IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-3, Issue-4 E-ISSN: 2347-2693

Designing a Knowledge Discovery of Clustering Techniques in Pharmaceutical Compounds

V. Palanisamy[1*]and A. Kumarkombaiya[2]

[1]*Department of Computer Science and Engineering, Alagappa University, India
[2]Department of Computer Science Chikkanna, Government Arts College, India

palanisamy.alagappa@gmail.com

www.ijcseonline.org

Received: April /02/2015 Revised: April/11/2015 Accepted: April/23/2015 Published: April/30/ 2015

Abstract— To develop data mining techniques to support decision making and discovery of functional group of the connectivity atom for drug effects by analyzing chemical compound data in the form of structured data. Existing studies in data mining mostly focus on hierarchical clustering techniques applied in large and small dataset of pharmaceutical compound and analyse its performance based on time accuracy. In this paper focuses to apply cluster techniques of partition method like Enhanced K-means algorithm and hierarchical method like Birch and Chameleon algorithm used in pharmaceutical compound specifically represented as atom number, atom name like carbon, hydrogen, nitrogen, oxygen with connected atoms. These dataset form a functional group of atoms by functioning in three phases. The performance can be experimented based on time taken to form the estimated cluster, also overall execution time can be reduced by improvement of Enhanced Kmeans algorithm when compared to chameleon and Birch algorithm.

Keywords— *Enhanced K-Mean algorithm; Chameleon algorithm; Birch algorithm*

I. INTRODUCTION

Cluster analysis is the process of partitioning a set of data objects into subset. Each subset is a cluster, such objects in a cluster are similar to one another, yet dissimilar to objects in other clusters [3] [9]. It means the quality of cluster may be represented by its diameter, the maximum distance between any two objects in the cluster. Some aspects are follow in clustering method are partitioning criteria, separation of cluster, similarity measure, Clustering space are used to form a grouping of data from dataset. In this technique various method are used for cluster such as partitioning methods, Hierarchical methods, Density based methods and Grid based methods. After finding clusters by these clustering algorithms, an Apriori algorithm can be easily applied on clusters of finding the pattern for mining association rules.

There are some clustering algorithms such as Enhanced K-means, and Birch algorithm [4] are discuss in this paper. Enhanced K-means algorithm is a centroid based partitioning method, here quality of cluster can be measured by the within cluster variation, which is the sum of squared error between all object in cluster and centroid. Chameleon and Birch algorithm is based on hierarchical method [7] [8]. The main focus of the study is to find the behavior of various algorithms which is suitable to form a cluster in chemical compound data set, since to find the pattern of atoms analysis on further use of association rule.

II. LITERATURE REVIEW

Chameleon is a hierarchical clustering algorithm that uses dynamic modeling to determine the similarity between pairs of clusters [5]. Here the similarity is assessed based on how well-connected objects are within a cluster and on the proximity of clusters. That is, two clusters are merged if their interconnectivity is high and they are close together. The merge process facilitates the discovery of cluster based on similarity function.

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [2] an algorithm is an agglomerative hierarchical clustering method which builds a dendogram 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. Birch algorithm is an agglomerative clustering technique which is suitable for very large databases [6].

III. PROPOSED RESEARCH METHODOLOGY

A. Data Collection in Proposed Method

In this research work, the data set is taken from Pubchem namely pharmaceutical compound of Saxagliptin, which is currently used in the treatment of diabetes, which improve glycemic control in adults with type 2 diabetes

mellitus. It is an orally active hypoglycemic (anti-diabetic drug) of the new
Dipeptidyl peptidase-4 inhibitor class of drugs.

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 available in Pubchem) are connected to gather to form different functional groups.

The research involves using the values represent by number of atoms, position, connectivity and distance between atoms of saxagliptin pharmaceutical compound structures are taken from Pubchem (Drug Bank).

TABLE I. COLLECTION OF ATOM DETAILS

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
13	C
14	C
15	C
16	C
17	C
18	C
19	C
20	C
21	C
22	C
23	C
24	H
25	H
26	H
27	H
28	H
29	H
:	:
:	:
47	H
48	H

From table I represent the collection of input atom details related Saxagliptin compound structure like oxygen, Nitrogen, and carbon to gather the connectivity of required atom which represent in Table II with bond type details. It is used to interconnectivity of atoms with each other to form chain details through functional group of element and its connected chain.

TABLE II. CONNECTED ATOM SET

SI. No.	Atom No	Connected Atom	Bond Type
1	1	7	1
2	1	46	1
3	2	21	1
4	3	17	1
5	3	21	2
6	3	22	1
7	16	4	1
8	4	47	1
9	4	48	1
10	5	23	3
11	6	10	1
12	6	11	1
13	6	12	1
14	6	16	1
15	7	10	1
16	7	13	1
17	7	14	1
18	8	12	1
:	:	:	:
:	:	:	:
51	22	23	1
52	22	45	1

B. Aim of the research

The main goal of this paper, to develop data mining techniques to support decision making and discovery of functional group of the connectivity atom for drug effects by analyzing chemical compound data in the form of structured data.

Developing the general process for grouping a pharmaceutical of collective atom that have similar functionalities by using clustering technique for grouping between the atoms in the molecular structure and found the searching performance by comparing the data mining techniques like BIRCH [5], Chameleon [5] and Enhanced K-Means Clustering algorithm. Finally, form the functional group of elements (atoms) to generate chain details.

C. *Proposed Work*

There are three phases to construct the functional group of chain details of required compound atom in the pharmaceutical structure.

In first phase, k-nearest-neighbor graph approach to construct a sparse graph, there exist edges between two vertices, if one object is among the *k*-most-similar objects to other. The edges are weighted to reflect the similarity between objects where each vertex of the graph represents a data object.

In second phase, it can be proposed that the similarity between each pair of clusters such as C_i and C_j by their relative interconnectivity, RI (Ci, Cj), and their relative closeness, RC (Ci, Cj) based on chameleon algorithm.

$$RI(C_i, C_j) = \frac{\left|EC_{\{C_i, C_j\}}\right|}{\frac{1}{2}\left(\left|EC_{c_i}\right|\left|EC_{c_j}\right|\right)} \qquad (1)$$

where $EC_{\{C_i, C_j\}}$ is the edge cut, defined as above, for a cluster containing both C_i and C_j. Similarly, EC_{Ci} (or EC_{Cj}) is the minimum sum of the cut edges that partition C_i (or C_j) into two equal parts.

The relative closeness, RC(C_i, C_j), between a pair of cluster is the absolute closeness between the two cluster has normalized with respect to the internal closeness of the two clusters. It is defined as

$$RC(C_i, C_j) = \frac{\bar{S}EC_{\{C_i, C_j\}}}{\frac{|C_i|}{|C_i| + |C_j|}\bar{S}EC_{C_i} + \frac{|C_j|}{|C_i| + |C_j|}\bar{S}EC_{C_j}} \qquad (2)$$

where $\bar{S}EC_{\{C_i, C_j\}}$ is the average weight of the edges that connect vertices in C_i to vertices in C_j, and $\bar{S}EC_{C_i}$ or($\bar{S}EC_{C_j}$)

is the average weight of the edges that belong to the min bisector of cluster C_i (or C_j).

In third phase, enhanced K-Means clustering algorithm [1] can be used to merge the sub-cluster it takes into account both the interconnectivity as well as the closeness of the clusters. Initially centroids are calculated [2] [10] [11] [12] and as the data are same, it results in same calculations, so the number of iterations remains constant and elapsed time is also improved. It can be compute the distance to the previously by nearest cluster. If the new distance is less than or equal to the previous distance, the point stays in its cluster, and there is no need to compute its distances to the other cluster, the time required to compute distances to *k*−1 cluster centers. This is the reason that proposed Enhanced K-mean clustering algorithm is efficient from basic K-mean algorithm.

To apply Enhanced K-means clustering algorithm for origin of clustering process of atoms to merging the functional group of related atom for analyzing the atom in different stage. Pseudocode of the enhanced Kmeans algorithm is given below,

```
D= {a_1, a_2,a_3,… a_n}        // Set of n number of atom
a_i = { x_1, x_2,x_3,… x_m}    // Set of connectivity bond of one
                                  atom point.
C                              // Number of desired clusters.
Ensure: Set of functional group of clusters
Steps
//assign each point to its nearest cluster
For i=1 to n
    Compute squared Euclidean distance
    d_2(x_i, Clusterid[i]);
    If (d_2(x_i, Clusterid[i])<=Pointdis[i])
        Point stay in its cluster;
    Else
        For j=1 to k
            Compute squared Euclidean distance
            d_2(x_i, m_j);
        endfor
        Find the closest centroid m_j to x_i;
        m_j=m_j+x_i; n_j=n_j+1;
        MSE=MSE+d_2(x_i, m_j);
        Clusterid[i]=number of the closest centroid;
        Pointdis[i]=Euclidean distance to the closest
        centroid;
endfor
```

```
For j=1 to k
mj=mj/nj
end for
```

IV. Analysis of Result and Evaluation

From table III, illustrate the phase 1 of analysis to partitioning the atom based on connected atom of each atom type from table I and table II based on k-nearest-neighbor graph approach. From group id is 1, atom 1 of O connects with atom 7 of C and atom 1 of O connects with atom 46 of H and bond type is one. These inputs can be partitioned as OC-OH, similarly the group of atoms can be formed up to 22 atoms respectively.

TABLE III. Phase 1: PARTITONING THE ATOM

Clustering of Connected atoms	
Group id is : 1 O-C O-H	Group : 12 Group id is : 12 CH C-H
Group : 2 Group id is : 2 O=C	Group : 13 Group id is : 13 CH C-H
Group : 3 Group id is : 3 N	Group : 14 Group id is : 14 CH C-H
Group : 4 Group id is : 4 NH N-H	Group : 15 Group id is : 15 CH C-H
Group : 5 Group id is : 5 NC	Group : 16 Group id is : 16 CC C-H
Group : 6 Group id is : 6 CC CC CC C-C	Group : 17 Group id is : 17 CC CC C-H
Group : 7 Group id is : 7 CC CC C-C	Group : 18 Group id is : 18 CC CC C-H
Group : 8 Group id is : 8 CC CC CC C-H	Group : 19 Group id is : 19 CH C-H
Group : 9 Group id is : 9 CC CC CC C-H	Group : 22 Group id is : 22 CC C-H
Group : 10 Group id is : 10 CH C-H	Group : 20 Group id is : 20 CC CH C-H
Group : 11 Group id is : 11 CH C-H	

After partitioning the relevant atom, it can be moves to phase 2 for interconnectivity of connected atoms based on relative closeness method between a pair of clusters which absolute closeness between C_i and C_j, normalized with respect to the internal closeness of the two clusters.

By the collection of the related atom connectivity an input items as shows in table II, it can be analyse the interconnectivity of each atom level by level. For example, from an input of connected atom, in first level first atom connects seventh atom and in second level the seventh atom connects with tenth atom, thirteenth atom and fourteenth atom. Finally in third level tenth atoms interconnect with twenty sixth and twenty seventh atom. But twenty sixth and twenty seventh atom does not interconnect with any atom. Hence an analysis can be terminated. Then the interconnectivity of chain details can be show 1-7-10 respectively. Similarly, the process can be continuous for forming the chain details as shown in Level 2.

PHASE 2: INTERCONNECTIVITY OF CHAIN DETAILS

1-7-10

3-17-18-19

3-22

4-16

6-10

6-11

6-12

6-16

7-10

7-13

7-14

8-11

8-13

8-15

9-12

9-14

9-15

17-18-19

17-19

18-19

18-20-22

20-22

Finally, By applying an Enhanced K-Means algorithm in phase 3,represents the functional group of chain details from the number of atoms based on three level, it can merge based on the centroid distance initially it can be calculated and form the functional group of elements and give the result in chain detail of Saxagliptin pharmaceutical compound.

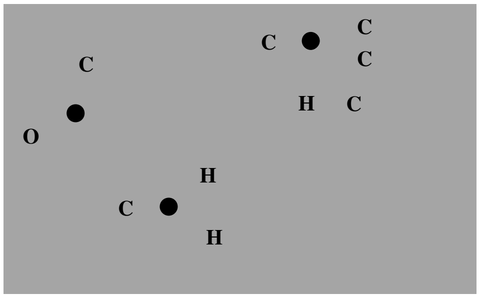

Fig1. Level 1 for Initial measuring the centroid distance

From fig.1, illustrates the level one for measuring the centroid distance of each atom and find the closest distance of atom to be formed for merging purpose.

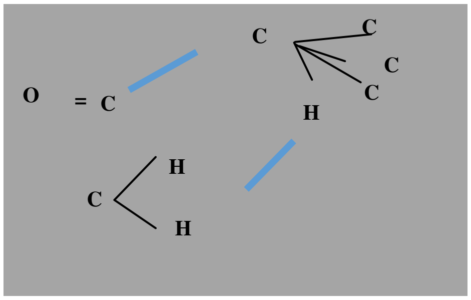

Fig2. Recalculating for merge the closeness of atom molecules

From fig.2, represents to merging the closest of atoms after forming the cluster. Finally to functional group of atom forming in chain detail efficiently as shown in fig.3.

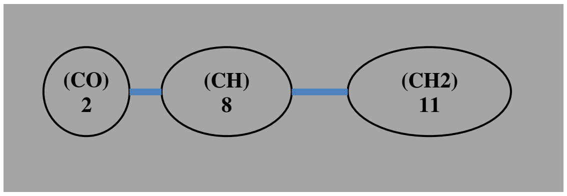

Fig3. Forming Functional group of Atom in chain details

PHASE 3: FUNCTIONAL GROUP CHAIN DETAILS

2-8-11

4-18-19-20

4-1

5-17

7-11

7-12

7-13

7-17

8-11

8-14

8-15

9-12

9-14

9-16

10-13

10-15

10-16

18-19-20

18-20

19-20

19-22-1

22-1

From fig.4, represents the accurate computing functionality of each algorithm performance experimented in MAT LAB. Here an Enhanced K-means algorithm can give 98% efficient result of performance at time of cluster when compared with Birch algorithm takes 95% and Chameleon algorithm takes 90% for clustering process.

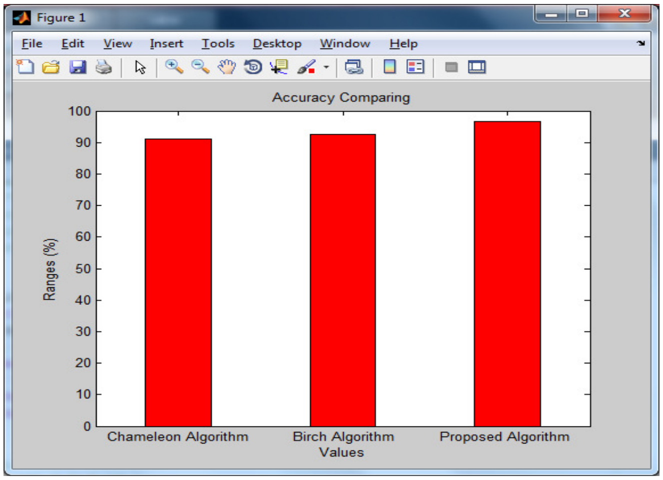

Fig 4. Accuracy for comparing Chameleon, Birch and Enhanced Kmeans Algorithm

V. Conclusion

In this paper , it can be presented an idea of enhanced Kmeans clustering technique instead of Agglomerative Hierarchical clustering technique forming the functional group of atoms in chain details for further analyse of molecule in pharmaceutical compound drug by given atom number, atom name and connected atom details. It can be functional group to form the chain details based on centroid distance. Also it can be improve the execution of the forming cluster when compare with other techniques.

REFERENCES

[1] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, “An Efficient enhanced k-means clustering algorithm,” Journal of Zhejiang University, 10(7):1626–1633, 2006.

[2] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, “A New Algorithm to Get the Initial Centroids,” Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pages 26–29, August 2004.

[3] Daxin Jiang, Chum Tong and Aidong Zhang, “Cluster Analysis for Gene Expression Data,” IEEE Transactions on Data and Knowledge Engineering, 16(11): 1370-1386, 2004.

[4] V. Palanisamy, A. Kumarkombaiya, “Analysing Pharmaceutical Compounds Based On Cluster Techniques”, International Journal of Computer Science Research & Technology, ISSN: 2321-8827, Vol. 1 (03).

[5] Jiawei Han and Micheline Kamber,” Data Mining: Concepts and Techniques”. Publication: ISBN-10: 0123814790 | ISBN-13: 9780123814791, Edition: 3

[6] 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.

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

[8] 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.

[9] M. R. Anderberg; Cluster Analysis for Applications: Academic Press, New York, 1973.

[10] D. Pelleg and A. Moore; X-means: Extending k-means with efficient estimation of the number of clusters: In Proceedings of the Seventeenth International Conference on Machine Learning, San Francisco, pp. 727- 734, 2000.

[11] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and AngelaY. Wu. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans.Pattern Anal. Mach. Intell., 24(7):881–892, 2002.

[12] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, “A New Algorithm to Get the Initial Centroids,” Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pages 26–29, August 2004.