


Product Review-Based Customer Sentiment Analysis Using an Ensemble of mRMR and Forest Optimization Algorithm (FOA)

Parag Verma, Uttarakhand University, Dehradun, India*

 <https://orcid.org/0000-0002-3201-4285>

Ankur Dumka, Computer Science and Engineering, Women Institute of Technology, Dehradun, India

Anuj Bhardwaj, Computer Science and Engineering, Chandigarh University, Punjab, India

Alaknanda Ashok, College of Technology, G. B. Pant University of Agriculture and Technology, India

ABSTRACT

This research presents a feature selection problem for classification of sentiments that uses ensemble-based classifier. This includes a hybrid approach of minimum redundancy and maximum relevance (mRMR) technique and forest optimization algorithm (FOA) (i.e., mRMR-FOA)-based feature selection. Before applying the FOA on sentiment analysis, it has been used as feature selection technique applied on 10 different classification datasets publicly available on UCI machine learning repository. The classifiers for example k-nearest neighbor (k-NN), support vector machine (SVM), and naïve Bayes used the ensemble based algorithm for available datasets. The mRMR-FOA uses the Blitzer's dataset (customer reviews on electronic products survey) to select the significant features. The classification of sentiments has improved by 12-18%. The evaluated results are further enhanced by the ensemble of k-NN, NB, and SVM with an accuracy of 88.47% for the classification of sentiment analysis task.

KEYWORDS

Classification Techniques, Feature Selection, Forest Optimization Algorithm, K-Nearest Neighbor, Minimum Redundancy and Maximum Relevance (mRMR) Technique, Sentiment Analysis, Supervised Learning

1. INTRODUCTION

Over the past few years, the dataset dimensionality has been increased in various domains like text-based sentiment analysis or bioinformatics.(Zhai et al., 2014) This reality has brought an intriguing challenge to the research field as much Artificial Intelligence (AI) or Machine Learning (ML) methods unable to manage high dimensional input data that involve products. Indeed, on the occasion that we examine the dimensionality of data posted in the well-known UCI repository and libSVM database,(Chang, 2001) we can see that the largest dimensionality of the dataset has expanded to over 30 million (approximately). Therefore, a part of these calculations is additionally when they face larger instance sizes. In this new situation, it is usual to manage information collection that is much larger than both the number of highlights and the number of tests, so current learning techniques must be adjusted.

DOI: 10.4018/IJAMC.2022010107

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

To address this issue, dimension reduction methods can be applied to reduce the number of features and to enhance the performance of the resulting learning process. One of the most frequently used dimensionality reduction processes is the feature selection (FS), which accomplishes dimensionality reduction by emptying abstracts and additional features. (Liu & Motoda, 1998) Since FS places the highlights first, it is particularly valuable for applications where model translation and information extraction are important. In any case, existing FS techniques are not expected to scale well when managing a large-scale problem (in both various highlights and cases), in such a way that their effectiveness may be fundamentally broken or they can also be insignificant.

An analysis of sentiments is a way of identifying and classifying the emotions or opinions stated in some piece of text, sentence specifically in order to determining polarity whether the writer's disposition towards a particular topic or artefact is positive, negative, or neutral. For this purpose sentiment analysis and classification uses machine learning (ML) systems and natural language processing (NLP) together. The prevalence of rapid growth on the online social media and electronic network based societies provides all possible outcomes for customers to express their perceptions and exchange their ideas about entirety, for example, social or political issues through any article, book and films and so on through web-based networked media. These are usually in the form of survey material such as Likert type scaling data or text. Nowadays organizations are very fast, they evaluate popular perceptions about their customers or their articles of Internet-based social content. (Parvathy & Bindhu, 2016) Specific online service provider organizations are hooked in the evaluation of social media data in blogs, online forums, tweets, comments, and product feedback surveys. Publically shared reviews on sites or articles are used to recognize a customer's continued perception of any product or services to maintain a good commercialization with their decision making or the nature of its services or product quality. (Stylios et al., 2014) The critical problem that arises when collecting information from a social media networking environment is that the reviews consists mostly a large amount of unwanted data, including of HTML tags, linguistic and spelling errors, and the data is usually so bulky that removing those errors is human typical and time consuming task. An efficacious approach required to solving this problem is to select the usually relevant and significant features from the dataset and dispense repetitive or immaterial features. There are some pre-processing data cleaning techniques that rely on the choice of features selection. In the data mining process for high-dimensional dataset feature selection works as a highly effective pre-preparation strategy. Taxonomy of methods of feature selection present in Figure 1.

In the case of mining in social networks dataset, the analysis of high-dimension data is even more common. Classification of such high-dimensional dataset with a reasonable computational cost has become a vital topic of research in recent years. Most of the proposed solutions for analysis the sentiments are based on prior data processing or classification techniques to improve classification accuracy. In the same sequence Genetic algorithms (GA) and particle swarm optimization (PSO) have been used to select feature subsets from Artificial Intelligence (AI) domain. GA and PSO-based solutions have upgraded classification accuracy, but these substitute solutions are computationally high expensive due to that it affects performance of the system. GA and PSO are meta-heuristic algorithms that use a population of primary solutions. They can be used for problem optimization. This research paper presents and evaluates an ensemble based classification technique for sentiment analysis by using a newly developed evolutionary algorithm, called a Forest Optimization Algorithm (FOA). (Ghaemi & Feizi-Derakhshi, 2014) Ghaemi et al. proposed feature optimization algorithm named Forest Optimization Algorithm (FOA) in 2014, (Ghaemi & Feizi-Derakhshi, 2014) while its improved version named Feature Selection using Forest Optimization Algorithm (FSFOA) in 2016. (Ghaemi & Feizi-Derakhshi, 2016) Sowing and limiting populations based on lifetime in the tree process is simulated in these algorithm. The FOA produces the best tree (subset of features) among all other trees based on performance. The FOA supersedes GA and PSO when applied to reference functions and has the problem of optimizing weighting features using constant weights.

19 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/product-review-based-customer-sentiment-analysis-using-an-ensemble-of-mrmm-and-forest-optimization-algorithm-foa/284578

Related Content

Matheuristics for Inventory Routing Problems

Luca Bertazziani and M. Grazia Speranza (2012). *Hybrid Algorithms for Service, Computing and Manufacturing Systems: Routing and Scheduling Solutions* (pp. 1-14).

www.irma-international.org/chapter/matheuristics-inventory-routing-problems/58514

Short Term Hydro-Thermal Scheduling Using Backtracking Search Algorithm

Koustav Dasgupta and Provas Kumar Roy (2020). *International Journal of Applied Metaheuristic Computing* (pp. 38-63).

www.irma-international.org/article/short-term-hydro-thermal-scheduling-using-backtracking-search-algorithm/262129

A Survey of Ant Colony Optimization Algorithms for Telecommunication Networks

Ilham Benyahia (2012). *International Journal of Applied Metaheuristic Computing* (pp. 18-32).

www.irma-international.org/article/survey-ant-colony-optimization-algorithms/67331

Exploring a Self Organizing Multi Agent Approach for Service Discovery

Hakima Mellah, Soumya Banerjee, Salima Hassas and Habiba Drias (2010). *Evolutionary Computation and Optimization Algorithms in Software Engineering: Applications and Techniques* (pp. 127-141).

www.irma-international.org/chapter/exploring-self-organizing-multi-agent/44373

An Optimal Production Plan for Cashew Nuts Community Enterprise Using Metaheuristic Algorithms

Apisak Phromfaity, Natita Wangsohand and Prayoon Surin (2022). *International Journal of Applied Metaheuristic Computing* (pp. 1-23).

www.irma-international.org/article/an-optimal-production-plan-for-cashew-nuts-community-enterprise-using-metaheuristic-algorithms/292514