7. B. Liu, Sentiment analysis and opinion mining, Morgan & Claypool, (2012).
8. S. Mehrkanoon, C. Alzate, R. Mall, R. Langone and J. A. K. Suykens, Multi-class semi-supervised learning based upon kernel spectral clustering, IEEE Trans. Neural Networks & Learning Syst., 26 (4), 720–733, (2015).
9. Y. H. Pao, S. M. Phillips and D. J. Sobajic, Neural-net computing and intelligent control systems, Int. J. Contr., 56, 263–289, (1992).
10. P. Raghavendra, Approximating NP-hard problems: efficient algorithms and their limits, PhD thesis, University of Washington, (2009).
11. C. Sutton, A. McCallum and K. Rohanimanesh, Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data, Journal of Machine Learning Research, 8, 693–723, (2007).
12. J. Tanha, M. Van Someren and H. Afsarmanesh, An AdaBoost algorithm for multiclass semi-supervised learning, Proc. 12th International Conference on Data Mining, 1116–1121, (2012).
13. B. Wang and J. Tsotsos, Dynamic label propagation for semi-supervised multi-class multi-label classification, Pattern Recognition 52, 75–84, (2016).
14. P. Wlodarczak, M. Ally and J. Soar, Opinion mining in social big data, http://dx.doi.org/10.2139/ssrn.2565426, (2015).
15. Z. Zhou and M. Li, Semi-supervised regression with co-training, Proc. 19th International Joint Conference on Artificial Intelligence, (2005).

© Copyright: All rights reserved.