IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Comparison of Performance of Feature Selection Methods

P. Karthika[1], Mrs. S. Nithya Roopa[2], S. Monisha[3]

Student, Computer Science, Kumaraguru College of Technology, Coimbatore, India[1, 3]

Assistant Professor, Computer Science, Kumaraguru College of Technology, Coimbatore, India[2]

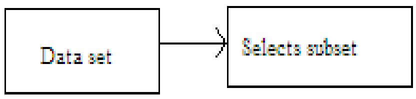
Abstract: Feature selection is a process of selecting the relevant subset of features among all the features in a dataset. The objective of the paper is to improve the performance level, understanding and reduce time complexity, which is also known as variable selection and attribute selection. In this paper, the best performance of feature selection methods has been examined using relief, fast clustering algorithm and ranked forward search algorithm. The results of performance of algorithms is compared and given in a graph.

Keywords: Feature Selection, Performance, Algorithms, Subset Features.

I. INTRODUCTION

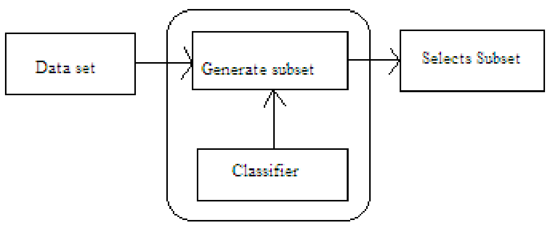
Data Mining is a process of analysing the data and requires only the necessary information. It is a tool for analysing data. Feature selection selects the subset of features which are relevant to cancer disease. It improves the accuracy of relevant features; reduce the complexity of selecting features.The resultant subset reduces the size of dataset. In this paper we propose feature selection using algorithms of the three methods. The performance of the algorithm is given in a graph. It distinguish into three methods, they are

Filter method
The filter approach is performed by giving the entire dataset as input, attributes has been selected using the algorithm only the relevant features as output.



Wrapper method
Wrapper method analyse the features and selects only the quality of features. The input of dataset is given to the classifier, the features are classified using backward elimination and unnecessary features are removed.

Hybrid method
It is a connection of filter and wrapper method, where the performance of hybrid is more than the other two approaches.



Here we use the cancer dataset for feature selection; the fields in the dataset represent the nuclei of a cancer cell. The dataset describes about the nucleus position, structure, and radius. The attribute diagnosis, where M – malignant and B – benign. The relief algorithm, fast clustering algorithm and ranked search algorithm has used to remove the subset of features.

Table 1: Information of dataset

Number of instances	569
Number of attributes	32
Dataset characteristics	Multivariate

II. ALGORITHMS

Relief algorithm
Relief algorithm is a weight based algorithm. The given dataset as S, dataset size m, t threshold relevance, random record as X, choose a random value at positive record as Z^+, and choose another random value at negative record as Z^-. Z^+ and Z^- nearest to X value. The threshold value within the range of 0 to 1.

Input: the entire dataset is given.
Output: The relevant feature is given as output.
Relief(S, m, t)
Separate S into S^+ = {positiveinstances} and
S^-= {negativeinstances}
w = (0, 0… 0)
Fori = 1 tom
Pick at random an instance X € S
Pick at random one of the positive instances closest to X, Z^+ € S^+
Pick at random one of the negative instances closest to X, Z^- € S^-
if(X is a positive instance)
then Near-hit = Z^+; Near-miss = Z^-
else Near-hit = Z^-; Near-miss = Z^+
Update-weight(W, X, Near-hit, Near-miss)
Relevance = (l/m)W
For i = 1 to p
if (relevance$_i$ $\geq$ t)

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

then f_i is a relevant feature
else f_i is an irrelevant feature
update-weight(W, X, Near-hit, Near-miss)
Fori= 1 top
$W_i = W_i - \text{diff}(x_i, \text{near-hit}_i)^2 + \text{diff}(x_i, \text{near-miss}_i)^2$

Relief uses near-hit and near-miss, where Z^+ value nearest to X value within same class is near-hit. Z^- nearest to X value within different class is near-miss. In the dataset contains triplets of features, calculate the weight as W and relevance is calculated by weight. The relevance value is compared with threshold value, if it is greater than t the feature is relevant feature.

Fast clustering algorithm

Fast clustering algorithm is used to remove the irrelevant feature using threshold value and remove the redundant feature using minimum spanning tree. Minimum spanning tree is calculated using prim's algorithm.

Input: D - the given data set
 Θ - T-Relevance threshold.
 C – target cluster class
Output: R- selects the feature subset.

1 for i=1 to m do
2 T-Relevance= SU(Fi, C)
3 if T-Relevance >θ then
4 R= S U {Fi};
5 G= NULL;
6 for each pair of features {F'i, F'j} ⊂ R do
7 F-Correlation = SU (F'i, F'j)
8 Add F'i and/ or F'j to G with F-Correlation as the weight of the corresponding edge;
9 minSpanTree = Prim(G);
10 Final= minSpanTree
11 for each edge Eij ϵ Final do
12 if SU(F'i, F'j) < SU(F'i, C) ∧ SU(F'i, F'j) < SU(F'j, C) then
13 Final = Final - Eij
14 R= ϕ
15 for each tree Ti ϵ Final do
16 F^j_r = argmaxF'k ϵ Ti SU(F'k, C)
17 R= S U { F^j_r};
18 return R

Symmetric Uncertainty- symmetric uncertainty is done by normalising the feature values and targets the most important features. It is calculated by,
$SU(A,B)=2\times \text{Gain}(A|B)\div(H(A)+H(B))$
 Where, $\text{Gain}(A|B) = (A) - H(A|B)$
$= (B) - H(A|B)$
$H(A) = -\sum f(a)\log_2 f(a)$
$H(A|B) = -\sum f(b) \sum f(a|b)\log 2f(a|b)$
Where,
f(a) is the probability density function
f(a|b) is the conditional probability density function.

Threshold Relevance- Threshold relevance between feature F_i and target class C. the relevance is calculated by SU (F_i ,C), if the value is greater than threshold value target class feature is taken.

Minimum Spanning Tree- Consider a graph G with k vertices and (k-1)/2 edges. Feature F_i,F_j is the vertices of graph and calculate the weight as edge. The minimum spanning tree constructed using prim's algorithm and find the weight of shortest path. We remove the edges which is smaller than the Threshold relevance of two vertices. After removing the each edge a new tree T is constructed.

Hybrid k-means Clustering Algorithm

Hybrid k-means algorithm refers to partitioning a group of data into smaller groups. We have n data points have to be grouped together in k clusters.The K-means algorithm uses the Euclidean distance, d(x,μi)=‖x−μi‖2
// Initialize the centre of the clusters
μ_i= some value ,i=1,...,k
// Attribute the closest cluster to each data point
C_i={j:d(xj,μi)≤d(xj,μl),l≠i,j=1,...,n}
//Set the position of each cluster to the mean of all data points belonging to that cluster
μi=1|c_i|∑j∈cixj,∀i
// Repeat steps 2-3 until convergence
|c|= number of elements in c
Where c_i is the set of points that belong to cluster i.
i = 1, 2,…..n
k-means algorithm finds the global solution of all data points.

III. RESULTS

We obtained the results, for relief algorithm only the relevant features are shown in subset.

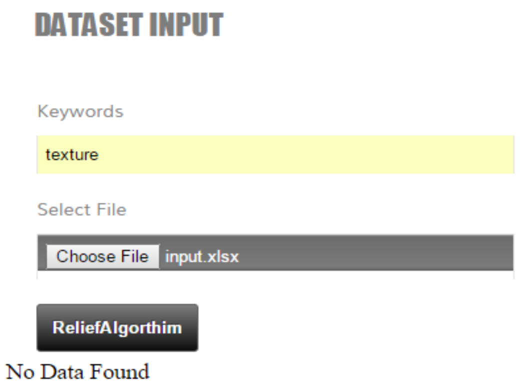


Fig.1. Input has been selected

If the attribute is said to be relevant feature, the values has been displayed or else it shows no data found.

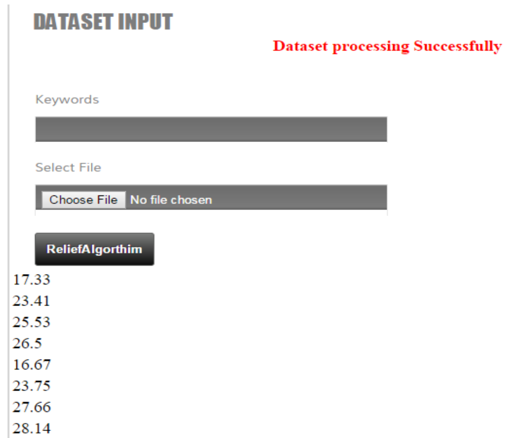


Fig.2. Relief algorithm display the relevant feature

For the fast - based clustering algorithm, the irrelevant features has been reduced and only the relevant subset of features has been shown.

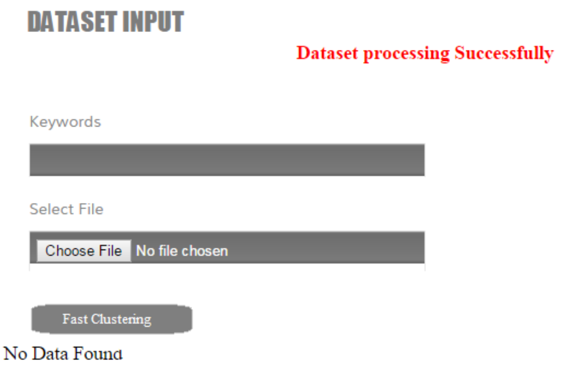


Fig. 3.Fast clustering removes irrelevant feature

The algorithm hybrid k-means clustering algorithm, group into small clusters. The algorithm stops when changes to next iteration.

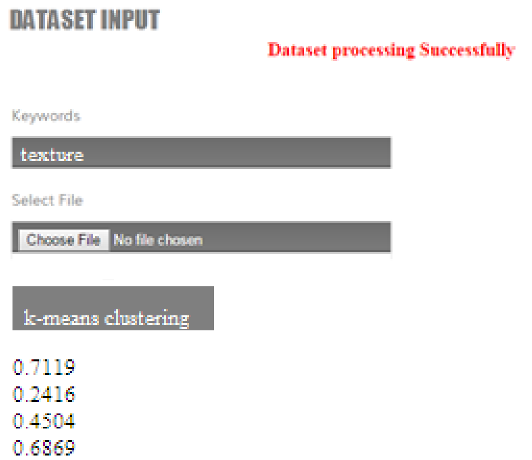


Fig. 4.K-means clustering selects subset feature

The below tabular column shows the final output of each algorithms, only the relevant features of data has been selected.

Table 2: Final Results

	Input Features selected	Output Features selected
Relief Algorithm	30	10
Fast clustering Algorithm	30	7
Hybrid k-means clustering Algorithm	30	6

IV. CONCLUSION

This paper provides an idea about the feature selection, where the performance of the algorithms is shown. There are many algorithms used, but here we take only three algorithms, one of each method in feature selection.The algorithms in each methods of feature selection are taken, where subset of features has been selected. A future development of reducing the execution time of the algorithms, and also the performance. It provides more efficient for a better understanding of selecting relevant features.

REFERENCES

[1] R.P.L.DURGABAI, "Feature Selection using ReliefF Algorithm"Vol. 3, Issue 10, October 2014
[2] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy" J.Machine Learning Research, vol.10,no. 5, pp. 1205-1224, 2004.
[3] Mr. Swapnil R Kumbhar, Mr.Suhel S Mulla, "Literature Review on Feature Subset SelectionTechniques" Volume 03, Issue 09, September 2014.
[4] MrsKomal Kate, Prof. S. D. Potdukhe"Fast Feature subset selection algorithm based on clustering for high dimensional data" Volume 2, Issue 6, October-November, 2014 ISSN 2091-2730.
[5] J.K.Madhavi, G.VenkateshYadav, "An Improved Fast Clustering method for FeatureSubset Selection on High-Dimensional Dataclustering" Volume 3, Issue 10, October 2014.
[6] Swapnil Sutar, "Feature Selection Algorithm Using Fast Clustering and Correlation Measure" Volume: 02 Issue: 07 | Oct-2015.
[7] L. Ladha, T.Deepa, "Feature Selection Methods andAlgorithms" Vol. 3 No. 5 May 2011.
[8] An Unsupervised Feature Selection Method Based On Genetic Algorithm, Nasrin Sheikhi, Amirmasoud Rahmani, Mehran Mohsenzadeh, Department of computer engineering, Islamic Azad University of Iran research and science branch, Ahvaz, Iran Reza Veisisheikhrobat, National Iranian South Oil Company(NISOC), Ahvaz, Iran, Vol 9, no.1 Jan,2011.
[9] t. liu, s. liu, z. chen, and w. ma,"an evaluation on feature selection for text clustering," proc. ieee int'l conf. machine learning (icml'03, pp. 488-495, 2003.
[10] S.Khalid, T.Khail, S.Nasreen, "A survey of feature selection and feature extraction techniques in machine learning" IEEE on Science and Information Conference, 27 – 29 Aug 2014.
[11] Yijun Sun, Sinisa Todorovic "Local Learning Based Feature Selection for High Dimensional Data Analysis" IEEE Transactions On Pattern Analysis And Machine Intelligence, VOL. 32, NO. 9 SEPT.2010.
[12] DanielaM. Witten and Robert Tibshirani **"**A framework for feature selection in clustering" 2010 Jun 1; 105(490): 713–726.
[13] A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time (1 + ϵ)-approximation algorithm for k-means clustering in any dimensions. In Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS), pages 454–462, 2004.
[14] Dan Pelleg and Andrew W Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In ICML, pages 727–734, 2000.