Divisive Hierarchical Clustering of Drugs Based on Chemical Compositions

Zin Mar Wai, Kalyar Win
Computer University (Taung Ngu)
zzzzzzin@gmail.com, dklw2009@gmail.com

ABSTRACT

Clustering is the process of grouping the data into classes or clusters. Objects within a class have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. In this paper we intend to cluster drugs based on their chemical composition so that users can know which drug on which cluster is composed of what chemicals by which composition. We will implement this system by using a hierarchical divisive monothetic clustering, called DIVICLUS_T. It allows becoming a decision tree of the hierarchy. This paper gives the valuable information of drugs for drug- researchers.

Keywords: Data mining, Clustering, Divisive, DIVCLUS_T

1. INTRODUCTION

The amount of data maintained in an electronic format has seen a dramatic increase in recent times. The amount of information doubles every 20 months, and the number of databases is increasing at an even greater rate. The search to determine significant relationships among variables in the data has become a slow and subjective process. As a possible solution to this problem, the concept of *Knowledge Discovery in Databases – KDD* has emerged. The process of the formation of significant models and assessment within KDD is referred to as data mining. Data mining is used to uncover hidden or unknown information that is not apparent, but potentially useful [1].

Clustering is the process of grouping the data into classes or clusters so that objects with a cluster have high similarity in comparison to another, but are very dissimilar to objects in other clusters.

There are various methods for clustering. A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical clustering method works by grouping data objects into a tree of clusters. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

The agglomerative hierarchical clustering is a bottom-up strategy that starts by placing each object in its own cluster. Then it merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied.

The divisive hierarchical clustering reverses the process of agglomerative hierarchical clustering, by starting with all objects in one cluster, and successively dividing each cluster into smaller ones.

In this paper, divisive hierarchical clustering is used for clustering of drugs by their chemical composition. DIVCLUS_T algorithm is used for this task. It provides bipartitions for each step until the desired number of clusters.

The paper is organized as follows. In the next section, Section 2, we will describe related work of Divisive Hierarchical Clustering. Section 3 presents detail explanation of Divisive Hierarchical Clustering and DIVCLUS_T algorithm. In section 4, we will present propose framework. Implementation of the proposed framework is discussed in section 5.Conclusion will be discussed in section 6.

2. RELATED WORKS

S.K.Tasoulis and D.K.Tasoulis proposed an improvement of the Principal Direction Divisive Partitioning algorithm. The proposed algorithm merges concepts from density estimation and projection-based methods towards a fast and efficient clustering algorithm, capable of dealing with high dimensional data. Experimental results showed improved partitioning performance compared to other popular methods. Moreover, they explored the problem of automatically determining the number of clusters that is central in cluster analysis [6].

N.M.Reddy, N.V.Ramana and K.R.Reddy presented a new methodology for the most complex unit commitment problem using agglomerative and divisive hierarchical clustering. Euclidean costs, which is a measure of difference in fuel cost and start-up costs of any two units are first calculated. Depending upon the value of Euclidean costs, similar type of units are placed in a cluster. This cost is also useful for preparing priority lists for the units in a cluster and forming the different clusters. They used two individual algorithms. While the load

is increasing, agglomerative cluster algorithm is proposed. Divisive cluster algorithm is used when the load is decreasing [4].

P.P.Rodrigues, J.Gama and J.P.Pedroso presented a time series whole clustering system that incrementally constructs a tree-like hierarchy of clusters, using a top-down strategy. The Online Divisive-Agglomerative Clustering (ODAC) system used a correlation-based dissimilarity measure between time series over a data stream and possesses an agglomerative phase to enhance a dynamic behavior capable of concept drift detection. Main features include splitting and agglomerative criteria based on the diameters of existing clusters and supported by a significance level. Only the leaves are updated, reducing computation of unneeded dissimilarities and speeding up the process every time thestructure grows [5].

3. CLUSTER ANALYSIS

Cluster analysis is a multivariate analysis technique where individuals with similar characteristics are determined and classified (grouped) accordingly. Through cluster analysis, dense and sparse region can be determined in the distribution, and different distribution patterns may be achieved.

The degree of similarity or dissimilarity may be determined from the recorded values for one or multiple characteristics for each individual in clusters. There are no dependent variables for cluster analysis.

Clustering procedures require that similarity be quantified. One quantitative measure for interval scale data is the distance between cases. Other measures may be used to create dissimilarity or distance matrix that can be used as the basis for creating clusters [3].

3.1 Hierarchical Clustering

Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone. This rigidity is useful in that it leads to smaller computation costs by not having to worry about a combinatorial number of different choices. However, such techniques cannot correct erroneous decisions. There are two approaches to improving the quality of hierarchical clustering: (1) perform careful analysis of object "linkages" at each hierarchical partitioning, such as in Chameleon, or (2) integrate hierarchical agglomeration and other approaches by first using a hierarchical agglomerative algorithm to group objects into microclusters, and then performing macroclustering on the microclusters using other clustering method such as iterative relocation, as in BIRCH [1].

3.2 Divisive Clustering

In divisive clustering, some methods are polythetic, whereas some others are monothetic. A cluster is called monothetic if a conjunction of logical properties, each one relating to a single variable, is both necessary and sufficient for membership in the cluster. A clustering method which builds, by construction, monothetic clusters is then monothetic. In divisive clustering, monothetic clusters are obtained by using, for each division, a single variable and by separating objects having specific variable values from those who do not. Monothetic divisive clustering methods are usually variants of the association analysis method and are designed for binary data.

In this system, monothetic divisive clustering is used [3].

3.3 Inertia Criteria

A general approach for splitting a set = {Ω,…, i,…,n } of n objects into k disjoint clusters involves defining a measure of adequacy of a partition P_k and seeking a partition which optimizes that measure. In this system, the inertia criterion is used.

$$\mathbf{Z} = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} 1 \cdots j \cdots p \\ \left[\begin{array}{c} \cdot \\ \vdots \\ \cdots\ z_i^j\ \cdots \\ \vdots \\ \cdot \end{array}\right. \end{array}, \quad \mathbf{w} = \left[\begin{array}{c} w_1 \\ \vdots \\ w_i \\ \vdots \\ w_n \end{array}\right]$$

In the calculation of the inertia criterion, an object i ∈ will be weighted by w_i and identified with the corresponding row of the matrix Z.

Ω

$$\mathbf{z}_i = \left(z_i^1 \cdots z_i^p\right)^t$$

The inertia of a cluster $C_\ell \subseteq$ is then defined by

$$I(C_\ell) = \sum_{i \in C_\ell} w_i\ d^2_{\mathbf{M}}(\mathbf{z}_i, \mathbf{g}(C_\ell)),$$

where w_i is the weight of the object i and $g(C_\ell)$ is the cluster centroid defined by:

$$\mathbf{g}(C_\ell) = \frac{1}{\sum_{i \in C_\ell} w_i} \sum_{i \in C_\ell} w_i \mathbf{z}_i$$

The distance d_M between the two vectors z_i and $z_{i'}$ of R_p is defined by

$$d^2_{\mathbf{M}}(\mathbf{z}_i, \mathbf{z}_{i'}) = (\mathbf{z}_i - \mathbf{z}_{i'})^t \mathbf{M} (\mathbf{z}_i - \mathbf{z}_{i'})$$

where M is a p × p positive definite matrix.
The sum of the inertias of all clusters is called the within-cluster inertia

$$W(P_k) = \sum_{\ell=1}^{k} I(C_\ell)$$

It is an heterogeneity criterion for the adequacy of a partition $P_k = (C_1, ...,C_k)$. Similarly, the inertia of the centroids $g(C_\ell)$, weighted by $\mu(C_\ell)$is called the between-cluster inertia

$$B(P_k) = \sum_{\ell=1}^{k} \mu(C_\ell)\, d^2_{\mathrm{M}}(\mathrm{g}(C_\ell), \mathrm{g})$$

Where,

$$\mu(C_\ell) = \sum_{i \in C_\ell} w_i$$

g = g(Ω), This is an isolation criterion for the adequacy of P_k.
Finally, because the total inertia of a set of Rp points can be partitioned into within and between-cluster inertia, the following formula is obtained.

$$I(\Omega) = W(P_k) + B(P_k)$$

and so minimizing W (the heterogeneity) is equivalent to maximizing B (the isolation).

3.3.1 Inertia Criteria for Numerical Data

For a numerical matrix X, the inertia criterion is calculated from the weighted data matrix (Z,w) with Z = X and w=m. Moreover, the matrix M used in the quadratic distance d_M is usually the identity matrix I or the diagonal matrix of the inverse of squared standard deviations: This latter distance is used when the variables are measured on very different scales. In this system, we will use the identity matrix for the matrix M.

3.4 DIVCLUS_T

DIVCLUS-T is a divisive hierarchical clustering algorithm based on a monothetic bipartitional approach allowing the diagram of the hierarchy to be read as a decision tree. it simultaneously provides partitions into homogeneous clusters anda simple interpretation of those clusters. So the output of the algorithm is a CLUstering-Tree. It provides a simple and natural interpretation of the clusters. It is based on the minimization of the inertia criterion. The bipartitional algorithm and the choice of the cluster to be split are based on the minimization of the within-cluster inertia or maximization of between-cluster inertia.

In the divisive hierarchical clustering algorithm, one recursively splits a cluster into two sub-clusters, starting from the set of objects $\Omega = \{1,.....,n\}$: given the current partition $P_k = (C_1,.....,C_k)$, one cluster C_l is split in order to find a partition P_{k+1} which contains k+1 clusters and optimizes the chosen adequacy measure, based on the inertia criterion (maximum between cluster inertia).

The DIVCLUS-T algorithm consists of two stages:

(1) Splits a cluster C_l into a bipartition (A_l, $\hat{A}_l$) of maximum between- cluster inertia.

(2) Chooses in the partition Pk the cluster Cl to be split in such a way that the new partition Pk+1 has maximum between- cluster inertia.

3.4.1 DIVCLUS_T Algorithm

At first stage, the DIVCLUS_T:

1. Create p(n-1) binary questions from the cluster, C_l.
2. Splits the cluster C_l into a bipartition (A_l, $\bar{A}_l$) for each binary question.
3. Calculate the centroids for the clusters A_l and $\bar{A}_l$.
4. Calculate the distance between two centroids of the clusters A_l and $\bar{A}_l$.
5. Calculate the between-cluster inertia for the clusters A_l and $\bar{A}_l$.
6. Repeat step 2 to 5 for all binary questions.
7. Choose maximum between-cluster inertia.
8. Split the cluster according to the binary question which cause the maximum between-cluster inertia.
9. At second stage, the DIVCLUS_T:
10. For the current partition $P_k = (C_1,..,C_k)$
11. Calculate the maximum between-cluster inertia for each cluster.
12. Choose the maximum of the k maximum between-cluster inertia for all clusters.
13. Choose that cluster to be split to get the partition P_{k+1}.

4. PROPOSED DESIGN

The proposed design is shown in figure ɔ. Firstly, the system takes raw data of drugs, chemical compositions as input. These chemical compositions are transformed in percentage form, and form the data table which has rows of n objects of drugs name and columns of p chemical variables.

Rows (objects) are represented by drug names and columns (variables) are represented by chemical composition of drugs and the name of symptoms they can be cured. Then, inertia criteria is used to split n objects into k disjoint clusters which optimize that measure.

To get the hierarchical clusters of drugs, a divisive clustering algorithm called DIVCLUS-T is used. It consists of two steps: Bipartition and choosing a cluster to be split. These two steps are repeated until a condition is satisfied. Finally, the divisive clustering tree is obtained as an output. This tree will represent the clusters of drugs which have

same properties to cure some symptoms of diseases.

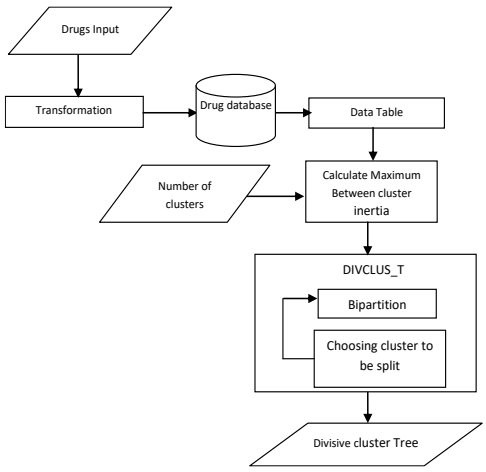

Figure 1. Proposed Design

5. IMPLEMENTATION

In the proposed system, we intend to cluster drugs based on their chemical composition so that users can know which drug on which cluster is composed of what chemicals by which composition.

In order to split optimally a cluster C_ℓ one has to choose the bipartition (A_l, $\bar{A}_l$) amongst the $2^{n_\ell-1-1}$ possible bipartitions of this cluster of n_ℓ objects.

5.1 Calculating Inertia of a Bipartition

Let (A_l, $\bar{A}_l$) be a bipartition of a cluster C_ℓ of with described by a numerical weighted data table (Z,w). Minimizing the within-cluster inertia W(A_l, $\bar{A}_l$) is equivalent to maximizing the between cluster inertia B(A_l, $\bar{A}_l$). Moreover we know that B(A_l, $\bar{A}_l$) can be written as a weighted distance between the centroids g(A_ℓ) and g($\bar{A}_l$):

$$\{c = \frac{x^j_{(i)} + x^j_{(i+1)}}{2},\ \ x^j_{(i)} \neq x^j_{(i+1)},\ \ i = 1, ..., n-1\}$$

5.2 Creating Binary Questions

The binary questions are formulated in terms of the initial data matrix X. A binary question Q on a numerical variable X_j is given by:

Is $X_j \leq c$?

This binary question, also denoted by Q = [$X_j \leq$ c], splits a cluster C_ℓ into two sub-clusters A_ℓ and $\bar{A}_l$ such that A_ℓ = {i ∈ C_ℓ | ≤ c} and $\bar{A}_l$= {i ∈ C_ℓ > c}.

The values c are defined as the midpoints between two consecutive observations:

$$B(A_\ell, \bar{A}_\ell) = \frac{\mu(A_\ell)\mu(\bar{A}_\ell)}{\mu(A_\ell) + \mu(\bar{A}_\ell)}\ d^2_{\mathbf{M}}(\mathbf{g}(A_\ell), \mathbf{g}(\bar{A}_\ell))$$

Thus there will be a maximum of $n_\ell - 1$ different bipartitions induced by the binary questions on X_j . The data flow diagram for calculating binary questions is shown in figure 3.

In the proposed system, the users can set the number of clusters as he desired. After the clusters have been formed, the user can analyze each cluster. The user can see which drugs are included in each cluster, and can check the binary questions that used in the formation of clusters. These are shown in figure 4 and 5.

In this experiment, the number of clusters is set as ๑. So the system chooses binary questions ๒ times and create ๑ clusters of drugs.

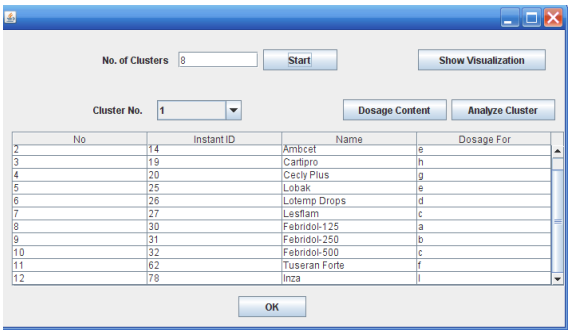

Figure 2. Cluseting of Drugs with no: of Clusters

Now, the cluster number 7 is chosen to see its general properties. The cluster 7 consists of 5 binary questions as follows:

folic acid > 0.5

Dextromethorphan ≤ 5.0

B6 ≤ 0.5

Pioglitazone ≤ 7.5

Paracetamol ≤ 100

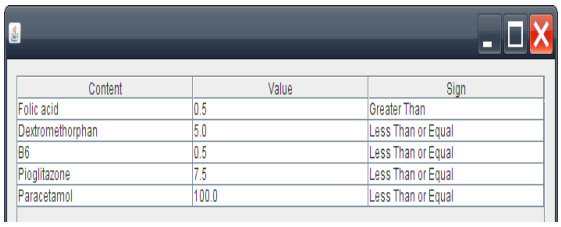

Figure 3. Cluster Properties

In figure 4, the distribution of dosage box for cluster 7 is shown. This box means that the current cluster contains what kinds of drugs. This box doesn't express specific disease, but can express the related diseases cured by the drugs contained in the respective cluster. In this experiment, cluster 7

contains drugs for diabete, liver and heart related diseases.

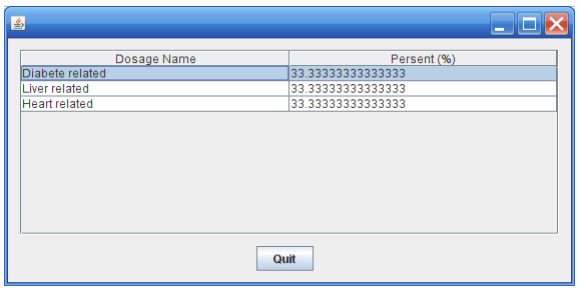

Figure 4. Dosage of Drugs

In this system, the user can see the result of clusters as a decision tree. In this experiment, we set number of clusters as 8. So we obtain the tree structure of 8 clusters together with binary questions. And the user can check the number of drugs in each cluster as shown in figure ๖.

Figure5. Clustering Tree

6. CONCLUSTION

In this proposed system, divisive hierarchical clustering technique is used to cluster various drugs. The clustering algorithm used in this system, DIVCLUS-T, is monothetic divisive clustering algorithm. By using this system , users can have knowledge of what kinds of drugs are available on the market and which drugs can be used for what disease of what symptoms of diseases. This proposed system will be very useful for pharmacy.

REFERENCES

[1] B.Mirkin "Clustering for Data Mining". *A Data Recovery Approach*. Chapman & Hall/CRC,2005.

[2] F.Murtagh, "A survey of recent advances in hierarchical clustering algorithms", *The Computer Journal*, 26, 329-340,1893.

[3] M.R Anderberg,*Cluster analysis for applications*. Academic Press, New York,1973.

[4] N.M.Reddy, N.V.Ramana and K.R.Reddy, "Unit Commitment Solution using Agglomerative and Divisive Cluster Algorithm", *An Effective New Methodolog*, ACTA press, 2009.

[5] P.P.Rodrigues, J.Gama and J.P.Pedroso, *ODAC: Hierarchical Clustering of Time Series Data Streams*, LIACC, University of Porto, 2004.

[6] S.K.Tasoulis and D.K.Tasoulis, *Improving Principal Direction Divisive Clustering*, Department of Mathematics, University of Patras.