Integrating Multiple Modalities Of Omics Data Specifically Proteomic Data Using Unsupervised Learning Algorithms

Madhan S Asst Prof
Department of computer science and engineering
University college of Engineering thirukkuvalai
Nagapattinam (dt)
Tamilnadu
India

Bhagyalaxmi P
Department of computer science and engineering
University college of Engineering thirukkuvalai
Nagapattinam (dt)
Tamilnadu
India

Kayalvizhi K
Department of computer science and engineering
University college of Engineering thirukkuvalai
Nagapattinam (dt)
Tamilnadu
India

Manjula S
Department of computer science and engineering
University college of Engineering thirukkuvalai
Nagapattinam (dt)
Tamilnadu
India

ABSTRACT

The multi-omics approach has become a great topic in the biomedical field, with researchers drawn to globally analyzing data integrated from multiple "omics" such as the genome, transcriptome or proteome. The rapid development of high-throughput technologies and computational frameworks enables the examination of biological systems. Multi-omics factor analysis (MOFA), a computational method for discovering the principle sources of variation in multiple data sets. Multiple kernel learning system has been implemented for predicting the particular disease by using the unsupervised learning algorithms like K-means and neural network algorithms. This can be achieved by implementing the process called preprocessing, clustering, segmentation, and processing of data through the genes and multiple omes like proteomes.

Keywords – Multi-omics, MOFA, Kernel Learing System, K-means Clustering Algorithm, Neural Network, clustering, segmentation, prediction, proteomes.

INTRODUCTION

Omic data is high dimensional and complex data, Integration of multiple technologies has emerged as an approach to provide a more comprehensive view of biology and disease. Accurate prediction of disease risk is needed for implementing personalized medicine.

This paper introduces a neural network architecture for integrating multiple modalities of omic datasets like proteomic data to predict the disease by their victims. The proteomic datasets are used to predict the disease by using the unsupervised learning algorithms like K-means clustering and Artificial Neural Networks. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses.

In K-mean Clustering, the training protein sequences are processed against the trained protein sequences to get the kernel parts. K-means clustering is an unsupervised learning algorithm. The main goal is to obtain high accuracy rate of prediction. In this case, you don't have labeled data unlike in supervised learning. The set of data that want to group into and to put them into clusters, which means objects that are similar in nature and similar in characteristics need to be put together. This is what k-means clustering is all about. The term K is basically is a number and it tells how many clusters in the system.The neural networks get the kernel parts and processed it to obtain the accurate disease.

PROCEDURE

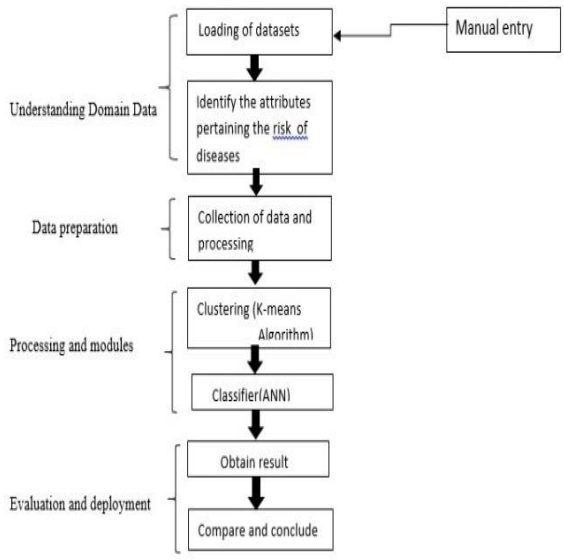
- Protein sequenece is given as a input.
- That input is proceed to the preprocessing.
- The preprocessing is used to remove the damaged data and the empty data, in the overall dataset. Tokenization is the process used in preprocessing.
- Cluster the error free input from the preprocessing step, To implementing the

- clustering process, K initial "means" are randomly generated within the data domain i.e, Randomly assign each data point to a cluster.
- That “means” values of the clustered protein sequence is Segmentized.
- Segmentation is the task which group a set of objects based on the similarities by the clusters. The similar data items are grouped together to get better performance of accessing the data.
- Thus the output is computed by the protein sequence given as a input by using the neural network algorithm.

PROPOSED SYSTEM

In this section we mentioned about the system architecture.fig 1 represents the overview of systems architecture. The core modules of the proposed system consist of :

- Understanding the input data and selecting the attribute related to disease.
- **Data Preparation:** Transformation and Pre-processing of missing data is carried out.
- **Processing Module:** It specifies about the algorithmic approach applied over the system to obtain high accuracy result. Pre-processing modules are separately discussed in upcoming section.
- **Evaluation and deployment:** Final Analysing modules provide information related to generated output. It compares and conclude about measurable resultant artefacts like sensitivity, accuracy etc.



ALGORITHMIC DESCRIPTION

K-means ALGORITHM

The main goal of using Kmeans clustering technique is that it organizes the data into classes such that there is

- high intra-class similarity
- low inter-class similarity

K-means algorithm [15] [16] is famous clustering algorithm widely used in data mining project. The main aim of this clustering is to find the positions μ_i, i=1...k within-cluster to minimize sum of squares distance from the centroid. K-means algorithm depends on k clusters, and it may stuck for different solutions. So to remove such dependency, modified or improved k-means was proposed. Kmeans is accompanied with Lloyd's algorithm to get rid of dependencies. Using this method the results show the quality of clusters is not compromised.

Steps for K-means algorithm are:

1. Initialize the center of the clusters from n data points x_i, i=1...n that have to be partitioned in k clusters
2. Attribute the closest cluster to each data point using Euclidean distance
3. Set the position of each cluster to the mean of all data points belonging to that cluster
4. Repeat steps 2-3 until convergence

In our system Kmeans algorithm plays a crucial role in order to obtain the appropriate number of data groups.Using this algorithm along with Euclidean distance centroids are calculated for different patient attribute. Mean value is taken into account for sample data and henceforth it is judgemental to predicate the patient status. If the mean value of the patient is nearest to the sample mean value, the patient more likely to be affected by heart disease.

ARTIFICIAL NEURAL NETWORK:

An artificial neural network (ANN), usually called neural network (NN). ANN is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. There are three input layers are present in ANN: input layer, hidden layer also called as intermediate layer and output layer. Hidden layers are present in between input and output layer.

Input layer: The input units present in this layer shows the raw information that is fed into the network.

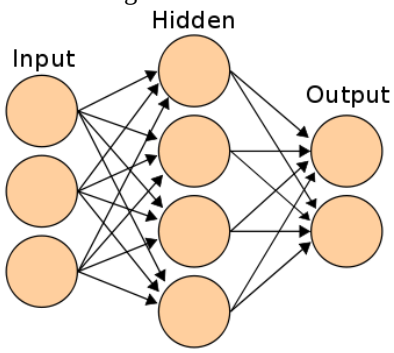
Hidden layer: The activity of each hidden unit is based on the activity of each input unit and weights on the connection between them.

Output layer: The activity of each output unit is based on the activity of each hidden unit and weights on the connection between them.

The ANN algorithm follows:

1. The data from input layer is given to hidden layer.
2. Input values from input layer are used and modified using some weight value and sent to output layer.
3. The value is again modified by some weights from connection between hidden and output layer.
4. This information is processed and output layer gives final output. Finally, this output is processed by activation function.

ANN follows trial and error method in order to get optimal solution. The structure of neural network is shown in Fig



The output is calculated by below function.

$$y_j = \sum_{f=1}^{k} w_{ij}\, x_i$$

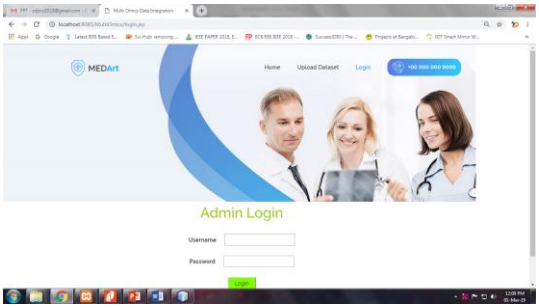
Where,

y_j represents output neuron.
x_i is input neuron
w_{ij} is the weight connecting x_i and y_j
Σ is sigmoidal function

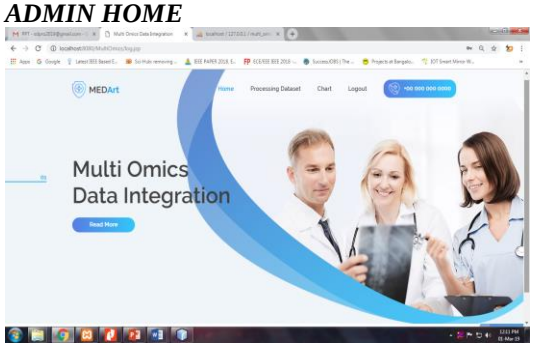
As mention in figure ANN consist of three layers input layer, hidden layer also called as intermediate layer and output layer. In his system former clustered normalized data groups are feed as input to neuron. The patterns vital to heart attack prediction are selected on basis of the computed significant weightage. Weightage are provided based on the range decided for the selected attribute from the dataset. The neural network is trained on dataset. The dataset is divided into two part 70% and 30%.

OUTPUT SCREENS

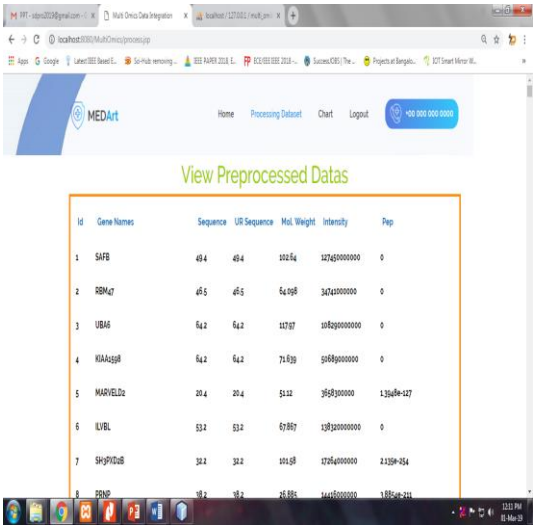
LOGIN PAGE



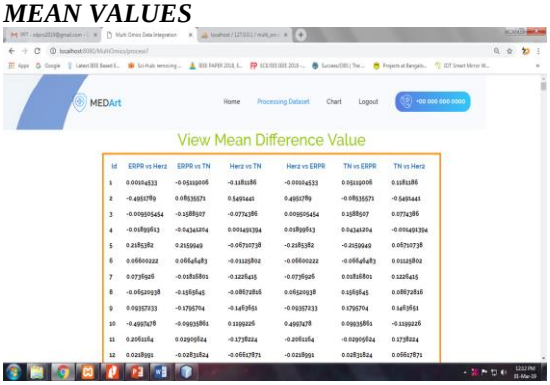
ADMIN HOME



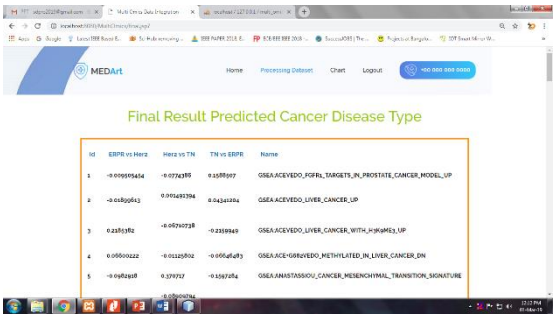
PREPOSSING

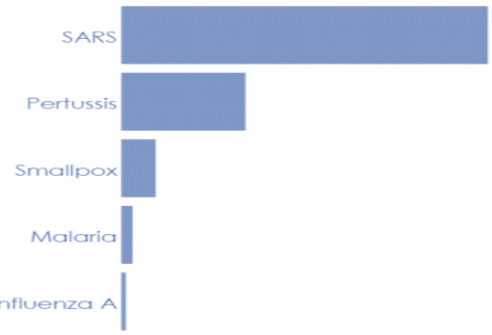


MEAN VALUES



RESULT PREDICTION





KERNEL LEARNING SYSTEM

The deep learning neural network is used for integrating the multi-omics data. Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The Module used to predict the disease are preprocessing of data, Clustering, Segmentation and network traversal in neural network.

PREPROCESSING

Missing value Handling – k-Means clustering cannot deal with missing values. Any observation even with one missing dimension must be specially handled. If there are only few observations with missing values then these observations can be excluded from clustering. However, this must have equivalent rule during scoring about how to deal with missing values. But care must be taken to ensure that missing imputation doesn't distort distance calculation implicit in k-Means algorithm.

The Preprocessing of proteomic data is most useful step to optimize the progress of the prediction process. Preprocessing is used to identify impossible data combinations, missing data, out of range value, and error data in the data set. The preprocessing is used to remove the damaged data, and the empty data in the overall dataset. Process to be used in preprocessing is Tokenization. To Tokenize all the data in the data set will give the more effective processing in other parts of algorithm. Elimination of useless data reduce the processing time of prediction.

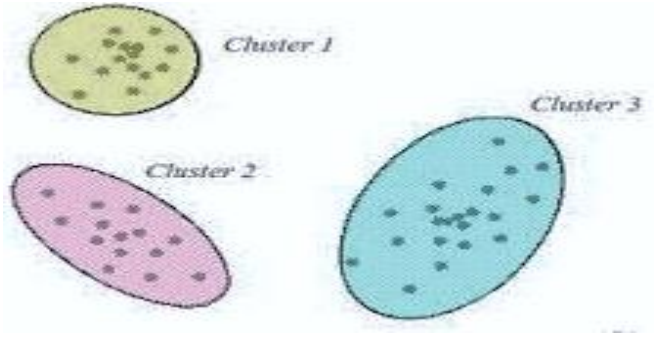
CLUSTERING

k-means clustering aims to find the set of clusters such that every data point is assigned to the closest center, and the sum of the distances of all such assignments is minimized. It is used to find data clusters such that each cluster has the most closely matched data. Clustering is the task of finding the similarities of between the data's to grouping the objects in such a way that objects in the same group are more similar to each other than to those in other groups.

To implementing the clustering process, K initial "means" are randomly generated within the data domain i.e, Randomly assign each data point to a cluster. K clusters are created by associating every observation with the nearest mean. The centroid of each of the K clusters becomes the new mean. These processes are repeated until convergence has been reached.

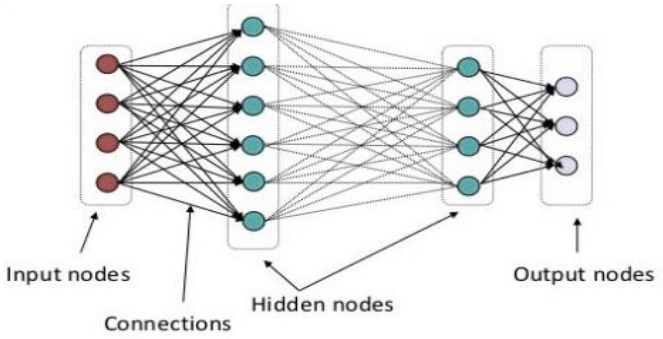
SEGMENTATION

Segmentation is the task of grouping a set of objects based on the similarities by the clusters. The similar data items are grouped together to get better performance of accessing the data. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze.



NEURAL NETWORKS

In a neural network we don't tell the computer how to solve our problem. Instead, it learns from observational data, figuring out its own solution to the problem at hand. An ANN is based on a collection of connected data which are in same segment called artificial neurons. The output of each artificial neuron is computed by some non-linear function of the sum of its inputs. Analysis of the weights and activations in the network can give us biological insights into understanding which data are most relevant for the decision process and how different types of omics influence each other. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold.



CONCLUSION

In this project we have analyzed the proteomic data on the basis of unsupervized learning algorithm specifically K-mean clustering and Artificial neural network and we experimented with that proteomic data to obtained the accurate diseases and the symtoms of the disease. We have found that the use of protein sequences gives the more sensitive results of prediction because it having the 20 attributes(20 amino acids) in the protein sequence. By comparing with the DNA sequences, the analysis of protein sequences has less number of junk data and having less anxiety to taking the protein sequences. Compared to several typical calculating algorithms, the scheming accuracy of our proposed algorithm reaches 94.8%.The report consists of possibility of occurrences of diseases.

REFERENCES

1. Shuang Li, K. Joeri van der Velde, Morris A. Swertz Machine Learning for multi-omics data integration and variant pathogenicity estimation.

2. 1.Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D (2010) An integrated approach to uncover drivers of cancer. Cell 143: 1005–1017 [PubMed]

3. 2.Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57: 289–300

4. Multi-omics data integration using cross-modal neural networks Ioana Bica, Petar Veli˘ckovi´c, Hui Xiao and Pietro Li`o

5. Transfer Learning for Molecular Cancer classification using Deep Neural Networks. Rahul Kumar Sevakula ; Vikas Singh ; Nishchal K. Verma ; Chandan Kumar.

6. Shinde, R., Arjun, S., Patil, P., & Waghmare, J. (2015). An Intelligent Heart Disease Prediction System Using K-Means Clustering International Journal of Computer Science and Information Technologies, 6(1), 637-639.