Chapter 61
Robustness and Predictive Performance of Homogeneous Ensemble Feature Selection in Text Classification

Poornima Mehta

https://orcid.org/0000-0002-8137-9504
Jaypee Institute of Information Technology, Noida, India

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.

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 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*

Related Content

A Web Database IR-PDB for Sequence Repeats of Proteins in the Protein Data Bank
Selvaraj Samueland Mary Rajathei (2017). *International Journal of Knowledge Discovery in Bioinformatics (pp. 1-10).*
www.irma-international.org/article/a-web-database-ir-pdb-for-sequence-repeats-of-proteins-in-the-protein-data-bank/190790

Human Oral Bioavailability Prediction of Four Kinds of Drugs
Aixia Yan, Zhi Wang, Jiaxuan Liand Meng Meng (2011). *Interdisciplinary Research and Applications in Bioinformatics, Computational Biology, and Environmental Sciences (pp. 141-154).*
www.irma-international.org/chapter/human-oral-bioavailability-prediction-four/48372

Using Blockchain Technology, Artificial and Natural Neural Network in Bioinformatics
Lubov A. Belyanina (2019). *International Journal of Applied Research in Bioinformatics (pp. 42-49).*
www.irma-international.org/article/using-blockchain-technology-artificial-and-natural-neural-network-in-bioinformatics/237200

Image Processing Tools for Biomedical Infrared Imaging
Gerald Schaeferand Arcangelo Merla (2010). *Biocomputation and Biomedical Informatics: Case Studies and Applications (pp. 187-197).*
www.irma-international.org/chapter/image-processing-tools-biomedical-infrared/39612

Resolving Sample Traces in Complex Mixtures with Microarray Analyses
George I. Lambrou, Eleftheria Koultouki, Maria Adamakiand Maria Moschovi (2013). *Bioinformatics: Concepts, Methodologies, Tools, and Applications (pp. 1025-1071).*
www.irma-international.org/chapter/resolving-sample-traces-complex-mixtures/76108