The Fisher Component-based Feature Selection Method

Abdul Baseer Buriro
Department of Electrical Engineering
Sukkur IBA University
Sukkur, Pakistan
abdul.baseer@iba-suk.edu.pk

Suresh Kumar
Department of Computer Systems Engineering
Sukkur IBA University
Sukkur, Pakistan
suresh@iba-suk.edu.pk

Received: 11 June 2022 | Revised: 23 June 2022 | Accepted: 26 June 2022

Abstract-A feature selection technique is proposed in this paper, which combines the computational ease of filters and the performance superiority of wrappers. The technique sequentially combines Fisher-score-based ranking and logistic regression-based wrapping. On synthetically generated data, the 5-fold cross-validation performances of the proposed technique were compatible with the performances achieved through Least Absolute Shrinkage and Selection Operator (LASSO). The binary classification performances in terms of F1 score and Geometric Mean (GM) were evaluated over a varying imbalance ratio of 0.1:0.9 – 0.5:0.5, a number of informative features of 1 – 30, and a fixed sample size of 5000.

Keywords-feature selection; regularization; dimensionality reduction; class imbalance

I. INTRODUCTION

With the advent of low-cost electronics, the dimensions and complexity of datasets, e.g. the Internet of Things (IoT), text classification, and medical imaging, are constantly increasing. High dimensional datasets, compared to their low dimensional counterparts, require more processing time and more space complexity, which is known as the curse of dimensionality. Besides, the presence of irrelevant and/or redundant features leads to less-interpretable and over-fitted learning models. The issue becomes more severe when the number of features is more than the number of instances/samples like gene selection in microarray data, where typically the data examples are fewer than 100 and the number of raw features ranges from 6,000 to 60,000 [1]. In such a situation, a trained model may become useless as it would result in infinitely high variance [2]. Furthermore, the covariance/correlation-based methods like Principal Component Analysis (PCA), Fisher Discriminant Analysis (FDA), and linear regression become ill-posed problems. A common way to address the issue is to select the informative and non-redundant features from a known feature set, which results in lower dimension datasets. The use of such datasets improves the generalized accuracy of a trained model, reduces computational time, and improves model interpretability.

Three categories of feature selection, i.e. filter, wrapper, and embedded methods, are widely been used. The filters measure the relevance (e.g., Fisher score, mutual information, correlation [3-5]) of a feature with the target. Filters are computationally fast, simple, easy to scale, and generally don't require a learning model [6]. A filter, however, generally ignores feature dependencies and does not consider combined discriminatory power, and consequently, can result in suboptimal feature selection [7]. Wrappers evaluate the usefulness of features' subsets by training and testing a model on them, and generally use a cross-validation method. Forward feature selection, backward elimination, and Recursive Feature Elimination (RFE) are the most common used wrapper methods. The optimal feature combination that maximizes the overall performance is determined by adding and/or removing features. Wrapper methods are, however, computationally expensive. Greedy (best subset) selection requires 2^p, whereas, both forward and backward feature selection techniques require at least $\frac{p(p+1)}{2}$ models/iterations [2]. Embedded (also known as shrinkage) methods intrinsically maximize the overall performance during the training/learning of a model, like a Least Absolute Shrinkage and Selection Operator (LASSO) or $l1$ regularization [4]. However, not all the models such as k-Nearest Neighbors (kNN), and decision tree incorporate $l1$ regularization during their training.

Contrary to feature selection techniques, dimensionality reduction techniques project the original input data into a lower-dimensional feature space and don't require learning/training. PCA is a classical data analysis and dimensionality reduction technique, which captures the maximum variability in data. PCA transforms the feature space into a new (meta) orthogonal feature space of the same dimension. Its effectiveness is, however, limited to unsupervised problems. FDA on the other hand is a traditional supervised dimensionality reduction technique, which respectively maximizes and minimizes the between-class distance and within-class mean distance. Contrary to PCA, the dimension of the FDA-transformed data is one less than the number of classes (i.e. $C-1$) [8, 9].

Authors in [4] ranked the features per the sum of absolute values of the coefficients of the first two principal components. The ranked features were then evaluated with a wrapper. They

Corresponding author: Abdul Baseer Buriro

compared the results with the ones from analysis of variance (ANOVA), absolute correlation, and classifier-based ranking, and achieved higher performance. Since PCA is unsupervised, their method may underperform in real datasets when the classification-related information is different from the direction of the maximum variance. Several feature selection techniques have been published and implemented in well-known machine learning libraries like Python's sklearn, and Matlab. These techniques, however, uniquely behave with varying datasets, like different models select different features in standard wrapper feature selection techniques. For class imbalanced datasets, most of the classification models [10] and, subsequently, the wrapper and embedded feature selection techniques, perform poorly, i.e. may not accurately determine all informative features. The identification of the most suitable features with varying class imbalance ratios is therefore of much importance and requires a solid solution.

In this paper, a feature selection technique is proposed, that combines the computational ease of filters and the performance superiority of wrappers. The technique sequentially combines Fisher-score-based ranking and logistic regression-based wrapping. In doing so, the proposed technique requires the training of a maximum of p models. On synthetically generated data, the proposed technique gave a very compatible mean of 5-fold cross-validation F1-scores and Geometric Means (GM) to LASSO. The binary classification performances were evaluated over a variable imbalance ratio of 0.1:0.9 – 0.5:0.5.

II. METHODS AND EXPERIMENTAL DESIGN

A simulation study was conducted to examine the effect of informative and non-informative features on feature selection techniques. Like [4], synthetic data of 30 features with varying imbalance rates and noise (i.e. non-informative features), were generated using the "make classification" library in python's scikit-learn.datasets. Each of the binary classes was a single cluster. The information features were drawn independently from the normal $(0,1)$ distribution for each class and then combined as random linear combinations within each cluster to add covariance [11]. The non-informative features were represented by random noise. The datasets were generated with varying number of information features from 1 – 30, and with a step of 2. The imbalance rates were changed from 0.1:90 – 0.5:0.5, and the sample size was fixed to 5000 samples. The model was validated using 5-fold cross-validation.

A. Feature Ranking

Feature ranking is to place the features in an order per their importance in order to identify the most important features. For this purpose, FDA was used. For binary classification, FDA results in one component. The features were ranked according to the absolute values of their respective coefficients in the FDA component. Eigenvalues and eigenvectors respectively indicate the explained variance of an FDA component and the linear combination of the original features, i.e. $FC = w_1x_1 + w_2x_2 + \cdots + w_px_p$, where x_j is j^{th} feature and w_j is corresponding weight of j^{th} feature. Eigenvalues and eigenvectors are computed from $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B$, where $\boldsymbol{S}_W$ and $\boldsymbol{S}_B$ are respectively the within-class and between-class scatter matrix, defined as:

$$\boldsymbol{S}_B = \sum_{j=1}^{c} n_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T \quad (1)$$

$$\boldsymbol{S}_W = \sum_{\boldsymbol{x}\in D_j}^{c} (\boldsymbol{x} - \boldsymbol{\mu}_j)(\boldsymbol{x} - \boldsymbol{\mu}_j)^T \quad (2)$$

where $\boldsymbol{\mu}_j$, n_j , and c are the mean vector, the number of instances (observations) for class j, and the total number of classes respectively.

B. Feature Selection

The ranked features were further sequentially evaluated using the logistic regression-based wrapper, shown in Figure 1. The top-ranked feature was first fed to the classifier and the classification performance due to class imbalance rate was recorded in terms of F1 score. Then iteratively, the next consecutively ranked feature was added. At each iteration, a feature was added to the subset if its combined mean 5-fold cross-validation classification performance was higher than the previous iteration. The feature was otherwise discarded. The value of C was kept high (i.e. 1×10^9) to avoid inherent LASSO or ridge regularization by the classifier.

C. LASSO

LASOO, defined as $\|\beta\|_1 = \lambda \sum_{j=1}^{p} |\beta_j|$, is a practical strategy to perform feature selection and classification simultaneously. $\|.\|_1$ and $\lambda > 0$ are the $l1$-norm and regularization parameter respectively. LASSO selects the features by generating a sparse weight vector (few non-zero coefficients). Consequently, LASSO reduces the variance and the intricacy of the model, and therefore, is an alternative feature selection technique [12]. Non-zero coefficients indicate the importance of the respective features, whereas, zero coefficients indicate irrelevant and redundant features. The number of non-zero coefficients are roughly controlled via λ. Larger the λ, sparser is the weight vector and vice versa. The optimal value of $C = \frac{1}{\lambda}$ was therefore selected from a range 10^{-3}- 10^3 with a step of 10, using 5-fold cross-validation via grid search method.

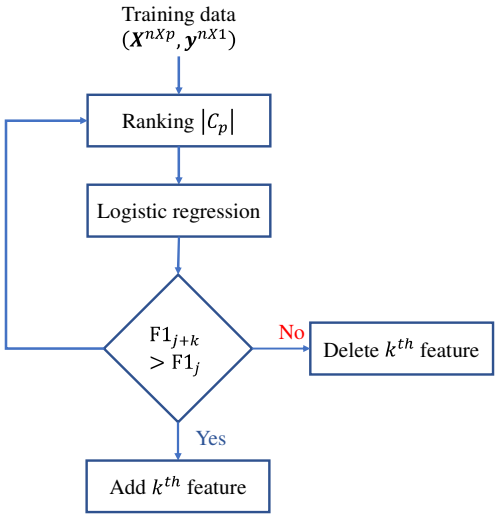

Fig. 1. The proposed feature selection technique. $\boldsymbol{X}^{nXp}$ is the training feature set, where n is the number of examples/instances and p is the number of features. $\boldsymbol{y}^{nX1}$ shows n binary training labels. C_p represents p coefficients/weights of the Fisher component.

D. Classification

A commonly used logistic regression classifier [13] was used to validate the efficacy of the proposed feature selection technique under varying imbalance rates and number of irrelevant features. Let a feature vector $\boldsymbol{x} \in \mathbb{R}^{nX1}$ with n rows, and an associate class labels/target $y \in \{0, 1\}$ be the training data for logistic regression classifier defined as:

$$p(y/X) = \frac{\exp(\beta_0 + \beta \boldsymbol{x})}{1 + \exp(\beta_0 + \beta \boldsymbol{x})} \quad (3)$$

where $p(y/\boldsymbol{x})$ is the conditional probability of a target given a feature $\boldsymbol{x}$, and β shows the classifier's parameters.

Class imbalance is generally handled by sampling the data or modifying the classifier. Standard classification algorithms generally use a default threshold of 0.5 to assign class membership. Any adjustments to such threshold change the class membership referred to as cost-sensitive learning. Simulation studies have shown that cost-sensitive learning can increase the sensitivity and decrease the specificity and vice versa. However, the accuracy more or less remains the same. The synthetically generated datasets were skewed or class-imbalanced. The cost-sensitive learning was therefore employed by assigning equal prior probabilities (0.5) to both classes to account for the class imbalance ratio in the training datasets. A generalized classification system was attained using stratified 5-fold cross-validation.

E. Evaluation of the Feature Selection Techniques

As the datasets were synthetically generated, we have the prior knowledge of informative and non-informative features. Each row of the matrix shown in Table I was used to describe the performances of the proposed and LASSO feature selection methods. Both feature selection techniques were evaluated using true positive rate, i.e. $TPR = \frac{CS}{IF}$, false positive rate, i.e. $FPR = \frac{IS}{IF}$, and true selected positive rate, i.e. $TSPR = \frac{CS}{TS}$. *CS*, *IS*, and *TS* are the number of correctly, incorrectly, and total selected features respectively. *IF* is the number of informative features used in the datasets. The performances of the proposed method and LASSO at selecting features were the average of 5-fold cross-validation with balanced datasets and varying number of informative features.

F. Evaluation of the Overall Classification Performance

For a binary-class problem, multiple confusion matrix-based performance metrics like sensitivity, specificity, and precision are commonly used. These metrics, however, by themselves are incomplete. Practically, combinations of these metrics are used to summarize the entire performance. For class imbalance datasets, F1 score $\left(\frac{(1+\beta^2) sensitivity \times precision}{\beta^2 \times sensitivity + precision}\right)$ and geometric mean (GM) $\left(\sqrt{sensitivity \times specificity}\right)$ have been the widely used measures of performance. The overall performances of the system under varying class imbalance ratios were therefore measured in F1 score and GM. The coefficient β shows the relative preference of sensitivity against precision [14]. In this study, both sensitivity and precision were given the same weight, and therefore, $\beta = 1$.

III. SIMULATION RESULTS AND DISCUSSION

Table I shows the mean 5-fold cross-validation feature selection performances of LASSO and the proposed method with varying number of informative features out of a total of 30 features. With balanced datasets, and irrespective of the number of informative features, LASSO compared to the proposed method had higher TPR but at the cost of higher FPR. This indicates that LASSO correctly selected more features than the total number of informative features, and can therefore result in an over-complex model. Besides computational requirements, such models are hard to interpret. Conversely, the proposed method, compared to LASSO, gave lower FPR and selected fewer features, and subsequently, can result in a simpler model.

TABLE I. MEAN 5-FOLD FEATURE SELECTION PERFORMANCES OF THE PROPOSED AND LASSO FEATURE SELECTION METHODS

IF	**LASSO**			**Proposed**		
(out of 30)	***CS***	***IS***	***TS***	***CS***	***IS***	***TS***
10	8	17	25	5	7	12
15	11	11	22	8	6	14
20	17	9	26	12	7	19
25	24	5	29	16	3	19

Figure 2 indicates the mean 5-fold performances of the proposed method and LASSO with logistic regression classifier and with the balanced datasets. The number of informative features out of the total 30 features varied from 1 to 30 features with a step of 2 features. The remaining features (i.e. total features – informative features) were the irrelevant features consisting of noise. On average, both LASSO and the proposed method gave similar results. Irrespective of the number of informative features and the feature selection technique, both GM and F1 score exhibited similar trends, indicating a balanced dataset. The fluctuations in both the performance measures are due to the number of informative features. Higher performances indicate that both feature selection techniques have correctly selected the informative features, whereas, lower performances indicate that some of the irrelevant features have also been selected, and informative features have been rejected by both methods as shown in Table I.

Figure 3 shows the mean 5-fold cross-validation performance of the proposed method and LASSO on the datasets with 20 informative and 10 irrelevant features. The class imbalance ratio varied from 0.1:0.9 to 0.5:0.5. Both LASSO and the proposed method gave similar results. However, the F1 score linearly increased with the decrease in the imbalance rates, and subsequently, the maximum values were achieved with the balanced datasets, whereas, the imbalance rate did not demonstrate much of its effect on the GM because it involves specificity and sensitivity, and does not tell about the number of false positives.

The proposed method, compared to LASSO, resulted in simpler models with comparable classification performance (see Table I, Figures 2-3).

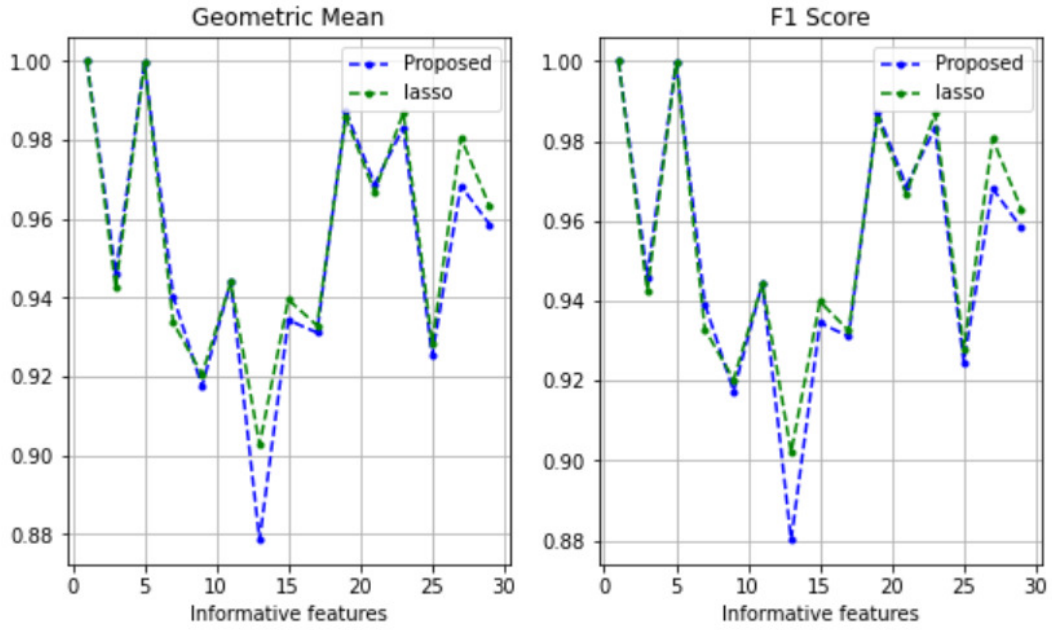

Fig. 2. Effect of number of informative features on classification performances on synthetically generated class-balanced dataset.

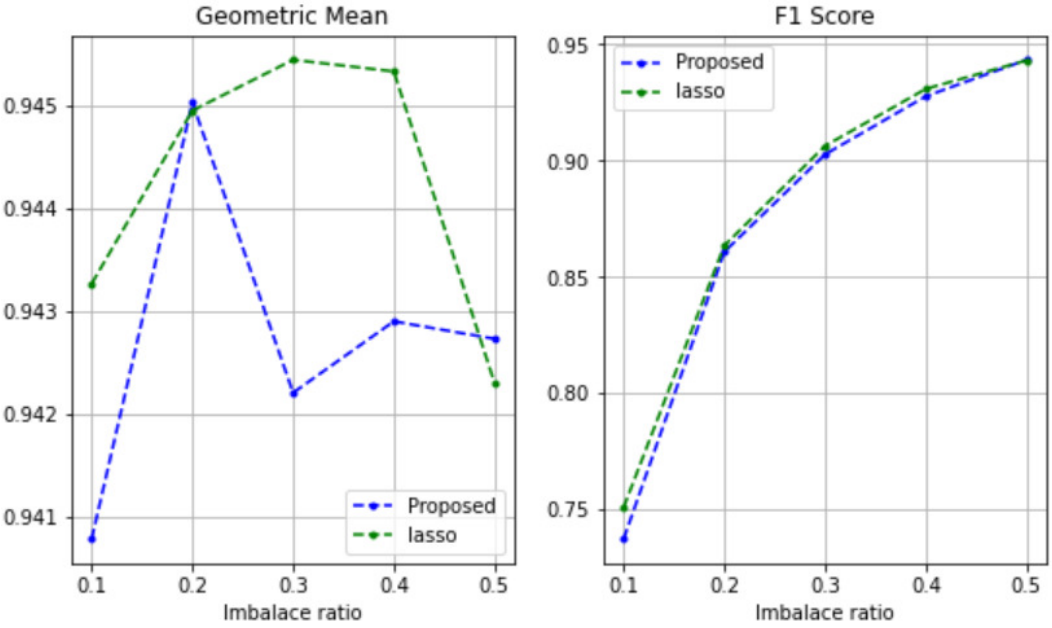

Fig. 3. Effect of imbalance ratio 0.1:0.9 – 0.5:0.5 on the classification performance using 20 informative and 10 irrelevant features.

IV. CONCLUSION

In this paper, an alternative feature selection technique is presented that uses the simplicity of a filter and the performance superiority of a wrapper. The proposed technique requires fewer models than its counterparts, i.e. forward and backward feature selection methods. Unlike regularization, the proposed method can easily be used with any classifier. Furthermore, the overall classification performance of the proposed method was comparable to the one of LASSO (a regularization technique), in terms of F1 score and GM.

REFERENCES

[1] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[2] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, New York, NY, USA: Springer. https://doi.org/10.1007/978-1-0716-1418-1.

[3] S. Nuanmeesri and W. Sriurai, "Thai Water Buffalo Disease Analysis with the Application of Feature Selection Technique and Multi-Layer Perceptron Neural Network," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 6907–6911, Apr. 2021, https://doi.org/10.48084/etasr.4049.

[4] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Assessing feature selection method performance with class imbalance data," *Machine Learning with Applications*, vol. 6, Dec. 2021, Art. no. 100170, https://doi.org/10.1016/j.mlwa.2021.100170.

[5] D. K. Singh and M. Shrivastava, "Evolutionary Algorithm-based Feature Selection for an Intrusion Detection System," *Engineering, Technology & Applied Science Research*, vol. 11, no. 3, pp. 7130–7134, Jun. 2021, https://doi.org/10.48084/etasr.4149.

[6] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007, https://doi.org/10.1093/bioinformatics/btm344.

[7] Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, Arlington, VA, USA, Apr. 2011, pp. 266–273.

[8] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Zolghadri Jahromi, "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds," *Pattern Recognition*, vol. 44, no. 7, pp. 1357–1371, Jul. 2011, https://doi.org/10.1016/j.patcog.2010.12.015.

[9] P. More and P. Mishra, "Enhanced-PCA based Dimensionality Reduction and Feature Selection for Real-Time Network Threat Detection," *Engineering, Technology & Applied Science Research*, vol. 10, no. 5, pp. 6270–6275, Oct. 2020, https://doi.org/10.48084/etasr.3801.

[10] J. Gong and H. Kim, "RHSBoost: Improving classification performance in imbalance data," *Computational Statistics & Data Analysis*, vol. 111, pp. 1–13, Jul. 2017, https://doi.org/10.1016/j.csda.2017.01.005.

[11] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Aug. 2011.

[12] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, Coimbatore, India, Jul. 2016, pp. 18–20, https://doi.org/10.1109/ICACA.2016.7887916.

[13] A. B. Musa, "A comparison of ℓ1-regularizion, PCA, KPCA and ICA for dimensionality reduction in logistic regression," *International Journal of Machine Learning and Cybernetics*, vol. 5, no. 6, pp. 861–873, Dec. 2014, https://doi.org/10.1007/s13042-013-0171-7.

[14] W.-J. Lin and J. J. Chen, "Class-imbalanced classifiers for high-dimensional data," *Briefings in Bioinformatics*, vol. 14, no. 1, pp. 13–26, Jan. 2013, https://doi.org/10.1093/bib/bbs006.