


Robustness and Predictive Performance of Homogeneous Ensemble Feature Selection in Text Classification

Poornima Mehta, Jaypee Institute of Information Technology, Noida, India

 <https://orcid.org/0000-0002-8137-9504>

Satish Chandra, Jaypee Institute of Information Technology, Noida, India

ABSTRACT

The use of ensemble paradigm with classifiers is a proven approach that involves combining the outcomes of several classifiers. It has recently been extrapolated to feature selection methods to find the most relevant features. Earlier, ensemble feature selection has been used in high dimensional, low sample size datasets like bioinformatics. To one's knowledge there is no such endeavor in the text classification domain. In this work, the ensemble feature selection using data perturbation in the text classification domain has been used with an aim to enhance predictability and stability. This approach involves application of the same feature selector to different perturbed versions of training data, obtaining different ranks for a feature. Previous works focus only on one of the metrics, that is, stability or accuracy. In this work, a combined framework is adopted that assesses both the predictability and stability of the feature selection method by using feature selection ensemble. This approach has been explored on univariate and multivariate feature selectors, using two rank aggregators.

KEYWORDS

Data Perturbation, Ensemble Feature Selection, Predictive Performance, Selection Stability, Sentiment Analysis, Sentiment Lexicon, SentiWordNet, Text Classification

INTRODUCTION

Feature selection helps reduce the dimensionality of a dataset by selecting a subset of attributes that are capable enough to represent the knowledge of the entire dataset. Methods used to carry out data classification gain a lot from feature selection once the redundant, noisy and irrelevant attributes are removed from the dataset. With the availability of so many feature selection techniques choosing a suitable technique for a given dataset is a challenge. Generally, the criterions used to gauge the effectiveness of an attribute selection method are the classification performance and algorithm complexity. Lately the stability of the attribute selection method has also become a metric for the same (Awada, Khoshgoftaar, Dittman, Wald, & Napolitano, 2012). This is due to the requirement in real world applications of choosing a similar set of attributes each time the method is used on the same dataset with a little perturbation (Kalousis, Prados, & Hilario, 2007). Stability is the measure

DOI: 10.4018/IJIRR.2021010104

of the robustness of the results, when the dataset composition is changed. Robustness implies that removing or adding a small percentage of data instances should not affect the set of features selected in a significant way. Unstable feature selection methods reduce the confidence of domain experts and confuse them. Researchers have been working in the area of stability and classification performance of feature selection algorithms in the genetic analysis (or bioinformatics) domain. However till date, there is lack of work on stability in the *text classification domain* and is an open area of research in future. Dittman, Khoshgoftaar, Wald, & Napolitano, 2013 studied the effect of dataset difficulty on the stability of feature selection method in the domain of gene selection from high dimensional micro-array datasets.

It has been demonstrated that there is no lone optimal attribute selection method and that quit a few subsets of attributes are capable of discerning the data just the same (Saeys, Abeel, & Van de Peer, 2008). Instead of using a particular attribute selection technique and using its output, *ensemble approach* may be used by combining the subsets obtained from various attribute selection methods. The various feature subsets obtained from different methods can be considered as local optima in the feature subset space, whereas the ensemble feature selection may give a more favourable subset (Saeys et al., 2008).

Creating a feature selection ensemble includes (i) Deciding the ensemble components that produce different subsets of features; (ii) Aggregation of the above subsets to a final feature subset based on some rank aggregation strategy.

The data perturbation approach for creating feature selection ensemble involves application of the same feature selector to different perturbed versions of the training data, obtaining different ranks for the same feature. This is followed by rank aggregation procedure to obtain a final rank for each feature.

Contribution

This work involves analysing the effect of ensemble attribute selection using data perturbation approach in the *text classification* domain. The aim is to study together, both the robustness and predictability of the ensemble model in the text classification domain. The effect on predictive performance and stability using the feature selection ensemble version has been studied on four feature selectors – Information Gain(IG), Maximum Relevance Minimum Redundancy(mRMR), ReliefF and oneR, using two rank aggregation functions - *mean* and *median*. The IG and oneR methods represent univariate methods whereas the ReliefF and mRMR represent multivariate methods. The authors have also compared the performance of the features selected through ensemble attribute selection of IG and mRMR methods to that of the features chosen using their earlier proposed feature selection method - *Normalized IG and CHI feature selection (NICFS)* (Mehta and Chandra, 2019). The performance results of NICFS were better than the results of the ensembles of IG and mRMR methods created. A diversity study among the four rankers was also conducted to ensure that the rankers chosen in this work are diverse in behaviour. The stability of the four feature selection methods was also compared in which IG turned out to be the most stable method.

LITERATURE SURVEY AND BACKGROUND

Based on the way they interact with the classifier feature selection methods may be classified into filter, wrapper and embedded (Guyon, Gunn, Nikravesh, & Zadeh, 2008). From the values of the attribute and the target class, filter methods calculate the score of the attribute, independent of the classifier. The output may possibly be the attribute weight, attribute rank or the attribute subset. The wrapper methods search through the space of subsets of attributes, guided by the performance of the associated classifier. Wrappers are computationally costly but yield better performance than filter methods. Lastly, embedded methods use the internal parameters of a classifier while performing feature selection. These obtain a good trade-off between computational cost and performance. The

13 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/robustness-and-predictive-performance-of-homogeneous-ensemble-feature-selection-in-text-classification/267123

Related Content

Measuring Followership

Paul Kaak, Rodney A. Reynolds and Michael Whyte (2013). *Online Instruments, Data Collection, and Electronic Measurements: Organizational Advancements* (pp. 245-253).

www.irma-international.org/chapter/measuring-followership/69744

Debunking Intermediary Censorship Framework in Social Media via a Content Retrieval and Classification Software

Baramée Navanopparatskul, Sukree Sinthupinyo and Pirongrong Ramasoota (2013). *International Journal of Information Retrieval Research* (pp. 1-26).

www.irma-international.org/article/debunking-intermediary-censorship-framework-in-social-media-via-a-content-retrieval-and-classification-software/93184

Latent Topic Model for Indexing Arabic Documents

Rami Ayadi, Mohsen Maraoui and Mounir Zrigui (2014). *International Journal of Information Retrieval Research* (pp. 57-72).

www.irma-international.org/article/latent-topic-model-for-indexing-arabic-documents/126329

Improving Search and Navigation User Experience by Making Use of Social Data

Mario Cataldi, Luigi Di Caro and Claudio Schifanella (2018). *Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications* (pp. 2132-2156).

www.irma-international.org/chapter/improving-search-and-navigation-user-experience-by-making-use-of-social-data/198640

Comparison Analysis of GLCM and PCA on Parkinson's Disease Using Structural MRI

Sanjana Tomer, Ketna Khanna, Sapna Gambhir and Mohit Gambhir (2022). *International Journal of Information Retrieval Research* (pp. 1-15).

www.irma-international.org/article/comparison-analysis-of-glcm-and-pca-on-parkinsons-disease-using-structural-mri/289577