HIERARCHICAL CLUSTERING FOR CATEGORICAL AND NORMAL ATTRIBUTES

Raed Ibraheem Hamad

Abstract.

The amount of data kept in computer files and databases is growing at a phenomenal rate. At the same time the users of these data are expecting more sophisticated information from them. The problem of data mining or knowledge discovery has become increasingly important in recent years. There is an enormous wealth of information embedded in large data warehouses. Alternatively the data mining has been called exploratory data analysis, data driven discovery, and deductive learning. The clustering algorithm which is one of the data mining algorithms is useful technique for grouping data points such that points within a single group/cluster have similar characteristics

(or close each other), while points in different groups are dissimilar. We used clustering by using the links concepts to measure the similarity /proximity between a pair of data points. Where the Rock algorithm that employs links and not distances as in the traditional clustering algorithms that use distances to measure the similarity between points. The Rock generates better quality clusters and exhibits good scalability properties which can used to perform the insert and search operations very fast in addition to presenting the results of work, we also applied the Rock over variety database.

Key words: Data Mining, Knowledge Discovery

, Clustering Algorithms

1. *Introduction*

The problem of data mining or knowledge discovery has become increasingly important in recent years. There is an enormous wealth of

information embedded in the large data warehouses maintained by retailers [7,8,10]. Clustering in data mining is useful to discover distribution patterns in the underlying data. Clustering is a division of data into groups of similar objects. Each group called cluster consists of objects that are similar objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer faster necessarily loses certain fine details (akin to lossy data compression), but chives simplification, if represents many data objects by few clusters, and hence it models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. Clustering is unsupervised learning of a hidden data concept data mining deals with large databases that impose on clustering analysis additional severe computational requirements. These challenges led to the emergence of powerful broadly applicable data mining clustering methods. From a practical perspective clustering plays on outstanding rote in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, web analysis, CRM, Marketing medical diagnostics. Computational biology and any others clustering algorithms categorization of clustering algorithms in neither straightforward nor canonical. In this paper the classification closely for these Algorithms with main topic (Rock algorithm) and its development.

1.1 Plan of Further Presentation

Traditionally clustering techniques are broadly divided in hierarchical and partitioning [2,8,9].

Hierarchical clustering is farther subdivided into agglomerative and divisive. The basics of hierarchical clustering include lance Williams's formula, idea of conceptual clustering, now classic algorithms SLINK, COBWEB, as well as newer algorithms CURE and CHAMELON. While hierarchical algorithms build clusters gradually, partitioning algorithms earn clusters directly.

In doing so, they either try to discover clusters by iteratively relocating points between subsets or try to identify clusters as areas highly populated with data. Algorithms of the first kind are explained the partitioning Relocation methods. They are further categorizing into probabilistic clustering K-medoids methods (algorithms PAM, CLARA, CLARANS and its extension) and -means methods (different schemes, initialization, optimization harmonic means, extensions). Such methods concentrate on how well points fit into their clusters and tend to build clusters of proper convex shapes some algorithms word with data indirectly by constructing summaries of data over the attribute, space subsets. They perform space segmentation these are Grid-Based methods. They frequently use hierarchical agglomeration as on phase of processing.

2. Hierarchical

Hierarchical clustering algorithms actually create a set of clusters. Hierarchical algorithm differ in how the sets are created [4,5,8,9]. A tree data structure called dendrogram, can be used to illustrate the hierarchical clustering technique and the sets of different clusters. The root in a dendrogram tree contains one cluster where all elements are together. The leaves in the dendrogram each consist of a single element cluster. Internal nodes in the dendrogram represent new clusters formed by merging the clusters which appear as its children in the tree. Each level in the tree is associated with distance measure which was used to merge the clusters together. All clusters created at a particular level were combined because the children clusters had a distance between them less than the distance value associated with this level in the tree.

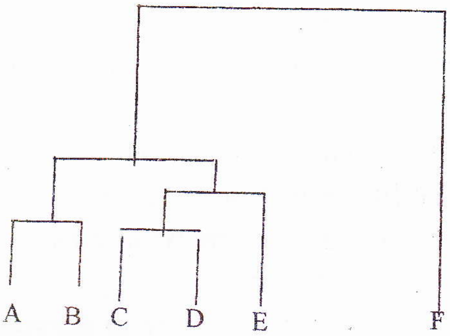
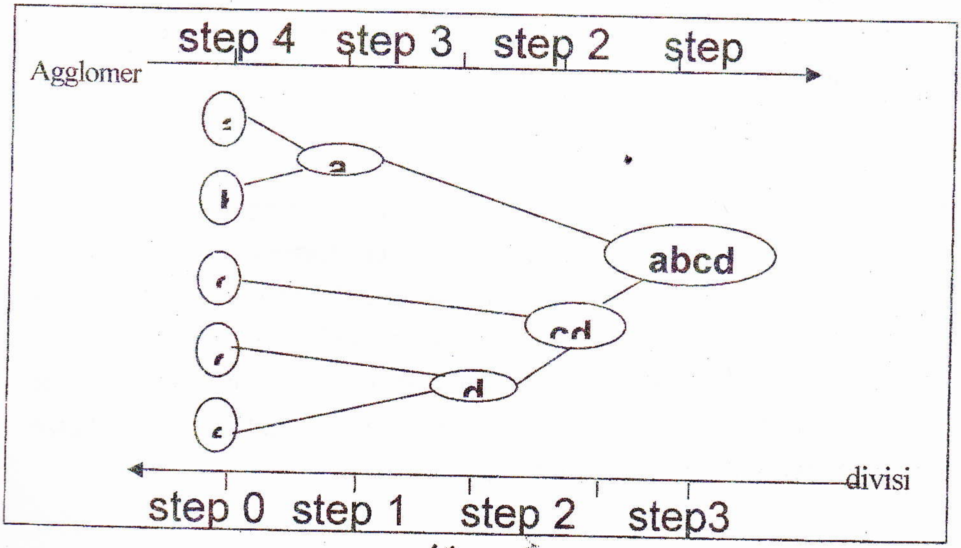


Figure (1) Dendrogram

outputting the dendrogram actually produces a set of clusters rather than just one clustering.

Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down). An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two one-point (singleton) clusters and recursively merges two or more most appropriate clusters.

A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved.



3. Clustering With Categorical Attributes

Traditional algorithms do not always work with categorical data example (3.1) illustrates some problem which exists when clustering categorical data. This example hierarchical based centroid algorithm to illustrate the problems. [3,4,10]. Example (3.1) consider an information retrieval system where documents can contain keywords {book, water, sun, sand, swim, read} suppose where are four documents where the first contains the words {book} the second contains {water, sun, sand, swim}, the third contains {water, sun, swim, read}, and the fourth {sand, read} using the following Boolean points (1, 0, 0, 0, 0, 0). (0, 1, 1, 1, 1, 0). (0 , 1, 1, 0, 1, 1). (0, 0, 0, 1, 0, 1) we can use the Euclidean distance to develop the following adjacency matrix of distances.

	1	2	3	4
1	0	2.24	2.24	1.73
2	2.24	0	1.41	2
3	2.24	1.41	0	1.73
4	1.73	2	1.73	0

Example (3.1) problems of categorical data

Categorical data the distance between point 2 and 3 is the smallest (1,41) and thus they are merged. When they are merged we get a cluster containing {(0, 1, 1, 1, 1, 0). (0, 1, 1, 0, 1, 1)} with a centroid of (0, 1, 1, 0.5, 1, 0.5). Notice that at this point we have a distance from this new cluster centroid to the original points 1 and 4 being 2.24 and 2 respectively while the distance between original points 1 and 4 is 1.73. Thus we next merge those points even though they have no keywords in common. So with K=2 we have the following clusters: {{1,4}. {2,3}}. Boolean attributes themselves are a special case of categorical attributes. The domain of categorical attributes is not limited to simply. True and false values, but could be any arbitrary finite set of values. An example of categorical attribute is color whose domain includes values such as brown, black, white etc, clustering in the presence of such categorical and normal attributes is the focus of this paper.

4. The Rock Clustering Algorithm

In this section described the Rock (RObust Clustering using linKs) clustering algorithm which belongs to the class of agglomerative hierarchical clustering algorithms. The steps involved in clustering using Rock are described in figure (4.1) after drawing a random sample from the database,

a hierarchical clustering algorithm that employs links is applied to the sampled points.

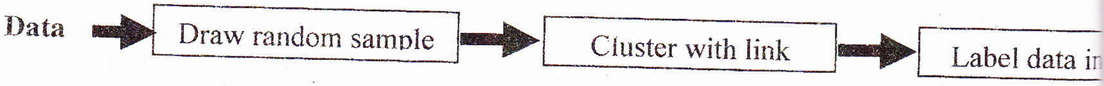


Figure (3) over view of Rock

Rock algorithm has some concepts:

1. Similarity measure between points.
2. User defined parameter 0.
3. Compare the similarity value with 0.
4. Compute the pair of points to be neighbors.
5. Compute the link for each point.

5. Clustering Paradigm

This section contain the new clustering model that is based on the notions of neighbors and links [2,5,9,10].

5.1 Neighbors

A point's neighbors are those points that are considerably similar to it. Let sim(pi, pj) be a similarity function that is normalized and capture the closeness between the pair of points pi and pj. We assume that sim assumes values between 0 and 1, with larger values indicating that the points are more similar given a threshold 0 between 0 and 1. Pair of points pi, pj are define to be neighbors it the following holds: $Sim(P_i, P_j) \ge \theta$ 1

In the above equation 0 is a user Defined parameter that can be used to control how close pair of points must be in order to be considering neighbors. The similarity between the two points (Transactions) Tl and T2 is compute by Jaccard coefficient.

$$Sim(T_1,T_2)=\frac{|T_1\cap T_2|}{|T_1\cup T_2|} \qquad \text{.........................}\ 2$$

Categorical data typically is of fixed dimension and is more structured than market basket data. However it is still possible that in certain records values may by missing for certain attributes. The proposal to handle categorical attributes with missing values by modeling each record with categorical attributes as a transaction.

Records

	F1	F2	F3	F4	F5	F6
R1	X	Y	Z	?	?	N
R2	X	Y	Z	H	O	?

Transactions

T_1	F1.X	F2.Y	F3.Z	F6.N	
T_2	F1.X	F2.Y	F3.Z	F4.H	F5.O

```
Procedure compute – neighbors
Begin
    For $I$= 1 to |Ds| do
    For $J$= $I$ +1 to |Ds| do {
    For each $X \in$ Ds ($i$) do
    For each $Y \in$ Ds ($j$) do{
    If ( $X.Atr = Y.Atr$ ) and ( $X.val = Y.val$) then
    $Trn = Trn + 1$
    $un = un + 1$
    else
    If $X.Atr = Y.Atr$ then
    $un = un + 2$
    }
    $sim = Trn / un$
    If $sim >= \theta$ then
    save the point to its nbrlist
    }
end
```

Algorithm for computing neighbors to categorical data

With normal data sets the proposed system deals with different dimensional records and also it less structure than categorical data there is no needing for modeling each record with categorical attributes as a transaction. "Normal" data are data that are drawn (come from) a population that has a normal distribution. This distribution is inarguably the most important and the most frequently used distribution in both the theory and application of statistics. If X is a normal random variable, the compare between each value in the first field from the first record with all values in the second record until the value found or the search into the record finished. This find the position of the value is not important as in categorical attributes. This algorithm illustrates the computing of neighbors for normal data. The similarity function proposed can then be used to compute the similarities between records with normal data rather than the corresponding transactions.

```
Procedure compute – neighbors
Begin
    For $I=1$ to $|Ds|$ do
    For $J= I+1$ to $|Ds|$ do {
    For each $X \in Ds(i)$ do
    For each $Y \in Ds(j)$ do{
    If $X = Y$ then
    $Trn = Trn + 1$
    }
    $un = sum(|Ds(i)|, |Ds(j)|) - Trn$
    $sim = Trn / un$
    If $sim >= \theta$ then
    save the point to its nbrlist
    }
End
```

Algorithm for computing neighbors to normal data

5.2 links

Clustering points based on only the closeness or similarity between them is not strong enough to distinguish two "not so well-separates" clusters because it is possible for points in different clusters to be neighbors.

We can define link (pi, pj) to be the number of common neighbors between pi and pj. From the definition of links, if follows that if link (pi, pj) is large, then if is more probable that pi and pj belong to the same cluster. For characterize the best clusters. If one could mathematically characterize the best clusters, then this would aid in the development of algorithms that attempt to find these good clusters, we present a criterion function. The best clusters are the ones that maximize the value of the criterion function. We would like to maximize the sum of link (p_q, p_r) for data point pairs (p_q, p_r) belonging to a single cluster and at the same time minimize the sum of link (p_q, p_r), for (p_q, p_r) in different clusters

this leads to the following criterion function that we would like to where cj denotes cluster i of size ni we use a measure similar to the criterion function in order to determine the best pair of clusters to merge at each step of Rock's hierarchical clustering (ci, cj) let link [ci, cj], store the number of cross links between clusters ci and cj that is link [ci, cj]. then we define the goodness measure g(ci,cj) for merging clusters(ci, cj) as follows.

$$g(C_i,C_j)=\frac{Link(C_i,C_j)}{(n_i+n_i)^{1+2f(\theta)}-n_i^{1+2f(\theta)}-n_i^{1+2f(\theta)}}$$

5.3 Rock Algorithm

Rock's hierarchical clustering accepts as input the set s of (n) sampled points be clustered, and the number of desired clusters K. The Procedure begins by computing the number of links between pairs of points in step. Initially each point is a separate cluster. For each cluster I, we build a local heal q[i]. And maintain the heal during the execution of the heal during the execution of the algorithm. q[i] contains every cluster such that link [I,j] is non-zero. The clusters in q[i] are ordered in the decreasing order of the goodness measure with respect to I, g(i,j). The algorithm also maintains on additional global heal Q that contains on additional global heal Q that contains all the clusters. Furthermore the clusters in Q are goodness measures. Thus g(j, max (q[j])) is used to order the various clusters j in Q where max (q[i]), the max element in q[j] is the best cluster to merge with cluster j. At each step the max cluster in Q and the max cluster in q[j] are the best pair of clusters to be merged. The while-loop in Step 5 iterates until only k clusters remain in the global heap Q. In addition, it also stops clustering if the number of links between every pair of the remaining clusters becomes zero. In each step of the while-loop, the max cluster u is extracted from Q by extract max and q[u] is used to determine the best cluster v for it. Since clusters u and v will be merged, entries for u and v are no longer required and can be deleted from Q. Clusters u and v are then merged
in Step 9 to create a cluster w containing (u + v) points.

```
procedure cluster(S, k)
begin
1. link := compute_ links(S)
2. for each s ∈ S do
    3. q[s] := build_ local_ heap(link, s)
4. Q := build _global _heap(S, q)
5. while size(Q) > k do {
    6. u := extract_ max(Q)
    7. v := max(q[u])
    8. delete (Q, v)
    9. w := merge(u, v)
    10. for each x ∈ q[u] U q[v] do {
```

```
link[x, w] := link[x, u] + link[x, v]
delete (q[x], u); delete(q[x], v)
insert (q[x], w, g(x, w)); insert(q[w], x, g(x, w))
update (Q, x, q[x])
}
insert (Q, w, q[w])
deallocate (q[u]); deallocate(q[v])
}
end
```

Clustering Algorithm

6 Concluding remarks

In this paper, we have presented a new approach to deals with normal data, and conclude the following:

1. We proposed a new concept of links to measure the similarity/proximity between a pair of data points with categorical and normal attributes.
2. Employs links and not distances for merging clusters are very encouraging
3. Methods naturally extend to non-metric similarity measures that are relevant in situations where a domain expert/similarity table is the only source of knowledge.
4. ROCK algorithm discovered almost pure clusters for each data base.
5. ROCK's performance to scale quite well for large databases.

REFERENCES

[1] R. Agrawal, K.-I. Lin, H.S. Sawhney, and K. Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In Proceedings of the VLDB Conference, Zurich, Switzerland, pp. 490-501 (1995).

[2] Anna Szymkowiak , Jan Larsen and Kars Hansen ," Hierarchical clustering for Data mining" , Technical university of Denmark http://www.eivind.imm.dtu.dk.html

[3] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. In Proceedings of the 19th Annual ACM Symposium on Theory of Computing (1987).

[4] T.H. Cormen, C.E. Leiserson, and R.L. Rivest. Introduction to Algorithms. The MIT Press, Massachusetts (1990).

[5] R.O. Duda and P. E. Hard. Pattern Classification and Scene Analysis. A Wiley-Interscience Publication, New York (1973).

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial database with noise. In International Conference on Knowledge Discovery in Databases and Data Mining (KDD-96), Portland, Oregon, pp. 226 -231, AAAI Press (1996).

[7] A. Hinneburg and D.A. Kein " An Efficient Approach to clustering in large multimedia Data bases with noise", In Proc. Int. Conf. Knowledge Discovery and Data mining , 1998.

[8] Margaret H. Dunham "Data miming Techniques and Algorithms" , Prentice Hall , 2000.

[9] Pavel Berkhim "Survey of clustering Data mining Techniques", Accrue Software , Inc, 2002.

[10] Sudipto Guha,Rajeev Rastogi, Kyuseok Shim,"Rock: A Robust Clustering Algorithm for categorical attributes", in proceedings of the 15th international conference on data engineering(2001).