Journal of Industrial Information Technology and Application *ISSN(Online): 2586-0852*
Vol.6 No.4

Ensemble Feature Selection Using the Evaluation based on Distance from Average Solution (EDAS) Method

A New Application of Multiple Criteria Decision Making to Binary Classification

Dharyll Prince M. Abellana [1, *], Robert R. Roxas [1], Demelo M. Lao[1] , Paula E. Mayol[1]

[1] Department of Computer Science, College of Science, University of the Philippines Cebu, Cebu City, 6000 Cebu, Philippines

Abstract. This paper investigates the applicability of multiple criteria decision making in ensemble feature selection. This paper adopts the evaluation based on distance from average solution (EDAS) method. Results show that the proposed ensemble FS algorithm was able to reduce the dataset without compromising the performance of the classifier. The findings in this study would contribute to the literature in several ways. For one, the paper is one of the very few works to demonstrate how MCDM can be used in feature selection. Moreover, this paper is the first to demonstrate the applicability of EDAS as an ensemble FS algorithm. As such, the findings in this paper could spark the cross-fertilization of feature selection and MCDM.

Keywords; ensemble feature selection; multiple criteria decision making; binary classification

1. Introduction

In data mining, classification is one of the most tackled machine learning tasks in several domains [1]. However, with datasets becoming high dimensional over the years, the curse of dimensionality poses several roadblocks for scholars and practitioners in the field [2]. Several scholars consider dimensionality reduction as the most straightforward approach in addressing the curse of dimensionality [3]. In the current literature, there are two major ways for performing dimensionality reduction: (i) feature projection (FP) and (ii) feature selection (FS) [3]. FP transforms data from high dimensional space to a lower dimensional space, while retaining the relationships

* Corresponding author: dmabellana@up.edu.ph
Received: Feb 6, 2022; Accepted: Feb 28, 2022; Published: Dec 30, 2022

between the original features [4]. For example, principal component analysis generates new features using the linear combinations of the original features [4]. While this technique can be useful for summarizing the original features using fewer variables, the generated features are usually not straightforward to interpret. Thus, when interpretation is crucial in the modelling process, it is usually not preferred. By contrast, FS selects a subset of the original features that best represents the original features [3]. For example, the features can be ranked according to their association with the class, and then retain only the top 10% in the reduced dataset. Because this approach only selects a subset of the original features, interpretation of the model would be relatively straightforward as no new features are created [3]. Thus, this approach is preferable in many modelling scenarios. Filter-based FS is a type of FS that relies on the characteristics of the training data to select features [3]. Unlike other FS algorithms, it does not require the training of a classifier to evaluate a feature [3]. This makes filter-based FS algorithms computationally tractable [3]. Due to this, however, its evaluation is largely limited to the extent of how a chosen filter captures the characteristics of the data. For example, information gain might be better than correlation for capturing characteristics of some features, while correlation might be better for the other features. Hence, the choice of filter is a crucial aspect for filter-based FS. It is then imperative to include as many filters as possible to better capture the characteristics of the data. In the current literature, scholars tackled this problem using ensemble feature selection. As such, this type of FS adopts an ensemble (or a group) of filters to evaluate the features. While efforts have been made in this area, a major gap is the lack of a comprehensive framework for combining different filters. For example, using a simple average of the ensembles is insensitive to skewed distributions. Contrariwise, a weighted average does not have a mechanism for handling contradicting filters (e.g., redundancy vs. relevance). Therefore, there is a compelling need for more systematic aggregation methods. In this study, the use of a multiple criteria decision making (MCDM) method is explored.

2. Methodology

This section presents the case background and the procedure for the proposed ensemble FS algorithm.

A. Case background

In this study, one dataset and one classification algorithm are used to test the performance of the ensemble FS algorithm. The dataset used is the "LSVT Voice Rehabilitation Dataset" from [5]. The dataset consists of 309 features and 126 instances. All features are continuous. To reduce noise in the data, each feature is discretized using kernel density estimation. Four filters: (i) feature relevance (using Kendall τ), (ii) feature redundancy (using Kendall τ), (iii) symmetric uncertainty, and (iv) Relief-F, are used to compose the ensemble. A decision tree (DT), with the classification and regression tree (CART) algorithm, is employed as the classifier for the experiment. The control group uses the original dataset, while three experimental groups use the filtered

dataset with different proportions retained: (i) 10% of the features, (ii) 15% of the features, and (iii) 25% of the features. Due to violations from normality, the Kruskal-Wallis test is used to determine the presence of significant differences between the experimental groups and the control group.

B. Evaluation based on Distance from Average Solution (EDAS)

EDAS is an MCDM method that evaluates the alternatives (i.e., features) using two measures: (i) positive distance from average (PDA) and (ii) negative distance from average (NDA) [6]. The best alternative has the highest PDA value or lowest NDA value. Let there be n features and m filters, then the decision matrix is $X = \left[x_{ij}\right]_{n\times m}$, where x_{ij} is the score of feature i under filter j. The average solution under each filter is stored in vector $AV = \left[AV_j\right]_{1\times m}$, where $AV_j = {\sum_{i=1}^{m} x_{ij}}/{n}$. The PDA and NDA are then constructed from the beneficial and cost criteria. The beneficial criteria are the filters that are to be maximized: (i) feature relevance, (ii) symmetric uncertainty, and (iii) Relief-F. The *non-beneficial* criteria are the filters that are to be minimized. Feature redundancy is the only *non-beneficial* criterion in this paper.

Let $PDA = \left[PDA_{ij}\right]_{n\times m}$ and $NDA = \left[NDA_{ij}\right]_{n\times m}$. The PDA_{ij} and NDA_{ij} are calculated depending whether or not filter j is a beneficial or cost criterion. If filter j is a beneficial criterion, then $PDA_{ij} = max\ (0,\ x_{ij} - AV_j)/AV_j$ and $NDA_{ij} = max\ (0,\ AV_j - x_{ij})/AV_j$. Otherwise, $PDA_{ij} = max\ (0, AV_j - x_{ij})/AV_j$ and $NDA_{ij} = max\ (0,\ x_{ij} - AV_j)/AV_j$. The final appraisal score (i.e., used for ranking the features) for feature i is calculated as follows:

$$AS_i = \frac{1}{2}\left(\frac{SP_i}{\max\limits_{i\in\{1,\dots,n\}} SP_i} - \frac{SN_i}{\max\limits_{i\in\{1,\dots,n\}} SN_i} + 1\right), 0 \le AS_i \le 1$$

The filters are ranked from highest to lowest AS_i value. A more detailed procedure for EDAS is provided by [6].

3. Results and Discussion

In this section, the results of the paper are presented. For brevity, the following coding scheme is used throughout the analysis: (i) *Naïve* pertains to the control group in which all features of the original dataset are used to train the classifier, (ii) *EDAS_10* pertains to using only the first 10% of the ranked features to train the classifier, (iii) *EDAS_15* pertains to using only the first 15% of the ranked features to train the classifier, and (iv) *EDAS_20* pertains to using only the first 20% of the ranked features to train the classifier. This coding scheme is used in Figure 1 and Table 1. The control group and the experimental groups are inspected by plotting each distribution on a violin plot as shown in Figure 1. The descriptive statistics are presented in Table 1. It can be seen in Figure 1 that each distribution is highly skewed and does not conform to a normal distribution. As such, using the analysis of variance (ANOVA) is not suitable.

Furthermore, the Shapiro-Wilk test on the residual of the fitted linear model resulted into a p-value equal to 0.03575, which provides evidence of the residual's non-normality at a significance level of $\alpha = 0.05$.

Table 1. Summary Statistics. Due to the skewness of the distributions, the median is used as the center and the interquartile range is used as the measure of dispersion in describing the distribution of AUC. In the *Mean Rank* column, the AUC are ranked across the groups from least to greatest considering ties that occurred.

Treatment and Control Group	Median (AUC)	Interquartile Range (AUC)	Mean Rank (AUC ranks)	n
Naive	0.792	0.253	21.55	10
EDAS_10	0.763	0.089	16.8	10
EDAS_15	0.764	0.147	22.55	10
EDAS_20	0.757	0.142	21.1	10

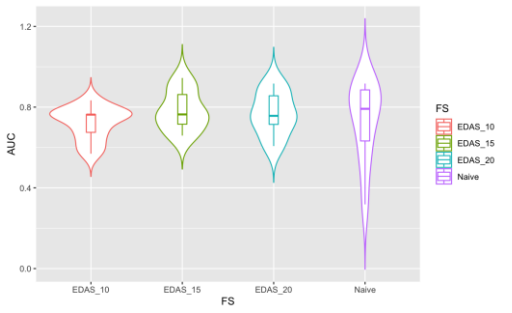

Figure 1. Violin Plot. The shape of each plot depicts the kernel density of the data in each group. A box plot is also presented within each violin plot. The vertical axis of the plot represents the area under the curve (AUC) of the classifier's receiver operating characteristic (ROC) curve.

Using the Kruskal-Wallis test in R, the test results into a Chi-square (χ^2) statistic equal to 1.4252 and a p-value equal to 0.6996, which implies that there is insufficient evidence to reject the null hypothesis that the groups are similar. Drawing inference from this result, it can be said that the proposed ensemble algorithm was able to reduce the dataset without significantly compromising the performance of the classifier in terms of AUC. While this finding is compelling, some limitations still exist. Firstly, since only the Kruskal-Wallis test is used to infer about the groups, the interpretation was limited to the mean ranks instead of the mean AUC. If a parametric test was performed, the result would have obtained higher statistical power. Secondly, only one dataset and one classifier were considered in the analysis. As such, the interpretation is limited only to the considered conditions. Finally, only a fixed factor was considered

in the analysis. Hence, it is not possible to infer beyond the levels adopted. For example, it is not possible to infer about the performance of the proposed ensemble algorithm when retaining only 12% of the features or any other proportion that is not considered in the analysis.

4. Conclusion and Future Works

In this study, an ensemble FS algorithm was explored using the EDAS method of MCDM. Results showed that proposed ensemble FS algorithm could select smaller subsets of the original features without significantly compromising classification performance. The findings in this paper would significantly contribute to the literature. Firstly, it is one of the very few papers that investigate the applicability of MCDM to ensemble FS. Secondly, it is the first to demonstrate the applicability of EDAS as an ensemble FS algorithm. While results in this study are compelling, some limitations are present. For one, the design of experiment uses only one dataset and one classifier for evaluating the performance of the proposed ensemble FS algorithm. Future works could expound on this and adopt additional classifiers and datasets. As such, additional factors may be incorporated in the analysis. For another, other MCDM models could be explored to address current limitations such as the fixation on the choice of weights. Finally, future works could evaluate the proposed ensemble FS algorithm using additional performance measures such as stability.

References

[1] Kang, M., & Jameson, N. J. (2018). Machine Learning: Fundamentals. Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things, 85-109.

[2] Debie, E., & Shafi, K. (2019). Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses. *Pattern Analysis and Applications*, *22*(2), 519-536.

[3] Guru, D. S., Suhil, M., Raju, L. N., & Kumar, N. V. (2018). An alternative framework for univariate filter based feature selection for text categorization. *Pattern Recognition Letters*, *103*, 23-31.

[4] Mahmoudi, M. R., Heydari, M. H., Qasem, S. N., Mosavi, A., & Band, S. S. (2021). Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries. *Alexandria Engineering Journal*, *60*(1), 457-464.

[5] Tsanas, A., Little, M. A., Fox, C., & Ramig, L. O. (2013). Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *22*(1), 181-190.

[6] Dhanalakshmi, C. S., Madhu, P., Karthick, A., Mathew, M., & Kumar, R. V. (2020). A comprehensive MCDM-based approach using TOPSIS and EDAS as an auxiliary tool for pyrolysis material selection and its application. *Biomass Conversion and Biorefinery*, 1-16.