Dimensionality Reduction by Cascading Mutual Correlation with Symbolic Approach

Veerabhadrappa
Department of Computer Science,
University College, Mangalore – 575 001, India

ABSTRACT
In this paper, we propose a novel cascading approach, by cascading the feature selection method using mutual correlation with this symbolic approach. In the symbolic approach, the new dimensionality reduction method through transformation of features into symbolic data using the property of collinearity and variance based cumulative sum of features is used. The feature values are transformed into line segments and thus reduced to two symbolic features namely, number of line segments and average slope of the line segments. In addition the first and last feature values are also considered to distinguish the samples with the same average slope values. In this proposed approach of cascading the feature selection method using mutual correlation with this symbolic approach, the entire feature set is reduced to only 4 features. Experimental results on the standard datasets WDBC, WBC, CORN SOYANEAN and WINE shows that the proposed methods achieve better classification performance with negligible time.

General Terms
Dimensionality Reduction, feature selection, symbolic approach.

Keywords
Symbolic features, mutual correlation, Extraction of lines, Cumulative sum of features.

1. INTRODUCTION
Feature selection is one of the fundamental problems in machine learning and pattern recognition. The role of feature selection is critical, especially in the applications involving many irrelevant features. If the feature selection is conducted independent of classifier, it is normally referred to as filter method. If the feature selection uses the classifier to evaluate the performance of each subset, it is normally referred to as wrapper method. Most of the feature selection algorithms rely on heuristic searching and thus cannot provide any guarantee of optimality. This is largely due to the difficulty in defining an objective function that can be easily optimized by some well-established optimization techniques. Generally the wrapper methods use nonlinear classifiers to evaluate the goodness of the selected feature subsets. Recently, several authors proposed hybrid approaches that take advantages of both filter and wrapper methods. Examples of hybrid algorithms include *t*-statistics and a Genetic Algorithm [10], a correlation based feature selection algorithm and a Genetic Algorithm [9], Principal Component Analysis and an Ant Colony Optimization algorithm [9], chi-square approach and a multi-objective optimization algorithm [8], mutual information and a Genetic Algorithm [1][3]. 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 number of features under consideration.

Using feature selection for the second time involves almost the same amount of computational complexity as that of the first feature selection method. Applying any symbolic approach on the original dataset to transform the features into line segments is time consuming. To overcome these difficulties, we propose to apply one symbolic on the reduced subset obtained from feature selection method to transform the features into line segments. To the best of our knowledge, no significant work has been done in cascading of feature selection with symbolic approach. Hence we made an attempt here to apply the feature selection approach followed by a symbolic approach. Instead of using either feature selection or symbolic approach to reduce the dimension, the combinations of feature selection and symbolic methods can be applied as cascading approach, i.e., apply one feature selection like Mutual correlation on the original feature set to reduce its dimension and then apply symbolic approach on this reduced feature set to further reduce its dimension.

The rest of the paper is organized as follows. Section 2 proposes novel method of cascading the feature selection method mutual correlation with symbolic approach. Experimental results are presented in section 3 followed by conclusion in section 4.

2. PROPOSED METHOD
Cascading mutual correlation with symbolic approach
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 coefficient is invariant to scaling and translation. Hence two features with different variances may have same value of this measure. The P-dimensional feature vectors of N number of instances is given by

$$\mathbf{X}_i = \left[{}^{i}\mathbf{x}_1, \ldots {}^{i}\mathbf{x}_P \right] \quad I = 1,..., N$$

The mutual correlation [2] for a feature pair x_i and x_j is defined as

$$r_{x_i,x_j} = \frac{\sum_k {}^{k}x_i \, {}^{k}x_j - N\bar{x}_i\bar{x}_j}{\sqrt{\left(\sum_k {}^{k}x_i^2 - N\bar{x}_i^2\right)\left(\sum_k {}^{k}x_j^2 - N\bar{x}_j^2\right)}} \quad (1)$$

where k = 1,...N

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.

$$r_{j,\delta} = \frac{1}{\delta}\sum_{i=1,i\neq j}^{\delta}\left|r_{x_i,x_j}\right| \qquad (2)$$

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 correlation, i.e.

$$r_{j,\delta-1} = \frac{\delta r_{j,\delta} - \left|r_{x_\alpha,x_j}\right|}{\delta-1} \qquad (4)$$

Thus P (P<<D) dimensional feature subset obtained is given as input to the symbolic approach to transform into line segments [6].

These P features are arranged based on their value of variance before fitting the line and to discard some least significant features based on their value of variance, i.e., arrange these P features in descending order of their variances and then take only those features with high variance by discarding the features with low variance. Suppose that the data items are represented by d[i,j] where i is the sample and j is the feature and $1\leq i\leq N$, $1\leq j\leq P$, where N is the number of samples and P is the number of features. Line segments that best approximate the feature values of a sample i are detected by taking 3 consecutive points A(k,d[i,k]), B(k+1,d[i,k+1]) and C(k+2,d[i,k+2]) and testing for collinearity. The points are collinear only when the sum of the distances between two pairs of points approximately equals the distance between the third pair of points. i.e., AB≈BC+AC or BC≈AB+AC or AC≈AB+BC. Here AB means the Euclidean distance between the points A and B. If the points A, B and C are found to be nearly collinear then the points B, C and D(k+3,d[i,k+3]) is checked for collinearity until a point at (j, d[i,j]) which is not on the line A, B, C, D,… is encountered. The features d[i,k], d[i,k+1],… d[i,j-1] is approximated by a line joining between the points (k,d[i, k]) and (j-1,d[i,j-1]). A new line that fits the features d[i,j], d[i,j+1],…, is determined similarly. Fig.1 shows three line segments that approximate eight feature values. Then we compute the slope and average slope of all the line segments. The transformed symbolic features are the number of line segments and the average slope of three line segments.

Algorithm 1: Cascading Feature selection based on mutual correlation with symbolic approach

Input: Original feature set X of size N x D, M the required reduced number of features

Output: Reduced feature set of size M (M<<D)

Method:

Stage1: Feature selection based on mutual correlation

1. Initialize δ= D.
2. Discard feature x_α for α determined by equation (3).
3. Decrement δ= δ-1, if δ<M return the resulting M dimensional feature set and stop.
4. Recalculate the average correlations by using equation (4).
5. Go to step 3.

Stage 2: Transforming M features into symbolic features

Input: Reduced set of feature values $f_1, f_2,\ldots, f_M$.

Output: Reduced set of features F_1, F_2, F_3 and F_4.

6. Arrange the M features in descending order of their variance and take only those features with high variance (say k) by discarding features with low variance.
7. Find the cumulative sum of these k features namely f_1, $f_1+f_2,\ldots,f_1+f_2+\ldots+f_k$.
8. Fit the lines for the feature values of a sample by considering a threshold value and check for collinearity.
9. Find the number of line segments say F_1.
10. Compute the slope of each line segment and thereby compute the average slope of the line segments say F_2.
11. Set F_3= the value of the first feature value and F_4= the value of the last feature value.
12. Repeat the steps 7 to 12 for all the samples.

Algorithm 1 ends

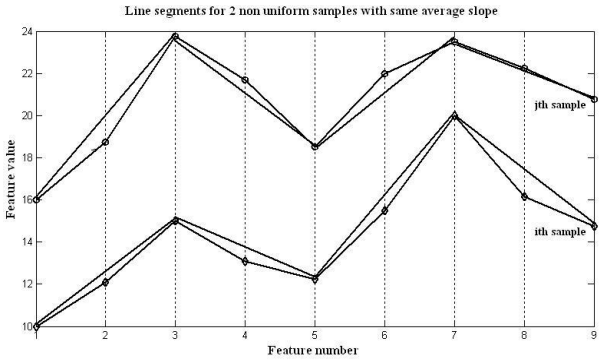

Fig.1: Line segments for two samples with same average slope.

The standard complete linkage clustering algorithm has been employed on these reduced feature set to obtain reliable clusters.

3. EXPERIMENTATION

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, Mutual correlation [2], and symbolic approach [6] 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. All precision, recall and F measure values are in percentage. Results of experiments performed on the standard datasets like WDBC (Wisconsin Diagnostic Breast Cancer), WBC

(Wisconsin Breast Cancer), CORN SOYBEAN and WINE datasets are shown in the following tables.

The superiority of the proposed model is established through the parameters precision, recall and F measure of the obtained clusters.

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:

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

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

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

3.1 Experimentation on WDBC dataset

The mammogram dataset of Wisconsin Diagnostic Breast Cancer (WDBC) consists of 569 instances each with 30 features. This contains two clusters having 212 malignant samples and 357 benign samples. Experimentation is conducted on this dataset and the results of Cascading Feature selection based on mutual correlation with symbolic approach (Algorithm 1) is tabulated in table 1 and the comparative results of all the 5 methods is tabulated in table 2. In the table, where we have compared all methods, the integer within the bracket stands for the number of features selected or extracted from the respective algorithm.

Table 1: Cluster results for the proposed method

Clusters	1	2
C_a	212	357
C_r	235	334
$C_a \cap C_r$	194	316
Precision	82.553	94.611
Recall	91.509	88.515
Average Precision = 88.582 Average Recall = 90.012 Average F Measure = 89.292		

Table 2: Comparison of 5 methods for WDBC dataset

Sl. No.	Methods	Average Precision	Average Recall	Average F measure
1	PCA(9)	87.947	75.663	81.344
2	LPP (9)	84.575	86.747	85.627
3	Mutual Correlation(18)	89.523	78.538	83.638
4	Symbolic(4)	89.743	88.664	**89.200**
5	Mutual Correlation(18) + Symbolic(4)	88.582	90.012	89.292

3.2 Experimentation on WBC dataset

The mammogram dataset of Wisconsin Breast Cancer (WBC) consists of 683 instances each with 16 features. This contains two clusters having 239 malignant samples and 444 benign samples. Experimentation is conducted on this dataset and the results of Cascading Feature selection based on mutual correlation with symbolic approach (Algorithm 1) is tabulated in the table.3, and the comparative results of all the 5 methods is tabulated in table 4.

Table 3: Cluster results for the proposed method

Clusters	1	2
C_a	239	444
C_r	234	449
$C_a \cap C_r$	220	430
Precision	94.017	95.768
Recall	92.050	96.847
Average Precision = 94.893 Average Recall = 94.449 Average F Measure = 94.670		

Table 4: Comparison of 5 methods for WBC dataset

Sl. No.	Methods	Average Precision	Average Recall	Average F measure
1	PCA (7)	89.454	78.114	83.400
2	LPP (8)	88.782	82.121	85.322
3	Mutual Correlation(10)	91.484	83.135	87.110
4	Symbolic	93.963	90.758	**92.333**
5	Mutual Correlation(10) + Symbolic	94.893	94.449	**94.670**

3.3 Experimentation on CORN SOYBEAN dataset

The Corn Soybean dataset consists of 61 samples and each sample is of 24 dimensions. The dataset contains two clusters C_1= 32 corn samples and C_2= 29 soybean samples. Experimentation is conducted on this dataset and the results of Cascading Feature selection based on mutual correlation with symbolic approach (Algorithm 1) is tabulated in the table 5, and the comparative results of all the 5 methods is tabulated in table 6.

Table 5: Cluster results for the proposed method

Clusters	1	2
C_a	32	29
C_r	32	29
$C_a \cap C_r$	32	29
Precision	100.00	100.00
Recall		
Average Precision = 100.00 Average Recall = 100.00 Average F Measure = 100.00		

3.4 Experimentation on WINE dataset

The WINE database contains results of chemical analysis of wines grown in the same region of Italy but derived from three different cultivators. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The database contains 178 instances categorized into three classes with 59, 71 and 48 instances bearing 13 features in each class, respectively. Experimentation is conducted on this dataset and the results of Cascading Feature selection based on mutual correlation with symbolic approach (Algorithm 1) is tabulated in the table 7, and the comparative results of all the 5 methods is tabulated in table 8.

Table 7: Cluster results for the proposed method

Clusters	1	2	3
C_a	59	71	48
C_r	70	58	50
$C_a \cap C_r$	59	58	48
Precision	84.286	100.000	96.000
Recall	100.000	81.690	100.000
Average Precision = 93.429 Average Recall = 93.897 Average F Measure = 93.662			

Table 6: Comparison of 5 methods for CORNSOYBEAN Dataset

Sl. NO	Methods	Average Precision	Average Recall	Average F measure
1	PCA (8)	98.054	98.151	98.103
2	LPP (11)	98.485	98.276	98.380
3	Mutual Correlation(20)	98.485	98.276	98.380
4	Symbolic	100.00	100.00	**100.00**
5	Mutual Correlation(20) + Symbolic	100.00	100.00	**100.00**

Table 8: Comparison of 5 methods for WINE dataset

Sl. No.	Methods	Average Precision	Average Recall	Average F measure
1	PCA(10)	89.845	88.749	89.294
2	LPP(11)	91.095	91.608	91.351
3	Mutual Correlation(11)	89.945	89.191	89.566
4	Symbolic	89.085	89.576	**89.330**
5	Mutual Correlation(11) + Symbolic	93.429	93.897	**93.662**

4. CONCLUSION

In this paper, a novel cascading approach of dimensionality reduction is proposed. In this method, the feature selection method based on mutual correlation followed by transforming the reduced features to symbolic type (line segments) is applied. Experiments are conducted on well-known datasets like WDBC, WBC, CORN SOYBEAN and WINE to demonstrate the superiority of the proposed model. From the table 2, it is clear that, for WDBC dataset the proposed cascading method achieve better performance in terms of F measure values when compared to PCA, LPP, Mutual Correlation methods, From the table 6, it is clear that, for WBC dataset the proposed cascading method achieve better performance in terms of F measure values when compared to PCA, LPP and Mutual Correlation methods. From the table 6 and 8, it is clear that, for CORN SOYBEAN and WINE dataset, the proposed cascading method achieve better performance in terms of F measure values when compared to PCA, LPP, Mutual Correlation, and symbolic method.

5. ACKNOWLEDGEMENTS

I thank UGC for granting the Minor Research Project to carry out this research work.

6. REFERENCES

[1] Fatourechi M, Birch G and Ward R K (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.

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

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

[4] Kira K and Rendell L A (1992), A practical approach to feature selection, Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, Scotland, UK, Morgan Kaufmann Publishers, San Mateo, pp. 249-256.

[5] Lalitha Rangarajan and Veerabhadrappa. (2010), Dimensionality reduction through transformation of features into line segments, International Journal of Recent Trends in Engineering and Technology(IJRTET), ACEEE, Vol.4, No.2,pp:91-95.

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

[7] Shazzad K M and Park J S(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.

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

[9] Yan Z, C Yuan (2004), Ant colony optimization for feature selection in face recognition, Lecture notes in Computer Science 307,pp. 221–226.

[10] Young D M, Odell P L and Marco V R (1985), Optimal linear selection for a general class of statistical pattern recognition models, Pattern Recognition Letters, pp 161-165.

7. AUTHOR PROFILE

Dr. Veerabhadrappa has obtained M.Sc. and Ph.D degrees in Computer Science from the University of Mysore, India, respectively in the years 1989 and 2011. He is Associate Professor and head of the Department of Computer Science, University College, Mangalore, Karnataka. He has authored 14 peer-reviewed papers in journals and conferences. He is the reviewer of the Journal of Neurocomputing and International Journal of Computer Applications (IJCA). He has delivered many lectures at various workshops, seminars and conferences. He is the member of various academic bodies of several universities. His area of research covers dimensionality reduction, face/object recognition and symbolic data analysis.