Combining Clustering with Classification: A Technique to Improve Classification Accuracy

Yaswanth Kumar Alapati*
Assistant Professor, Dept. of Information Technology,
R.V.R. & J.C. College of Engineering, Guntur, A.P
Email: alapatimail@gmail.com

Korrapati Sindhu
Assistant Professor, Dept. of Computer Science & Engineering,
R.V.R. & J.C. College of Engineering, Guntur, A.P

Abstract - Most of the Real-World datasets are High-Dimensional. To find patterns in High-Dimensional datasets it would be useful to be applying modern methods of classification such as support vector machines. These methods are computationally expensive. To find useful patterns in High-Dimensional data Feature Selection Algorithms can be used. Results show that clustering prior to classification is beneficial. For efficient results it is better to apply feature selection algorithms for dimensionality reduction.The results also show that for each dataset it is important to choose a clustering method carefully.

Keywords-Clustering, Classification, Feature Selection

I. INTRODUCTION

Data Mining is a wide area that integrates techniques from various fields including machine learning, artificial intelligence, statistics and pattern recognition for the analysis large volumes of data. There have been a large number of data mining algorithms embedded in these fields to perform different data analysis tasks. Data mining has attracted lot of attention in the research industry and in society as a whole in recent years, due to enormous availability of large amount of data and the need for turning such data into useful information and knowledge. Data Mining is the field of discovering new and potentially useful information from huge databases [3].

Classification technique is capable of processing a wider variety of data than regression and is growing in popularity [1]. Classification is the process of finding a model that describes and distinguishes classes and concepts. Classification is a form of data analysis that extracts models describing important data classes. Many classification methods have been proposed by researchers in machine learning, pattern recognition and statistics. Most algorithms are memory resident typically assuming small datasets.

Classification has numerous applications including fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis. Data classification is two-step process, consisting of learning step (where a classification model is constructed) and classification step (where the model is used to predict the class label for the given data).

The learning step is also called as training phase, where a classification algorithm builds the classifier from a set of tuples from the database. In the classification step model is used to predict the class label of given data. Clustering can be used in the fields like Biology, Information Retrieval, Medicine, Capturing information of spatial cells [2].

Feature Selection is the process of selecting a subset of relevant features for use in model construction. In machine learning and statistics feature selection is also known as variable selection, attribute selection or variable subset selection [4]**.** Dimensionality reduction is one of the most popular techniques to remove noisy (i.e. irrelevant) and redundant features. Dimensionality reduction techniques can be categorized mainly into feature extraction and feature selection. Feature extraction approaches project features into a new feature space with lower dimensionality and the new constructed features are usually combinations of original features.

For the classification problem, feature selection aims to select subset of highly discriminant features. In other words, it selects features that are capable of discriminating samples that belong to different classes. For the problem of feature selection for classification, due to the availability of label information, the relevance of features is assessed as the distinguishing different classes. For example, a feature Fi is said to be relevant to a class Cj if Fi and Cj are highly correlated.

In the past thirty years, the dimensionality of the data involved in machine learning and data mining tasks has increased explosively. Data with extremely high dimensionality has presented serious challenges to existing learning methods [6], i.e., the curse of dimensionality [5]**.** With the presence of a large number of features, a learning model tends to over fit, resulting in their performance degradation.

II. PROPOSED METHOD

It is not always feasible to apply classification algorithms directly on dataset. First the data has to be pre-processed. Pre-processing may also involve dimensionality reduction. The proposed method proves that the classifier works well with clustered data that is before applying any classification algorithm on dataset cluster the data and then apply classification algorithm there by the accuracy of classifier is improved. For High-Dimensional datasets first apply Feature Selection Algorithm. For each dataset it is important to choose a clustering method carefully.

Different phases in the proposed framework are

- Feature selection
- Clustering
- Classification

1) Feature Selection

Feature selection is the process of identifying a subset of the most useful features that produces compatible results as the original entire set of features. Features provide the information about the data set. In high-dimensional data representation each sample is described by many features. The data sets are typically not task specific, many features are irrelevant or redundant and should be pruned out or filtered for the purpose of classifying target objects. Given a set of features the feature selection problem is to find a subset of features that "maximizes the learner's ability to classify patterns".

The feature selection algorithms used in proposed framework are Correlation-based Feature Selection (CFS), Relief-F

2) Clustering

After reducing the dimensionality of a dataset apply clustering algorithm on reduced dataset. After clustering add the cluster id to the dataset. The clustering algorithms used in the proposed frame work are k-means and hierarchical clustering

3) Classification

Apply the classification algorithm on clustered data. The classification algorithms used in the proposed framework are Naive Bayes Classifier and Neural Network Classifier

III. EXERIMENTAL SETUP

A number of experiments on benchmark datasets have been conducted to verify the strength of the proposed approach. A summary of the datasets is presented in Table I. 10-fold cross validation is used for reporting the classification results for all the datasets

TABLE I DATASETS

Dataset	**#Instances**	**#Attributes**	**#Classes**
Lung Cancer	32	57	3
Coil2000	5822	86	2
Mfeat-Fourier	2000	77	10
Arrhythmia	452	280	16

K-means and Hierarchical Clustering algorithms are used for clustering the datasets. The neural networks are trained using tan sigmoid activation functions for the neurons and Levenberg-Marquardt back propagation method for learning of the weights.

IV. EXPERIMENTAL RESULTS

The following results are discussed

1. Impact of applying clustering prior to classification.
2. Impact of using different Feature subset selection algorithms

A. Impact of applying clustering prior to classification

Table II shows the accuracy of a classifier without applying any clustering or feature subset selection algorithms.

TABLE II Accuracy of Different Classifiers with Different Datasets

SNO	Dataset	Accuracy	
		Naïve Bayes Classifier	Neural Network Classifier
1	Lung Cancer	84.37	71.87
2	Coil2000	78.71	91.61
3	Mfeat-Fourier	75.75	66.70
4	Arrhythmia	62.38	59.29

The accuracy of a classifier can be improved by applying clustering technique before applying classification algorithms on dataset. Table III shows the accuracy of a classification algorithm by applying clustering prior to classification.

TABLE III Accuracy of Different Classification Algorithms with clustering

SNO	Dataset	Accuracy of Naïve Bayes Classifier with		Accuracy of Neural network Classifier with	
		K-means Clustering(%)	Hierarchical Clustering (%)	K-means Clustering (%)	Hierarchical Clustering (%)
1	Lung Cancer	87.5	96.87	96.87	96.875
2	Coil2000	94.59	97.2	91.618	95.4
3	Mfeat-Fourier	90.7	98.95	97.6	99
4	Arrhythmia	69.25	97.52	98.3	99

B) Impact of using different Feature subset selection algorithms

High-Dimensional data may slow down the mining process and reduce the accuracy. Accuracy is defined as the ration of number of instances for which the outcome is correct to the total number of tests made. Accuracy of a classifier can be improved by applying Feature Subset selection Algorithms and also the classification time can be reduced.

Table IV shows the accuracy of a classifier by applying feature subset selection algorithm prior to k-means clustering algorithm

TABLE IV Accuracy of classification algorithm with Feature Selection and Clustering

SNO	Dataset	Accuracy of Naïve Bayes classifier (%)	
		CFS	Relief F
1	Lung Cancer	90.63	93.8
2	Coil2000	91.7	94.58
3	Mfeat-Fourier	99.2	97.55
4	Arrhythmia	69.25	63.94

V. CONCLUSION

A novel method is proposed to improve the accuracy of a classification algorithm. The evidence from the experimental results shows that applying clustering technique prior to classification algorithm is beneficial. Experimental Results also shows that Accuracy of a classifier can be improved by applying Feature Subset Selection Algorithms.

REFERENCES

[1] Jiawei Han and MichelineKamber Data Mining: Concepts and Techniques,2nd edition
[2] J. A. Hartigan and M. A. Wong (1979) "A K-Means Clustering Algorithm", Applied Statistics, Vol. 28, No. 1, p100-108
[3] L.Arockiam, S.Charles,Arulkumar et.al(2010), "Deriving Association between Urban and Rural Students Programming Skills", International Journal on Computer Science and Engineering Vol. 02, No. 03, pp 687-690.
[4] https://en.wikipedia.org/wiki/Feature_selection
[5] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer, 2001
[6] H. Liu and H. Motoda. Computational Methods of Feature Selection. Chapman and Hall/CRC Press, 2007