UGC Minor Research Project on

<u>Dimensionality reduction using feature selection with symbolic approach</u>

- Dr.Veerabhadrappa

Executive Summary

The measurements or properties used to classify objects/patterns are called dimensions. In case of real world data, the size of the feature set is very high. Therefore, scientists are forced to explore high-dimensional data more often than ever, since the information impacting on daily life is growing enormously. Moreover, dimensionality reduction is necessary for successful pattern recognition in order to curb the effect known as the 'curse of dimensionality' when the dimension of the data is much larger than the limited number of patterns in a training set. This effect manifests itself in decreasing the recognition accuracy as the dimensionality grows. On the other hand, a part of the information is lost during dimensionality reduction; hence it is important that the resulting low-dimensional data to preserves a meaningful structure and the relationships among the original high-dimensional space. Many statistical learning problems involve some form of dimensionality reduction either explicitly or implicitly. Much importance has been attributed to the process of dimensionality reduction which is the most fundamental and one of the important stages in the field of Pattern Recognition. The dimensionality reduction is concerned with the problem of mapping data points that lie in a high dimensional data space to a low dimensional embedding space. Common methods that perform dimensionality reduction are feature selection (in which a subset of features is selected) and feature transformation (in which linear/nonlinear combination of original features is found and later few transformed features are selected). In this project work, we have proposed dimensionality reduction for cascading of dimensionality reduction methods.

In order to enhance the performance, we have presented multilevel dimensionality reduction methods in which the feature selection methods are combined with another feature selection method or feature extraction method. Applying one approach like feature selection on the original feature set to reduce its dimension followed by another approach (again either feature selection or feature extraction) on this reduced feature set results in further reduction of the dimension.

The dimensionality reduction methods which transform the features into symbolic type exploits the property of collinearity of feature values, cumulative feature sum and variance based cumulative feature sum. The feature values are transformed into line segments and thus reduced into 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 number of line segments and same average slope. Hence the entire feature set is reduced to only 4 features namely, number of line segments, average slope of the line segments, first feature value and last feature value.

In this report, two novel cascading approaches of dimensionality reduction are proposed.

- **First one is the feature selection method based on mutual correlation followed by transforming the reduced features to symbolic type (line segments).**
- **Second one is the feature selection method based on feature quality measure followed by transforming the reduced features to symbolic type (line segments).**

The superiorities of the newly proposed models of dimensionality reduction are established experimentally by an extensive comparative analysis with other well-known existing methods. Experiments are conducted on well-known datasets like

WDBC, WBC, CORN SOYBEAN and WINE to demonstrate the superiority of the proposed models. The F measure performance of these methods for all the 4 datasets is compared. For all the data sets, the proposed cascading methods achieve better performance in terms of F measure values when compared to PCA, LPP, Mutual Correlation and FQ measure methods, but little less when compared to the symbolic method. For CORN SOYBEAN and WINE dataset the proposed cascading methods achieve better performance in terms of F measure values when compared to PCA, LPP, Mutual Correlation, FQ measure and symbolic method.

Based on these two methods two papers are published in the following reputed journals:

- Veerabhadrappa, Lalitha Rangarajan, "Fusion of Feature Selection with Symbolic Approach for Dimensionality Reduction", **International Journal of Computer Systems (IJCS), pp: 267-270, Volume 3, Issue 3, March 2016 *(ISSN: 2394-1065)*.** The link to download the paper is http://www.ijcsonline.com/IJCS/IJCS_2016_0303015.pdf
- Veerabhadrappa, "Dimensionality Reduction by Cascading Mutual Correlation with Symbolic Approach", **International Journal of Computer Applications (IJCA), pp: 5-8, Volume 140-No.7, April 2016 *(ISSN: 0975-8887)*.** The link to download the paper is **http://www.ijcaonline.org/archives/volume140/number7/24604-2016909362**

Conclusion:

We can explore the possibility of combining more than one feature selection method with symbolic approach to reduce the dimension further.