Available online at www.sciencedirect.com

ScienceDirect

Materials Today: Proceedings 4 (2017) 8827–8836

www.materialstoday.com/proceedings

ICAAMM-2016

Artificial Intelligence approach for Classifying Molecular Dataset using Density based technique with appropriate Euclidean Distance measure

K. Sai Prasad*[a], O. Subhash Chander[b], G. Prabhakar Reddy[a], S. Gururaj[b,*]

[a]*Asst. Professor, Department of Computer Science,MLR Institute of Technology, JNTU Hyderabad, India*
[b]*Asst. Professor, MCA Department, Nizam College, O.U, Hyderabad, India*

Abstract

Data mining is the process of extracting hidden analytic information from large databases using multiple algorithms and techniques. Artificial Intelligence is concerned with improving algorithms by employing problem solving techniques used by human beings. In this paper, we implemented a methodology to classify a huge data set into clusters using a standard mathematical equation. A new clustering algorithm is proposed which involuntarily evolves the number of clusters as well as proper partitioning from a data set with artificial intelligence techniques. Here for the assignment of points to different clusters a standard distance calculation measure Euclidean distance formula is used which is intended to overcome point symmetry based clustering. An idea to implement this methodology on chemical/molecular data came from a software program Fast Rigid Exhaustive Docking which generates interactions of new molecules whose biological activity is unknown. The molecular data thus obtained as scores are tabulated as values. This is considered as raw data, which is further preprocessed and hence from one point to other point distance is calculated using Euclidean distance which forms a multi dimensional database with several combinations. Later on clusters are analyzed and pattern recognition is performed by using an algorithm to obtain results. Density based clustering technique is performed for better results on molecular database.

In the field of chemistry there is much need of data mining for identifying the best molecule whose biological activity value is unknown. Using Artificial intelligent techniques we predicted the biological activity of a molecule when a dependent core training set is provided and leads to different case studies of test set molecules for which data mining techniques are utilized. As data mining is a wide research area that has many applications so new algorithms like this can create a boon in research areas and hence evaluating best results.

Selection and Peer-review under responsibility ofthe Committee Members of International Conference on Advancements in Aeromechanical Materials for Manufacturing (ICAAMM-2016).

Keywords: Clustering, Classification, Density based clustering, Euclidean distance, Artificial Intelligence, FRED Docking

2214-7853© 2017 Elsevier Ltd. All rights reserved.
Selection and Peer-review under responsibility ofthe Committee Members of International Conference on Advancements in Aeromechanical Materials for Manufacturing (ICAAMM-2016).

1. Introduction

In short our idea to work on chemical/molecular data came from a software program Fast Rigid Exhaustive Docking which generates interactions of new molecules. The molecular data obtained as scores will be tabulated as values. These values are considered as raw data, which is further preprocessed and hence from point to point distance is calculated using Euclidean distance which forms a multi dimensional database with several combinations.

Cluster analysis is based on measuring similarity between objects by computing the distance between each pair. There are a number of methods for computing distance in a multidimensional environment. Distance is a well understood concept that has a number of simple properties.

- Distance is always positive
- Distance from point x to itself is always zero
- Distance from point x to point y cannot be greater than the sum of the distance from x to some other point z and distance from z to y.
- Distance from x to y is always the same as from y to x.

There are number of distance measures such as Euclidean distance, Manhattan distance and Chebychev distance. Measures of distance have always been a part of human history. Euclidean Distance is one such measure of distance and its intransience is a demonstration to its simplicity and endurance. Euclidean distance of the difference vector is most commonly used to compute distances and has an intuitive appeal but the largest valued attribute may dominate the distance. Overcoming the other distance measures Euclidean Distance remains the conventional routine to measure distance between two objects in space, when that space meets the prerequisites.

Euclidean distance function is:

$\boldsymbol{Dist}\,(\mathrm{x}, \mathrm{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$ **n** is the number of the features for point object x and y, x_i and y_i are the ith feature of point object x and y. [8]

1.1. Clustering and Classification

Clustering:

Clustering is the most significant and unsupervised process of machine learning application of data mining. Like classification where classes are previously defined, in clustering the classes are divided according to class variable. Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. [6]

Classification: Classifying data by using classification techniques in data mining is a very distinctive task. The main aim of classification is to properly calculate the value of each class variables. There are different types of classification methods used in data mining such as bayes, functions, rules, trees etc. This classification process is divided into two steps.

- The first step is to build the model from the training set, i.e. randomly samples are selected from the data set.
- In the second step the data values are assigned to the model and verify the model’s accuracy. [1]

In contrast to machine learning, clustering is a method of unsupervised learning and tries to combine sets of objects having relationship among them, while classification is a method of supervised learning. This approach concentrates on huge datasets with high dimension. Clustering process is not one time task but is continuous and an iterative process of knowledge discovery from huge quantities of raw and unorganized data. [5]

Neural Network: Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. [7] The hidden layers then link to an 'output layer' where the answer is output. Our work of density based clustering and classification technique can be applied to any data as in field of neural networks where in lakhs of data need to be clustered for recognizing patterns.