Gradient Based Efficient Feature Selection

Syed Zulqarnain Gilani Faisal Shafait
Ajmal Mian
School of Computer Sciences and Software Engineering
The University of Western Australia
zulqarnain.gilani@uwa.edu.au

Abstract

Selecting a reduced set of relevant and non-redundant features for supervised classification problems is a challenging task. We propose a gradient based feature selection method which can search the feature space efficiently and select a reduced set of representative features. We test our proposed algorithm on five small and medium sized pattern classification datasets as well as two large 3D face datasets for computer vision applications. Comparison with the state of the art wrapper and filter methods shows that our proposed technique yields better classification results in lesser number of evaluations of the target classifier. The feature subset selected by our algorithm is representative of the classes in the data and has the least variation in classification accuracy.

1. Introduction

Feature selection, the process of selecting a subset of relevant and non-redundant features, is critical to developing robust, supervised or unsupervised, machine learning models. Supervised learning models are typically presented with a set of training observations $\mathbf{X} = \{\boldsymbol{x_1}, ..., \boldsymbol{x_M}\}$ ($\mathbf{X} \in \mathbb{R}^{N \times M}$), where N observations are characterized by M vectors of features (or attributes) and class labels c. The problem of feature selection, thus, is to learn a reduced space $\mathbf{F} \in \mathbb{R}^{N \times m}$ that can best classify a future *unseen* observations $\mathbf{q}$ into one of the classes c.

Given a set of observations comprising of hundreds or thousands of features it is possible that a large number of features are either irrelevant or redundant with respect to the set of classes and hence are less likely to classify a test instance correctly [31, 32]. Moreover, even with a subset of "good" features, it is not necessary that their combination would lead to a good classification performance, meaning thereby, that the m best features may not be the best m features [2, 13].

For decades now, feature selection has been a dynamic field of research in machine learning [19, 27], statistics [11], data mining [4, 14] and statistical pattern recognition [21]. It has found success in many applications like image retrieval [8, 29], genomic microarray analysis [31, 34], intrusion detection [17], text categorization [9, 32] and customer relationship management [22]. Feature selection is known to have many advantages like alleviation of the *curse of dimensionality* to reduce the computational cost, reduction of irrelevant and redundant features to improve the classification accuracy and selection of features that have a physical interpretation to help identify and monitor the target diseases or function types [24].

Although there seems to be more focus on dimensionality reduction techniques in the field of computer vision, feature selection has seen vast applications in this field. Researchers have effectively used feature selection methods in problems like land use classification based on satellite images [12], object tracking [28], pose classification [23] etc.

Approaches to feature selection can be divided into two main categories: filter methods and wrapper methods [15, 16]. Whatever the approach, the main objective is to find features that are relevant to the set of class labels and at the same time have less mutual redundancy. Filter methods employ statistical techniques and make use of the intrinsic information within the features to attain this objective while wrappers target some classification algorithm for this task.

Given two feature vectors $\{\boldsymbol{x_i}, \boldsymbol{x_j}\}$ from the set of observations $\mathbf{X}$ and a class c from the set of classes C, $\Gamma(\boldsymbol{x_i}, c)$ is defined as the statistical relevance of the feature $\boldsymbol{x_i}$ with class c, while $\Gamma(\boldsymbol{x_i}, \boldsymbol{x_j})$ is defined as the redundancy between the two features $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. In filter methods, the function Γ can take on one of the many statistical parameters like the mutual information [25], correlation [10] or standard deviation [20].

The Fast Correlation Based Filter (FCBF), proposed by Yu *et al.* [34, 35], removes irrelevant and redundant features. Features with a symmetrical uncertainty according to their class below a given threshold are removed because they are considered as irrelevant. Additionally, only fea-

tures that do not have any approximate Markov blanket in the current set of remaining features are kept in order to reduce redundancy. Similarly, Minimal-Redundancy-Maximal-Relevance (mRMR), a filter based approach proposed by Peng *et al.* [6, 25], performs three tasks: (1) it looks for features with maximum relevance using the information content within each feature (2) next it looks for features with minimum redundancy using the mutual information between the features (3) finally, the algorithm simultaneously maximizes the relevance and minimizes the redundancy by either taking their difference (Mutual Information Difference (MID)) or the quotient (Mutual Information Quotient (MIQ)). The authors take $I(x, y)$ as the mutual information which is derived from information theory.

Contrary to the filter methods, wrappers use a target classification algorithm to select a subset of features. Training data is used to train a classifier and learn a subset of relevant and non-redundant features. This learnt classifier is then employed to classify a query instance. However, this method is computationally very expensive and the success in obtaining high quality features depends on the number of feature subsets that are tried in the training phase to search for the best feature subset. Typically the total number of possible subsets of M features is given by $2^M - 1$. To balance the tradeoff between classification accuracy and computational cost, different search strategies such as complete, heuristic, and random search have been proposed [5]. A complete search, often known as Brute Force (BF), evaluates all possible feature subsets to select the one that gives best results. Hill Climbing Forward Search (HCFS) algorithm, proposed by Kohavi *et al.* [15], first looks for the best feature from within the M features and adds it to the selected feature set $\mathbf{F}$. It then iteratively keeps adding a single features to $\mathbf{F}$ until the addition of a feature does not improve the accuracy. The Best-First Forward Selection (BFFS) [15] wrapper is similar to HCFS algorithm, except that it does not stop if a new feature from the feature set M reduces the classification accuracy. Instead it discards that feature and restarts the search. The algorithm is stopped if classification accuracy has not improved or if no new feature has been added to the set F in last k iterations. Both HCFS and BFFS are greedy algorithms and do not guarantee to find an optimal feature subset.

Filter methods do not consider the target classification algorithm and hence either suffer from the classifier-specific issues of that algorithm or do not benefit from its advantages. The reduced feature subset selected by such methods remains the same regardless of the type of learning algorithm used. Furthermore, the feature subset selected by filter methods are often refined using the wrapper approach. In contrast, while the wrapper methods evaluate the actual target classifier, they suffer from high computational complexity. Even the greedy sequential search which reduces the search space from $\mathcal{O}(2^M - 1)$ to $\mathcal{O}(M^2)$ can become very inefficient for high-dimensional data.

Table 1. List of symbols.

Symbols	Desscription
$\boldsymbol{X}_{N\times M}$	Feature matrix with N observations and M features
$\boldsymbol{\Omega_i}$	Feature matrix in i^th iteration $(i = 1, .., k)$
$\boldsymbol{S_i}$	Feature subsets of $\boldsymbol{\Omega}$ $(i = 1, .., M)$
$\nabla(.)$	Application of a classifier (e.g. LDA, SVM)
Λ_i	Classification result by applying $\nabla(.)$ to $\boldsymbol{S_i}$
γ_i	Gradient between Λ_{i+1} amd Λ_i $(i = 1, .., M - 1)$
R_i	Maximum Λ in each iteration $(i = 1, .., k)$
$\boldsymbol{F}_{N\times m}$	Required reduced feature space with m features

We propose a novel search method for forward selection wrapper approach in order to select more representative physically interpretable features. A gradient based feature selection algorithm is proposed that is able to search the feature space more efficiently and effectively. While reducing the number of evaluations of the target learning algorithm and the number of representative features, our proposed approach also results in high accuracy. We evaluate our proposed algorithm on seven public datasets and compare the results with three filter and three wrapper feature selection methods. We show that compared to the wrapper methods, our algorithm yields better classification accuracy with a reduced set of features and in case of filter methods it outperforms them in terms of classification accuracy and number of features selected.

2. Proposed algorithm

An overview of our proposed algorithm is given in Figure 1. Given a feature matrix $\mathbf{X} \in \mathbb{R}^{N\times M}$, with N observations and M features, we apply the classifier $\nabla(.)$ to classify the features x_i using one feature at a time. This is the first step and is performed only once. For huge dimensional datasets, in the interest of reducing the computational cost, it can be replaced by obtaining the features using a filter approach. The features are now arranged in descending order of their classification accuracy and the feature matrix $\boldsymbol{\Omega}_1 = \{\boldsymbol{S}_1, ..., \boldsymbol{S}_M\}$ is obtained such that

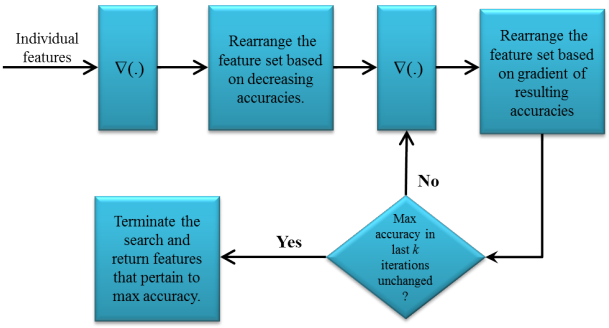

Figure 1. An overview of our Gradient based Efficient Feature Selection algorithm (GEFS).

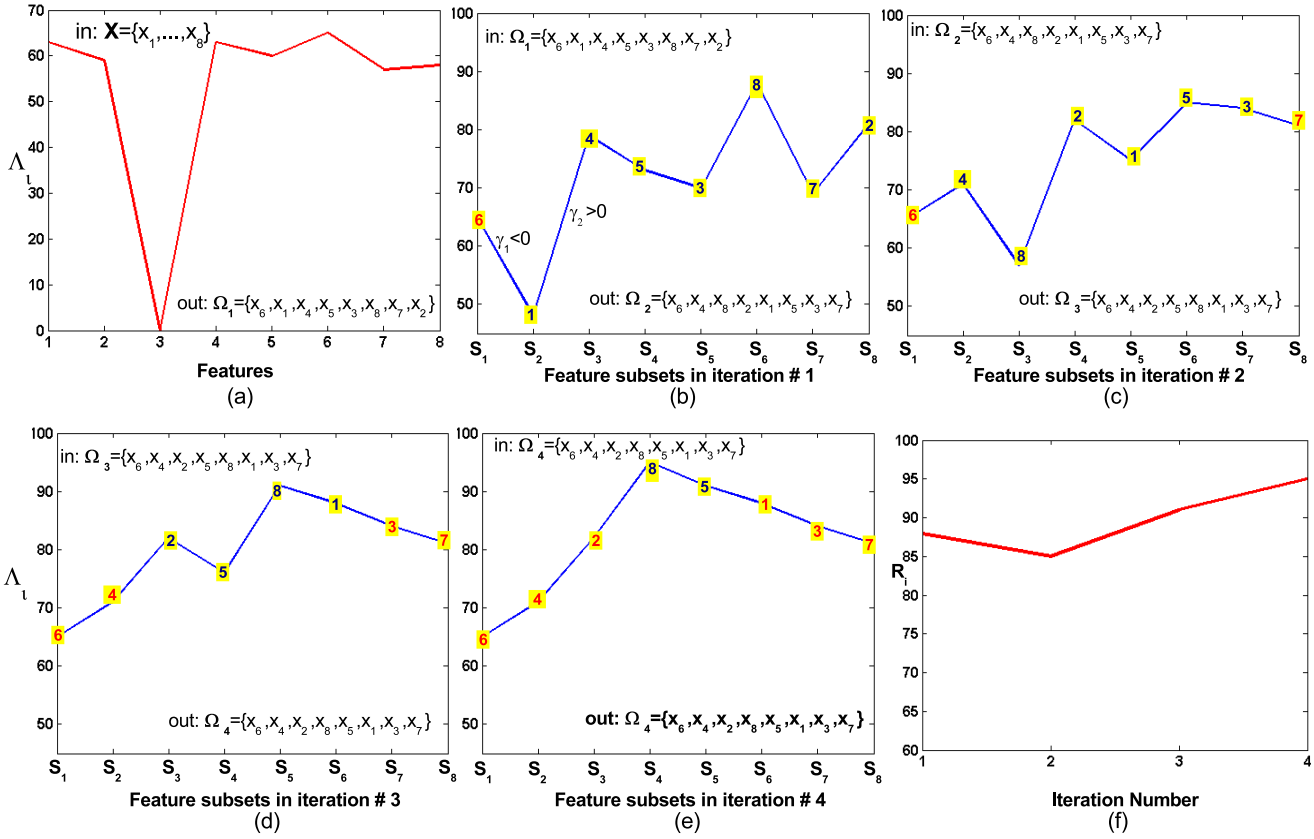

Figure 2. A graphical description of our proposed gradient based feature selection algorithm. (a) The first step of classifying the individual features. (b-e) Successive iterations of classifying and reordering the features. Only subsets with features in blue are evaluated at each iteration (f) Plot of maximum accuracy in each iteration. Notice that R_i will converge after the last iteration.

$\mathbf{S}_1 \subset \mathbf{S}_2 \subset ...\mathbf{S}_{M-1} \subset \mathbf{S}_M$. Once again the classifier $\nabla(.)$ is applied to these feature subsets to obtain the classification accuracy Λ_i of each subset. The maximum accuracy of each iteration $R_i = max(\Lambda_i, .., \Lambda_M)$ is also recorded as a heuristic to stop the search. Starting from Λ_1 we find the gradient γ of the classification result of every two consecutive subsets in the feature matrix $\mathbf{\Omega}_i$ such that $\gamma_i = \Lambda_{i+1} - \Lambda_i$. Next we rearrange $\mathbf{\Omega}_i$ such that all features that contributed to a positive gradient are moved up while those that resulted in a negative gradient are moved down in the search order. In this way we obtain $\mathbf{\Omega}_2$ for the next iteration and the process continues iteratively. As the process continues the search starts stabilizing and some of the features do not change their order since the Λ_i associated with their subsets stays the same over successive iterations. Such features are excluded from further evaluation in order to reduce the computational cost.

The feature selection search stops when R_i, the maximum accuracy of all subsets in each iteration does not change for k consecutive iterations. Notice that this is one of the heuristics used to stop the BFFS in [15]. This state can be reached in two cases; (1) The order of features in the feature matrix $\mathbf{\Omega}_i$ does not change even after rearranging based on the gradients γ_i. (2) The maximum accuracy R_i does not change even after rearranging the features in $\mathbf{\Omega}_i$. At convergence the feature subset with the maximum Λ is returned as the reduced most relevant non redundant feature set $\mathbf{F}$.

The proposed algorithm applied to a hypothetical feature set is graphically illustrated in Figure 2. The first step of applying $\nabla(.)$ to classify individual features x_i is shown in Figure 2(a). Notice that none of the features has a classification accuracy of more than 70%. Figures 2(b-e) show the details of the iterative procedure. The features after being sorted in the order of their accuracy are arranged to form subsets $\boldsymbol{S}_i$. These subsets are classified and the result Λ_i is shown in (a). Gradient γ_i is calculated for consecutive Λ_i and the features are once again reordered. The output sequence of one iteration becomes the input of the next. As the process starts stabilizing, notice that the feature subsets in the beginning and at the tail are excluded from further evaluation. These are shown in red colour. Finally, in (e)

there is no change in the order of features even after rearranging based on the gradient γ_i. The maximum accuracy in each iteration R_i is shown in Figure 2(f). R_i will converge after the fourth iteration and hence the feature selection search stops after k iterations. The feature subset $\boldsymbol{S}_4$ of the fourth iteration, having the highest classification accuracy Λ_4 is returned as the resultant feature set $\mathbf{F}$.

3. Evaluation

3.1. Datasets and algorithms for comparison

We have tested our proposed algorithm on two large, three medium and two small datasets. Details of these datasets are given below while a summary is presented in Table 2.

- Bankruptcy data from StatLib [30] and Vertebral Column (VerCol) datasets from UCI machine learning repository [1] comprise the medium sized datasets.
- Breast Tissue, Glass Identification and Wine datasets from UCI machine learning repository [1] are amongst the small size datasets.
- Face Recognition Grand Challenge ver 2 (FRGC) [26] and Binghamton University 3D Facial Expression (BU-3DFE) [33] are the two large datasets which have strong applications in computer vision. FRGC v2 dataset contains 4007 3D face scans of 466 individuals. We use 182 distances between landmarks that have been automatically detected by Creusot *et al*. [3] as features to perform gender classification. BU-3DFE face database consists of 2500 scans of 100 subjects. There are 25 scans of each subject in seven different expressions. The dataset comes with 83 feature points as ground truth. We select 20 landmarks critical to gender classification and extract 49 distances between them as features.These datasets have been used to evaluate the performance of our proposed technique in gender classification, an application related to computer vision. Figure. 3 shows some of the features used for gender classification on FRGC dataset.

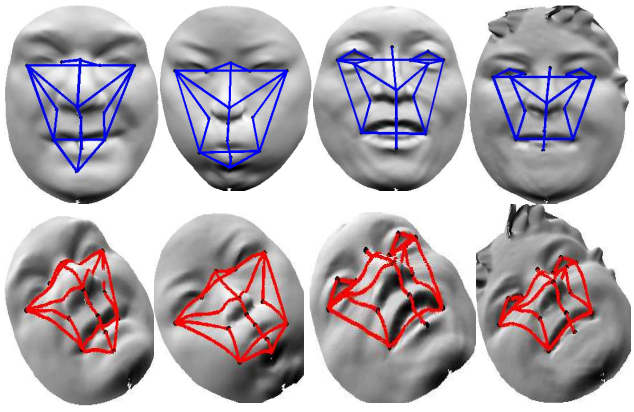

Figure 3. Examples of some of the distances extracted as features for gender classification on FRGC dataset.

Table 2. Summary of datasets

Name	Features	Instances	Classes
FRGC v2	182	4007	2
BU-3DFE	49	2500	2
Bankruptcy	5	50	2
VerCol	6	310	2
Breast Tissue	9	106	4
Glass Identification	9	214	7
Wine	13	178	3

Based on the technique used to search the feature space we name our proposed algorithm Gradient based Efficient Feature Selection (GEFS). On small and medium sized datasets we compare our proposed algorithm with three wrapper and three filter methods apart from using all features. Wrappers include the Brute Force (BF), Hill Climbing Forward Search (HCFS) and Best-First Forward Selection (BFFS) [15] while filter methods are Fast Correlation Based Filter (FCBF) [35], Minimal-Redundancy-Maximal-Relevance (mRMR) [25] and TTest [18]. On the large datasets it is not possible to use the Brute Force (BF) since the search space exceeds 10^{15} subsets.

3.2. Evaluation criteria

There is no fixed criteria in the literature to evaluate the quality of a feature selection algorithm. The most commonly used metrics are Mean Classification Accuracy (MCA) yielded by the output reduced feature set and the number of evaluations of the target algorithm performed to select these representative features. Both, FCBF and MRMR have a parameter to control the number of reduced features selected by these algorithms [25, 35]. For a fair comparison, we select all features and test them using j features at a time in the order in which the feature selection algorithm has selected them. It is evident that for filter methods the number of evaluations is equal to the number of features in the data. Amongst the wrapper methods, only BFFS algorithm has a parameter of k iteration as a heuristic to stop the feature space search. Again for a fair comparison we set $k = 3$ for the BFFS as well as our proposed algorithm. The target algorithm used to evaluate all feature selection methods is the Linear Discriminant Classifier (LDA) [7]. We perform all evaluations with 10-fold cross validation.

4. Results and Analysis

Comparative results are shown in Figure 4. The MCA of Brute Force method is the gold standard as it tests the com-

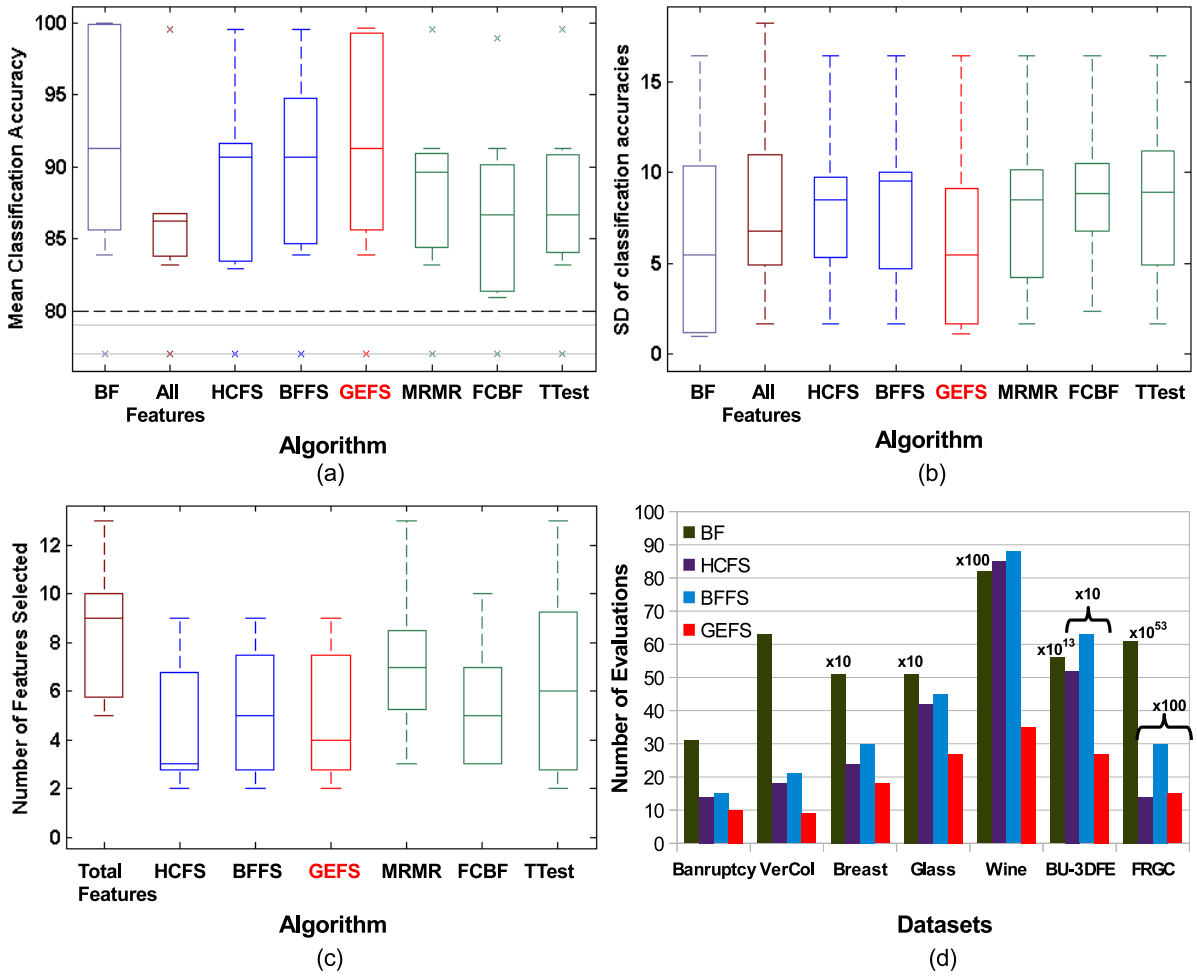

Figure 4. Comparison of our proposed method with seven other methods using (a) Mean Classification Accuracy of LDA (b) standard devaition of accuracies over 10-folds and (c) number of features finally selected, as criteria. (d) Comparison with wrapper approaches over the number of evaluations of LDA. Notice that for small and medium datasets, the performance of our proposed technique is equal to that of Brute Force.

plete feature space. Figure.4(a) shows that our proposed algorithm GEFS performs at par with BF. The classification results are consistently better than all other feature selection algorithms mentioned. MCA of all algorithms for glass dataset is approximately 66% and shown as an outlier in the graph. Note that Brute Force classification could not be performed on BU-3DFE and FRGC datasets. The MCA included in the analysis was set to 100% for fairness in comparison. Figure.4(b) shows the standard deviation of the classification accuracies. It is evident from the graph that along with improved accuracy our classification results also have less standard deviation over 10-folds. The final number of features selected is of high interest for practical applications. For example in gene phenotype classification, when a small number of genes are selected, their biological relationship with the target diseases is more easily identified [6]. Figure.4(c) shows the final number of features m selected by each algorithm as well as the total number of features available in the small and medium datasets. The only algorithm that selects lesser number of features is HCBF, but it does so at the cost of accuracy. Being a greedy algorithm it terminates when there is no improvement in the results, but the feature subset selected is not necessarily the set of representative features.

Finally the number of evaluations of the target algorithm are depicted in Figure.4 (d). For all filter methods the number of evaluations is equal to the number of features in the dataset and hence we did not find it prudent to include these algorithms on this particular comparison. It is obvious from the figure that our proposed algorithm, GEFS, requires less number of evaluations of the target algorithm. It is thus an evidence of efficient search in the feature space to se-

Table 3. A comparison between our proposed method and the state of art on two evaluation criteria for each dataset. Note that we have not mentioned the target evaluation runs of filter methods as it is understood that this number is equal to the total number of features in the dataset. * Stands for all features, ** and $^{\#}$ The exact figures are 56.2×10^{13} and 61.3×10^{53} respectively.

Datasets	Mean Classification Accuracy								Number of Evaluations			
	BF	All Fe*	HCFS	BFFS	GEFS	MRMR	FCBF	Ttest	BF	HCFS	BFFS	GEFS
Bankruptcy	91.3	85.7	91.3	91.3	91.3	91.3	91.3	91.3	31	14	15	10
VerCol	83.9	83.2	82.9	83.9	83.9	83.2	82.9	83.2	63	18	21	9
Breast	90.7	86.8	90.7	90.7	90.7	87.9	86.7	89.5	511	24	30	18
Glass	66.6	65.2	65.2	65.7	65.7	65.2	63.8	66.1	511	42	45	27
Wine	99.5	99.5	99.5	99.5	99.5	99.5	98.9	99.5	8191	85	88	35
BU-3DFE	100.0	86.7	91.7	95.9	99.6	89.8	86.7	86.7	56.2**	522	630	274
FRGC	100.0	86.2	85.1	87.0	98.6	89.6	80.9	86.5	61.3$^{\#}$	1428	2958	1562

lect an optimal feature set. Note that we did not perform Brute Force evaluation on BU-3DFE and FRGC datasets. The number of evaluations given here is the size of the total search space for feature selection for these datasets, i.e.$\mathcal{O}(2^M - 1)$.

Table 3 gives a detailed picture of comparative results. As mentioned before, we do not compare our proposed algorithm with filter methods on the criteria of number of evaluations of LDA. For Bankruptcy dataset the performance of all algorithms is at par with each other, however, only GEFS achieves this accuracy in less number of LDA evaluations. The striking advantage of GEFS is evident as the number of features increase. With medium and large datasets it is not feasible to run the Brute Force technique. HCFS gives suboptimal classification accuracy as it is a greedy algorithm. BFFS yields better accuracy than HCFS at the cost of increased evaluations of the target classification algorithm. Our proposed method, GEFS gives better accuracy results with lesser number of LDA evaluations.

5. Conclusion

We have presented a gradient based feature selection algorithm (GEFS) that uses a wrapper technique to select an optimal feature subset. We also showed that our proposed approach is efficient in searching the feature space and can converge in less number of iterations. The proposed algorithm was tested on small and medium as well as large datasets which have applications in computer vision. Results were compared with state of the art wrapper and filter approaches. Our analysis shows that GEFS yields better accuracy in lesser number of evaluations of the target classification algorithm. It selects a reduced feature set which can make it easy for researchers and analysts to identify their relationship with the target class.

Acknowledgment

Syed Zulqarnain Gilani is funded by the International Postgraduate Research Scholarship (IPRS). This research was also supported by ARC grant DP110102399 and the UWA FECM grant.

References

[1] K. Bache and M. Lichman. UCI machine learning repository, 2013. 4

[2] T. M. Cover. The best two independent measurements are not the two best. *Systems, Man and Cybernetics, IEEE Transactions on*, (1):116–117, 1974. 1

[3] C. Creusot, N. Pears, and J. Austin. A machine-learning approach to keypoint detection and landmarking on 3D meshes. *International Journal of Computer Vision*, 102(1-3):146–179, 2013. 4

[4] M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature selection for clustering-a filter solution. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 115–122. IEEE, 2002. 1

[5] M. Dash and H. Liu. Consistency-based search in feature selection. *Artificial intelligence*, 151(1):155–176, 2003. 2

[6] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005. 2, 5

[7] R. Duda, P. Hart, and D. Stork. *Pattern Classification and Scene Analysis 2nd ed.* 2001. 4

[8] J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, and A. M. Aisen. Unsupervised feature selection applied to content-based retrieval of lung images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(3):373–378, 2003. 1

[9] G. Forman. An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research*, 3:1289–1305, 2003. 1

[10] M. A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999. 1

[11] T. Hastie, R. Tibshirani, and J. J. H. Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001. 1

[12] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158, 1997. 1

[13] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37, 2000. 1

[14] Y. Kim, W. N. Street, and F. Menczer. Feature selection in unsupervised learning via evolutionary search. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 365–369. ACM, 2000. 1

[15] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997. 1, 2, 3, 4

[16] P. Langley et al. *Selection of relevant features in machine learning*. Defense Technical Information Center, 1994. 1

[17] W. Lee, S. J. Stolfo, and K. W. Mok. Adaptive intrusion detection: A data mining approach. *Artificial Intelligence Review*, 14(6):533–567, 2000. 1

[18] H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics Series*, pages 51–60, 2002. 4

[19] H. Liu, H. Motoda, and L. Yu. Feature selection with selective sampling. In *ICML*, pages 395–402, 2002. 1

[20] H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In *Tools with Artificial Intelligence, 1995. Proceedings., Seventh International Conference on*, pages 388–391. IEEE, 1995. 1

[21] P. Mitra, C. Murthy, and S. K. Pal. Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):301–312, 2002. 1

[22] K. Ng and H. Liu. Customer retention via data mining. *Artificial Intelligence Review*, 14(6):569–590, 2000. 1

[23] M. H. Nguyen and F. De la Torre. Optimal feature selection for support vector machines. *Pattern recognition*, 43(3):584–591, 2010. 1

[24] F. Nie, H. Huang, X. Cai, and C. H. Ding. Efficient and robust feature selection via joint 2, 1-norms minimization. In *Advances in Neural Information Processing Systems*, pages 1813–1821, 2010. 1

[25] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(8):1226–1238, 2005. 1, 2, 4

[26] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 947–954. IEEE, 2005. 4

[27] M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003. 1

[28] J. Shi and C. Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 593–600. IEEE, 1994. 1

[29] D. L. Swets and J. J. Weng. Efficient content-based image retrieval using automatic feature selection. In *Computer Vision, 1995. Proceedings., International Symposium on*, pages 85–90. IEEE, 1995. 1

[30] P. Vlachos. Statlib datasets archive, 1998. 4

[31] E. P. Xing, M. I. Jordan, R. M. Karp, et al. Feature selection for high-dimensional genomic microarray data. In *ICML*, volume 1, pages 601–608, 2001. 1

[32] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997. 1

[33] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3D facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006. 4

[34] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, volume 3, pages 856–863, 2003. 1

[35] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004. 1, 4