Bi-level dimensionality reduction methods using feature selection and feature extraction

Veerabhadrappa[1, 2]
[1]Department of Computer Science, University College, Mangalore- 575 001, India

Lalitha Rangarajan[2]
[2]Department of Studies in Computer Science, University of Mysore, Manasagangothri, Mysore- 570 006, India

ABSTRACT

Variety of feature selection methods have been developed in the literature, which can be classified into three main categories: filter, wrapper and hybrid approaches. Filter methods apply an independent test without involving any learning algorithm, while wrapper methods require a predetermined learning algorithm for feature subset evaluation. Filter and wrapper methods have their drawbacks and are complementary to each other. The filter approaches have low computational cost with insufficient reliability in classification while wrapper methods tend to have superior classification accuracy but require great computational effort. The methods proposed in this paper are bi-level dimensionality reduction methods that integrate filter method and feature extraction method with the aim to improve the classification performance of the features selected. In the two approaches proposed, in level 1 of dimensionality reduction, feature are selected based on mutual correlation and in level 2 selected features are used to extract features using PCA or LPP. To evaluate the performance of the proposed methods several experiments are conducted on standard datasets and the results obtained show superiority of the proposed methods over single level dimensionality reduction techniques (feature selection based on Mutual correlation, PCA and LPP).

Keywords: Mutual Correlation, dimensionality reduction, feature selection/extraction.

1. INTRODUCTION

In recent years, improvement in data acquisition capacity, lower cost of data storage and development of database and data warehousing technology have led to the emergence of high dimensional dataset. Many of these features are irrelevant and redundant which increase the search space size resulting in difficulty to process the data further. This curse of dimensionality is a major obstacle in machine learning and data mining. Hence dimensionality reduction is an active research area in the field of pattern recognition, machine learning, data mining and statistics. The purpose of dimensionality reduction is to improve the classification performance through the removal of redundant or irrelevant features. Dimensionality reduction can be achieved in two different ways namely feature transformation and feature selection.

Feature extraction/transformation is a process through which a new set of features is created. The feature transformation may be a linear or nonlinear combination of original features.

Feature selection is a process, through which no new set of features will be generated, but only a subset of original features is selected and feature space is reduced. The choice between feature selection and feature extraction depends on the application domain and the specific data set. Feature selection methods are classified into 3 categories namely filter, wrapper and hybrid approaches.

Filter approaches select features using characteristics of individual features. Wrapper approaches use a specific machine learning algorithm/classifiers such as decision tree or SVM and utilize the corresponding classification performance to select features. The filter approaches have low computational cost with low reliability of classification while wrapper approaches tend to have superior classification accuracy with high computational cost. Advantages of the filter-based techniques are that they can easily scale up to high-dimensional datasets and that they are computationally fast and independent of the learning algorithm. For application of large datasets, filter based approaches have proven to be more practical than wrapper approach because of their speed. A common disadvantage, however, is that the interaction with the classifier and the dependence among features are ignored, which leads to varied classification performance when the selected features are applied to different classification algorithms. On the other hand, the advantage of wrapper approaches is that they have a high probability of producing classifiers with better classification performances than the filter approaches as they take into account the feature dependencies and their collective contribution to model generation. A common drawback, however, is that the wrapper approaches have a higher risk of over-fitting and can be very computationally intensive when processing a large number of features. Hybrid approach is a recent technique which exploit the advantages of both filter and wrapper approach. A hybrid approach employs both an independent test and performance evaluation function of the feature subset.

Instead of using either feature selection or feature extraction technique to reduce the dimension, the combinations of feature selection and feature extraction methods can be applied as two level approach, i.e., apply one approach (either feature selection or feature extraction) on the original feature set to reduce its dimension and then apply another approach (again either feature selection or feature extraction) on this reduced feature set to further reduce its dimension. These cascading approaches can be categorized as follows:

a) Feature selection algorithm followed by another feature selection algorithm.
b) Feature selection algorithm followed by feature extraction algorithm.
c) Feature extraction algorithm followed by another feature extraction algorithm.

In this paper, we propose two novel combination of feature selection and feature extraction methods namely

(i) Feature selection based on mutual correlation followed by PCA. This is basically a filter approach, but takes care of feature dependencies as in a wrapper approach even though we are not using any classifier and hence low in computational cost. PCA has

proven record of high success in reducing dimensions. So in this method we have extracted features using PCA and again retaining only first few principal components.

(ii) Feature selection using mutual correlation followed by feature extraction using LPP.

PCA is a global transformation and LPP is useful where local information is at most important. In our second method, features are extracted using LPP. It may noted that feature extraction methods PCA/LPP are used on reduced set of features selected in the first step and hence the time required for transformation is far less than using PCA/LPP on the original set of features.

This paper is organized into 5 sections. A brief review of the related work is given in section 2. The proposed combined methods are introduced in section 3. Experimental results are presented in section 4 followed by summary and conclusion in section 5.

2. RELATED WORK

Several feature selection techniques have been proposed in the literature and survey of feature selection algorithms may be found in Molina et al. [1] and Guyon and Elisseeff [2]. Many researchers are involved in the study of goodness of a feature subset in determining an optimal one [2-8, 14, 26]. The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subset. A serious drawback about this method is the higher computational cost [4]. The filter model selects feature that are independent of any learning algorithm and it relies on various measures of the general characteristics of training data such as distance, information dependency and consistency [3]. In these methods, the relevance of each feature is evaluated individually and a score is given to each of them. The features are ranked by their scores and the ones with a score greater than a threshold are selected. According to the availability of class labels, there are feature selection methods for supervised learning [5, 6] as well as for unsupervised learning [7, 8]. Existing feature selection methods mainly exploit two approaches: individual feature evaluation and subset evaluation [2]. Individual evaluation methods rank features according to their importance in differentiating instances of different classes and can only remove irrelevant features as redundant features with similar rankings. Methods of subset evaluation search for a minimum subset of features that satisfies some goodness measure and can remove irrelevant features as well as redundant ones [9].

In filter approaches, features are scored and ranked based on certain statistical criteria and the features with highest ranking values are selected. Frequently used filter methods include, chi-square test [10], mutual information [11] and Pearson correlation coefficients [12]. Filter methods are fast but lack in robustness against interactions among features and feature redundancy. In addition, it is not clear how to determine the cut-off point for rankings to select only truly important features and exclude noise.

In the wrapper approach, features election is "wrapped" in a learning algorithm. The learning algorithm is applied to subsets of features and tested on a hold-out set and prediction accuracy is used to determine the feature set quality. Generally, wrapper methods are more effective than filter methods. Since exhaustive search is not computationally feasible, wrapper methods must employ a search algorithm to search for an optimal subset of features. SBS (Sequential Backward Selection) starts with the set of all features and progressively eliminates the least promising ones. SBS stops if the performance of learning algorithms drops below a given threshold due to removal of any remaining features. SBS relies heavily on the monotonicity assumption [13]. This states that

prediction accuracy never decreases as the number of features increases. In reality, the predictive ability of a learning algorithm may decrease as the feature subspace dimensionality increases after a maximum point due to a decreasing number of samples for each feature combination. When faced with high-dimensional data, SBS often finds difficulties in identifying the separate effect of each explanatory variable on the target variable. Because of this, good predictors can be removed early on in the algorithm (in SBS, once a feature is removed, it is removed permanently), where as SFS (Sequential Forward Selection) starts with an empty set of features and iteratively selects one feature at a time starting with the most promising feature until no improvement in classification accuracy can be achieved. In SFS, once a feature is added, it is never removed. SBS is robust to feature interaction problems but sensitive to multicollinearity. On the other hand, SFS is robust to multicollinearity problems but sensitive to feature interaction. The problem with SFS and SBS is their single track search. Hence, Pudil et al. [14] suggest floating search methods (SFFS, SFBS) that performs greedy search with provision for backtracking. However recent empirical studies demonstrate that sequential floating forward selection (SFFS) is not superior to SFS [15] and sequential floating backward selection (SFBS) is not feasible for feature sets of more than about 100 features [16]. Stochastic algorithms developed for solving large scale combinatorial problems such as ant colony optimization(ACO),genetic algorithm (GA), particle swarm optimization(PSO) and simulated annealing (SA) are at the forefront of research in feature subset selection [13,17–19]. These algorithms efficiently capture feature redundancy and interaction and do not require the restrictive monotonicity assumption. However, these algorithms are computationally expensive. Recently, several authors proposed hybrid approaches taking advantages of both filter and wrapper methods. Examples of hybrid algorithms include *t*-statistics and a GA [20], a correlation based feature selection algorithm and a genetic algorithm [21], principal component analysis and an ACO algorithm [22], chi-square approach and a multi-objective optimization algorithm [23], mutual information and a GA [24,25]. The idea behind the hybrid method is that filter methods are first applied to select a feature pool and then the wrapper method is applied to find the optimal subset of features from the selected feature pool. This makes feature selection faster since the filter method rapidly reduces the effective number of features under consideration. Advocates of hybrid methods argue that the risk of eliminating good predictors by filter methods is minimized if the filter cut-off point for a ranked list of features is set low. However, hybrids of filter and wrapper methods may suffer in terms of accuracy because a relevant feature in isolation may appear no more discriminating than an irrelevant one in the presence of feature interactions [26].

3. PROPOSED METHODS

3.1 Feature Selection based on Mutual Correlation method followed by PCA

Correlation is a well known similarity measure between two random variables. If two random variables are linearly dependent, then their correlation coefficient is close to ± 1. If the variables are uncorrelated the correlation coefficient is 0. The correlation coefficint is invariant to scaling and translation. Hence two features with different variances may have same value of this measure. Suppose the number of instances is N and the p-dimensional feature vectors are

$$X_i = \left[{}^i x_1, \ldots {}^i x_p \right] \quad i = 1, \ldots, N$$

The mutual correlation [27] for a feature pair x_i and x_j is defined as

$$r_{x_i, x_j} = \frac{\sum_k {}^k x_i {}^k x_j - N \bar{x}_i \bar{x}_j}{\sqrt{(\sum_k {}^k x_i^2 - N \bar{x}_i^2)(\sum_k {}^k x_j^2 - N \bar{x}_j^2)}}$$

(1)

If two features x_i and x_j are independent then they are also

uncorrelated, i.e. $r_{x_i, x_j} = 0$. Let us evaluate all mutual correlations

for all feature pairs and compute the average absolute mutual

correlation of a feature over δ features

$$r_{j, \delta} = \frac{1}{\delta} \sum_{i=1, i \neq j}^{\delta} \left| r_{x_i, x_j} \right|$$

(2)

The feature which has the largest average mutual correlation

$$\alpha = \arg \max_j r_{j, \delta} \qquad (3)$$

will be removed during each iteration of the feature selection algorithm. When feature x_α is removed from the feature set, it is also discarded from the remaining average correlations i.e.

$$r_{j, \delta-1} = \frac{\delta r_{j, \delta} - \left| r_{x_\alpha, x_j} \right|}{\delta - 1}$$

(4)

Algorithm

Level I:

    Input: Original feature set X of size N x p
    Output: Reduced feature set X of dimension
        D (D<<p)
    Method:
        1. Initialize $\delta = p-1$
        2. Discard features x_α for α determined by equation (3).
        3. Decrement δ= δ-1, if δ<D return the resulting D dimensional feature set and stop Otherwise.
        4. Recalculate the average correlations by using equation (4).
        5. Go to step 2.

Level II:

    Input: Feature set X of dimension D.
    Output: New transformed feature set Y
        of dimension f (f<<D).
    Method:
        1. Apply PCA to X=[$x_1, \ldots, x_D$] to obtain $(y_1, \ldots, y_f) = T(x_1, \ldots, x_D)$.

3.2 Feature Selection based on Mutual Correlation method followed by LPP

Apply the level I of the above algorithm to get the reduced feature set X of dimension D and the then apply LPP in level II to obtain the transformed feature set of dimension f (f<<D).

The standard complete linkage clustering algorithm has been employed on the reduced feature matrix for clustering the samples/dataset.

4. EXPERIMENTAL RESULTS

In this section, we present the experimental results to corroborate the success of the proposed model. The well known existing dimensionality reduction techniques such as PCA, LPP, and correlation based feature selection have been considered for comparative study. The superiority of the proposed model is established through the parameters precision, recall and F measure of the obtained clusters. Results of experiments performed on the standard WINE, ZOO and WDBC (Wisconsin Diagnostic Breast Center) datasets are shown in tables 2, 3 and 4. The summary of these datasets is given in the Table 1.

Table 1: Summary of datasets used in this paper

Dataset	Instances	Features	Classes
WINE	178	13	3
ZOO	101	16	7
WDBC	569	30	2

4.1 Experimentation on WINE dataset

The WINE dataset consists of 178 samples and each sample is of 13 dimensions. The WINE data is having 3 clusters $C_1 = 59$, $C_2 = 71$ and $C_3 = 48$. Comparative analysis is carried out with mutual correlation based feature selection, PCA, LPP and the proposed two methods: mutual correlation based feature selection with PCA and mutual correlation based feature selection with LPP.

To measure the accuracy of the clusters obtained precision, recall and F measure parameters are computed. The precision, recall and F measure are defined as follows:

$$\text{Precision} = \frac{C_a \cap C_r}{C_r}, \quad \text{Recall} = \frac{C_a \cap C_r}{C_a} \quad \text{and}$$

$$\text{F Measure} = \frac{2 * precision * recall}{precision + recall}$$

where C_a is the actual number of elements in the cluster and C_r is the number of elements in the clusters obtained.

The average precision, average recall and average F measure for WINE dataset are tabulated in the Table 2.

The number of features selected/extracted mentioned in the table is the optimal values obtained through experimentation for all the methods. From the Table 2, it is clear that feature selection using mutual correlation gives the F measure of 89.566 for 11 features of 13 original features. On these selected 11 features, when PCA is applied only 6 dimensions of feature vectors are sufficient to obtain the higher F measure value of 92.950. Similarly when LPP is applied, 8 dimensions of feature vectors are required to get F

measure value of 95.148. From Figure 1, it is clear that the F measure for the methods Correlation with PCA and LPP shows better result for WINE dataset when compared to other methods.

4.2 Experimentation on ZOO dataset

The ZOO dataset consists of 101 instances of animals each with 16 features out of which 15 are of type Boolean and 1 is of type numeric. Animals are categorized into 7 classes as Mammals, Fish, Bird, Invertebrate, Insect, Reptiles and Amphibians with number of instances in each class 41, 13, 20, 10, 8, 4 and 5 respectively. Comparative analysis is carried out with mutual correlation based feature selection, PCA, LPP, mutual correlation based feature selection with PCA and mutual correlation based feature selection with LPP. The average precision, average recall and average F measure are tabulated in Table 3. From the table 3, it is clear that LPP alone results in lower value of F measure, but when it applied on the reduced feature set obtained from correlation gives better result. Figure 2 shows that F measure for the method Correlation with LPP shows better result for ZOO dataset when compared to other methods.

4.3 Experimentation on WDBC dataset

The mammogram dataset of Wisconsin Diagnostic Breast Center (WDBC) consists of 569 instances each with 30 features. This contains two clusters having 212 malignant samples and 357 benign samples. Comparative analysis is carried out with mutual correlation based feature selection, PCA, LPP, mutual correlation based feature selection with PCA and LPP. The average precision, average recall and average F measure are tabulated in Table 4. From Figure 3, it is clear that the F measure for the methods Correlation with PCA and LPP shows better result for WDBC data-set when compared to other methods.

5. CONCLUSION

In this paper, a bi-level dimensionality reduction approach based on Mutual Correlation with PCA and Mutual Correlation with LPP are introduced. The proposed models achieve better performance in terms of F measure with reduced feature set. Experiments on well known datasets like WINE, ZOO and WDBC are conducted to demonstrate the superiority of the proposed models. As a future work, we are exploring the possibility of multi-level dimensionality reduction approaches by cascading more than two methods.

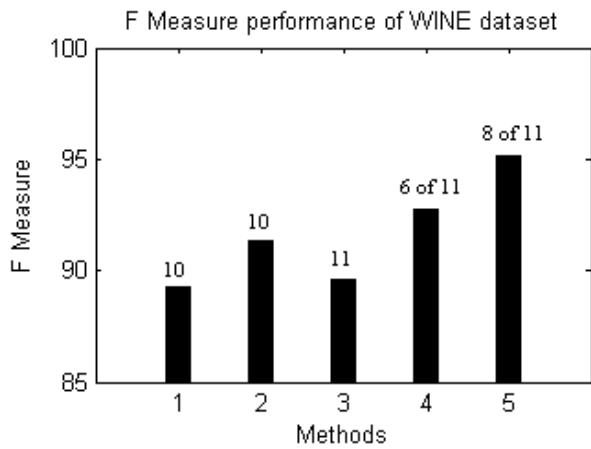**Table 2**: **Comparison of various methods for WINE dataset**

	PCA (number of features extracted=10)	LPP (number of features extracted =10)	Mutual Correlation (number of features selected=11)	Mutual Correlation +PCA (number of features selected=11 and number of features extracted =6)	Mutual Correlation +LPP (number of features selected=11 and number of features extracted =8)
Average Precision	89.845	91.095	89.945	**92.458**	**94.715**
Average Recall	88.749	91.608	89.191	**93.447**	**95.584**
Average F Measure	89.294	91.351	89.566	**92.950**	**95.148**

Table 3: Comparison of various methods for ZOO dataset

	PCA (number of features extracted=8)	LPP (number of features extracted =9)	Mutual Correlation (number of features selected=12)	Mutual Correlation +PCA (number of features selected=12 and number of features extracted =8)	Mutual Correlation +LPP (number of features selected=12 and number of features extracted =10)
Average Precision	80.553	73.995	82.315	**81.043**	**90.924**
Average Recall	87.561	74.111	82.334	**89.686**	**92.892**
Average F Measure	83.911	74.053	82.325	**85.146**	**91.898**

**Table 4: Comparison of various methods for WDBC dataset**

	PCA (number of features extracted=16)	LPP (number of features extracted =18)	Mutual Correlation (number of features selected=20)	Mutual Correlation +PCA (number of features selected=20 and number of features extracted =6)	Mutual Correlation +LPP (number of features selected=20 and number of features extracted =10)
Average Precision	87.947	89.161	89.755	**87.185**	**90.396**
Average Recall	75.663	82.363	78.302	**86.962**	**88.598**
Average F Measure	81.344	85.627	83.638	**87.073**	**89.752**

Figure 1: F measure for WINE dataset
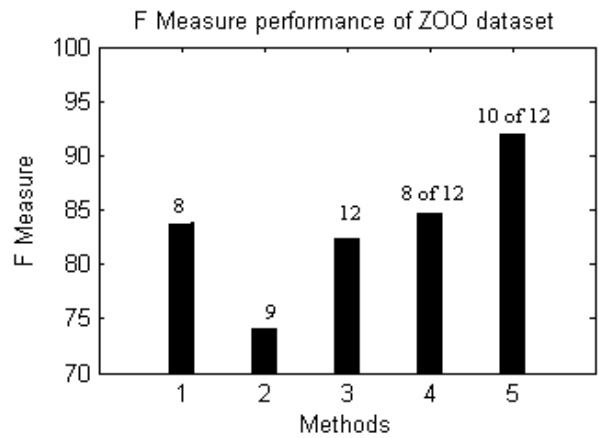
Figure 2: F measure for ZOO dataset
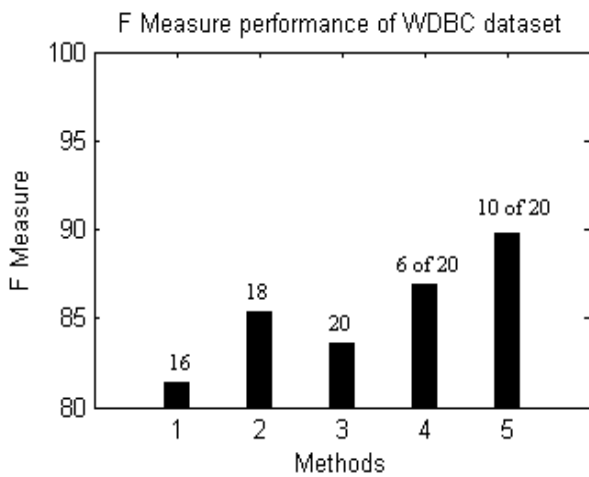
Figure 3: F measure for WDBC dataset

In all these graphs, the methods 1,2,3,4 and 5 at X axis corresponds to PCA, LPP, Mutual Correlation, Mutual Correlation followed by PCA and Mutual Correlation followed by LPP respectively.

6. REFERENCES

[1]. Molina, L.C., Belanche, L., Nebot, A., 2002, Attribute Selection Algorithms: A survey and experimental evaluation. Proceedings of 2nd IEEE's KDD, pp 306-313.

[2]. Guyon, I., Elisseeff, A.2003, An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3, pp 1157-1182.

[3]. Liu, H.A., Setiono, R., 1998, Incremental feature selection. Applied Intelligence 9, pp 217-230.

[4]. Kohavi, R., John, G.H, 1997, Wrappers for feature subset selection. Artificial Intelligence 97, pp 273-324

[5]. Yu, L., Liu, H, 2003, Feature Selection for High-Dimensional Data: A Fast Correlation-based Filter Solution. Proc. Int.Conference, ICML2003, pp 856-863.

[6]. Dash, M., Liu, H, 1997, Feature selection for classification. Intelligent Data Analysis: An International Journal 1, pp 131-156.

[7]. Park.J.S, Shazzad.K.M, Kim.D.S, 2005, Towards modeling lightweight intrusion detection system through correlation-based hybrid feature selection. Information Security and Cryptology, Proceedings 3822, pp 279-289.

[8]. Dash, M., Liu, H., Motoda, H, 2000, Consistency based feature selection. Knowledge Discovery and Data Mining, Proceedings 1805, pp 98-109.

[9]. T.Djatna, Y.Morimoto, 2008, A Novel feature selection Algorithms for Strongly Correlated Attributes using Two Dimensional Discriminant Rules, 6th IETCE Data Engg. Workshop,

[10]. X.Jin, A.Xu, R.Bie, P.Guo, 2006, Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles, Lecture Notes in Computer Science, 3916, pp 106–115.

[11]. C.Liao, S.Li, Z.Luo, 2007, Gene selection using Wilcoxon rank sum test and support vector machine for cancer, Lecture Notes in Computer Science ,4456, pp 57–66.

[12]. H.Peng, F.Long, C.Ding, 2005, Feature selection based on mutual information criteria of max-dependency, max-relevance ,and min redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27, pp 1226–1238.

[13]. J.Yang,V. Honavar,1998, Feature subset selection using a genetic algorithm, IEEE Intelligent Systems and their Applications 13, pp 44–49.

[14]. P.Pudil, J.Novovicov, J.Kittler, 1994, Floating search methods in feature selection, Pattern Recognition Letters, 5(11), pp 1119–1125.

[15]. M.Bensch, M.Schroder, M.Bogdan, W.Rosenstiel, P.Czerner, R. Montino,G. Soberger, P.Linke,R.Schmidt, 2005, Feature selection for high-dimensional industrial data ESANN, pp 27–29.

[16]. H.T.Ng, W.B.Goh, K.L.Low, 1997, Feature selection, perceptron learning, and a usability case study for text categorization, 20th Annual International Conference on Research and Development in Information Retrieval, Philadelphia pp 67–73.

[17]. S.M.Vieira, M.C.Sousa, T.A.Runkler, 2007, Ant colony optimization applied to feature selection in fuzzy classifiers, Lecture notes in computer science, 4529, pp 778–788.

[18]. X.Wang, J.Yang, X.Teng, W.Xia, J.Richard, 2007, Feature selection based on rough sets and particle swarm optimization, Pattern Recognition Letters, 28, pp 459–471.

[19]. M.Ronen,Z.Jacob. 2006, Using simulated annealing to optimize feature selection problem in marketing applications, European Journal of Operational Research 171, pp 842–858.

[20]. F.Tan, X.Fu, H.Wang, Y.Zhang, A.Bourgeois, 2006, A hybrid feature selection approach for micro array gene expression data, Lecture Notes in Computer Science 3992, pp 678–685.

[21]. K.M.Shazzad, J.S.Park, 2005, Optimization of intrusion detection through fast hybrid feature selection Proceedings of the Sixth International Conference on Parallel and Distributed Computing, IEEE Computer Society, Washington, DC, USA, pp 264–267.

[22]. Z.Yan, C.Yuan , 2004, Ant colony optimization for feature selection in face recognition, Lecture notes in Computer Science 3072, pp 221–226.

[23]. K.M.Osei-Bryson, K .Giles,B. Kositanurit, 2003, Exploration of a hybrid feature selection algorithm ,Journal of the Operational Research Society 54 , pp 790–797.

[24]. M.Fatourechi, G.Birch, R.K.Ward, 2007, Application of a hybrid wavelet feature selection method in the design of a self-paced brain interface system, Journal of Neuro engineering and Rehabilitation 4,

[25]. J.Huang,Y.Cai,X.Xu, 2006, A wrapper for feature selection based on mutual information, 18th International Conference on Pattern Recognition,vol.2, pp 618–621.

[26]. Iffat A..Gheyas, Leslie S.Smith, 2010, Feature subset selection in large dimensionality domains, Pattern Recognition, 43, pp 5–13.

[27]. M.Haindl, P.Somol, D.Ververidis, C.Kotropoulos, 2006, Feature Selection Based on Mutual Correlation, Proceedings of Progress in Pattern Recognition, Image Analysis and Application, 4225, pp 569-577.

7. ABOUT THE AUTHORS

Veerabhadrappa obtained his B.Sc., and M.Sc., degrees in Computer Science and Technology from the University of Mysore, India, respectively, in the years 1987 and 1989. He is a faculty in the Department of Computer Science, University College, Mangalore, India and currently he is on Teacher Fellowship to carry out Ph.D in the Department of Studies in Computer Science, University of Mysore, India. He authored 8 peer-reviewed papers in journals and conferences. His area of research covers dimensionality reduction, face/object recognition and symbolic data analysis.

Dr. Lalitha Rangarajan is currently a Reader at the Department of Computer Science, University of Mysore, India. She has two master degrees one in Mathematics from Madras University, India and the other in Industrial Engineering (specialization: Operations Research) from Purdue University, USA. She has taught mathematics for five years in India soon after the completion of masters in mathematics. She is associated with the Department of Studies in Computer Science, University of Mysore, soon after completion of masters at Purdue University. She completed her doctorate in Computer Science in 2004 and since then doing research in the areas of image processing, retrieval of images, bioinformatics and pattern recognition. She has more than 30 publications in reputed conferences and journals to her credit.