Feature Subset Selection Using Agglomerative Clustering

Priyansh Jain[1], Dr. Anuradha Purohit[2]
[1] *Department of Computer Engineering*
[2]*Department of Computer Technology and Applications*
(E-mail: priyanshjain648@ymail.com)

***Abstract*—** The goal of classification is to accurately predict the target class for each sample present in the dataset according to features present in it. Performance of a classifier can be degraded if the dataset contains large number of redundant or irrelevant features. For handling this problem, feature selection technique is used. Feature Selection is the process of selecting the most useful features from the given dataset. There are various methods used for Feature Selection of which the clustering based methods like hierarchical clustering, density based, partitional clustering etc. are popularly used. In this paper, an agglomerative clustering based feature selection approach is proposed. In the proposed work, features are divided into clusters by using Ward's method. On the basis of clustering of features, dendrogram is created. The dendrogram created is sectioned at k number of clusters. Further the Representative Feature of each cluster is selected using Distance Correlation. Features in different clusters are relatively independent; the proposed work has a high probability of producing a subset of relevant features. The approach has been tested on various benchmark datasets like wine, WDBC, lung cancer, sonar, WPBC, spambase. The results obtained are outperformed when compared with other clustering approaches.

Keywords— Feature Selection; Ward's method; Distance Correlation; Classification; Representative Feature.

I. INTRODUCTION

In the digital era, handling the massive data is a challenging task which may contain irrelevant and the redundant features. If data mining tasks like clustering and classification are performed on this type of data, then undesirable outputs may be generated. For solving the problem of less useful features, feature selection and feature extraction techniques are used, especially when there are too many features in problem space. Feature subset selection is the process of selecting a subset of relevant features [1]. It is also called as variable selection or attribute selection. Feature subset selection includes and exclude attributes present in the data without changing them. Two popular methods used for feature selection are wrapper and filter.

Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy [1].

Filter techniques assess the relevance of features by looking at the intrinsic properties of the data. In filter criteria, all the features are scored and ranked based on certain statistical criteria. The features with the highest ranking values are selected and the low scoring features are removed [1]. In filter method clustering based feature selection is also a popular technique. Clustering based feature selection is the process which divides similar features in to clusters using distance or similarity criteria. Various methods of clustering are hierarchical, partitioning and density based clustering etc. Agglomerative clustering is popularly used technique for feature selection.

In this paper, a feature subset selection approach using agglomerative clustering has been proposed. The proposed approach uses Ward's method to find relevant features from datasets. This paper is organized in different sections. Section II describes the related work done in the field of feature selection. The proposed feature subset selection approach using agglomerative clustering is discussed in section III. In section IV, the experimental results on various datasets are presented, discussed and compared with other clustering approaches. The conclusion of the paper has been presented in section V.

II.RELATED WORKS

Feature Selection is one of the techniques that reduces the dimensionality of the data. The small subset of relevant features is selected that maximizes relevance of the features to the target class labels and minimizes redundancy [2]. The need of performing feature subset selection is to improve the performance of classification and clustering by improving the predictive accuracy. A lot of work has been reported in the literature to perform feature selection. Arauzo et al. [3] proposed a feature selection algorithm based on attribute estimation known as relief algorithm. The algorithm finds similarity of feature to target class by assigning a relevance value to every feature. On the basis of relevance value it select features which are similar to target class. H Park et al. [4] proposed an extended Relief algorithm which is inspired by nearest-neighbours and it performs better specifically for similar types of induction algorithms. Since Relief randomly samples instances and their neighbours from the training set, the answers it gives are unreliable without a large number of samples. Butterworth et al. [5] proposed an hierarchical clustering based feature selection algorithm. The algorithm cluster features using a special metric of Barthelemy-Montjardet distance, and then makes use of dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. Song Q et al. in [6] proposed a clustering based feature selection algorithm which is applied in two steps. In

the first step, clustering of features are done with the help of graph-theoretic clustering method. In the second step, the representative feature of each cluster is chosen as the feature having maximum mutual information with the target class. Zeinab Dehghan et al. [7] proposed an approach for feature subset selection using bottom up clustering. In this algorithm clustering is done with the help of representative features of clusters using mutual information. Representative feature in each cluster is the feature with the maximum mutual information against other features in that cluster.

With the rapid increase of dimensionality of data such as genomic micro-array data, text categorization and digital images, feature selection has become an important issue, though it is still considered to be an intractable problem in machine learning and data mining. In many applications, the data is usually represented by a huge number of features (attributes), and the raw data often contain many uninformative (irrelevant and redundant) ones which may largely degrade the learning performance and compromise the quality of data mining task. Therefore, to retain important features and remove irrelevant and redundant ones, various new techniques are evolving for feature subset selection.

III. PROPOSED APPROACH

An approach for performing Feature Subset Selection using Agglomerative Clustering has been proposed. The proposed approach works in two steps. In the first step, features are divided into clusters by using Wards method. In the second step, the representative feature of each cluster is selected using Distance Correlation. These steps are carried out using following procedure:

- Clustering of Features using Ward's method.
- Creating Dendrogram of Features.
- Obtaining k Number of Clusters.
- Selecting Representative Feature using Distance Correlation.
- Testing obtained Feature Subset using SVM Classifier

The steps are explained in detail as follows:

A. Clustering of Features using Ward's Method

Ward's method is a hierarchical clustering method which is used to calculate error sum of squares (ESS) between features or clusters. On the basis of minimum ESS, Ward's method merge features and form clusters. Firstly the input from the dataset is selected and each feature is considered as an individual cluster. Then the error sum of square between each combination of features is calculated using Ward's method as given in equation (1).

$$e^2_x = \Sigma_{i=1}\Sigma_{j=1}[Y_{ij} - M_j]^2 \quad \text{.....(1)}$$

Where,

e^2 e^2x is the error sum of square at x cluster
n is the number of columns
p is the number of rows
Yij is the feature
M is the mean of attributes.

On the basis of minimum value of error sum of square two features are merged from different clusters and form a new cluster. In the next iteration the new cluster and other remaining feature repeats the process until all feature merge in a single cluster.

B. Creating Dendrogram of Features

The clustering of features has been done using Ward's method. The dendrogram of the features obtained is created. A dendrogram is a tree like structured graph which is used to represent the clustering sequence of features. The result of clustering is presented either as the distance or the similarity between the clustered rows or columns depending on the selected distance measure. By visualizing graph it is easy to understand and predict the number of clusters. The example dendrogram is shown in Fig. 1.

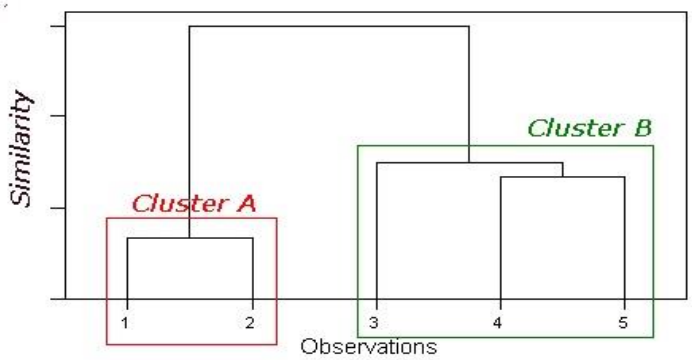

Fig. 1 Dendrogram

C. Obtaining k Number of Clusters

The dendrogram created is then cut at various levels according to the threshold value as per the requirements from the user. Hence, k clusters will be obtained after the dendrogram is sectioned. The k clusters along with their features are stored.

D. Selecting Representative Feature using Distance Correlation

The clusters along with their features are obtained by sectioning dendrogram at various levels. Then the Representative Feature of each cluster is selected using Distance Correlation as given in equation (2).

$$dcor = sqrt(dcov/sqrt(dvarx * dvary)) \quad \text{.....(2)}$$

where,

dcor is the distance correlation between features.
dcov is the distance covariance between features.
dvarx, dvary are the distance variance of features x and y respectively.

Now dcor value as obtained will be used for finding Representative Feature of each cluster. Hence, the feature with highest Distance Correlation value will be selected from each cluster as Representative Feature. The Representative Features that are obtained are stored in feature subset. The modified

dataset is then obtained which contains representative feature of clusters.

E. Testing the Feature Subset Obtained using SVM Classifier

The feature subset obtained is tested using SVM classifier by tenfold cross-validation method. That is, each modified dataset is divided into tenfold partitions where nine partitions are used for training set and one partition is used for test set. The same procedure is performed iteratively nine times after exchanging the role of each partition so that all partitions are used for test once.

IV.EXPERIMENTAL SETUP, RESULTS AND ANALYSIS

The proposed work is implemented using MATLAB r2013a. To evaluate the results obtained using proposed approach six benchmark datasets are taken from UCI Machine Learning Repository. These datasets are Wine, Breast Cancer Wisconsin Diagnostics (WDBC), Sonar, Lung Cancer, Spambase and Breast Cancer Wisconsin Prognostic (WPBC). Wine and Spambase is multiclass and integer/real valued, Sonar, WDBC and WPBC are real valued and multiclass datasets, Lung Cancer is integer valued and multiclass datasets. A brief description of these datasets is summarized in Table I.

TABLE I. DATASETS USED FOR EXPERIMENTATION

Datasets	Number of Features Present	Number of Classes	Number of Instances	Attribute Type
Wine	13	3	178	Integer, Real
WDBC	32	2	569	Real
Sonar	60	2	208	Real
Lung Cancer	56	3	32	Integer
Spambase	57	2	4601	Integer
WPBC	34	2	198	Real

TABLE II. RESULTS OBTAINED FOR THE PROPOSED APPROACH WITHOUT FEATURE SELECTION

Datasets	Number of Features Present	Average Number of Features Obatined	Classification Accuracy (in %)
Wine	13	13	95
WDBC	32	32	96.90
Sonar	60	60	64
Lung Cancer	56	56	43.3
Spambase	57	57	60
WPBC	34	34	71

To evaluate the accuracy of selected features obtained using agglomerative clustering, SVM classifier has been used. The accuracy of SVM classifier is obtained by tenfold cross-validation method. In 10-CV method, each dataset is divided into tenfold partitions that nine partition are used for training set and one partition is used for test set. The same procedure is performed nine times after exchanging the role of each partition so that all partitions are used for test once.

TABLE III. RESULTS OBTAINED FOR THE PROPOSED APPROACH WITH FEATURE SELECTION

Datasets	Number of Features Present	Average Number of Features Obatined	Classification Accuracy (in %)
Wine	13	11	95.88
WDBC	32	24	97.50
Sonar	60	7	64
Lung Cancer	56	20	60
Spambase	57	30	60.73
WPBC	34	23	71.57

As per the results obtained using proposed approach are shown in Table 3. 95.88% classification accuracy is obtained for wine dataset, 97.50%, 60.00%, 64.00%, 60.73%, 71.57% classification accuracy is obtained for WDBC, lung cancer, sonar, spambase and WPBC datasets respectively.

The results obtained are compared with the result presented in Accuracy of Unsupervised Attribute Clustering Algorithm [8] and Accuracy of Attribute Clustering Algorithm [9].The comparison chart for the datasets taken is as shown in table 3.

TABLE IV. COMPARISON BETWEEN PROPOSED APPROACH AND PREVIOUS APPROACHES

Datasets	Accuracy of Unsupervised Attribute Clustering Algorithm Approach [8] (in %)	Accuracy of Attribute Clustering Algorithm Approach [9] (in %)	Accuracy of Proposed Approach (in %)
Wine	92.59	92.70	95.88
WDBC	92.55	93.50	97.50
Sonar	72.33	74.04	64
Lung Cancer	-	43.70	60

It is clear from table 4 that our proposed approach outperforms with other approaches in terms of classification accuracy because distance correlation easily detects linear and non linear relationships among features.

V.CONCLUSION

In this work, an approach for feature subset selection using agglomerative clustering is proposed to perform feature selection. The proposed work create cluster of similar features using Ward's method. The Representative Feature from each cluster is selected using Distance Correlation. The proposed work reduces redundancy, storage requirements and improves classification accuracy because only one feature is selected from each cluster and the feature selected is more similar to target class. For evaluating classification accuracy of the obtained feature subset, Support Vector Machine classifier is used. Overall accuracy of the classifier for obtained feature subset is evaluated using ten-fold cross validation method.

Benchmark datasets like wine, WDBC, sonar, lung cancer, Spambase and WPBC are used to test the proposed work and comparable results are obtained.

VI.REFERENCES

[1] https://machinelearningmastery.com/an-introduction-to-feature-selection/

[2] Isabelle Gyuon, Andre Elisseeff: An Introduction to variable and Feature Selection, Journal of Machine Learning Research, 1157-1182 (2003).

[3] A.Arauzo-Azofra, J.M.Benitez, J.L.Castro: A Feature Set Measure Based on Relief, Proc., 5^{th} International Conference on Recent Advances in Soft Computing, 104-109 (2004).

[4] H. Park, H. Kwon: Extended Relief Algorithms in Instance-Based Feature Filtering," Proc., 6^{th} International Conference on Advanced Language Processing and Web Information Technology, 123- 128 (2007).

[5] R. Butterworth, G. Piatetsky-Shapiro, D.A. Simovici:On Feature Selection through Clustering, Proc., IEEE 5^{th} International Conference on Data Mining, 581-584 (2005).

[6] Song Q, Ni J, Wang G.:A fast clustering-based feature subset selection algorithm for high-dimensional data, IEEE Transaction on Knowledge Data Engineering 25(1): 1-14 (2013).

[7] Zeinab Dehghan, Eghbal G. Mansoori: Feature subset selection using bottom up clustering, Springer, London, 1-10 (2016).

[8] P.Y. Zhou, K.C.C. Chan:An Unsupervised Attribute Clustering Algorithm for Unsupervised Feature Selection, Proc., IEEE International Conference on Data Science and Advanced Analytics, Paris, France, 1-7 (2015).

[9] Elham Chitsaz, Mohammd Taheri, Mansour Z.:An improved fuzzy feature clustering and selection based on chi-squared-test, Proc., International Multiconference of Engineers and Computer Scientists, IMECS, 35-40 (2009).