Constraint based Clustering in Feature Subset Selection Algorithm

S. Aswini
PG scholar(CSE)
S.K.P Engineering College
Tamil nadu

A. Kumaresan
HOD In Dept of CSE
S.K.P Egineering College
Tamil nadu

D. Murali
Asst.Prof.CSE
S.K.P Engineering College
Tamil nadu

ABSTRACT
Constraint based clustering that satisfies set of user-defined constraint. Constraint-based clustering is an example of a mining task where flexibility is desirable. It is a generalization of standard clustering in which the user can impose constraints on the clustering to be found, such as similar and dissimilar constraints. Feature selection involves recognizing a subset of the most constructive features that produces companionable outcomes as the innovative intact set of traits. Feature selection, as a data for preliminary considered. Features are processing step is effectual in reducing the spatial property, confiscate irrelevant data, mounting learning accuracy. We proposed an algorithm for the Region of Influence (ROI). The algorithm is numb to the order in which the pairs are divided into cluster by using relative neighborhood graphs (RNGs).

Keywords
Relative neighborhood graphs, Region of influence, Feature subset selection.

1. INTRODUCTION
Thus, in many appliance, it is needed to have the clustering process take user inclination and into constraints into contemplation. Examples of such information include the expected number of clusters, the minimum or maximum cluster size, weight for dissimilar items[1]. Moreover, when a grouping a task involves a rather high dimensional space, it is very intricate to engender meaningful cluster by relying only on the grouping parameters. We proposed constraint based clustering finds groups that persuade user-specified inclination or condition. Depending on the nature of the conditions, constraint-based clustering may embrace rather atypical approaches.

Here are a few categories of constraints.

1. Constraint on individual objects.
2. A constraint on the selection of clustering parameters.
3. User specified constraints on the properties of individual groups.

Constraint on individual objects user can identify on the attribute to be group. For example in hospital monitor the patient health condition. To produce yearly record that will be contain particular age of people affect by diabetes i.e.,age in the range 35-45 affected by diabetes. Another example age in 50-60 range they will most affect by cholesterol.

Constraints on the selection of clustering parameter: People like to situate craving for each clustering parameter. Grouping parameters are actually somewhat specific to given grouping algorithm. Examples such as k-means algorithm .

User specified constraints on the properties of individual groups: A client may perhaps like to identify character of the end result of the clusters. For example courier service center identify the location for *n* service location metropolis .The center has catalog of clients that chronicle the client name, location. Center find the particular client based on the name and location they specified.

Feature selection is frequently used as a data for preliminary considered step . It is a procedure of choosing a separation of original description so that the feature space is best possible to reduced according to a firm evaluation measure [2]. Irrelevant, unnecessary, or noisy data, and brings the abrupt effects for diligence.

Hustling up a data mining algorithm, increasing mining recital such as prognostic accuracy and result lucidity. Feature selection has been a lush field of data mining and widely applied to many fields. Example: Text cataloguing, Icon Repossession, Punter relationship administration, Intrusion detection and genomic analysis[3]. **Demerits** of existing algorithm such as minimum spanning tree (MST). The algorithm works well in many cases where the clusters are well separated. A problem may occur when a “large” edge e has another “large” edge as its neighbor. In this case, e is likely not to be characterized as inconsistent and the algorithm may fail to unravel the underlying clustering structure correctly.

Subsisting methods for semi-supervised clustering plummet into two broad category first one is constraint-based and metric-based. In constraint-based tactics , the clustering algorithm itself is mutated so that punter-provided labels or brace wise conditions are used to steer the algorithm towards a more suitable data rifting .This is complete by mutating the clustering intention function so that it take account of satisfaction of constraints.

In constraint based clustering, given a set of data D ,object n. distance function of two different clustering data dc: $D \times D \rightarrow L$, L is a positive integer ,and to find set of constraint on data . Find L clustering (C_{K1} ,…….,C_{Kl}) that means minimize the cluster data in the form of

$\sum_{i=1}^{L}$ ($C_{ki,}$char$_i$) and each cluster satisfies the constraints c denote as C_{kl}=C[9].

We proposed an algorithm such as the Region of Influence (ROI) be part of a identify region with its own inimitability claim to be a regional power . Wield vital influence on the geographic point of the region as well as on its creed construction. Its main advantages is the inimitable of the "regions" .Wherein each spot under swot up has its own cluster of passable similar spots that form the starting point for the transmission of information on tremendous to the spot of fascination.Proposed a technique called RNG (Relative Neighborhood Graph).In RNG graph has set of 'Q' points in the define region as follows: two points 'r' and 's' in the set 'Q' define an edge of RNG(Q) when

$$\text{Dis}(r,s) \leq \max\{\text{dis}(r,t), \text{dis}(q,t)\}$$

For all points t in Q, where 'dis' is the distance fuction.

2. FRAMEWORK OF THE PROPOSED FEATURE SUBSET SELECTION WITH CONSTRAINT

Extract the data from large database in first step remove the irrelevant feature. Based on attribute selection measure we can select a relevant feature and remove irrelevant objects [10]**. We** can select attribute based three attribute selection measures are follow: **Information gain, Gain ratio, gini index.** Information gain is one of the attribute selection measures. The trait with the utmost information gain is preferred. Information gain quantify is predisposed toward tests with many upshots. It's only applicable in a large number of values.

Gain ratio is extension of information gain, which attempts to overcome this predisposition. It pertain a kind of conventional to information gain[5]. Next we can identify redundant features, i.e., any duplicate information exist. This step leads to very erudite one. The attribute with the utmost gain ratio is split up attribute. **Gini index** calculate the adulteration of attribute 'A'. The Gini index mull over a binary rip for each tuples. First mull over the case where 'A' is discrete value tuples having *l* different values,{A_1,A_2,….,A*l*} . Finish the attribute selection next step to be remove irrelevant feature that attribute will not suitable to particular data base[8] . We can use one algorithm for reducing costs such that ROI. And using one technical graph such as a Relative neighborhood graph.

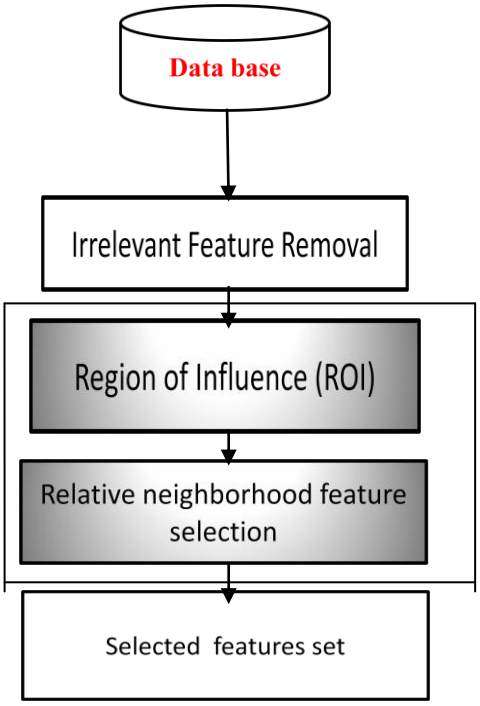


Fig. 1: Framework of the proposed feature subset selection with constraint

3. ALGORITHM FOR INFLUENCE

After removing irrelevant feature from database , client can identify same attribute to be cluster based on algorithm Region Of Influence(ROI).In these algorithm calculate the attributes in two different region in that they will find similar attribute.

3.1 Definition of algorithm

The region of influence of two different cluster $y_i, y\underline{x}_j \in Y$ is labeled as:

$RL(y_i, y\underline{x}_j)=\{y: \text{cond}(d(y,y_i), d(y,y_j), d(\underline{y}_i,y_j)), \underline{y}_i \neq \underline{y}_j\}$

where cond($d(y,y_i)$, $d(y, y_j)$, $d(y_i, y_j)$) may be labeled as:

a) $\max\{d(y,\underline{y}_i), d(y, y_j)\} < d(y_i, y_j)$,

b) $d^2(y,y_i) + d^2(y, y_j) < d^2(y_i,y_j)$,

c) $(d^2(y, y_i) + d^2(y, y_j) < d^2(y_i, y_j))$ *OR* $(\alpha \min\{d(y, y_i), d(y, y_j)\} < d(y_i,y_j))$,

d) $(\max\{d(y, y_i), d(y, y_j)\} < d(y_i, y_j))$ *OR* $(\alpha \min\{d(y, y_i), d(y, y_j)\} < d(y_i, y_j))$,

where α affects the size of the ROI defined by $\underline{x}_i$, $\underline{x}_j$ and is called relative edge consistency.

ALGORITHM: REGION OF INFLUENCE

/*calculate the first cluster start from first element to last M element */

For i=1 to M

/* identify the next group to be cluster */

For j=i+1 to M

/*Determine the same attribute with help of ROI*/

Ascertain the region of influence are (y_i, y_j)

/* yi and yj close to each one the they will be put in same cluster*/

If R (y_i, y_j) ∩ (y-{y_i, y_j}) =Ø then

Add the edge connecting y_i, y_j.

End if

End for

End for

3.2 Explanation of algorithm

Calculate the connected modules of the end resulted graph and recognize them as clusters. In words:

1) The edge between y_i and y_j is added to the graph if no other vector in *Y* lies in $RL(\underline{x}_i,\underline{x}_j)$.

2) Since for y_i and y_j close to each one to be expected that $RL(y_i,y_j)$ restrains no other vectors in *Y*, it is predictable that nearby each other points will be delegated to the identical cluster. ROI has main advantages in that algorithm is uncaring to the sequence in which the pairs are mull overed.

4. CONSTRUCTION OF RNG GRAPH

In computational geometry, the relative neighborhood graph is an undirected graph identify on a rest of points in the Euclidean flat surface by joining two points a and b by an edge whenever there does not subsist a third point q that is nearer to both a and b than they are to both erstwhile[4].

RNG graph based on the graph theoretical clustering

To erect the relative neighborhood graph proficiently in O (*n* log *n*) time. It can be calculated in O (*n*) accepted time.The relative neighborhood graph can be computed in linear time from the point set, it is identified only in the conditions of the aloofness flanked by points

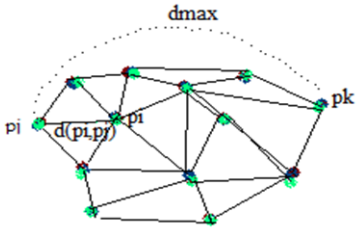


Fig 2: The RNG of 13 casual indicates in a unit square

4.1 Computing Relative Neighborhood Graph :

They will consider set of points 'S' in that graph .Calculate the group of well disconnected pairs of subsets of S[7].

d_{max} =max[d(pi, pk),d(pj, pk)] for each point
pk=(k=1,2,…..,n.k≠i, k≠j)
if d(pi,pj)>dmax
then
(pi,pj)-ϵ iff [d(pi,pj)>max[d(pi,pk),d(pj,pk)]
if d(pi,pj)<dmax
then
the set of points alienated by
dmax -ϵ iff [d(pi,pj)>max[d(pi,pk),d(pj,pk)]
if d(pi,pj) ≤ dmax for any pk(i≠k,j≠k)
then
(pi,pj) ϵ-iff d(pi,pj) ≤ d[(pi,pk),d(pj,pk)]

4.2 Performance Analysis of RNG

Above step of algorithm running time $O(n^2)$ time ,computation of algorithm at once these explanation leads to first step.
Second step:runs O(n) time ,computation of algorithm at least N times where N is the number of edges in RNG graph. This algorithm is apply for n-dimensional data and any distance appraise.

5. EXPERIMENTAL STUDY

5.1 Data Source

The reason for calculating the performance and effective of our proposed ROI algorithm, validating whether or not the technique is prospectively useful in practice. Collecting set data from web that leads to categories the disease. The data sets cover a variety of application realms such as medical, image and bio- microarray data classification.

5.2 Experiment setup

Researchers collect data based on the disease affected by people normal way and also they provide particular range i.e based age calculation they will infected by some diseaase. The clustering data list below table 1:

Table 1. Summary of dataset about disease affect by people

Name	Type	Width	Label	value
Age	Numeric	4	Age in years	None
Agecat	Numeric	4	Age in category	None
Gender	Numeric	8	Gender	{0, female}...
Diabetes	Numeric	4	History of diabetes	{0, No}...
Bp	Numeric	4	Blood pressure	{0, Hypotension}...
Smoker	Numeric	4	Smoker	{0, No}...
Arsenic sis	Numeric	8	arsenic sis	{0, no}...
Diarrhoea	Numeric	8	diarrhoea	{1, affected}...
Factdep	Numeric	8	fact about depression	{1, below 18 age}...
Hepatitis	Numeric	8	Hepatic-tis	{0, no}...
Jaundice	Numeric	8	Jaundice	{0, no}...
Malaria	Numeric	8	malaria fever	{0, no}...
Scabies	Numeric	8	Scabies	{0, no}...
Trachea	Numeric	8	lung cancer	{0, no}...
Trachoma	Numeric	8	affect by children	{0, no}...
Tuberculasis	Numeric	8	tuberculasis virus	{0, no}...
Typhoid	Numeric	8	Fever	{0, no}...
Proc	Numeric	8	Surgical treatment	{-2, Died before surgery}...
Comp	Numeric	4	Surgical complications	{-3, No surgery performed}...
Result	Numeric	4	Surgery result	{-3, No surgery performed}...
Los	Numeric	4	Length of stay	{-2, Died before surgery}...
Cost	Numeric	8	Treatment costs	{-1.00, D.O.A.}...

Above table list group of disease ,such disease affected by people normally in that table list name of the disease category of the disease then the value that hold. In fourth column label of the disease i.e, description about that disease. Finally check whether the disease affect or not denoted by numeric variable example '0' no disease and '1' yes infection are there.

5.3 Experiment Procedure

Our main aim to be clustering the disease based age category. We can cluster disease for example age range {35-45} means they will mainly affected by Diabetes.

Table 2. Frequency table
Gender

		Frequency	%	Valid%	Cumulative %
Valid	female	53	.5	65.4	65.4
	male	28	.3	34.6	100.0
	Total	81	.8	100.0	
Missing	System	9919	99.2		
Total		10000	100.0		

Above frequency table cluster the gender based on male patient and female patient. From the frequency table female frequency 53 and there percentage is .5 and valid percentage 65.4 and then cumulative percentage 65.4.

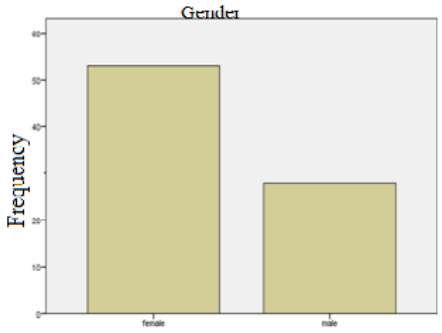


Fig 3:Averge frequency of table 1

Explanation above figure is normally female entrant more disease affected than male entrant. In yearly calculation based on given data set large number of female people affected by lung cancer, particularly breast cancer .in that range {30-40}

Table 3:Blood pressure

		Frequency	%	Valid %	Cumulative %
Valid	Hypotension	1207	12.1	12.1	12.1
	Normal	6134	61.3	61.3	73.4
	Hypertension	2659	26.6	26.6	100.0
	Total	10000	100.0	100.0	

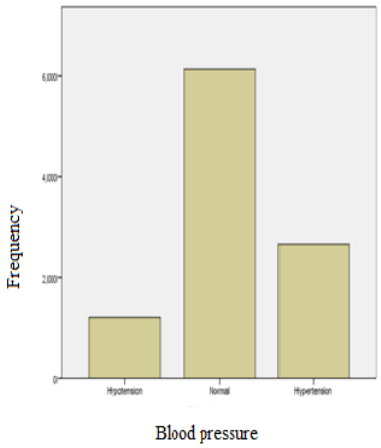


Fig 4:Blood Pressure Frequency

Table 4:Fact about depression

		Frequency	%	Valid %	Cumulative %
Valid	0	9	.1	27.3	27.3
	below 18 age	20	.2	60.6	87.9
	{20-45}	3	.0	9.1	97.0
	above 75	1	.0	3.0	100.0
	Total	33	.3	100.0	
Missing	System	9967	99.7		
Total		10000	100.0		

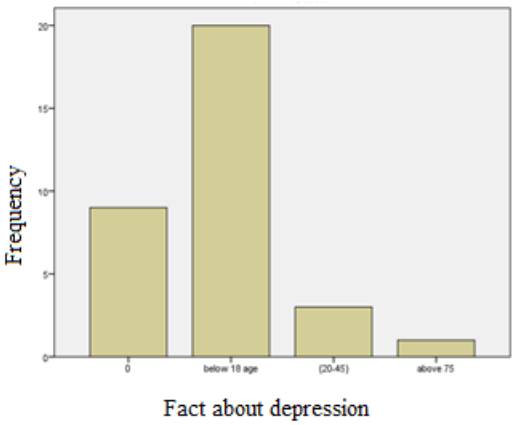


Fig 5:Fact about Depression

Table 5:Surgery result

		Frequency	%	Valid %	Cumulative %
Valid	Well	3537	35.4	81.7	81.7
	Stable	524	5.2	12.1	93.9
	Critical	195	2.0	4.5	98.4
	Died	71	.7	1.6	100.0
	Total	4327	43.3	100.0	
Missing	No surgery performed	3671	36.7		
	Died before surgery	631	6.3		
	D.O.A.	1371	13.7		
	Total	5673	56.7		
Total		10000	100.0		

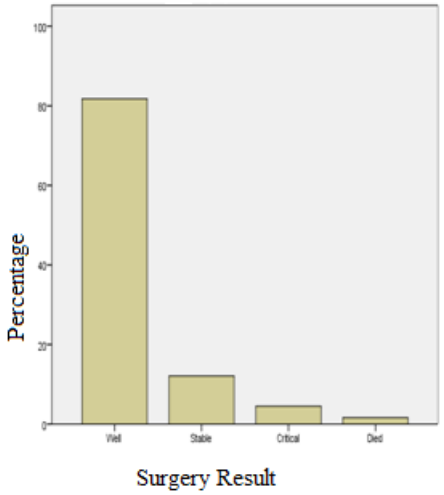
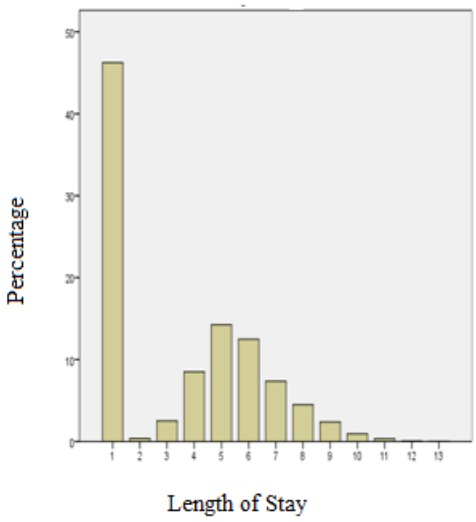
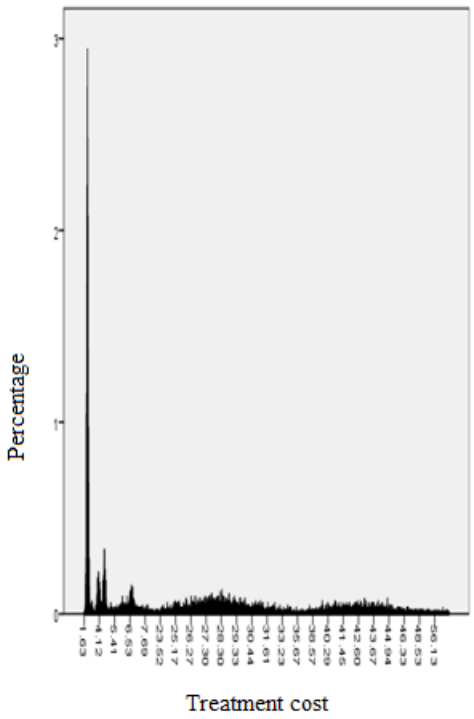


Fig 6: Surgery Result

In Surgery Result they produce four type of result i.e., well ,Stable ,Critical, Died based on our clustering data high percentage in well surgy,10% is stable ,5% is Critical, and finally in rear case they lead to died. In above surgery table to provide result about frequency, percentage valid percentage, cumulative percentage. They also category three field i.e., valid field, missing data , total value.

In missing data they have number of surgery , D.O.A, in hospital they missing data about patient .





5.4 Result and Analysis

To increase the clustering efficiency, data can first be preprocessed, thereby avoiding the processing of all of points individually. Object movement, deadlock detection and constraint satisfaction can be tested, where reduces the number of points to be computed. This methodology ensures the effective clustering can be performed in large data sets under the user-specified constraints with better efficiency and scalability.

5.4.1 Proportion of Selected Features

Selected features based on the data set we are collected i.e, patient record collection dataset in that they have name, age, disease ,gender in that based on the gender who are affected more frequently. They can produce result such as frequency table corresponding bar chart clearly identify the clustering. For example fact of depression in that they cluster the patient the age range in between 15 to 30 age people affect high amount of depression.

5.5 Discussion & Conclusion

A first advantages of the region of influence are insensitive to the order in which the pairs are considered. Also work well as most existing algorithm such as minimum spanning trees, The algorithm works well in many cases where the clusters are well separated. A problem may occur when a "large" edge *e* has another "large" edge as its neighbor. Second advantages is that customer tells the system what data to group, and then what constraint, without saying that which grouping method must be used. For example k-means algorithm. In future work, we plan to explore different types of constraint measures, and study some formal acreages of feature selection.

6. REFERENCES

[1] http://forum.jntuworld.com/ Data-Warehousing-and-Data-Mining

[2] Yu L. and Liu H., Feature selection for high-dimensional data: a fast correlation-based filter solution, in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856-863, 2003.

[3] Toward Integrating Feature SelectionAlgorithms for Classification and Clustering Huan Liu, Senior Member, IEEE, and Lei Yu, Student Member, IEEE

[4] Jaromczyk J.W. and Toussaint G.T., Relative Neighborhood Graphs and Their Relatives, In Proceedings of the IEEE, 80, pp 1502-1517, 1992

[5] Construction of Decision Tree : Attribute Selection Measures R. Aruna devi[1], Dr. K. Nirmala

[6] http://www.icsd.aegean.gr

[7] . Algorithm for computation of Relative Neighborhood Graph Electronics Letters 3rd july 1980 vol.16,no.14

[8] Good Approximation for the Relative Neighborhood Graph Diogo Vieirer Andrade , Luiz Henrique de Figueiredo

[9] Constraint Based Clustering in Large Database Anthony K.H.Tong, jiaweihan ,Laks V.S. Lakshmanan]

[10] A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data Qinbao Song, Jingjie Ni and Guangtao Wan