CLUSTERING BASED FEATURE SUBSET SELECTION FOR SEMI-SUPERVISED DATA

M. Kumaresa Babu
Assistant Professor,
Department of Computer Science and Engineering
Tagore Engineering College, Chennai
mbabu1984@gmail.com

Abstract :- *The curse feature subset selection magnifies when the training data are semi-supervised. The proposed framework uses both Laplacian score and fisher score to rank the features. To enhance the performance of the subsequent machine learning algorithm, the framework incorporates correlation based maximum spanning tree to eliminate redundancy among the selected subset of features performs better than 3 other feature subset selection algorithms in terms of both classification accuracy, evaluated using 4 different datasets.*

Keywords*:Correlation based maximum spanning tree, Feature subset selection,*

1.INTRODUCTION

Dimensionality reduction is the major issue with high-dimensional data. It is classified into two techniques namely Feature Selection and Feature extraction. Their aim is to improve the Learning performance. Feature Selection is the process of selecting relevant features for high-dimensional data. Its selects the subset of feature from the original data without any transformation and maintains the originality of the feature. Feature selection is applied for Supervised and Unsupervised learning. Supervised feature selection is performed for data whose target is known. There are several techniques which performs supervised feature selection like Relief and Relief [1], Fisher score[2], based on Mutual information such as mRMR (minimum Redundancy –Maximum Relevance) [3], and FCBF (Fast Correlation-Based Filter) [4]. Unsupervised feature selection becomes a tedious process

because of the absence of target labels. Some of the techniques which performs unsupervised feature selection are variance score [5], Laplacian Score [6], PCA (Principal Component Analysis) [7], MiCi (Maximum information Compression index) [8].In most of the real world data consists of combination of both supervised and unsupervised data. Thus we move on to semi-Supervised feature selection technique.

In this work we focus on semi-supervised feature selection for dimensionality reduction. Initially redundant features are eliminated for both supervised and unsupervised data. Then clusters are formed for unsupervised data. Thus the redundant features are removed by selecting a single Feature from each cluster with the help of fisher score. Thus the obtained features would be relevant. The performance can be evaluated by using Classification (Accuracy) and Redundancy Analysis (relevance).

2. RELATED WORKS

In this section we discuss the two scores on the basis of the score function. We

Received on : 26.08.2020
Revised on : 26.09.2020
Accepted on : 03.08.2020
Published on :04.08.2020

Corresponding Author: **M.Kumaresa Babu** [mbabu1984@gmail.com]

describes Fisher Score [10] and the Constraint Score [9] with their uses. The Fisher Score was used for supervised feature selection. However it selects feature independently according to their score under fisher criterion, which leads to optimal subset of features. The feature with high quality assigns similar values to instances in the same class and different values to instances from different classes.

The score for i^{th} feature S_i will be calculated by fisher score as follows [10]:

$$S_i = \frac{\sum n_j(\mu_{ij}-\mu_i)^2}{\sum n_j \rho_{jj}^2} \qquad (1)$$

Where, μ_{ij} represents the mean of i^{th} feature in j^{th} class, ρ_{jj} represents variance of i^{th} feature in j^{th} class,n_j represents number of instances in the j^{th} class and μ_i represents mean of i^{th} feature.

The fisher score evaluates each feature individually and it cannot handle feature redundancy.

The constraints provides desired partition and makes possible for unsupervised feature selection to increase the performance [9].For any pair of constraints the different types of constraints generated are:

- Must-Link Constrain(ML)
- Cannot-Link Constraint(CL)

It guides feature selection according to the pair wise selected constraints which can be classified into two sets are given above which can be expressed as Ω_{ML} (Must-Link constraint) and Ω_{CL}(C

$$\mathbf{C_r} = \frac{\sum_{(xi,xj)} \in \Omega_{ML} (fri\text{-}frj)^2}{\sum_{(xi,xj)} \in \Omega_{CL} (fri\text{-}frj)^2} \qquad (2)$$

By using this it utilizes only few labelled data and also this score has several disadvantages that the score depends on the selected constraint subset and it leads to decrease the performance of the feature selection process.

3.PROPOSED APPROACH

In the following section, we present two algorithms for feature selection Fisher Discriminant Analysis for supervised data set and Laplacian Score for unsupervised data set. The both Algorithm is used for calculating scores for each features in the semi-supervised context. It is a greedy algorithm that finds minimum spanning tree for connected weighted graph. Which means it finds subset of edges the forms a tree includes each vertex, where the total weight of all edges in the tree is minimized.

Algorithm 2 Kruskal's Algorithm

Input: N×N Weight matrix in Adjacency form

Output: N×N Adjacency Matrix of Minimum Spanning Tree

1. Let G= (V, E) be the graph

{

Initialise graph as T = (V, Ø)

2. Arrange Edge E in increasing order cost

3. Select the Edge

For (i=1; i<=n-1; i++)

{

Select the smallest cost Edge
If (edge connects two different connected components) add Edge to T

4.EXPERIMENTAL STUDY

In this section, we evaluate the performance of our framework and compare it with other feature selection methods. This comparison concerns the classification accuracy results that we present in Fig. 1, Fig.2shows the performance of the proposed algorithm(Laplancian Fisher cluster based feature selection-LFC) for the test data by classification accuracy (NN) versus different datasets and feature selection algorithms. The speciality of the proposed algorithm is that it automatically converges to select the number of features required for classification. Whereas, for other feature subset selection techniques the number of features is decided by the user or the optimization algorithm. The decision of optimization algorithm increases the time complexity since it involves more iteration in selecting the optimal number of features. Table 3 shows the number of features selected by LPC algorithm, and the same number of features is used to test the performance of other algorithms. And it is found that results (selection of number of features) are same as that of optimization algorithm

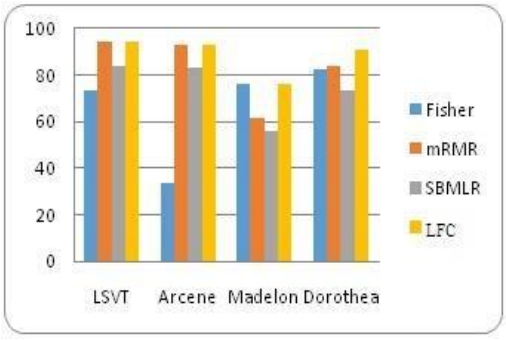

Fig. 1. Classification accuracy for test data on different datasets.

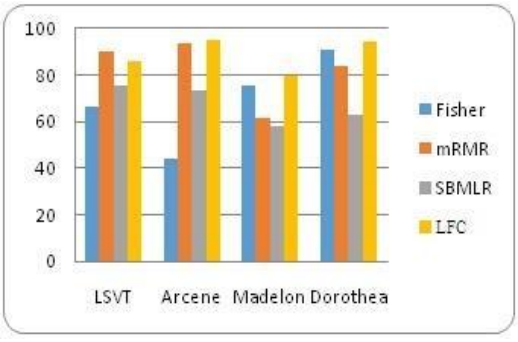

Fig. 2. Overall Classification performance of different datasets

5 . CONCLUSION

In this paper, we proposed a frame work for feature selection based on scores and redundancy is eliminated for selecting subset of relevant features from the semi-supervised data set. A score function was developed to evaluate the scores to find the relevant feature on semi-supervised data on both locally geometrical structure of

unlabeled data and labeled data. The proposed work has several advantages:

- It handles feature redundancy with an approach is used to eliminate redundant features from the relevant ones.
- It improves performance of the selected feature set when compared with original feature set.

REFERENCES

[1]M.Robnik-Sikonja and I.kononenko, "Theoritical and empirical analysis of relief and reliefF," *Mach. Learn.*, vol. 53, no. 1-2, pp. 23-69, 2003

[2]P. E. H. R. O. Duda and D. G. Stork, *Pattern Classification*. Wiley-Interscience Publication, 2001

[3] Peng, H.; Fulmi Long; Ding, C., "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol.27, no.8, pp.1226,1238, Aug. 2005

[4] Lei Yu and Huan Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," *j.Mach. Learn. Res.*, vol. 5, pp.1205-1224, oct. 2004

[5]C.M. Bishop, *Neural Networks for Pattern Recognition.* New York, NY, USA: Oxford University Press, 1995.

[6]Bo Liao; Yan Jiang; Wei Liang; Wen Zhu; LijunCai; Zhi Cao, "Gene Selection Using Locality Sensitive Laplacian Score," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* , vol.11, no.6, pp.1146,1156, Nov.-Dec. 1 2014