Analysing Pharmaceutical Compounds Based On Cluster Techniques

V. Palanisamy[1], A. Kumarkombaiya[2]
[1]*Professor and Head[1], Department of Computer Science and Engineering, Alagappa University*
[2]*Assistant Professor, Department of Computer Science, Chikkanna Government Arts College*

Abstract

Data mining is the practice of automatically searching large stores of data to discover relations that connect variables in a database. The major issues in data mining research partitioning them into several groups: mining methodology, user interaction, efficiency and scalability. In this paper focus on a performance of clustering techniques quite broadly to refer comparative between BIRCH, CURE, Chameleon algorithm method that allows us to gain insights into the biological actions of chemicals by analyzing large amounts of data.
Keyword: BIRCH, CURE and Chameleon algorithm.

1. Introduction

Cheminformatics is an area of application which was found the molecular structure of drugs contains groups of atoms like carbon, hydrogen, oxygen and nitrogen (pharmaceutical drug discovery, databases of available chemicals) are connected to gather to form different functional groups. Main goal of this work is grouping of chemical compound structures that have similar functionalities by using clustering technique for grouping between the atoms in the molecular structure and found the searching performance of BIRCH, CURE and Chameleon algorithm.

2. Clustering Techniques

Clustering is a task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups [8]. This is done by classifying the objects. Owing to huge amount of data collected in databases, cluster analysis becomes a highly active topic in data mining research. The overall process of clustering involves the following steps:

1. Generate appropriate descriptors for each object in the data set.
2. Select an appropriate similarity measure.
3. Use an appropriate clustering method to cluster the data set.
4. Analyze the results.

In this paper, refer Fig: 1 Structure of Saxagliptin is a proposed compound drugs data set for analysis grouping of data and searching the performance using BIRCH, CURE and chameleon algorithm.

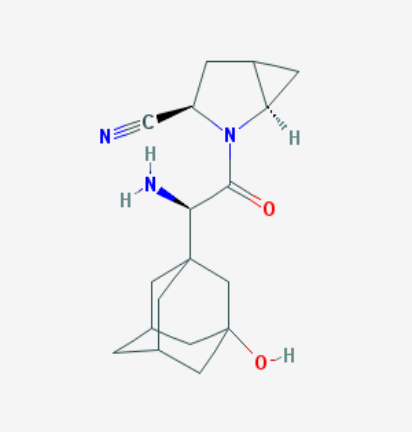

Fig 1 : Chemical Structure of Saxagliptin

3. Comparing Cluster Method

Agglomerative hierarchical clustering algorithms such as CURE, ROCK, Chameleon to analyse the cluster to merge the grouping of data **[1]**. CURE measures the similarity of two clusters based on the similarity of the closest pair of the representative points belonging to different clusters, without considering the homogeneity of the two clusters. CURE ignores the information about the aggregate inter-connectivity of objects in two clusters [1][6]. ROCK observed the similarity function tends to merge clusters which have disjoint set of attributes. After drawing a random sample from the database, the hierarchical clustering algorithm that employs links is applied to the sample objects and it starts with singleton object as an individual class and progressively merges the clusters based on the goodness criteria determined by the link structure [1][6].

Chameleon uses a graph partitioning algorithm to partition graph k-nearest graph into a

large number of relatively small sub clusters. The cluster C is partition into sub clusters C_i and C_j. Then it uses an agglomerative hierarchical clustering algorithm that iteratively merges sub clusters based on their similarity. Chameleon determine similarity between each pain of clusters C_i and C_j according to their relative connectivity, RI (C_i , C_j) and their relative closeness RC(C_i, C_j). It was not applied to high dimensions [6][7].

4. Birch Algorithm

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [2] an algorithm is an agglomerative hierarchical clustering method which builds a dendrogram called clustering feature tree (CF tree) while scanning the data set to condense information about sub-cluster of points. It contains two key phases such as i) Scans the database to build an in-memory tree and ii) Applies clustering algorithm to cluster the leaf nodes. Birch handles the task in a very novel manner. It maintains a set of Cluster Features (CF) of the sub-cluster. The criteria for merging two sub-clusters are taken from the information provided solely by the set of CFs of the respective sub-cluster. Each entry in the CF tree represents a cluster of objects and is characterized by a triple feature as (N, LS, SS), where N is the number of data objects in the cluster and LS is the linear sum of the data object and SS is the square sum of the data object in the cluster.

$$|C| = n;$$

$$\sum_{i \varepsilon C} X_i = LS \quad \text{and}$$

$$\sum_{i \varepsilon C} X^2_i = SS$$

Since the objects are multi dimensional, the summations and exponentiation in the above expressions are dealt with component-wise. The dimensions of LS and SS are the same as that of the object in Cluster.

BIRCH assumes that the distance functions between clusters are so defined that they can be expressed in terms of the parameters encoded as (N, LS, SS) respectively. If CF1 = (N1, LS1, SS1), and CF2 = (N2 , LS2, SS2) are the CF entries of two disjoint sub-clusters.

The CF entries of the sub-cluster formed by merging the two disjoin sub-clusters is:
CF1 + CF2 = (N1 + N2 , LS1 + LS2, SS1 + SS2)

4.1 CF Tree

CF tree based on some threshold value and distant measurement (branching factor) to make a cluster from the tree. Since the parameters of this algorithm include centroid, average distance of all entries, each leaf node can have two pointers previous and next, which can be used to chain all leaf node together for an efficient scan. A leaf node represents a cluster made up of all sub-clusters represents by its entries. All entries in a leaf node must satisfy the threshold requirements with respect to threshold value, the diameter of the sub-cluster must be less than threshold value. A CF tree is built dynamically and incrementally as new data objects are inserted. The main task of this algorithm is performance analysis right time and memory space.

Phase 1: Loading the input data, Phase 2: Condensation of CF tree by redefining the threshold, tree size become manageable. Phase 3: Global clustering and Phase 4: Cluster refining – this is optional, and requires more passes over the data to refine the results.

It make cluster from this tree based on some threshold value and distant measurement. The next step of the algorithm is to take two random points and find the portion of the clusters cover by the area. The main tasks of algorithm are analysis the performance and cluster the given data set into merged one.

The output file contains the parameters along with Centroid Euclidean inter-cluster distance measure for the two clusters C_1 and C_2 with cluster Centroid[3] as

$$D_0(C_i,C_j)=\left[\sum_i (X^i_{centroid,i} - X^i_{centroid,j})\right]^{\frac{1}{2}}$$

The tree will be stored in the format- root of the tree which is a special node with two pointers- left and right. It also contains the intermediate node information along with leaf nodes which mainly contains the original data points from the database. BIRCH is a scalable clustering algorithm with respect to the number of objects and good quality of clustering of the data.

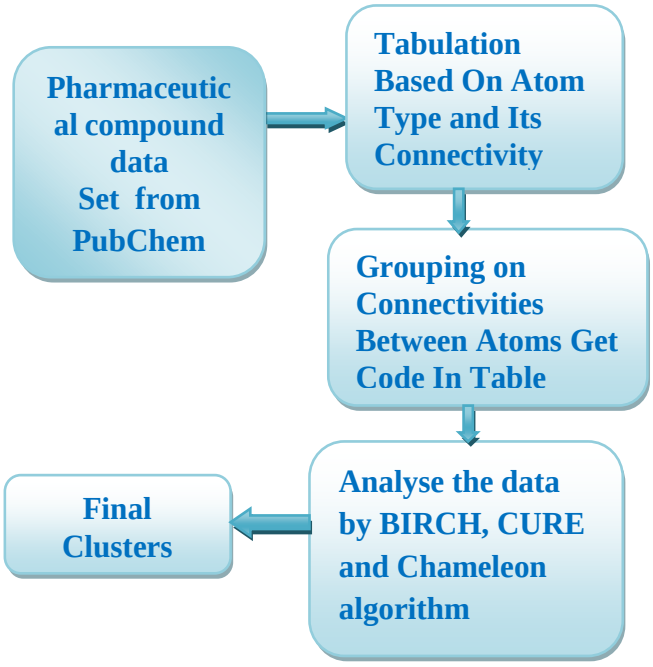

Fig 2: Finding cluster from given data

Saxagliptin chemical compound structure is taken as input into those algorithms [9]. The input to the system is matrix which contains atom number, atom type, and connected atom with bonding.From Fig: 2 represent the methodology of clustering techniques to form the group of atoms in respective way.

From Table 1, the data set contains atom number and atom type then to denote the connected atoms as in single or double bond respectively refer from Table 2. In paper it takes atoms from dataset and pairs are connected atom into the bond as given below.

Atom No	Atom Type
1	O
2	O
3	N
4	N
5	N
6	C
7	C
8	C
9	C
10	C
11	C
12	C
:	:

Table 1: Data Set contain Atom type

Atom 1	Atom 2	Connected Bond
1	7	1
1	46	1
2	21	1
3	17	1
3	21	1
3	22	1
4	16	1
4	47	1
4	48	1
5	23	1
6	10	1
6	11	1
:	:	:
:	:	:

Table 2: Data set contains Atoms with Connected Bonds

5. Grouping of Cluster

From Fig: 3, after grouping the similar codes founded by cluster from the given data set which is based on atom number by connected atoms. This process is continuing until the completion of end of data set.

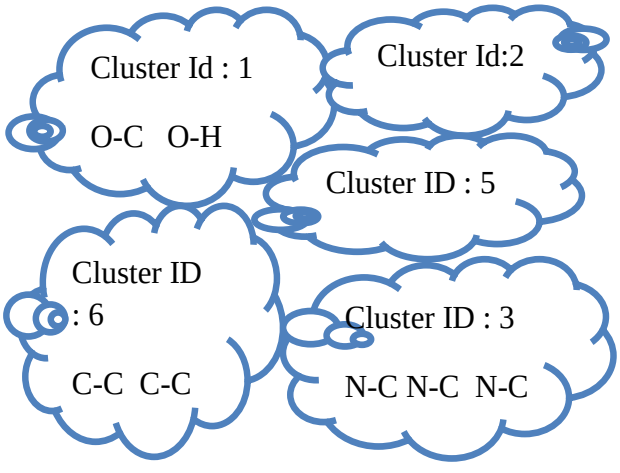

Fig: 3 Grouping of Objects

6. Performance Factor

The performance of hierarchical BIRCH clustering algorithms is presented in this section. Pharmaceutical data set has been taken into consideration and grouping the data. From the fig: 4 represent an analysis of cluster performance and time taken for required BIRCH algorithm with other CURE and Chameleon algorithm respectively.

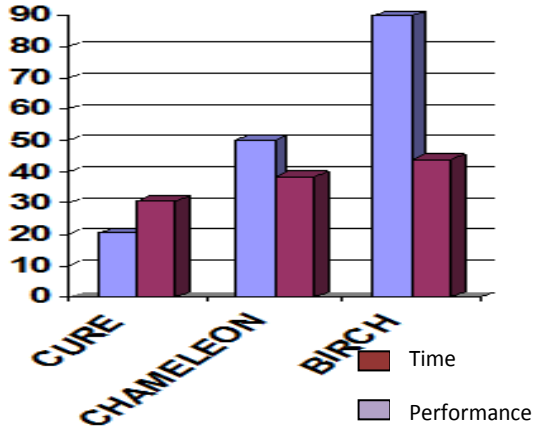

Fig 4: Analysis the performance and Time taken for cluster techniques

7. Conclusion

Clustering is one of the techniques in Data mining for grouping similar objects. In this paper we compared the Hierarchical clustering algorithm to merging the data set from PubChem. The input data is Saxagliptin chemical compound dataset for grouping the data and merge in efficient manner. BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points for producing the best quality clustering with a single scan of the data and to find better accuracy results for searching the data rather than the other ones for given datasets.

8. Reference

[1] Arun K Pujari, "*Data mining techniques*", University Press (India) Private Limited.

[2] T. Zhang, R. Ramakrishnan and M. Livny: *BIRCH : "An Efficient Data Clustering Method for Very Large Databases".* SIGMOD '96 6/96 Montreal, CanadaIQ1996ACM0-89791-794-4/96/0006

.**[3]** Daniel T. Larose , *Data Mining Methods and Models*, Copyright © 2006 John Wiley and Sons, Inc.

[4] David Hand, Heikki Mannila & adhraic Smyth, "*Principles of Data Mining*", MIT Press, 2001

[5] Arun K Pujari, "*Data mining techniques"*, University Press (India) Private Limited

[6] Marjan Kuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emami Chukanlo, "*A survey of hierarchical clustering algorithms",* The Journal of Mathematics and Computer Science Vol .5 No.3 (2012) 229-240.

[7] Jiawei *Han* and Micheline *Kamber," Data Mining*: *Concepts* and Techniques". Publication: ISBN-10: 0123814790 | ISBN-13: 978-0123814791, Edition: 3

[8] https://sites.google.com/site/dataclusteringalgorithms/hierarchical-clustering-algorithm

[9] http://www.ncbi.nlm.nih.gov/pccompound/?term=saxagliptin.