www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3, Issue 10 October, 2014 Page No. 8937-8941

Semantic Search Based Feature Subset Selection for Multi Dimensional Data

M. Srinu[1], *K. Aruna Bhaskar*[2]
[1] Student Dept.of CSE, Sri Sai Aditya Institute of Science & Technology, Surampalem,
Kakinada, E.G D.t, A.P, India
srinu4gate@gmail.com
[2] Sr. Asst.Professor, Dept.of CSE, Sri Sai Aditya Institute of Science & Technology, Surampalem,
Kakinada, E.G D.t, A.P, India.
lettertoarunbhaskar@gmail.com

Abstract: ***Cluster analysis is one of the prominent unsupervised learning techniques widely used to categorize the data items based on their similarity. Mainly off-line and online analysis through clusters is more attractive area of research. But, high dimensional big data analysis is always introducing a new dimension in the area of data mining. We have different variable selection methods for clustering of data like density based, model based and criterion based variable selection methods. Because high dimensional cluster analysis is giving less accurate results and high processing time when considering maximum dimensions. To overcome these issues dimensionality reduction techniques have been introduced. Here, a million dollar questions are, which dimensions are to be considered? , what type of measures have to be introduced? And how to evaluate the cluster quality based on those dimensions and measures? Proposed approach effectively answers these questions by introducing Ensemble feature subset selection measure along with Extend leader follower algorithm to justify the proposal with experimental evaluations.***

Keywords: feature subset, clustering, dimensionality reduction, filter method.

1. INTRODUCTION

Feature selection is generally used as a preprocessing step to machine learning task. It is a process of selecting a best subset of original features so that the feature space is optimally reduced based on a certain evaluation criterion. Feature selection has been an active field of research and development since 1970's and shown very effective in removing irrelevant and redundant features, increasing efficiency in learning tasks, improving learning performance like predictive accuracy, and enhancing clarity of learned results.

In recent years, data has become gradually larger in both rows (i.e., number of instances) and columns (i.e., number of features) in lots of applications such as genome projects, text categorization [3], image retrieval, and customer relationship management. This multi dimensional data analysis may cause serious challenge to many machine learning algorithms with respect to scalability and learning performance. For example, high dimensional data (i.e., data sets with hundreds or thousands of features), can contain high degree of irrelevant and redundant features which may greatly degrade the performance of learning algorithms. Therefore, feature selection becomes a need for machine learning tasks when considering multi dimensional data nowadays.

Because high dimensional cluster analysis is giving less accurate results and high processing time when considering maximum dimensions [1]. To overcome these issues dimensionality reduction techniques have been introduced. Here, a million dollar questions are, which dimensions are to be considered? , what type of measures have to be introduced? And how to evaluate the cluster quality based on those dimensions and measures?

Feature selection algorithms can generally fall into either the filter model or the wrapper model [9]. The filter model relies on the basic characteristics of the training data to select some features without considering any learning algorithm; therefore it does not inherit any bias of a learning algorithm. The wrapper model requires one predestined learning algorithm in feature selection and uses its performance to calculate and determine which features are selected. As for each new subset of features, the wrapper model requires to find a hypothesis (or a classifier). It tends to give better performance as it finds features, those are appropriates to the predestined learning algorithm, but it is computationally expensive [7]. When the number of features becomes larger, then the filter model is typically a choice due to its less computational complexity.

To include the advantages of both models, algorithms in a hybrid model have been proposed to deal with high dimensional data. In these algorithms, first, a goodness measure of feature subsets based on data characteristics is used to choose best subsets for a given cardinality, and then, cross validation is exploited to decide a final best subset across different cardinalities. These algorithms mainly focus on combining filter and wrapper algorithms to achieve best possible performance with a particular learning algorithm at the same time complexity of filter algorithms.

When you submit your paper print it in two-column format, including figures and tables. In addition, designate one author as the "corresponding author". This is the author to whom proofs of the paper will be sent. Proofs are sent to the corresponding author only.

2. EXISTING SYSTEM

The existing system is a novel clustering based feature subset selection algorithm for high dimensional data. The algorithm involves i) removal of irrelevant features, ii) constructing a minimum spanning tree from relative ones and iii) partitioning Minimum Spanning Tree and selecting representative features. In this algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is reduced.

2.1 LIMITATION

Any single feature set selection algorithm is not suppose enough to decide its result is more efficient than other because it may fit or may not fit to heterogeneous nature of data. So there is no exemption to the existing system also. There is a need to find an adaptive and optimal solution for feature subset selection for various data mining tasks such as classification, clustering etc.

3. PROPOSED SYSTEM

The proposed system is a novel ensemble method of variable selection. In this model more than one feature selection method is used to get feature subsets. After selection of feature subsets from multiple methods either majority voting or intersection of all feature subsets to get more relevant features will be selected for next data mining applications such as clustering or classification.

3.1 ADVANTAGES

- It is ensemble method of technique; hence we will get best feature subset for clustering.
- Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.
- It efficiently and effectively deals with both irrelevant and redundant features, and obtains good feature subset.

4. SYSTEM ARCHITECTURE

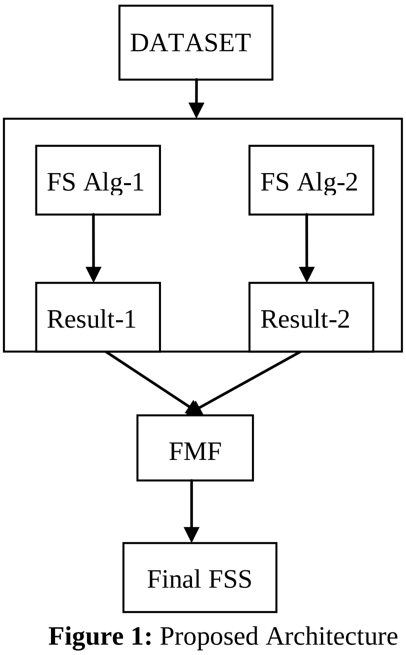


Figure 1: Proposed Architecture

Where
FS- Feature Selection
FMF- Feature Merge Function
FSS- Feature Subset

The proposed approach built on two feature selection algorithms called as FAST and Relief-F+kNN. The FAST algorithm works in three steps. In the first step, it removes the irrelevant features with the target class by using a threshold value [6]. In the second step, features are separated into clusters by using graph-theoretic clustering methods. In the third step, the most representative feature that is stoutly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of features.

4.1 Algorithm-1: FAST

```
Input: D($F_1$,$F_2$, ..,$F_m$, C) – the given dataset
       $\theta$ – the T-Relevance threshold.
Output: S – selected feature subset
//=== Part1: Irrelevant Features Removal ===
for i=1 to m do
    T-Relevance = SU ($F_i$, C)
    if T-Relevance > $\theta$ then
        S= S U {$F_i$}
//===Part2:Minimum Spanning Tree Construction===
G = NULL; // G is a complete graph
for each pair of features {$F_i'$, $F_j'$} $\subset$S
  do
    F-Correlation = SU ($F_i'$, $F_j'$)
    add $F_i'$ and / or $F_j'$ to G with F-Correlation as the
        weight of the corresponding edge;
minSpanTree = Prim (G); // Using Prim`s
  Algorithm to generate minimum spanning tree
//===Part3:Tree Partition and Representative Feature
      selection===
Forest = minSpanTree;
for each edge $E_{ij}$ $\in$ Forest do
    if SU($F_i'$,$F_j'$) < SU($F_i'$,C) $\wedge$ SU ($F_i'$,$F_j'$)<SU ($F_j'$,C) then
        Forest = Forest - $E_{ij;}$
S = $\Phi$
for each tree $T_i$ $\in$ Forest do
    $F^j_R$ = argmax $_{F_k' \in T_i}$ SU ($F_k'$, C)
    S = S U { $F^j_R$}
return S
```

RELIEF was originally proposed as an online feature selection algorithm based on some heuristical intuitions [2]. Relief estimates are improved than usual statistical feature estimates, like correlation or covariance because it considers attribute interrelationships. Later it was extending as Relief-F [8]. The original Relief can deal with nominal and numerical attributes. However, it cannot deal with incomplete data and is limited to two-class problems. Its extension, which solves these and other problems, is called Relief-F.

The Composition of Relief-F + kNN gives good results for feature sub set selection. This model randomly select instance from given train dataset and then it will find k nearest hit set called H and k nearest miss set called M. Based on the these sets it updates the weight of the each feature by considering the average value of the difference among Hit set (H) and Miss set (M) with the current instance and current attribute using any of the dissimilarity measure such as

Euclidian or Manhattan distance [10]. The algorithm details are as follows.

4.2 Algorithm-2: Relief-F + kNN

Input: for each training instance a vector of attribute values and the class value

Output: the vector W of estimations of the qualities of attributes

```
1. set all weights W[A] := 0.0;
2. for i=1 to m do begin
   a. randomly select an instance $R_i$
   b. find k nearest hits $H_j$;
   c. for each class $C \neq class(R_i)$ do
      i. from class C find k nearest misses $M_j(C)$;
   d. for A =1 to all_attributes do
      $W[A] := W[A] - \sum_{j=1}^{k} diff(A,R_i,H_j)/(m.k)$
            $+ \sum_{c \neq class(Ri)} [P(C)/(1-(class(R_i)))$
            $+ \sum_{j=1}^{k} diff(A,R_i,M_j(C))]/(m.k)$
3. end;
```

Now proposed Ensemble Feature Selection Algorithm is described as follows

4.3 Algorithm-3: Ensemble Feature Selection (EFS)

Input: Dataset,

Output: Feature Subsets FSS

```
1. FS1=KnnReliefF(Dataset, k)
2. FS2=FAST(Dataset)
3. Find FSS={FS1}∩{FS2}
4. return FSS
```

In the first step of algorithm some subset of features (FS1) are selected from Algorithm-1 and other sub set of features (FS2) are selected from Second algorithm. This algorithm filters both of the results by finding the common set of attributes from both the results as final feature subset. This final set can be used further for any data mining task;

In this paper these feature subset selection is used for clustering of data items. Original Leader Algorithm has been proposed in [4]. But there is a limitation in that approach. In original algorithm every data instance is assigned to nearest leader if the distance between current object and leader less than threshold. But here already assigned objects don't know about new leaders. In that case if already assigned objects are more near to any of a new leader then there should be provision to re-assignment of that instance to the new leader and remove it from current cluster [5]. Here is the algorithm with that extension.

4.4 Algorithm-4: Adaptive Leader Election Algorithm

Input: Dataset, similarity threshold th

Output: Clusters

```
1. FS=EFS(Dataset)
2. Extract first data instance and mark it as a leader and
   append to leaders' list
3. do for all instances Di = 2 to n
4. {
      a. Calculate the distance with all leaders
      b. Find the nearest leader
      c. if(distance with nearest leader <  th)
      d. {
            i. Assign it to the nearest leader
           ii. Mark the cluster number
          iii. Add it to member list of this cluster
           iv. Increment member count of this cluster
      e. }
      f. else
      g. {
            i. Add it to leader list
           ii. Increment leader counter, L=L+ 1
      h. }
5. }

6. Now again compare all the data instances with the
   leaders excluding leader instances and leaders formed
   before current instance to find the new nearest leader if
   any when compared to the current leader.
7. If new leader is nearer than current one then remove the
   instance from current cluster and assigned to new
   cluster.
8. Evaluate the cluster quality.
```

Furthermore modified algorithm cannot be executed for the entire leader set for each instance i.e. each data instance is compared with those leaders which are created after current home leader. So time complexity will be gradually decreased based on the data instance creation timestamp. The proposed method can also be applied for data in motion and it can handle concept drift in such data streams also due to dynamic creation of leaders. Thus it can be used for both online and offline analysis. In this algorithm pre-processing is done at 3rd step by selecting the relevant feature subset related to underlying concept. Due to dimensionality reduction this algorithm can efficiently used for high dimensional datasets also. To justify the proposed approach quality of each cluster is evaluated.

5. EXPERIMENTAL EVALUATIONS

Experiments are conducted on datasets downloaded from UCI Machine Learning Repository. Test Datasets includes Breast Cancer, 9Tumeors and KDDCup99. Those datasets are selected in such a way so that some contain large number of instances with low dimensions and some contain high dimension with less instances and finally high dimensions with large dataset size.

6. SCREENS

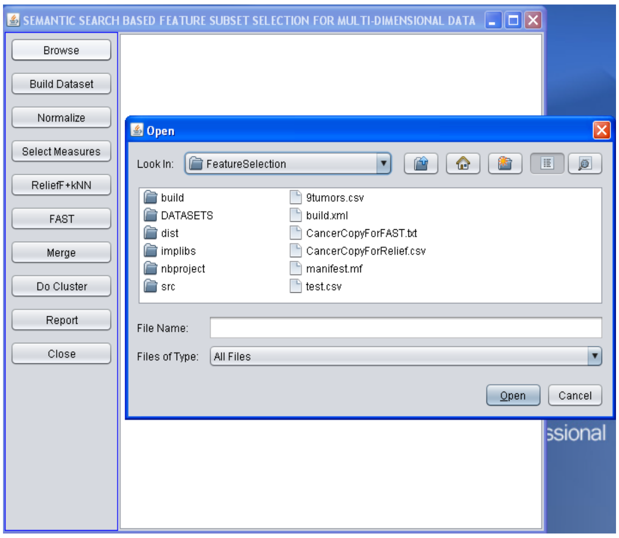
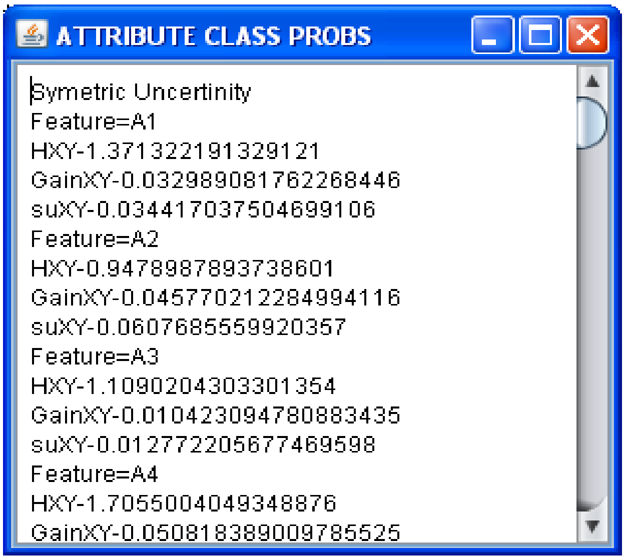
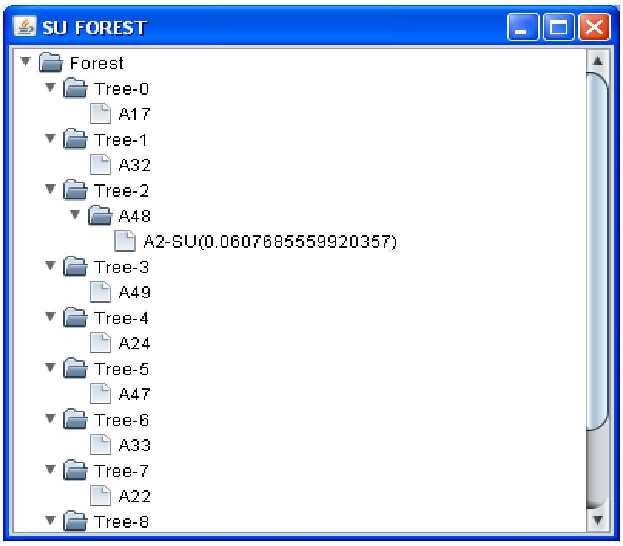


Figure 2- : Loading of Dataset



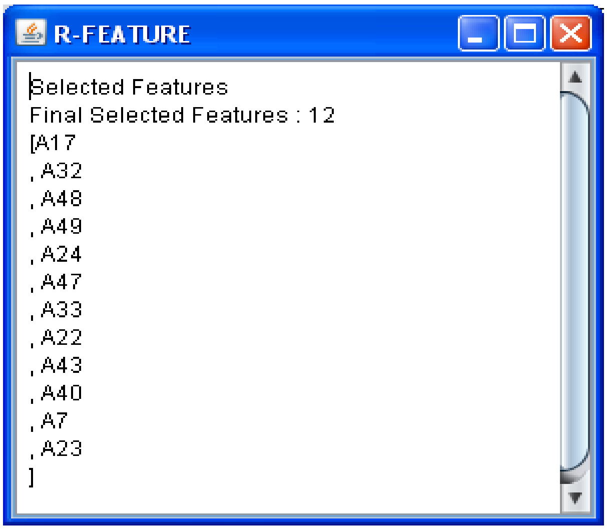
```
Symetric Uncertinity
Feature=A1
HXY-1.371322191329121
GainXY-0.0329890817622684446
suXY-0.034417037504699106
Feature=A2
HXY-0.9478987893738601
GainXY-0.045770212284994116
suXY-0.0607685559920357
Feature=A3
HXY-1.1090204303301354
GainXY-0.010423094780883435
suXY-0.012772205677469598
Feature=A4
HXY-1.7055004049348876
GainXY-0.050818389009785525
```

Figure 3- : Gain & SU Calculation



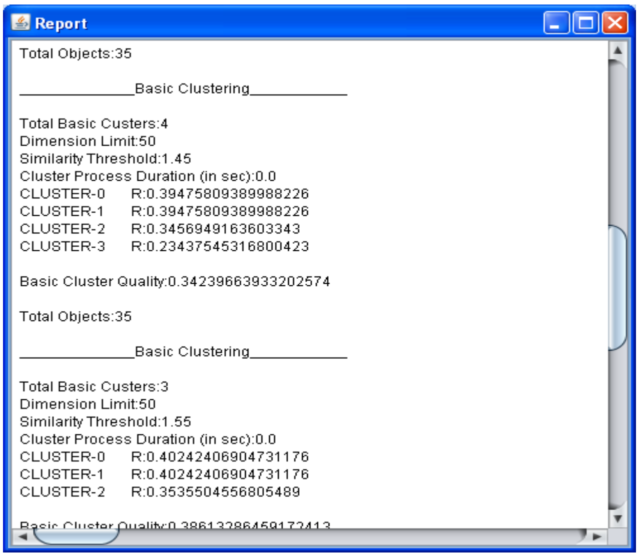
```
Forest
  Tree-0
    A17
  Tree-1
    A32
  Tree-2
    A48
      A2-SU(0.0607685559920357)
  Tree-3
    A49
  Tree-4
    A24
  Tree-5
    A47
  Tree-6
    A33
  Tree-7
    A22
  Tree-8
```

Figure 4- : Minimum Spanning Tree Partition



```
Selected Features
Final Selected Features : 12
[A17
, A32
, A48
, A49
, A24
, A47
, A33
, A22
, A43
, A40
, A7
, A23
]
```

Figure 5- : Relevant Features



```
Total Objects:35

_______________Basic Clustering_____________

Total Basic Custers:4
Dimension Limit:50
Similarity Threshold:1.45
Cluster Process Duration (in sec):0.0
CLUSTER-0      R:0.39475809389988226
CLUSTER-1      R:0.39475809389988226
CLUSTER-2      R:0.3456949163603343
CLUSTER-3      R:0.23437545316800423

Basic Cluster Quality:0.34239663933202574

Total Objects:35

_______________Basic Clustering_____________

Total Basic Custers:3
Dimension Limit:50
Similarity Threshold:1.55
Cluster Process Duration (in sec):0.0
CLUSTER-0      R:0.40242406904731176
CLUSTER-1      R:0.40242406904731176
CLUSTER-2      R:0.3535504556805489

Basic Cluster Quality:0.38613286459172413
```

Figure 6- : Cluster Report

7. CONCLUSION

Thus proposed approach effectively handles high dimensional online and offline data with Ensemble feature subset selection measure along with Adaptive Leader follower algorithm to justify the proposal with experimental evaluations. Feature work includes concurrent execution of these algorithms with multi-core systems as well as cloud computing.

REFERENCES

[1] U. Maulik and S. Bandyopadhyay, “Performance evaluation of some clustering algorithms and validity indices,”IEEE PAMI, vol. 24, pp. 1650–1654, 2002

[2] K. Kira and L.A. Rendell, “A Practical Approach to Feature Selection,” Proc. Ninth Int’l Workshop Machine Learning (ICML ’92), pp. 249-256, 1992 .

[3] Shuang-Hong Yang and Bao-Gang Hu,”Discriminative Feature Selection by Nonparametric Bayes Error Minimization” IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 8, August 2012.

[4] P.A.Vijaya, M. Narasimha Murty and D. K. Subramanian" Leaders-subleaders: an efficient hierarchical clustering algorithm for large data sets" Pattern Recognition Letters archive Volume 25 Issue 4, March 2004 Pages 505 - 513 Elsevier Science Inc. New York, NY, USA.

[5] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, no. 1, pp. 53–65, 1987.

[6] Hall, M. (1999). "Correlation based feature selection for machine learning". Doctoral dissertation, University of Waikato, Dept. of Computer Science.

[7] Langley, P. (1994). Selection of relevant features in machine learning. Proceedings of the AAAI Fall Symposium on Relevance. AAAI Press.

[8] C, I. (1994). Estimating attributes: Analysis and extension of RELIEF. Proceedings of the European Conference on Machine Learning (pp. 171– 182). Catania, Italy: Berlin: Springer-Verlag.

[9] 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).

[10] Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004

AUTHOR PROFILE

Mr. M. Srinu is a student of Sri Sai Aditya Institute of Science & Technology, SuramPalem. Presently he is pursuing his M.Tech [CSE] from this college and he is received his MCA degree from Aditya Engineering College, Affiliated to JNTUK University in the year 2009. His area of interest includes Web Technologies and Object Oriented Programming Languages, all current trends and technologies in Computer Science.

Mr. K. Aruna Bhaskar completed his M.Tech Degree in Computer Science and Engineering from Sri Vasavi Engineering College, Tadepalligudem, affiliated to JNTUK University, Kakinada, and Andhra Pradesh, recognized by UGC and approved by AICTE. He is working as Sr. Asst. Professor in CSE Dept., Sri Sai Aditya Institute of Science and Technology, Surampalem, India. He is currently doing research work on Mobile Computing. He has one International Conference and two National Conferences to his credit. His research interests include Neural Networks and Neuro Fuzzy Systems and Medical Image processing.