International Journal of Innovation, Management and Technology, Vol. 1, No. 2, June 2010
ISSN: 2010-0248

Analysing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes

Parul Agarwal, M. Afshar Alam, Ranjit Biswas

Abstract—**Clustering ,an important technique of data mining, groups similar objects together and identifies the cluster to which each object of the domain being studied belongs to. In this paper we provide in depth explanation of implementation adopted for k-pragna, an agglomerative hierarchical clustering technique for categorical attributes.**

Index Terms—**bottom up hierarchical clustering,categorical attributes, implementation**

I. INTRODUCTION

Clustering a technique of data mining is gaining importance over the last few years. It discovers interesting patterns in the underlying data. It groups similar objects together in a cluster(or clusters) and dissimilar objects in other cluster(or clusters). In this paper we shall provide with the details of implementation of our algorithm for categorical attributes Most of the algorithms suggest the measure used for calculating the similarity but do not provide necessary information for its implementation or Data structures that have been used.

The remainder of this section is organized as follows. Section II has a discussion on related work. In Section III we present the implementation details of the algorithm.Experimental results and Analysis are presented in Section IV and Section V respectively.. We then present our future work in the last section.

II. RELATED WORK

Traditionally, clustering methods are broadly divided into hierarchical and partitioning.Hierarchical clustering is further subdivided into agglomerative and divisive. A hierarchical clustering is a nested sequence of partitions. This method works on both bottom up and top down approaches .The agglomerative hierarchical technique works on bottom up approach.The general approach of hierarchical clustering is in using an appropriate metric which measures distance between 2 tuples and a linkage criteria which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets The linkage criteria could be of 3 types single linkage ,average linkage and complete linkage.

In single linkage(also known as nearest neighbour), the distance between 2 clusters is computed as:
D(Ci,Cj)= min {D(a,b) : where a Є Ci, b Є Cj.

Thus distance between clusters is defined as the distance between the closest pair of objects, where only one object from each cluster is considered. i.e. the distance between two clusters is given by the value of the shortest link between the clusters.

In average Linkage method (or farthest neighbour), Distance between Clusters defined as the distance between the most distant pair of objects, one from each cluster is considered.

In the complete linkage method, ***D(Ci,Cj)*** is computed as
D(Ci,Cj) = Max { *d(a,b)* : a Є Ci,b Є Cj.}

the distance between two clusters is given by the value of the longest link between the clusters.
Whereas,in average linkage
D(Ci,Cj) = { *d(a,b)* / (l1 * l2): a Є Ci,b Є Cj. And l1 is the cardinality of cluster Ci,and l2 is cardinality of Cluster Cj.

And d(a,b) is the distance defined.}

Here the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group.

In agglomerative clustering,each object is represented by a single cluster. The single clusters are merged to make larger cluster and the process of merging continues until all the singular clusters are merged into a big cluster that consists of all the objects. The divisive method works on the top down approach .In this method all the objects are arranged within a big single cluster and the large cluster is continuously divided into smaller clusters until each cluster consists of a single object.

The basics of hierarchical clustering include Lance-Williams formula, conceptual clustering,SLINK[1], COBWEB[2] as well as newer algorithms like CURE[3] and CHAMELEON[4]. SLINK is a single-link hierarchical algorithm that can be carried out using arbitrary dissimilarity coefficients. In this algorithm a pointer representation , a compact form of dendogram has been introduced.In COBWEB,a clasiification tree that characterizes each cluster with a probabilistic distribution has been suggested..CURE(Clustering using Representatives) an algorithm that handles large databases employs a combination of random sampling and partitioning.A random sample is drawn from the data set and then partitioned and each partition is partially clustered. The partial clusters are then clustered in a second pass to yield the desired clusters CURE is effectively able to handle outliers..CHAMELEON adopts a two phase approach wherein a graph partitioning algorithm divides a dataset into a set of individual clusters.In the second phase an agglomerative hierarchical mining algorithm merges the clusters.

Whereas,in the partitioning method, a partitioning algorithm arranges all the objects into various groups or partitions,, where the total number of partitions(k) is less than

the number of objects(n).i.e. a database of n objects can be arranged into k partitions ,where k < n.Each of the partition thus obtained by applying some similarity function is a cluster. The partitioning methods are subdivided as probabilistic clustering[5] (EM ,AUTOCLASS), algorithms that use the k-medoids method (like PAM[6], CLARA[6], CLARANS[7]), and k-means methods (differ on parameters like initialization, optimization and extensions).The EM algorithm is based on expectation-maximization (EM) algorithm and it is used finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. AUTOCLASS, is a powerful unsupervised Bayesian classification system which mainly has application in biological sciences. It is able to handle the missing values.

The algorithm PAM(Partitioning around medoids) first computes *k* representative objects, called *medoids*. A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal.Its main features include a graphical display silhouette plot,robust and computation of dissimilarity matrix.

CLARA (Clustering Large Applications) deals with large data sets. It also tries to find *k* representative objects that are centrally located in the cluster.It considers data subsets of fixed size, so that the overall computation time and storage requirements become linear in the total number of objects.CLARANS identifies spatial structures present in the data.

Partitioning algorithms are also denity based i.e. try to discover dense connected components of data, which are flexible in terms of their shape.Several algorithms like DBSCAN[8], OPTICS have been proposed.. The advantage is that they are less sensitive to outliers and can discover clusters of irregular shapes. They usually work with low-dimensional data of numerical attributes,known as spatial data. Some algorithms construct summaries of data over the attribute space subsets and perform space segmentation and then aggregate appropriate segments. These are the grid based methods like BANG[9].There also exist evolutionary methods such as Simulated Annealing,Genetic Algorithms[10].Several scalabitlity algorithms e.g. BIRCH[11], DIGNET[12] have been suggested in the recent past to address the issues associated with large databases .There also exist the category of algorithms that consider the high dimensional data and work on Subspace Clustering, Projection Techniques, Co-Clustering Techniques. High dimensional data may go upto a a thousand of dimensions.Examples of such dataspaces generally occur in the field of medicine Multiple dimensions pose problems in their visualisation and enumeration due to their innumerable values that they take. Subspace clustering [13]is the task of detecting *all* clusters in *all subspaces*. This means that a point might be a member of multiple clusters, each existing in a different subspace. Subspaces can either be axis-parallel or affine. The problem with subspace clustering is ,that with d dimensions there exist 2^d subspaces.Projected clustering[14] assigns each point to a unique cluster, but clusters may exist in different subspaces. The general approach is to use a special distance function together with a regular clustering algorithm.Co-Clustering or Bi-Clustering[15] is simulataneous clustering of rows and columns of a matrix i.e. of tuples and attributes.

The techniques of grouping the objects are different for numerical and categorical data owing to their separate nature. The real world databases contain both numerical and categorical data.Thus, we need separate similarity measures for both types. The numerical data is generally grouped on the basis of the inherent geometric properties like distances(most common being Euclidean, Manhattan etc) between them. Whereas for categorical data the attribute values that they take is small in number and secondly, it is difficult to measure their similarity on the basis of the distance as we can for real numbers. There exist two approaches for handling mixed type of attributes. Firstly, group all the same type of variables in a particular cluster and perform separate dissimilarity computing method for each variable type cluster. Second approach is to group all the variables of different types into a single cluster using dissimilarity matrix and make a set of common scale variables. Then using the dissimilarity formula for such cases, we perform the clustering.

There exist several clustering algorithms for numerical datasets. The most common being K-means, BIRCH, CURE, CHAMELE ON. The k-means algorithm fixes k clusters apriori. It randomly selects k points as centres and then assigng the objects to be clustered to belong to these on the basis of the distance.It then recomputes the k centres.It continues the process till the centres don't move.K-means was further proposed as fuzzy k-means and also for categorical attributes.The original work has been explored by several authors for extension and several algorithms for the same have been proposed in the recent past. In[16], Ralambondrainy proposed an extended version of the k-means algorithm which converts categorical attributes into binary ones. In this paper the author represents every attributes in the form of binary values which results in increased time and space incase the number of categorical attributes is large

A few algorithms have been proposed in the last few years which cluster categorical data. A few of them listed in[17-19]. Recently work has been done to define a good distance (dissimilarity) measure between categorical data objects[20-21].For mixed data types a few algorithms[22-24] have been written. The author presents k-modes and k-prototypes which cater to categorical data and mixed data types. In the k-modes algorithm the author considers the number of mismatches between categorical attributes as the measure for performing clustering. In k-prototypes , the distance measure for numerical data is weighted sum of Euclidean distances and for categorical data ,a measure has been proposed in the paper.

III. IMPLEMEMENTATION DETAILS

In most of the papers ,we do not find the platform,tools used and implementation details .In this paper ,we provide with all the implementation details of the algorithm introduced in [25]

IACSIT International Association of Computer Science and Information Technology WWW.IACSIT.ORG

A. Platform used:

C language.

When the algorithm will be tested for real data containing thousand of tuples ,we can implement the same using Matlab Toolbox[26].The drawback with C Language is that the similarity matrix computation requires a lot of memory space and Languages like C/C++ do not provide this space.But we can overcome it by performing sequential partitioning of database or by randomly sampling it.The formulas used for calculation are described in [25].

B. Pre Processing Steps:

Remove tulpes with missing values .

C. Data Structures Used:

1) A domain array Dom[m][n] where m: No. of tuples and n: No. of Attributes.This array is initialized to the values of the data text file.
2) A Similarity matrix which stores the similarity between tuples or clusters ,an array sim[m][m].
3) A one dimensional array Cluster[m] where its values are inititialized ranging from 0 to m-1.The values of this array change when logical merge has been performed.

We initialize values as cluster[0]=0 and so on.

Either we can keep an array which records each merge or a linked list.Since ,we will be elaborating via a small real dataset,we explain by using an array.

D. Main Modules Defned:

1) Main ()
2) Calculate_sim() //calculates the similarity
3) Find_large_merge() // finds the largest value in similarity matrix using just the upper triangular matrix of sim[m][m] and find the largest value in it and merge logically. Merging results in no change in the sim matrix but the values of cluster array change.
4) Digit(int posn,int num)// identifies the digit at ith place in a number
5) length(int cluster[m]) .//finds the length of cluster[m]
6) cal_lines() // calculates the number of tuples
7) Cal_attr() //calculates the number of attributes.
8) display() // displays the contents of cluster [m].

E. Program structure defined:

1. Main Function:

1.a: calculate the number of lines(m) from text file.

1.b calculate the number of attributes(n).

1.c For n-k times ,where k is the number of clusters desired perform the following:

1.c.a Cal_sim()

1.c.b find_large_merge()

1.c.c Display the contents of each cluster.

We explain here 4 modules which are length(int mun),digit(int posn,int num), cal_sim() and find_large_merge()

```
//the module to find the length of a number
Length(int num)
{
if (num==0)
return 1;
while(num!=0)
{num=num/10;
++i;}
return i;}
```

```
//the module to find digit specified by posn of a number.
digit(int posn,int num)
{
len=length(num);
while(len!=posn)
{
num=num/10;
len--;
}
while(len!=1)
{
a=pow(10,(len-1));
num=num%a;
l--;
}
return num;
}
```

```
//the module for calculating similarity
Cal_sim()
{
//For the first time calculation of similarity flag =0
For i=0 to m-1
For j=i+1 to m-1
For k= 0 to n-1
{
If dom[i][k]==dom[j][k] then
sim[i][j]=sim[i][j] +0.5
} set flag=1
//for subsequent similarity
For i= 0 to m-1
For j=i+1 to m-1
{
If ((cluster[i] && cluster[j]) != 0)
{l1=length(cluster[i];
l2=length(cluster[j];
for ii = 1 to l1{
i1=digit(ii,cluster[i]);
for j=1 to l2{
j1=digit(jj,cluster[j]);
for k=0 to n-1
if(dom[i1][k]==dom[j1][k])

sim[i][j]=sim[i][j]+1/(l1+l2);
}}
```

We would like to elaborate on the logical merging.

Assume, cluster 0,cluster 2,cluster 3 have to merged.and 1,4th cluster have already been merged then
cluster[0]=320
cluster [1]=14
cluster[2]=0
cluster[3]=0
cluster[4]=0

find_large_merge() consists of the following
{//identify the largest element in upper triangular matrix
//i.e. if maxi,maxj correspond to the position of (largest value //found in upper triangular matrix) element in sim[m][m]

l=length(cluster[maxi]);
X=pow(10,l);

Table I

Cluster no.	1	2	3	4
0	7.50	6.5	9.5	7.0
1		8.5	6.5	9.5
2			7.0	8.0
3				8.0

Table II

Cluster no.	1	2	3	4
0	13.66662	13.66662	0	13.6662
1		7.5	0	7.5
2			0	6.5

```
if(maxi==0)
{
cluster[maxi]=maxj*X + cluster[maxi];
cluster[maxj]=0;
}
else
 {
cluster[maxi]=cluster[maxi]*X + cluster[maxj]);
cluster[maxj]=0;
}
```

IV. Experimental Results

A. Real Data Set:

1. Mushroom Dataset:

The mushroom dataset has 8124 tuples and 22 attributes..Each tuple has been labeled as poisonous or edible in the dataset provide for our identification .The datset has been taken from UCI Machine Learning Repository.[27] To elaborate our implementation we are providing here each step of our algorithm with 5 tuples and all attributes considered. Thus n=5,k=3 no.of clusters deired.

The loop would thus execute n-k times i.e. 2 times.
The first similarity matrix is provided in Table I
As seen in Table I, the highest similarity has been found between cluster 0 and cluster 3 with a value 9.5.Though,its clear that cluster 1 and cluster 4 also have 9.5 as the similarity value but in case of a tie, the (i,j)pair with the lowest i and lowest j would be picked up for merging.So we merge 0,3.
The contents of cluster array are as follows:

cluster[0]=30
cluster[1=1
cluster[2]=2
cluster[3]=0
cluster[4]=4.

The next time the loop executes we have the similarity matrix as shown in Table II.
From Table I the highest value is between cluster 0 and 1
Which implies that the contents of cluster array are as follows:
cluster[0]=130
cluster[1]=0
cluster[2]=2
cluster[3]=0
cluster[4]=4
our loop now terminates with the following results:
cluster 0,1,3 merged.
cluster2
cluster 4

The mushroom dataset as described has labels which identify the tuple's class.
We had taken the first 5 tuples for calculation .
In the text file it was found that actually cluster 0,cluster3 are poisonous and cluster 1,2 4 were edible.
K=3

Table I

Cluster 1:No of poisonous: 2
No. of edible: 1
Cluster 2: no .of poisonous :0
No. of edible:1
Cluster 3: no. of poisonous: 0
No. of edible:1
Thus, accuracy is approximately 0.95 .

V. ANALYSIS

1) If there are n tuples in the dataset, then the similarity matrix can be computed in $O(n^2)$.
2) Before the identification of which pair of clusters have to be merged at the kth merge step, there are n-k rows of the similarity matrix with n+1-k clusters remaining. At the kth stage, the search involves n-k-1 comparisons. So, total number of comparisons for n-1 stages is $O(n^2)$.
3) At the kth stage, once the most similar pair has been idendified,we need to update the n-k matrix entries corresponding to the cluster resulting from the merger. For n-1 stages, the number of comparisons and the updates is again of the order of $O(n^2)$.
4) Finally it is necessary to search for a new row maximum

IACSIT International Association of Computer Science and Information Technology WWW.IACSIT.ORG

in any row where the cluster corresponding to the maximum is involved in the merger. In a typical case it would be expected that one row would be updated at each stage resulting in $O(n^2)$.

5) But since the loop executes for n-k times so we have the complexity as $O(n^3)$

The hierarchical clustering algorithm have complexity which is $O(n^3)$

VI. Future Work

We will enhance this algorithm to work for both numerical and categorical data. This would increase its efficiency and usage multifold..The furture works of this algorithm would also involve testing it on several more real datasets with thousands of tuples and comparing the results with existing algorithms.It may also include if needed the modification of formula for calculating similarity for categorical data for performance analysis.The real datasets come with noise and outliers.We shall try to remove in the algorithm that we had proposed

References

[1] SLINK: An optimally efficient algorithm for the single link cluster method. The Computer Journal, 16(1):30–34.
[2] Fisher, Douglas H. (1987). "Knowledge acquisition via incremental conceptual clustering". Machine Learning 2: 139–172
[3] Sudipto Guha,Rajeev rastogi,Kyuseok Shim:"CURE": A Clustering Algorithm for large Databases. Proc. of the ACM_sigmod Int'l Conf. .
[4] Clustering Algorithm using dynamic Modelling.IEEE Computer ,/*Vol. 32.No. 8,68-75,1999.
[5] T. Mitchell, Machine Learning, McGraw Hill, 1997.
[6] Kaufman, L. and Rousseeuw, P. (1990). Finding Groups in Data—An Introduction to Cluster Analysis. Wiley Series in Probability and Mathematical Statistics. NewYork: JohnWiley& Sons, Inc.
[7] Ng, R. and Han, J. (1994). Efficient and effective clustering methods for spatial data mining.In Proceedings of the 20th international conference on very large data bases, Santiago,Chile, pages 144–155. Los Altos, CA: Morgan Kaufmann.
[8] Ester, M., Kriegel, H., Sander, J., and Xu, X. (1996). Adensity-based algorithm for discovering clusters in large spatial databases with noise. In Second international conference on knowledge discovery and data mining, pages 226–231. Portland, OR: AAAI Press.
[9] Schikuta, E. and Erhart, M. (1997). The BANG-clustering system: Grid-based data analysis.In Liu, X., Cohen, P., and Berthold, M., editors, Lecture Notes in Computer Science, volume 1280, pages 513–524. Berlin, Heidelberg: Springer-Verlag.
[10] Goldberg, D. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning.Reading, MA: Addison-Wesley.
[11] T.Zhang,R.Ramakishnan,M.livny:" BIRCH:An Efficient data Clustering method for very large databases.Proc. of the ACM_sigmod International Conference on Management of data,1996,pp.103-114
[12] R. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In VLDB-94, 1994.
[13] Kriegel, Hans-Peter; Kröger, Peer; Renz, Matthias; Wurst, Sebastian (2005), "A Generic Framework for Efficient Subspace Clustering of High-Dimensional Data", Proceedings of the Fifth IEEE International Conference on Data Mining (Washington, DC: IEEE Computer Society): 205–25
[14] Aggarwal, Charu C.; Wolf, Joel L.; Yu, Philip S.; Procopiuc, Cecilia; Park, Jong Soo (1999), "Fast algorithms for projected clustering", ACM SIGMOD Record (New York, NY: ACM) 28 (2): 61–72,
[15] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," IEEE/ACM Trans. Computat. Biol. Bioinformatics, vol. 1, no. 1, pp. 24–45, Jan. 2004.
[16] H.Ralambondrainy,"Aconceptual versionof the kmeans al gorithm"
[17] 4rnRecogn. Lett., Vol. 15, No. 11, pp. 1147-1157, 1995.
[18] Sudipto Guha,Rajeev rastogi,Kyuseok Shim:"ROCK":A robust clustering algorithm for categorical attributes".In Proc. 1999 International. Conference of . data Engineering,pp.512-521,Sydney,Australia,Mar.1999.
[19] Yi Zhang,Ada Wai-chee Fu,Chun Hing Cai,peng-Ann Heng:"Clustering Categorical Data.In Proc. 2000 IEEE Int. Conf. Data Engineering,San Diego,USA,March 2000.
[20] David Gibson,Jon Klieberg, Prabhakar Raghavan,"Clustering Categorical Data :An Approach based on Dynamic Syatems.Proc. 1998. Int. Conf. On Very Large Databases,pp. 311-323,New York,August 1998.
[21] U.R. Palmer,C.Faloutsos.Electricity based External Similarity of Categorical Attributes.In Proc. Of the 7th Pacific Asia Conference on Advances in Knowledge Discovery and Data Mining PAKDD'03,pp. 486-500,2003.
[22] Z. Huang, "Extensions to the k-means algorithm for clustering large datasets with categorical values", Data Mining Knowl. Discov., Vol. 2, No. 2
[23] C.Li,G. Biswas. Unsupervised learning with mixed Numeric and nominal data.IEEETransactions on Knowledge and Data Engineering,2002,14(4):673-690.
[24] G. Lee,D-K. Yun.Clustering categorical and Numerical data: A New procedure using Multi Dimensional scaling.International Journal of Information Technology and Decision Making, 2003,2(1) :135-160.
[25] pagarwal,aalam,rbiswas, a Hierarchical Clustering algorithm for Categorical Attributes, in the Proc. Of International conference of knowledge discovery ICKD 2010,March ,2010,Bali,Indonesia ISBN :978-0-7695-3982-9 Vol. 2,pp 365-369
[26] MATLAB. User's Guide. The Math-Works, Inc., Natick, MA 01760, 1994-2001.http://www.mathworks.com/access/helpdesk/help/techdoc/matlab.shtml
[27] P. M. Murphy and D. W. Aha. UCI repository of machine learning databases, 1992. www.ics.uci.edu/_mlearn/MLRepository.html.