CHAPTER VIII

CONCLUSION

In this research work, it can be concluded that the data mining techniques support decision making in the analysis of the pharmaceutical structure of compound data. The clustering process in data mining can understand chemical properties, their relationship to structures and make inferences in an efficient manner.

- The advantage of Enhanced K-means Algorithm is to find better initial centroids and provide an efficient way of assigning the data points to the suitable clusters. This approach does not require any additional inputs like threshold values. The proposed algorithm produces more accurate unique clustering results.

- In the proposed research, with respect to the initial centroid distance, the functional groups of pharmaceutical compounds based on the interconnectivity atoms in the pharmaceutical structure data base, can be identified.

- There are three phases to construct the functional group of chain details of the required compound atom in the pharmaceutical structure, similar to the chameleon algorithm. Enhanced K-means clustering algorithm can be used, based on the initial calculated Centroid distance instead of Agglomerative hierarchical clustering.

- In the first phase, K-nearest neighbor graph approach is used to construct a sparse graph, the edges between two vertices exists, if one object is among the \( k \)-most-similar objects to other.

- In the second phase, the similarity between each pair of clusters such as \( C_i \) and \( C_j \) by their relative interconnectivity, \( RI(C_i, C_j) \), and their relative closeness, \( RC(C_i, C_j) \) is observed and Enhanced K-means clustering algorithm is applied, instead of
Agglomerative clustering technique for origin of clustering process of atoms to merge the functional group of related atom for analyzing the atom in different stages.

- From the efficient result, the performance can be experimented in MATLAB; an Enhanced K-means Clustering algorithm takes 49 seconds for analysing the functional group of clustered atoms, when compared to the Birch and chameleon algorithms.

- After identifying the Functional group of chain details, the string searching algorithm is used to match the chemical compound which is represented in numeric data; it can then be identified as the correct matching molecules in string format based on Enhanced Knuth-Morris-Pratt Algorithm, and the searching performance can be increased.

- The time for pattern analysis takes 0.6 seconds to convert the compounds into string; accuracy of pattern matching algorithm is nearly 95% with Enhanced Knuth-Morris-Pratt algorithm rather than other algorithms like Boyer Moore algorithm and Knuth-Morris-Pratt algorithm. An output of the chain detail of the compound can give the result of 2-8-9-11=12-17 as C20 - CH2 - C - CH=CH-CH respectively.

- To identify the frequent item set of Etoposide pharmaceutical compound, generate the candidate generations of different frequent item sets with respect to the support value of 2 and increment one by one, until the final frequent item is matched and OC-CH-OH-CH2 achieved; Piloglitzone pharmaceutical compound item generates the candidate generations of frequent item as OC-CH-CH2 and CH-CH2-C.
• The frequent item set of Etoposide pharmaceutical compound, generates the candidate generations of second frequent item as OC-CH, OC-OH, OC-CH2, CH-OH, CH-C2, CH-CH2, CH-C, C2O –OH, CH2 -COH and OH-CH2 with respect to the support value of 2 and confident level of these items are 100%, 66%, 100%, 50%, 50%, 100%, 60%, 33%, 50%, 50% respectively. Then finally the frequent item is matched and OC-CH-OH-CH2 achieved in the fourth frequent item of candidate generation.

• In Pioglitazone pharmaceutical compound of functional group, let us assume the support value of 2 of frequent item set of candidate generations (C2); items like OC-CH, OC-CH2, CH-CH3 occurred and the confident level is 66%, 100%, 66% and when candidate generations (C3) of third frequent item is OC-CH-CH2 and CH-CH2-C then confident level is 15% and 26%.

• The structure searching involves an exact-match search of a chemical database for a specific query structure, to retrieve the acceptor and donor’s results and the ring formation details associated with a particular structure of pharmaceutical compound based on K-nearest neighbor algorithm.

• From the Etoposide pharmaceutical structure, the hydrogen bond acceptor is measured as 12, hydrogen bond donor is 3 and Benzene ring formation is 7 and in Pioglitazone pharmaceutical structure the hydrogen bond acceptor is measured as 3, hydrogen bond donor is 1 and Benzene ring formation is 3.

• The structure of such a molecule is completely defined by the bond length, the distance between the centers of the bonded nuclei. Distance between the two points, depending on the three dimensions of molecule distance can be calculated, based on Euclidean distance, based on the coordinate position of each atom
• Bond angle can be identified in between the two vector of molecules in functional group which can be calculated by the direction of vector A (a, a', a'') and line b with the direction vector B (b, b', b'').

• The strength of the bond can be measured by minimum bond length and more angles between the bonds respectively. Finally, the percentage of compound contained in pharmaceutical structure which represents the number of times the compound occurs in merged cluster is evaluated.

FUTURE WORK

In Future, we can extend and evaluate the key problem of bioinformatics as to how we can computationally compare the complex biomedical observations, such as comparison of molecular structure, gene expression patterns and protein networks based on Artificial Intelligence Neural Network.