International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

An Extended Laplacian Score Algorithm for Unsupervised Feature Selection

K.Sutha, J. Jebamalar Tamilselvi

C

Abstract: ***Experts from various sectors, utilize data mining techniques to discover most useful information from the huge amount of data, to improve their quality of outcomes. The Presence of irrelevant and redundant features affects the accuracy of mining result. Before applying any mining technique, the data need to be preprocessed. Feature selection, a preprocessing step in data mining provides better mining performance. In this paper, we propose a new two step algorithm for unsupervised feature selection. In the first step Laplacian Score is used to select the important features. And in the second step, Symmetric Uncertainty is used to remove redundant features. The experimental results show that the proposed algorithm outperforms the Laplacian Score algorithm.***

Index Terms: Feature Selection, unsupervised, Clustering, classification.

I. INTRODUCTION

In this digital world, the size of data increases drastically day by day. Data grows in terms of both in the numbers of instances as well as features. This increasing dimensionality degrades the performance of algorithms in Data mining, Machine Learning, Pattern Recognition, and Image Processing. High Dimensional data possess noisy, irrelevant, and redundant along with most useful information. To overcome this "Curse of Dimensionality", Dimensionality Reduction (DR) technique has to be applied to the huge amount of data, before discovering the hidden useful information. Feature Selection and Feature Extraction are the two DR techniques [1].

Feature Selection (FS) involves in selecting a subset of features from the original data set. This selected optimal feature subset provides better mining result as compared with complete set of features. FS reduces the computational cost, provides better model interpretability, and improves mining performance. Feature Selection methods are broadly classified into filter, wrapper and hybrid models [2]. Filter model feature selection depends on the intrinsic properties of the data during the selection process without using any mining algorithm. Wrapper model feature selection uses a predefined mining algorithm in the selection process and selects a feature subset depending upon the performance of mining algorithm. It is computationally expensive than the filter model algorithms [3]. Combination of Filter and Wrapper model forms the Hybrid model [4], useful for handling large data sets. It takes the advantages of both models.

FS algorithms are categorized as supervised, semi-supervised and unsupervised algorithms [5]. Supervised feature selection algorithm utilizes the class information to select the relevant features. Unsupervised feature selection involves in finding the best feature subset for unlabeled data. Semi-supervised feature selection handles partially labeled dataset.

This paper deals with unsupervised feature selection which deals with unlabeled data. Unsupervised feature selection algorithm selects a subset of most useful features in absence of class information. Hence it is a challenging task when compared with supervised and semi-supervised feature selection. Some of the popular Unsupervised feature selection algorithms are, Unsupervised Feature Selection using Feature Similarity measure (FSFS) [6], Laplacian Score for Feature Selection (LSFS) [7], Spectral analysis based feature selection [8], Multi-Cluster Feature Selection (MCFS) [9], Variance Score [10]. FSFS [6] initially groups all the similar features into clusters using Maximal Information Compression Index. One representative feature from each group is selected to form the reduced feature subset. LSFS [7] utilizes the local characteristics of the original data to assess the importance of features but it ignores redundancy. MCFS [9] use spectral analysis techniques to measure the feature correlations. It treats unsupervised feature selection as an optimization problem which involves an iterative process that is time consuming. Feature ranking based unsupervised feature selection algorithms, measure feature importance independently without considering the correlation between those features.

II. BACKGROUND STUDY

Laplacian Score[7], an unsupervised feature ranking algorithm ranks feature by computing its locality preserving power[11]. A feature is considered as an important one if its data points, x_i and x_j are closer to each other, having edge between them with smaller Laplacian Score. LS algorithm is stated as follows:

Suppose there are n features and m data points,

1. **Constructing a nearest neighbor graph** G having m nodes. An edge is placed between two nodes i and j, if its corresponding data points x_i and x_j are closer i.e. x_i is among k nearest neighbors of x_j or x_j is among k nearest neighbors of x_i.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
K.Sutha* Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India.
Dr.J. Jebamalar Tamilselvi, Professor, Department of MCA, Jaya Engineering College, Chennai, Tamil Nadu, India.

2. **Assigning the weights,** The edge of nodes i and j is assigned with $Wgt_{ij} = e^{-\frac{\|xi-xj\|^2}{t}}$. If there is no edge, assign $Wgt_{ji}=0$, where t is a constant.
3. **Computing the Laplacian Score,**
   Laplacian Score for r^{th} feature is computed as ,

$$L_r = \frac{\tilde{\mathbf{f}}_r^T L \tilde{\mathbf{f}}_r}{\tilde{\mathbf{f}}_r^T D \tilde{\mathbf{f}}_r}$$

Where **f**r = [fr1, fr2, · · · , frm]^T ,

D is the diagonal matrix $D_{ii} = \sum_j Wgt_{ji}$,

L = D – Wgt,

$$f_r = f_r - \frac{f_r^T D 1}{1^T D 1} 1, \quad 1 = [1, \ldots, 1]^T$$

The more important the feature is, the feature with smaller Laplacian Score. Laplacian Score algorithm ignores feature redundancy [12]. Symmetry is the best property in measuring correlation between features. Symmetric Uncertainty (Press et al., 1988) is calculated as follows,

$$SU=2\left[\frac{I(X|Y)}{En(X)+En(Y)}\right]$$

Where I(X|Y) is Information gain. Information Gain [13][14] is one of the feature ranking methods, which works by selecting the most informative features. (Quinlan, 1993), stated that Information Gain is the amount by which the entropy of X decreases reflects additional information about X provided by Y. It is calculated as I(X|Y) = En(X) – En(X|Y), E(X) is the entropy of X and En(X|Y) is the entropy of X after observing Y. Entropy(En) measures the uncertainty associated with a random variable X. Entropy, $En(X) = -\sum_{x\in X} p(x) log_2(p(x))$ and En(X|Y) is the entropy of X after observing values of another variable Y is given by $En(X|Y) = -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y) log_2(p(x|y))$,

P(x) - prior probabilities for all X values P(X|Y) - posterior probabilities of X given the values of Y. SU of value 0, shows that both the features X and Y are independent. SU of value 1, shows that both the features are correlated.

III. PROPOSED ALGORITHM

Laplacian Score is one of the popular feature ranking based feature selection algorithms. Literature survey revealed that Laplacian Score algorithm cannot handle redundancy. The occurrence of redundant features in the feature subset affects the mining performance. So there is a need to eliminate those redundant features. Redundant features are those features whose removal does not affect the mining performance. Motivated by this negative side of Laplacian Score, we extended the LS algorithm with redundancy removal phase. The proposed algorithm works as depicted in Fig.1. The important features are selected first and then the redundant features are removed, to obtain the optimal feature subset.

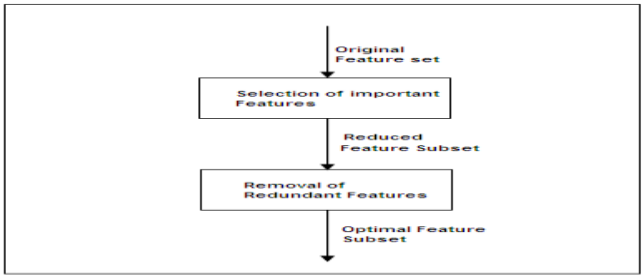


Fig 1. Proposed Method for selecting an optimal feature subset

The proposed algorithm is presented below.

Input : DS($F_1, F_2, F_3, \ldots\ldots F_N$) // The original set of features
λ - A Threshold value
Output: Op($F_1, F_2, F_3, \ldots\ldots F_M$) // Optimal Feature Subset
Procedure:

Step 1: Compute Laplacian Score all the features in DS, the original set as follows. LS for r^{th} feature is computed as

$$L_r = \frac{\tilde{\mathbf{f}}_r^T L \tilde{\mathbf{f}}_r}{\tilde{\mathbf{f}}_r^T D \tilde{\mathbf{f}}_r}$$

Step 2:Arrange all the features in ascending order of Laplacian Score.
//The feature with lowest score is the important feature.
Step 3: Select top N features.
Step 4: Calculate correlation between all features pairs using Symmetric Uncertainty,

4.1. Calculate the Entropy of feature X, $En(X) = -\sum_{x\in X} p(x) log_2(p(x))$

4.2. **Calculate the Entropy of entropy of X after observing values of another variable Y**,
$En(X|Y) = -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y) log_2(p(x|y))$

4.3. Compute I(X|Y) = En(X) – En(X|Y)

4.4 Calculate Symmetric Uncertainty as,
$SU=2\left[\frac{I(X|Y)}{En(X)+En(Y)}\right]$

Step 5: Discard one of the features with pair-wise correlation greater than the Threshold value λ.
Step 6: Repeat step 4 and 5 until the optimal feature subset (Op) is obtained.
Step 7: End.

In the first step, Laplacian Score is used to measure the importance of features. All the features are sorted in ascending order since features with lowest Laplacian Score is the important one. The top N features are selected as the important features. The next task is to eliminate the redundant features. Elimination of redundant features is done with the help of Symmetric Uncertainty (SU). SU measures the pair-wise correlation between all top k features obtained from the first stage. If the correlation between the feature pairs is greater than a predefined threshold SU value, one of strongly correlated features is considered as a redundant one and removed. The remaining features are selected as an optimal feature subset.

IV. EMPIRICAL STUDY

Experimental study is conducted on 6 UCI datasets, listed in Table1. The effectiveness of proposed algorithm is evaluated and compared with Laplacian Score algorithm. The number of features selected by proposed algorithm, and its classification accuracy is compared against Laplacian algorithms in Table 2 and 3.

Table 1: Datasets used for experimentation.

Datasets	Features	Instances	Class
One Hundred Plant Shape	65	1600	100
Dbworld Bodies	4703	64	2
Lungcancer	56	32	3
Madelon	501	2600	2
Cardiotocography	36	2126	10
Mfeat	217	2000	10

The number of features selected by Laplacian Score and proposed algorithms are presented in Table 2 and Fig 2. It reveals that the proposed algorithm selects minimum no. of features when compared with Laplacian Score (LS) algorithm. As the LS algorithm selects only the important features along with redundant features. A good feature selection algorithm selects feature subset as small as possible. The performance of mining algorithms is severely affected by the presence of redundant features. The proposed algorithm removes one of the strongly correlated features and selects least number of features as an optimal feature subset. It improves the efficiency and the effectiveness of mining algorithm.

Table 2: Comparison of No. of features selected by Laplacian and proposed algorithm.

Datasets	No of features in original dataset	Laplacian Score Algorithm	Proposed Algorithm
One Hundred Plant Shape	65	58	**48**
Dbworld Bodies	4703	579	**520**
Lungcancer	56	33	**7**
Madelon	501	81	**56**
Cardiotocography	36	18	**8**
Mfeat	217	59	**48**

Table 3: Comparison of Accuracy given by Laplacian and the proposed algorithm

Datasets	Laplacian Score	Proposed Algorithm
One Hundred Plant Shape	40.37	**48.58**
Dbworld Bodies	90.31	**91.4**
Lungcancer	56.71	**64.53**
Madelon	61.85	**63.94**
Cardiotocography	36.07	**59.05**
Mfeat	66.28	**71.64**

The accuracy given by the proposed and Laplacian Score algorithms are listed in Table 3 and Fig .3. The results reveal that the proposed algorithm gives better accuracy when compared with Laplacian Score. The proposed algorithm shows better performance for all the datasets, due to the elimination of redundant features. The better performance for each dataset is highlighted in boldface.

Fig.2. Comparison of no. of features selected by Laplacian Score and proposed algorithms.

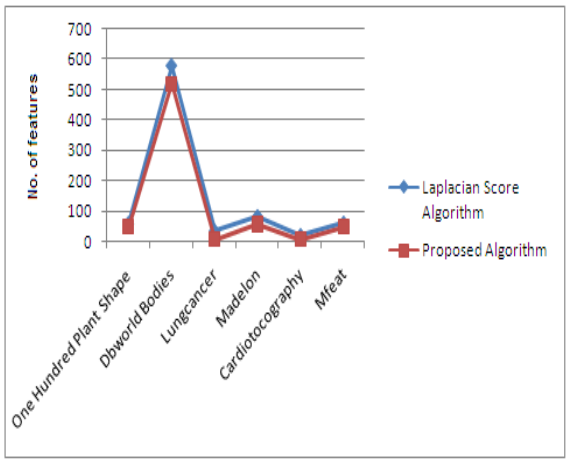
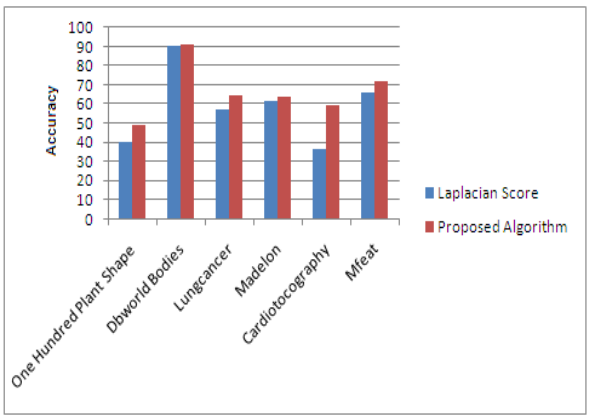


Fig.3. Comparison of accuracy given by the proposed algorithm and Laplacian Score algorithm



Experimental results (fig.2 and fig. 3) reveal that the combination Laplacian Score and Symmetric Uncertainty works well in selecting best feature subset with minimal number of features and also removes redundant features. As expected, it achieves better mining results.

V. CONCLUSION

In Data Mining, Feature Selection selects best feature subset without any noise, irrelevant and redundant features. We proposed a new two step algorithm for unsupervised feature selection. Laplacian Score ranks the important features in the first step. And then the redundant features are eliminated, using Symmetric Uncertainty, which measures the feature-feature correlation. The experimental results revealed that the proposed algorithm gives better accuracy and selects minimum number of features when compared with Laplacian Score algorithm.

REFERENCES

1. Pudil, P.; Novovicova, J., "Novel Methods for Feature Subset Selection with Respect to Problem Knowledge", Feature Extraction, Construction and Selection.p. 101,1998.
2. I.Guyon, A.Elisseeff, " An Introduction to Variable and Feature Selection", *Journal of .Mach. Learning Research*,Vol.3,pp.1157-1182,2003.
3. R.Kohavi, G.H John," Wrappers for Feature Subset Selection", Arti. Intell. Vol. 97, pp.273 – 324, 1997.
4. S.Das, "Filters, Wrappers and a Boosting-based Hybrid for Feature Selection", Proc. 18th ICML, pp.74-81, 2001.
5. Z Zhao, H Liu, "On Similarity Preserving Feature Selection", IEEE Trans. on Knowledge and Data Eng. Vol. 25, No.3, p. 619-632, 2013.
6. Mitra, P., C. A. Murthy, S. K. Pal. Unsupervised Feature Selection Using Feature Similarity. – IEEE Trans. Pattern Anal. Mach. Intell., Vol. **24**, No 3, pp. 301-312,2000.
7. X He, D Cai, P Niyogi," Laplacian score for feature selection", Adv. in Neural Info. Processing Sys. vol. 17, 2005.
8. Zhao Z., H. Liu. Spectral Feature Selection for Supervised and Unsupervised Learning. – In: Proc. of ICML'07, 2007, pp. 1151-1157.
9. Cai, D., C. Zhang, X. He. Unsupervised Feature Selection for Multi-Cluster Data. – In: Proc. of 16th Int'l Conf. on KDD'10, 2010, p. 333.
10. Bishop, C. M. Neural Networks for Pattern Recognition. Clarendon Press, 1995.
11. X He and P Niyogi," Locality Preserving Projections", Adv. in Neural Info. Processing Sys, 2004.
12. Z Zhao, Fred, Sharma, Salem, Aneeth, H.Liu, " Advancing Feature Selection Research – ASU Feature Selection Repository".
13. P Bermejo, L D L Ossa, J A Gamez, J M Pureta, " Fast Wrapper Feature Selection in High Dimensional datasets by means of Filter re-ranking", 2011.
14. L Yu, H Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Proc. of 20th ICML, 2003.

AUTHORS PROFILE

K.Sutha is a research scholar at Bharathiar University, Coimbatore, Tamilnadu. She received her MCA degree from Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India. Her area of interest includes Data Warehousing, Data Mining, and Big Data.

Dr. J. Jebamalar Tamilselvi received her Ph.D. in 2009 from the Department of Computer Applications at Karunya University, Coimbatore, INDIA. She received her B.Sc. (Computer Science) from Manonmanium Sundaranar University of Tamil Nadu, INDIA in 2003 and MCA Degree from Anna University, Coimbatore, Tamil Nadu, INDIA in 2006. Her area of interest includes Data cleansing approaches, Data Extraction, Data Integration, Data Warehousing and Data Mining. She is a life Member of International Association of Engineers (IAENG), International Association of Computer Science and Information Technology (IACSIT), and the Society of Digital Information and Wireless Communications. Reviewer and Member of International Journal of Engineering Science and Technology (IJEST) Member and Convergence Information Technology (JCIT). Her research has been accepted and published in 17 international journals, and 12 national and international conferences. She had been awarded the P.K Das Memorial Best Faculty Award in 2014 by the Nehru Group of Institutions, Coimbatore and the Education and Research Award in 2015 by the Karunya University, Coimbatore.