International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 2 Issue 11, November 2014

Feature Subset Selection Algorithm for Clustering Using Modified Prim's Algorithm

R. Balakrishnan[1], M. Soundarya[2]

[1]Assistant Professor, Department of IT, Dr. N.G.P. Arts and Science College, Coimbatore, India

[2]Research Scholar, Department of IT, Dr. N.G.P. Arts and Science College, Coimbatore, India

Abstract: ***In the proposed framework the feature selection subset is mainly used for machine classification and data mining. It is used for the high dimensional data reduction. This data reduction and best feature selection are handled through removal of irrelevant data's in the dataset. Best feature subset selections are obtained through modified prim's algorithm. In prim's algorithm although there is continuous developing forest of connected nodes but initially root node is chosen randomly which may or may not be the part of minimum edge of the graph. But proposed modified prim's algorithm, along with continuous developing connected spanning tree, root node is also forming minimum edge of the graph. The implementation analyzes the results through 10 different bench mark datasets and two classification algorithms.***

Keywords: classification, feature subset, dataset, modified prims

1. Introduction

Good feature selection and subset is the main concept in the data mining, this is used to prove their concepts in their data mining. It is one of the important and efficient way for reducing the dimensionality and remove the irrelevant data's in the dataset and improve their accuracy with their improvement in their learning machine [1].

In machine applications, feature selection is an important and it is divided into four different categories that are embedded, wrapper, filter and hybrid approaches [2].

Even though large number of feature subset selection methods has been proposed earlier, but in this methods a single algorithm can didn't prove the well in their feature selection. In [3, 4] provide some experiments and they study from the results, different feature selection subset have different significant factor on their classification accuracy.

Feature selection is used to face the high dimensional data, based on this the enormity of both size and their dimensional also provide several challenges due to this feature selection, and also it is used to handle the large number of instances for dealing the high dimensional dataset [5 and 6].

Assigns the weights based on features weighting algorithm and rank the features based on the relevance to the target concept. There are different type of relevance are given in [7 and 8].

2. Related Work

To identify and removing the unwanted and some irrelevant data's from the dataset, feature subset selection are used. Based on this removal better accuracy can be predicted [9].

Filter methods are used to attain the perfect accuracy and is more effective than tradition feature selection methods, and these filter selection techniques is one of the application of cluster analysis. To remove the words through preprocessing is done by several algorithms in distributional clustering [10, 11]. Another one word clustering was implemented in [12] for improvement of feature subset selection and provide better accuracy in text classification.

3. Proposed Framework

Accuracy of the machine learning is based on the relevant features, suppose if some unwanted data's are present in the features then it affects the classification accuracy. So, in this paper improved the accuracy based on the following steps:

1. Remove the irrelevant features.
2. Remove the redundant features.
3. For redundant feature selection, constructing the graph based on the minimum spanning tree.
4. To form the cluster with the use of minimum spanning tree, that partitioning the minimum spanning tree into a forest. In this each tree can be treated and identified as cluster.
5. Then from the cluster select the representative features.

Based on the following steps, this paper provides the effective and efficient for removal of irrelevant and redundant features and provides the better feature selection method. Block diagram of the proposed framework is given in figure 1.

International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 2 Issue 11, November 2014

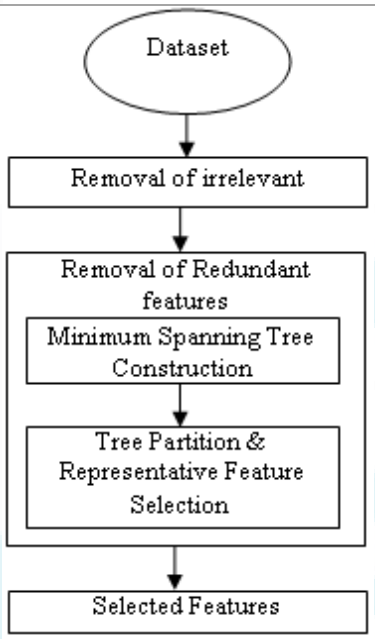

Figure 1: Block Diagram for Proposed Framework

In this paper let us assume the dataset D with large number of features that is indicated as n features fe $fe_1, fe_2, \ldots, fe_n$ and class C.

Algorithm for the proposed framework

Input: Given dataset
$(fe_1, fe_2, \ldots, fe_n$ *and class* $C)$
Output: S-Selected feature subset
Step 1: remove the irrelevant features
for i=1 to m
T-Relevance=$SU(fe_i, C)$;
If T-Relevance >θ then
$S = SU\{fe_i\}$;
Step 2: Minimum Spanning Tree
G=NULL;
F-Correlation=$SU(fe_i', fe_j')$;
Add fe_i' and/ or fe_j' to G *with* F-Correlation *as the weight of the corresponding edge*;
Minimum span Tree=modified Prim G); using modified prim algorithm
Step 3: Tree Partition and Representation
Feature Selection
Forest=Minimum Span Tree
for each edge $E_{ij} \in Forest$ *do*
If $SU(fe_i', fe_j') < SU(fe_i', C) \wedge SU(fe_i', fe_j') < SU(fe_j', C)$
Forest=Forest-E_{ij}
$S = \emptyset$
for each tree $T_i \in Forest$ *do*
$fe_R^j = argmax_{fe_k' \in T_i} SU(fe_k', C)$
$S = SU\{fe_R^j\}$
return S

Step 1: First step compute the T-Relevance $SU(fe_i, C)$ value, in the value for each feature fe denotes the $fe_i (1 \leq i \leq n)$. Suppose if the value is greater than the predefined threshold θ then comprise the target-relevant feature subset $fe' = \{fe_1', fe_2', .., fe_k'\}$ where $(k \leq n)$.

Step 2: Calculate the F-Correlation $SU(fe_i', fe_j')$ value for each pair of features fe_i', fe_j' as $fe_i', fe_j' \in fe' \wedge i \neq j$. Then both the features are combined and formed as vertices, weight of the edge between the vertices are denoted or represented by $SU(fe_i', fe_j')(i \neq j)$. Based on this weight complete graph is formed as $G = VE$. In this equation V represented as $V = \{fe_i' | fe_j' \in fe' \wedge i \in [1, k]\}$.

Correlation between the targets features are given by the complete graph G. suppose if the graph has k vertices than the edge can be calculated from $(k-1)/2$. Then build the minimum spanning tree for graph G to connect all the vertices such that the sum of the weights of the edges is the minimum, these calculation are done by using modified prim's algorithm.

Step 3: Modified Prim algorithm:

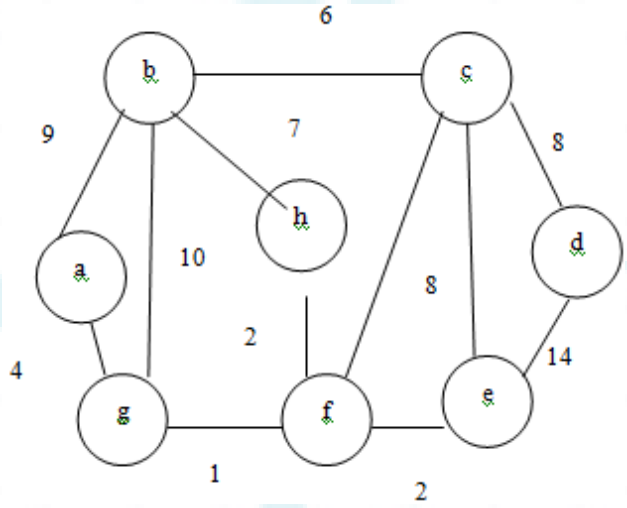

Figure 2: Graph

1) Jet value of all the vertices are initialized with infinity and parent value of every node is initialized as NIL. Now minimum edge is extracted from E [G] i.e. (g, f).Now vertex g is set as s root vertex and its Jet [g] is initialized as 0. Now put this edge i.e. (g, f) back to the set of edges E [G].Copy all the vertices of V [G] into minimum priority queue Q. Now vertex g is extracted from Q as jet [g] is zero. Now vertices that are adj [g] are extracted i.e. vertices a, b, f are extracted their Jet value is replaced with their respective edge weights.
2) In second step, selected node is f as edge (g, f) is light weight edge as well as safe edge.
3) After step 2, h node is selected. Here we have two light weight edges i.e. (f, h) and (f, e) and both are safe edge as well so algorithm can select any of the edge. Here selected is (f, h) so node h is selected.
4) Now selected node is e as edge (f, e) is light weight edge as well as safe edge.
5) Now selected node is a as edge (a, g) is light weight edge as well as safe edge.
6) Now selected node is a as edge (a, g) is light weight edge as well as safe edge

7) Now selected node is d as edge (c, d) is light weight edge as well as safe edge.

Step 4: Remove all the edges, that edges are smaller than both of the T-relevance$SU(fe_i', C)$ and $SU(fe_j', C)$. Then the removal results provides the two disconnection tress T_1 and T_2.

Step 5: Finally forest is obtained after removing all the edges. In that forest each tree is represented as a cluster and it is denoted as $V(T_j)$.

4. Experimental Results

Experimental proves the accuracy of proposed framework. To evaluate the performance of proposed algorithm and compare it with other feature selection algorithms.

The proposed framework algorithm is compared with existing algorithm [9], the proposed framework is slightly differ from existing algorithm, here best feature is selected by modified prim's algorithm. Two different types of classification algorithms are employed to classify data sets before and after feature selection. They are;

- The probability-based Naive Bayes (NB),
- The tree-based C4.5,

In this paper analysis the proposed framework by using 10 benchmark dataset, that are;

- Chess,
- mfeast-fourier,
- coil2000,
- elephant,
- arrhythmia,
- rs-noew,
- colon,
- fbis-wc,
- AR1oP and
- PIE1oP.

These 10 datasets are containing in the domain of text, image, face, microarray, Bio. Then finally average the accuracy, runtime based on the average of the domain.

Table 1: Feature Selection

Dataset	*Existing*	*Proposed*
Chess	16.22	14.21
Mfeast-fourier	19.48	17
Coil2000	3.49	2
Elephant	0.86	0.67
Fqs-nowe	0.31	0.11
Colon	0.30	0.28
Fbis.we	0.80	0.74
AR10P	0.21	0.17
PIE10P	1.07	1.00
Average (Image)	3.59	2.89
Average (Microarray)	0.71	0.68
Average (Text)	2.05	2
Average	1.82	1.79

Table 1 gives the feature selection between existing and proposed algorithms. Two algorithms are significantly reduction of dimensionality by selection of original features. Proposed framework obtains the best feature selection compared to existing algorithms that is 1.79%.

Table 2: Comparison of Run time

Dataset	*Existing*	*Proposed*
Chess	105	108
Mfeast-fourier	1472	1481
Coil2000	866	860
Elephant	783	790
arrhythmia	110	100
Fqs-nowe	977	980
Colon	166	169
Fbis.we	14761	14759
AR10P	706	710
PIE10P	678	650
Average (Image)	1520	1515
Average (Microarray)	1468	`1460
Average (Text)	6989	6980
Average	3573	3489

Table 2 provides the comparison of runtime between the existing and proposed algorithm. From the average conclusion proved that the proposed method proves the faster than the existing algorithm.

Table 3: Comparison of Accuracy for C4.5 algorithm

Dataset	*Existing*	*Proposed*
Chess	94.02	94
Mfeast-fourier	71.25	69
Coil2000	94.03	94
Elephant	99.90	99
arrhythmia	71.53	72
Fqs-nowe	69.81	69
Colon	90.40	90
Fbis.we	69.41	69.11
AR10P	77.69	77.31
PIE10P	84.13	84
Average (Image)	81.70	82
Average (Microarray)	83.77	84
Average (Text)	81.38	81.99
Average	82.44	83

Table 3 provides the classification accuracy of C4.5. From the above table observed that the proposed method proves the better classification accuracy for C4.5 than the existing method.

Table 4: Comparison of Accuracy for Naive Bayes algorithm

Dataset	*Existing*	*Proposed*
Chess	92.92	93
Mfeast-fourier	80.05	81
Coil2000	94.04	94.32
Elephant	99.47	99
arrhythmia	73.01	74
Fqs-nowe	69.81	69
Colon	95.08	95.12
Fbis.we	70.04	71
AR10P	69.23	69
PIE10P	96.83	96.99

Average (Image)	85.49	86
Average (Microarray)	91.38	91.89
Average (Text)	82.62	82.78
Average	86.84	86.99

Table 3 provides the classification accuracy of Naive Bayes algorithm. From the above table observed that the proposed method proves the better classification accuracy for Naive Bayes algorithm than the existing method.

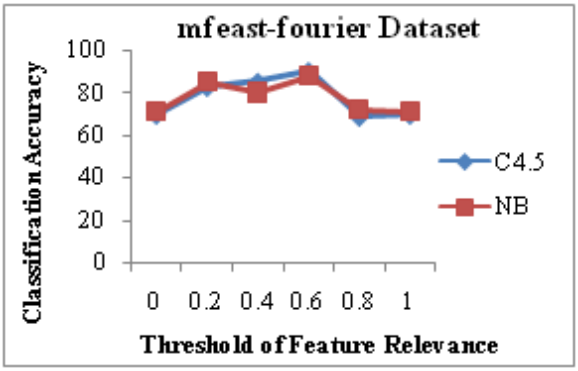

Figure 3: Accuracies of the two classification algorithms with different θ values

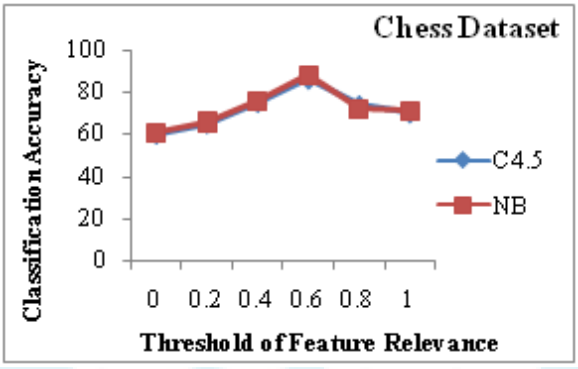

Figure 4: Accuracies of the two classification algorithms with different θ values for mfeast-fourier dataset

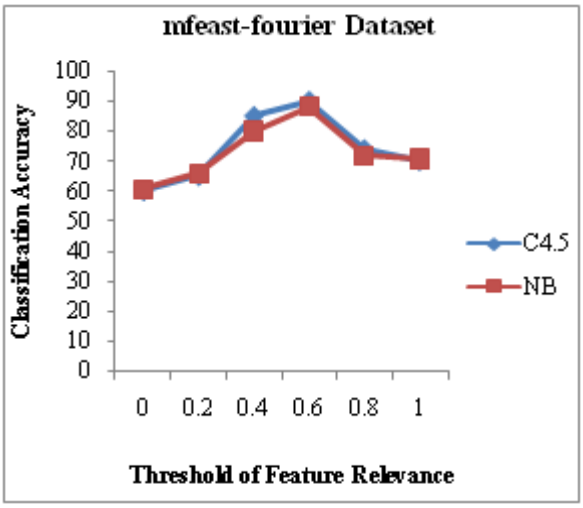

Figure 5: Accuracies of the two classification algorithms with different θ values for coil2000 dataset.

The above figure from 3 to 5 illustrates the classification accuracy for two algorithms for datasets from the benchmark dataset.

5. Conclusion

In this proposed work a novel based clustering algorithm for feature subset selection for high dimensional data. The proposed work analysis in three steps, first remove or eliminate the irrelevant and redundant data, second construct the minimum span tree using modified prim's algorithm, then third partition the minimum span tree and find or select the feature subset through cluster analysis and simultaneously reduce the dimensional. Analyzing the proposed framework through ten benchmark dataset through two classification algorithms NB and C4.5. From the experimental results it can be observed that the proposed work proves in better feature subset selection, faster run and better in accuracy.

References

[1] Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96-103, 1998.
[2] Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97, 245-271.
[3] Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. Proceedings of the Eighteenth International Conference on Machine Learning (pp. 74{81).
[4] Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res., 3, pp 1265-1287, 2003.
[5] Guangtao Wang, Qinbao Song, Heli Sun, Xueying Zhang, Baowen Xu and Yuming Zhou, "A Feature Subset Selection Algorithm Automatic Recommendation Method", Journal of Artificial Intelligence Research 47 (2013) 1-34.
[6] Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. Ph.D. thesis, The University of Waikato
[7] Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. Ph.D. thesis, The University of Waikato.
[8] Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97, 273–324.
[9] Liu H., Motoda H. and Yu L., Selective sampling approach to active feature selection, Artif. Intell., 159(1-2), pp 49-74 (2004).
[10] Liu, H., Motoda, H., & Yu, L. (2002b). Feature selection with selective sampling. Proceedings of the Nineteenth International Conference on Machine Learning (pp. 395 - 402).
[11] Pereira F., Tishby N. and Lee L., Distributional clustering of English words, In Proceedings of the 31st Annual Meeting on Association For Computational Linguistics, pp 183-190, 1993.
[12] Qinbao song, jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-dimensional Data, IEEE Transaction on knowledge and data Engineering, vol. 25, no.1,2013.

References

R. Balakrishnan received M.Phil degree from M.S. University. He is currently working as Assistant professor and working towards Ph.D in data mining at M.S. University, his area of interest is DBMS. He presented many research papers in National and International conferences.

M. Soundarya received M.Sc (IT) degree from Bharathiar University. She is currently doing her M.Phil degree in Bharathiar University in data mining. Her area of interest is networking and DBMS. She attended many National and International conferences.