HIGH DIMENSIONAL UNSUPERVISED CLUSTERING BASED FEATURE SELECTION ALGORITHM

Ms.Barkha Malay Joshi

M.E. Computer Science and Engineering, Parul Institute Of Engineering & Technology, Waghodia. India
Email: Barkha_ce@yahoo.co.in

Prof. G.B. Jethava
Information Technology Dept., Parul Institute Of Engineering & Technology, Waghodia
Email: g.jethava@gmail.com;

Prof. Hetal B.Bhavsar

Information Technology Dept, SVIT, Vasad
Email: het_bhavsar@yahoo.co.in;

ABSRACT: Feature selection is a process which selects the subset of attributes from the original dataset by removing the irrelevant and redundant attribute. Clustering is the technique in data mining which group the similar object in to one cluster and dissimilar object into other cluster. Some clustering technique does not support high dimensional dataset. By applying the feature selection as a preprocessing step for the clustering make it possible to handle the high dimensional dataset. Feature selection reduce the computational time greatly due to reduced feature subset and also improve clustering quality. Feature selection methods are available for supervised and unsupervised learning. This paper is related to working of feature selection method which is applied on different feature selection algorithm. The result proved that Feature selection through feature clustering algorithm is reduced the more attributes than the standard feature selection algorithm like relief and fisher filter.

KEYWORDS: Unsupervised feature selection, Feature similarity measure clustering based feature selection.

Introduction

Real world data may contain hundreds of attributes, many of which may be irrelevant to the mining or redundant. Keeping such attributes causing contention for mining algorithm employed, can result in discovered pattern of poor quality as well as slow down the mining process. Feature selection is the important and frequently used technique in data pre-processing for data mining[11]. Feature selection or dimensionality reduction is a pre-processing technique which selects the relevant feature from the dataset and removes the redundant and irrelevant feature from the dataset. The goal of feature selection is to find a minimum set of attributes such that the resulting probability distance of the data classes is as close as possible to the original distance obtained using all attributes[2][3].

Clustering is an interesting topic in the many research field which include data mining, statistics, machine learning, pattern recognition and biology.

Clustering is the technique that put the object of high similarity into one cluster and objects having less similarity into different cluster. So that the same cluster data objects are most similar than the different clusters. There are so many clustering algorithms like k-means, k- medoids which does not support the high dimensionality and feature sparseness. Therefore it is highly desirable to reduce the feature space dimensionality. There are mainly two techniques deal with this problem feature selection and feature extraction.

The available methods for the feature selection can be classified as Supervised and unsupervised feature selections. The supervised and unsupervised feature selection methods, like documents frequency(DF), term strength(TS), can be easily applied to clustering[3]. The supervised feature selection methods using information

gain(IG), $\chi 2$ statistic(CHI) can be applied for text classification where the class labels of the documents are available[1][3].

Sometimes, various data mining analysis (supervised and unsupervised learning) may need to apply to the same data. A subset selected by a supervised feature selection method may not be a good one for unsupervised learning and vice versa. So it is very desirable to have a feature selection method which works well for both. The clustering based feature selection does not need the class label information in the data set and is suitable for both supervised learning and unsupervised learning.

This paper is related to the various methods of feature selection. The experiment results have been demonstrated for the relief filer, fisher filter and Feature selection through feature clustering algorithm apply on cancer dataset.

The rest of this paper is organized as follows: Section II covers Unsupervised feature selection with different approaches, section III covers the feature similarity measure, section IV demonstrate the clustering based feature selection, Dataset characteristics included in section V section VI demonstrate the experiment results and graph and section VII shows the conclusion, future work and references cover in the section VIII and IX.

I. Unsupervised Feature Selection

Unsupervised feature selection (UFS) received some attentions in data mining and machine learning as clustering high dimensional data sets becomes an essential and routine task in data mining. Unsupervised feature selection is becoming an essential pre-processing step because it can not only reduce computational time greatly due to reduced feature subset but also improve clustering quality because no redundant features that could act as noises are involved in unsupervised learning. The following are the approaches used for unsupervised feature selection.

1. Wrapper approach

The wrapper approach uses the induction algorithm for the estimating the feature subset. This approach is provide the better performance than the filter approach. Figure 2 shows the basic process of the wrapper approach[7][12].

The wrapper approach divides the task into three components: (1) feature search, (2) clustering algorithm, and (3) feature subset evaluation.

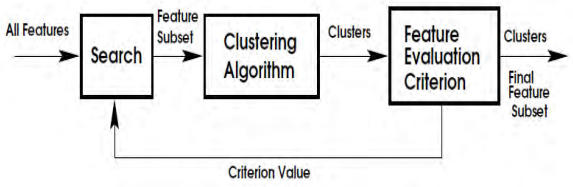


Figure 2 Wrapper approach.

(1) Feature search

Perform the feature search using the greedy approach like sequential forward and backward elimination. Forward search start with zero feature and added one feature at a time while backward elimination start with all the features and removes the worst attributes remaining in the set. The procedure may employ a threshold on the measure to determine when to stop the attribute selection process.

(2) Clustering algorithm

Different clustering algorithms are use for the generation of the accurate cluster. K-means ,K-medoid and EM are the example of clustering algorithms.

(3) Feature subset evaluation

Select the features from the generated cluster Which remove the redundant and irrelevant attributes. This method is use for selecting the interesting features from the clusters. Using this method we can get the quality of feature attributes.

2. Filter approach

Filter approach uses intrinsic properties of data for feature selection. This is the unsupervised feature selection approach. This approach performs the feature selection without using induction algorithms which is display in the figure 3[11]. This method is used for the transformation of variable space. This transformation of variable space is required for the collation and computation of all the features before dimension reduction can be achieved.

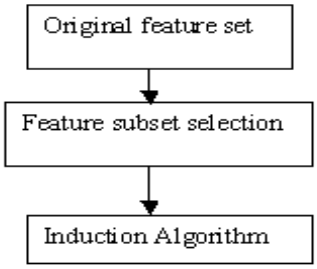


Figure 3:Filter approach

This approach is computationally simple, fast and scalable. Various feature selection based techniques already available like information gain (IG), Relief-F (RF), t-test (T) and chi-squared test (CS),Fisher filter.

II. Feature Similarity Measure

Unsupervised algorithm which used feature similarity for redundancy reduction but requiring no feature search. Feature similarity between two random variables based on the linear dependency.

Feature similarity approach classify into two categories like non parametrically test the closeness of probability distributions of the variable and functional dependency of variable.

Non parametrically test the closeness approach is sensitive of location and dispersion of the distribution. Functional dependency of variable is linearly dependant which is easily remove the dependency.

Linearly dependant variable of the attribute can calculate the feature similarity using the three techniques.

a) Correlation coefficient
b) Least Square Regression Error
c) Maximum Information Compression Index (MICI).

a) Correlation coefficient

Correlation coefficient can be measures by ρ. Correlation coefficient ρ between two random variables x and y is define as

$$\rho(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}}$$

Where cov= covariance , var = variance.

$\rho(x,y) = 1 \text{ or } -1$, then completely correlated

$\rho(x,y) = 0$ then uncorrelated

This technique is not desirable as variance has high information content and also sensitive to rotation is also not

desirable in many applications.

b) **Least Square Regression Error**

Least square regression error can be measure by the e. Least square regression error e between two variable x and y is obtain by the mean square error

$$e(x,y)^2 = \frac{1}{n}\sum (e(x,y))^2$$

Where $e(x,y) = var(y)(1-\rho(x,y)^2)$

$e(x,y) = 0$ then complete correlated

$e(x,y) = var(y)$ then uncorrelated

e^2 is known as the residual variance.

This technique is also sensitive to rotation of the scatter diagram in x-y plane.

c) **Maximum Information Compression Index**

MICI method is use for the finding the similarity of the attributes. This technique does not required the feature search. Maximum information compression index can be measure by λ_2.

$\lambda_2(x,y) =$ smallest eigenvalue of Σ i.e.

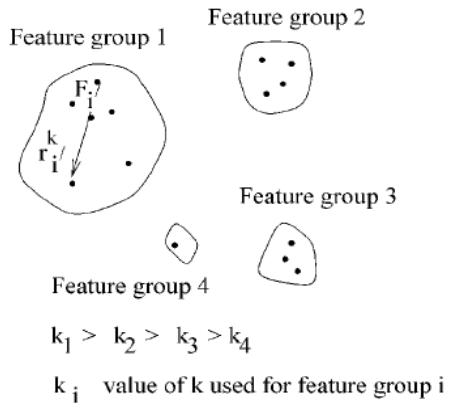
$$\lambda_2(x,y) = var(x) + var(y) - \sqrt{(var(x)-var(y))^2 + 4cov(x,y)^2}$$

This technique is a measure of minimum amount of information loss. MICI generate the single variable pair from the large amount of attributes.

III. Clustering Based Feature Selection

The absence of class label information makes unsupervised feature selection challenging and so it is hard to distinguish relevant features from the irrelevant features. Also, some real time application may require to apply various data mining functionality on same data. So, it is desirable to have some feature selection method that can be apply to supervised and unsupervised learning. The Clustering based feature selection algorithm, does not depend on the class label information in the data set, and hence it is unsupervised feature selection but it works for both supervised and unsupervised learning. It does not required any feature search and reduced the redundant and irrelevant data from the dataset.

Feature cluster



Feature cluster is done in two steps:

i. Features are partition into different clusters based on the similarity of feature
ii. Select representative feature from each cluster.

Partition of the feature is done based on the k-NN principle using one of the similarity measures.

Let original features be D and original feature set be O ={F_i, $i = 1,2 \ldots D$}. Dissimilarity between features F_i and F_j by S(F_i , F_j). r_i^k represent the dissimilarity between feature F_i and its kth nearest neighbour feature in R then[9],

1) Choose $k \leq D-1$. Reduced feature subset R to original feature set O.
2) Compute r_i^k. For each $F_i \in R$
3) Find minimum r_{it}^k of feature F_i . Retain the feature R and discard k nearest features of F_{it} .
4) Let $\epsilon = r_{it}^k$ If k≥ cardinality(R)-1.
5) If $k = 1$ than return feature set R as reduced set.
6) $r_{it}^k > \epsilon$
   k=k-1 then $r_i^k = inf_{F_i \in R} r_i^k$.
7) K=1 then return feature set R as reduced set.
8) Go to step 2.

Feature selection through feature clustering (FSFC) algorithm steps:

1. Select the dataset which contain the large number of features.

2. Calculate the MICI for each cluster.
   MICI calculate using
   $$var(x) + var(y) - \sqrt{(var(x) - var(y))^2 + 4cov(x,y)^2}$$
   Where var= variance , cov= covariance between two variables x,y.

3. Select the minimum C(Si,Sj) and merge that Si and Sj. Process continue until all objects feature information into the single cluster.

4. Select the top k cluster in the hierarchical cluster tree.

MICI is maximal information compression index which is used for feature selection. This algorithm is generic in nature and has a capability of multi scale representation of the data..

The above process stops until all the clusters are merged into one cluster. Complexity for this algorithm is $O(D^2)$. where D indicate the dimension. Feature selection approach based on hierarchical clustering which is very easy to generate the features attribute without much computation overhead.

Feature selection through feature clustering(FSFC) does not required feature search and feature selection occurs only once so the performance time will be reduce and generates accurate clusters with little overhead. This algorithm is fast compare to traditional unsupervised feature selection algorithms.

IV. Dataset characteristics

We used the 4 public cancer dataset. This dataset is available on www.upo.es/eps/bigs/datasets.html.

1. Colon Dataset: This is the Alon et al.'s Colon cancer dataset which contains information of 62 samples on 2000 genes. The samples belong to tumor and normal colon tissues.

2. Leukaemia: The total number of genes to be tested tested is 7129, and number of samples to be tested is 72, which are all acute leukaemia patients, either acute lymphoblastic leukaemia (ALL) or acute myelogenous leukaemia (AML).

3. Lymphoma: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.

4. Breast cancer: The training data contains 78 patient samples and number of genes is 24481.

Dataset	No of Instance	No of Original Features	Class
Colon	62	2000	2
Leukemia	38	7129	2
Lymphoma	45	4026	2
Breast Cancer	78	24481	2

Table:1 Dataset Characteristic

V. Experimental result

Simulation on cancer dataset generates the results for the different feature selection methods. There are many standard feature selection algorithms available in the simulation tool. In this paper we have simulated the relief and fisher filter algorithms on the cancer dataset and generate the reduced features and compare the result with Feature selection through feature clustering(FSFC) algorithm

Relief is the most successful algorithm for reduce the features. This algorithm does not consider the redundancy of the input attributes. The precondition for the relief filter is at least two attribute must be available and target attribute must be discrete. Relief filter is used the ranking search method for the reduce the original attribute process.

Fisher filter is also the available standard algorithm based on the filter approach. In this algorithm process selection is independent from learning algorithm. The precondition for this algorithm is at least one discrete attribute and one or more continuous attributes must be available. Feature search also done based on the ranking of the attributes.

Relief and fisher filter feature selection methods are applied on the cancer data set and try to generate the reduced features. The table 2 shows the results for Relief, fisher filter and feature selection through feature clustering (FSFC) algorithm. The result shows that, FSFC algorithm reduced the more attributes other than the standard algorithms like relief and fisher filter.

Sr. No	Dataset	No of Original Attributes	Feature Selection Algorithm	No of Reduced Attributes
1	Colon	2000	Relief	551
			FisherFilter	53
			Feature selection through feature clustering	11
2	Leukemia	7130	Relief	2000
			FisherFilter	286
			Feature selection through feature clustering	7
3	Lymphoma	4027	Relief	724
			FisherFilter	180
			Feature selection through feature clustering	15
4	Breast Cancer	24482	Relief	4547
			FisherFilter	86
			Feature selection through feature clustering	25

Table:2 Simulation of Feature selection

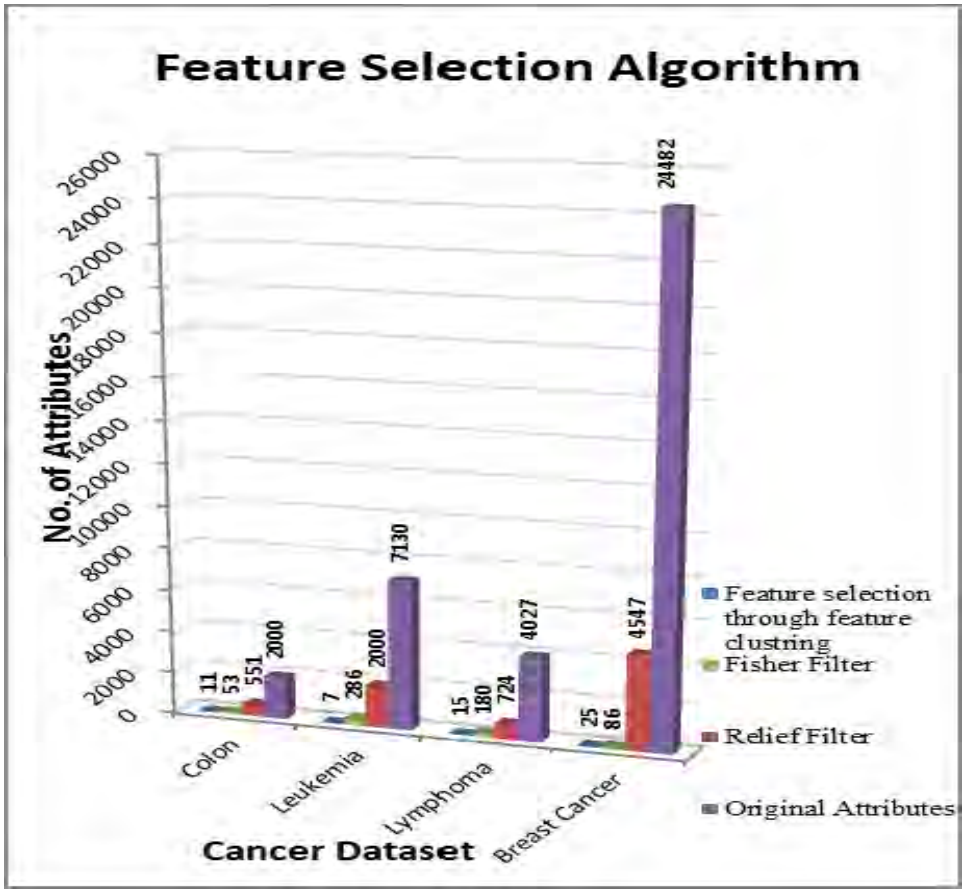


Figure :4 Feature selection algorithm on cancer dataset

VI. Conclusion

Feature selection is the pre-processing step which reduces the feature subset and thus supports high dimensional dataset. In this paper I have discussed many feature selection algorithm along with the Feature selection through feature clustering algorithm. The available algorithms like relief, fisher filter and proposed algorithm Feature selection through feature clustering are applied on cancer data and proved that Feature selection through feature clustering reduce more number of attributes compared to relief filter and fisher filter.

VII.Future Work

The relief and fisher filter generate the less number of attributes but this algorithms does not remove the redundant data. Clustering based feature selection algorithm remove the redundancy from the attributes and also provide the reduced or required attributes from the original attribute set. Clustering based feature selection algorithm support the high dimensional data set. In my future work I am planning to implement Clustering based feature selection algorithm on other clustering method which does not support high dimensional dataset and try to prove that the Clustering based feature selection algorithm generates more accurate clusters with minimum errors.

VIII.References

[1] Y. Yang and J. O. Pedersen, “A Comparative Study on Feature Selection in Text Categorization,” Proc. of Int’l Conf. on Machine Learning, pp. 412–420, 1997.
[2] Jiawei Han and Micheline Kamber “Data Mining: Concepts and Techniques”, 2nd ed.
[3] T. Liu, S. Liu, Z. Chen, and W. Ma, “An Evaluation on Feature Selection for Text Clustering,” Proc. of Int’l Conf. on Machine Learning, 2003.
[4] J. R. Quinlan, “Induction of Decision Trees,” Machine Learning, vol. 1,pp. 81–106, 1986.
[5] M. Dash and H. Liu, “Feature Selection for Classification,” Intelligent Data Analysis, vol. 1, no. 3, pp. 131–156, 1997.
[6] Yanjun Li, Congnan Luo, and Soon M. Chung, Member, IEEE,” Text Clustering with Feature Selection by Using Statistical Data,” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. XX, NO. YY, 2008.
[7] Jennifer G. Dy, Jennifer G. Dy. “Feature selection for unsupervised learning” Journal of Machine Learning Research 5 (2004) 845–889.

[8] Guangrong Li1, 2, Xiaohua Hu3, 4, Xiajiong Shen4, Xin Chen3, Zhoujun Li5 "A Novel Unsupervised Feature Selection Method for Bioinformatics Data Setsthrough Feature Clustering", Granular Computing, 2008. GrC 2008. IEEE International Conference on 26-28 Aug. 2008

[9] P. Mitra, C. A. Murthy, and Sankar K. Pal, "Unsupervised Feature Selection Using Feature Similarity", IEEE Transactions on Pattern Analysis and MachineIntelligenc, vol. 24, pp. 301-312, 2002.

[10] Y. Yang, "Noise Reduction in a Statistical Approach to Text Categorization, "Proc. of Annual ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 256–263, 1995.

[11] Asha Gowda Karegowda, M.A.Jayaram, A.S. Manjunath "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning" 2010 International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 7.

[12] Barkha H.Desai, Hetal B.Bhavsar, Nisha V. Shah "Supervised and Unsupervised Feature Selection based Clustering Algorithms" NCTM-2012,Visnagar.