Comparative Study on Clustering Techniques towards Virtual Screening

Naga MadhaviLatha Kakarla[1], Dr. G. Rama Mohan Babu[2]

[1]Research Scholar, Dept. of Computer Science and Engineering,AcharyaNagarjuna University*

[2]Professor, Dept. of Information Technology, RVR & JC College of Engineering

Guntur, Andhra Pradesh, India.

Abstract - Virtual Screening is a computational method used in Drug Discovery to search for the chemical structures which have particular properties. Virtual Screening uses molecular docking as a core. Most of the researchers who work on computational drug design and prediction use DrugBank as raw data and converts it into useful subsets. Clustering technique of data mining is primary necessity to proceed with Virtual screening. Clustering is a data mining technique for data analysis used in many fields. It is a process of grouping similar objects of similar properties. This review paper focuses on analyzing Clustering techniques and their usefulness in Virtual Screening.

Keywords -Clustering, Chemical Structure, Data mining, DrugBank, Drug Discovery, Molecular docking, Virtual Screening.

I. INTRODUCTION

Discovering a new drug is an expensive and time taking process. Industry faces huge attrition if it fails. It is always better to explore hidden uses of existing drugs which are used for cure of other disease/disorders. This minimizes the expenses and saves time for industry. Several works in literature examined and reported experimental evidence in this context which suggests the fact that certain drugs might possess dual inhibitory properties against specific and non-specific diseases. It is not certain that both the disease states should exist at once. A new study determined the use of diuretics, anti-hypertensive drugs in reducing the risk of Alzheimer’s disease dementia patients. Further, Canakinumab, drug that fights inflammation, can reduce the risk of heart attacks and strokes in people who have already had a heart attack.

Virtual Screening is a computation approach used in drug discovery to identify the structure that are capable of binding to drug targets that are protein receptors and enzymes from libraries of small molecules ligand based. The main aim of Virtual Screening is to search molecules of chemical structure that bind to the molecular target. There are two types of Virtual Screening techniques: ligand-based virtual screening and structure-based virtual screening [1].

A ligand-based virtual screening is to scan molecules data base against one or more active ligand structure with the use of 2D chemical similarity analysis methods. It is based on searching molecules with shape similar to that of known actives. There are many prospective applications of this virtual screening approach in the literature[2][3].

A structure-based virtual screening is to estimate the likelihood that the ligand will bind to the protein with high affinity by applying a scoring function. It involves docking ligands into a protein target[4][5][6].The computational complexity of many screening programs is O (N^2), where N indicates atom’s size. The computing infrastructure varies from ligand-based approach to stricter-based approach. A ligand-based virtual screening requires a single structure comparison operation whereas a structure-based virtual screening method requires a parallel computing infrastructure such as a cluster of systems. It requires the input from large libraries that can be queried by the parallel cluster.

Molecular docking is a computational method which is used to predict the ligand which has good interaction energy to bind to a receptor for an enzyme. It is used for structure-based method. The main component in molecular docking is a scoring function. A scoring function may bring docking application to new stage [7]. The use of scoring function is to indicate the correct poses from incorrect poses or binders from inactive compounds in a reasonable computation time. It estimates the binding affinity between the protein and ligand. There are three types of scoring functions: force field-based, empirical and knowledge-based.

Data mining is a technique that deals with large amount of data. It converts raw data into a meaningful and understandable structure which can be used contently for future purpose. The tasks in data mining involve anomaly detection, association rule learning, classification, clustering, regression and summarization. Two types of learning sets are used in data mining: Supervised Learning and Unsupervised Learning. Supervised Learning algorithm analyzes the training data and gives an inferred function. The training data consist of a set of training examples. In this, each example is a set of an input vector and a desired output value or supervisory signal. Classification data mining technique comes under supervised learning. Unsupervised Learning algorithm tries to find hidden structure in unlabeled data. In this, each example given to the learner are unlabeled, there is no reward signal to evaluate a potential function. Clustering data mining technique comes under unsupervised learning. Most of the researchers who work on computational drug design and prediction use DrugBank as raw data and converts it into useful subsets. Clustering technique of data mining is primary necessity to proceed with Virtual screening. Clustering is a data mining technique for data analysis used in many fields like Bioinformatics, Voice mining, Image processing, Text mining, Web Cluster engines, Pattern recognition, Whether report analysis.

II. CLUSTERING

Clustering is a for statistical data analysis technique[8]. Clustering is a data mining technique for data analysis used in many fields like Bioinformatics, Voice mining, Image processing, Text mining, Web Cluster engines, Pattern recognition, Whether report analysis. Cluster analysis is the process of maximizing the intraclass similarity and minimizing the interclass similarity. Cluster analysis is used to find similarities between data based on properties found in the data. Clustering is an unsupervised learning process. Clustering is the most important unsupervised learning problem. It deals with finding a structure in a collection of hidden data. A cluster is a collection of objects which are having similar properties between them and are dissimilar to the objects belonging to other clusters. Hard clustering is the process of clustering in which every data point either is belonging to a cluster totally or not. In hard clustering, clusters do not overlap. Soft Clustering is the process in which the data points are not placed in separate clusters instead, the probability of the cluster is assigned.

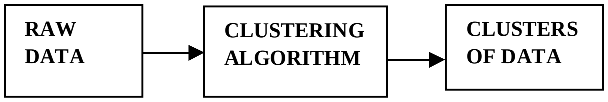

Figure 1: Process of Clustering

Clustering algorithms can be hierarchical-based algorithms and partition-based algorithms. Hierarchical algorithms find successive clusters using previously established clusters, whereas partition-based algorithms determine all clusters at time. Hierarchical algorithms can be agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

A. Types of Clustering: Clustering is a collection of data objects. Where the two types of similarities of clustering's:

- Intraclass similarity **-** Objects are similar to objects in same cluster
- Interclass dissimilarity **-** Objects are dissimilar to objects in other clusters.

The process of maximizing the intraclass similarity and minimizing the interclass similarity. Clusters are formed so that the object belonging to the same cluster which contains similar data and the objects with dissimilarity are placed in different clusters.

Classification of clustering -Clustering is classified into following subgroups:

1. Hierarchical clustering
2. Partition clustering
3. Exclusive Clustering
4. Overlapping Clustering
5. Fuzzy Clustering
6. Complete clustering

1. Hierarchical clustering - Hierarchical clustering exists as a cluster in a bigger cluster to form a tree. As a result, the hierarchical clustering is also known as nested clustering [9].

2. Partition clustering - The process of dividing the set of data objects such that each object should consists of exactly one subset. In partition clustering the clusters will not overlap.

3. Exclusive Clustering -Exclusive clustering deals with the assignment of each value to only one cluster.

4. Overlapping Clustering -Overlapping clustering is used to shrine up the aspect that an object can concurrently belong to more than one group.

5. Fuzzy clustering -In fuzzy clustering, the concept of membership weight comes into existence. Here every object will be a part of every cluster. The membership weight that goes between 0: if it utterly doesn't belong to cluster and 1:if it utterly belongs to the cluster[10].

6. Complete clustering-The task of performing the hierarchical clustering using a set of dissimilarities on 'n' objects that are being clustered is called complete clustering. They tend to find dense clusters of an approximately equal diameter.

B.Applications of Data mining techniques using Clustering- Various application areas of clustering techniques are:

- Medical imaging
- Business and marketing
- World Wide Web

- Computer science
- Analysis of Social networks
- Educational data mining
- Climatology
- Image segmentation

III. CONCLUSION

In this study, a systematic effort was made to identify clustering data mining techniques for data analysis. In this paper a brief review was done for the importance of virtual screening, molecular docking and several data mining clustering techniques. Discovering a new drug is an expensive and time taking process. Industry faces huge attrition if it fails. It is always better to explore hidden uses of existing drugs which are used for cure of other disease/disorders through employment of virtual screening and clustering techniques.

IV. REFERENCES

[1]. [McInnes C (October 2007). "Virtual screening strategies in drug discovery". Current Opinion in Chemical Biology. 11 (5):494-502. doi:10.1016/j.cbpa.2007.08.033. PMID 17936059

[2]. [Li H, Leung KS, Wong MH, Ballester PJ (July 2016). "USR-VS: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques". NucleicAcids Research. 44 (W1):W436-41. doi:10.1093/nar/gkw320.PMID 27106057.

[3]. Sperandio O, Petitjean M, Tuffery P (July 2009). "wwLigCSRre: a 3D ligand-based server for hit identification and optimization". Nucleic Acids Research. 37(Web Server issue): W504–9. doi:10.1093/nar/gkp324. PMC 2703967 . PMID 19429687.

[4]. Kroemer RT (August 2007). "Structure-based drug design: docking and scoring". Current Protein & Peptide Science. 8 (4): 312–28. doi:10.2174/138920307781369382. PMID 17696866.

[5]. Jump up^ Cavasotto CN, Orry AJ (2007). "Ligand docking and structure-based virtual screening in drug discovery". Current Topics in Medicinal Chemistry. 7 (10): 1006–14. doi:10.2174/156802607780906753. PMID 17508934.

[6]. Jump up^ Kooistra AJ, Vischer HF, McNaught-Flores D, Leurs R, de Esch IJ, de Graaf C (2016). "Function-specific virtual screening for GPCR ligands using a combined scoring method". Scientific Reports. 6: 28288. doi:10.1038/srep28288. PMC 4919634. PMID 27339552.

[7]. Meng XY[1], Zhang HX, Mezei M, Cui M Molecular docking: a powerful approach for structure-based drug discovery.curr Comput Aided Drug Des. 2011 Jun;7(2):146-57.

[8]. https://en.wikipedia.org/wiki/Cluster_analysis

[9]. Cheng-Ru Lin, Chen, Ming-SyanSyan , “Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging” IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 2,pp.145-159, 2005.

[10]. Guohua Lei, Xiang Yu, et.all, “An Incremental Clustering Algorithm Based on Grid”,IEEE 8 th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp.1099-1103, 2011.