IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.6, Issue.10, Oct. 2018 E-ISSN: 2347-2693

An Approach to Find the Serotypes of Rotavirus Using Self-Organizing Feature Map

R. Vijayalakshmi[1*], S. Isabella[2]

[1,2]Dept. of Computer Science, Sengamala Thayaar Educational Trust Women's College, Mannargudi, India

[*]*Corresponding Author: rvaviji@gmail.com, Tel.: +91-9965470439*

Available online at: www.ijcseonline.org

Accepted: 7/Oct/2018, Published: 31/Oct/2018

***Abstract*—** Self-organizing maps (SOM) are different from other artificial neural networks in the sense that they use a neighbourhood function to preserve the topological properties of the input space. It converts complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display. As it thereby compresses information while preserving the most important topological and metric relationships of the primary data items on the display, it may also be thought to produce some kind of abstractions. In this works, classifying the virus type using the SOM Toolbox. Self-organizing feature maps (SOFM) learn to classify input vectors according to how they are grouped in the input space. They differ from competitive layers in that neighboring neurons in the self-organizing map learn to recognize neighbouring sections of the input space.

Keywords— SOFM, ANN, Neurons, Cluster, Classification, Quantization, Visualization.

I. INTRODUCTION

SOM represents clustering concept by grouping similar data together. Therefore it can be said that SOM reduces data dimensions and displays similarities among data. With SOM, clustering is performed by having several units compete for the current object.

Once the data have been entered into the system, the network of artificial neurons is trained by providing information about inputs. During the training stage, the values for the input variables are gradually adjusted in an attempt to preserve neighbourhood relationships that exist within the input data set. As it gets closer to the input object, the weights of the winning unit are adjusted as well as its neighbours.

Although the term "Self-Organizing Map" could be applied to a number of different approaches, we shall always use it as a synonym of Kohonen's Self Organizing Map [1], or SOM for short. TeuvoKohonen writes "The SOM is a new, effective software tool for the visualization of high-dimensional data. It converts complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display. As it thereby compresses information while preserving the most important topological and metric relationships of the primary data items on the display, it may also be thought to produce some kind of abstractions [2]."

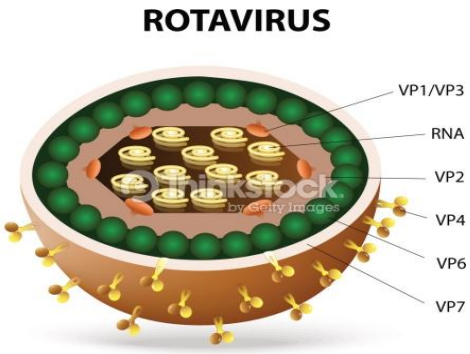

Fig-1:Rotavirus structure

There are five species of rotavirus, referred to as A, B, C, D and E. Humans are primarily infected by species A, B and C, most commonly by species A. All five species cause disease in other animals. Within rotavirus A there are different strains, called serotypes. As with influenza virus, a dual classification system is used based on two proteins on the surface of the virus. The glycoprotein VP7 defines the G serotypes and the protease-sensitive protein VP4 defines P serotypes. Because the two genes that determine G-types and P-types can be passed on separately to progeny viruses, different combinations are found.

Section I contains the introduction of Rotavirus using SOFM, Section II contain the proposed work, Section III contain the some measures of data collection methodology, Section IV

describes results and discussion, Section V concludes research work with future directions.

II. PROPOSED WORK

This paper proposed using self-organizing map [3] to cluster the virus types. The basic work of this paper is data collection. In that four types of virus sequences are collected. Each type contains twenty five sequences and also each type of virus contains five sequences for training. First we use the hundred sequences to cluster the virus types. Though that we find the quantization error and also the topographic error. After that we use the training data to find the position of the virus type. Using NPR tool the data are trained, through that we get the confusion matrix and also we find the error rate using the network clustering tool. Rasmol tool is used to visualise the structure of the virus. We can see the structure of the virus in different mode for example ribbon and group mode, stick and group mode. Final works of this paper contain to cluster the virus type using the self -organizing map [4, 5].

III. METHODOLOGY

Data Collection

The basic work of this paper is data collection. In that rotavirus types are collected. They are rotavirus A, rotavirus B, rotavirus C, rotavirus D. In each of that type has the complete genome sequence data set that is AGTC (Adenine, Guanine, Thymine and Cytosine). All sequence files are loaded in matlab coding. Through that code, the data can be counted and the counted data are written in excel, after that the excel file converted into data file that is .dat format.

The four types of data files are created with the combination of AGTC.

Singlets:

A,GT,C.

Doublets:

AA,AG,AT,AC,GG,GA,GT,GC,TT,TA,TG,TC,CC, CA,CG,CT.

Triplets:

AAA,CAA,TAA,GAA,ACA,CCA,TCA,GCA,ATA ,CTA,TTA,GTA,AGA,CGA,TGA,GGA,AAC,CAC,TAC,G AC,ACC,CCC,TCC,GCC,ATC,CTC,TTC,GTC,AGC,CGC, TGC,GGC,AAT,CAT,TAT,GAT,ACT,CCT,TCT,GCT,ATT ,CTT,TTT,GTT,AGT,CGT,TGT,GGT,AAG,CAG,TAG,GA G,ACG,CCG,TCG,GCG,ATG,CTG,TTG,GTG,AGG,CGG, TGG,GGG.

Quartets:

AAAA,CAAA,TAAA,GAAA,ACAA,CCAA,TCA A,GCAA,ATAA,CTAA,TTAA,GTAA,AGAA,CGAA,TGA A,GGAA,AACA,CACA,TACA,GACA,ACCA,CCCA,TCC A,GCCA,ATCA,CTCA,TTCA,GTCA,AGCA,CGCA,TGCA ,GGCA,AATA,CATA,TATA,GATA,ACTA,CCTA,TCTA, GCTA,ATTA,CTTA,TTTA,GTTA,AGTA,CGTA,TGTA,G GTA,AAGA,CAGA,TAGA,GAGA,ACGA,CCGA,TCGA,G CGA,ATGA,CTGA,TTGA,GTGA,AGGA,CGGA,TGGA,G GGA,AAAC,CAAC,TAAC,GAAC,ACAC,CCAC,TCAC,G CAC,ATAC,CTAC,TTAC,GTAC,AGAC,CGAC,TGAC,GG AC,AACC,CACC,TACC,GACC,ACCC,CCCC,TCCC,GCC C,ATCC,CTCC,TTCC,GTCC,AGCC,CGCC,TGCC,GGCC, AATC,CATC,TATC,GATC,ACTC,CCTC,TCTC,GCTC,AT TC,CTTC,TTTC,GTTC,AGTC,CGTC,TGTC,GGTC,AAGC ,CAGC,TAGC,GAGC,ACGC,CCGC,TCGC,GCGC,ATGC, CTGC,TTGC,GTGC,AGGC,CGGC,TGGC,GGGC,AAAT,C AAT,TAAT,GAAT,ACAT,CCAT,TCAT,GCAT,ATAT,CT AT,TTAT,GTAT,AGAT,CGAT,TGAT,GGAT,AACT,CAC T,TACT,GACT,ACCT,CCCT,TCCT,GCCT,ATCT,CTCT,T TCT,GTCT,AGCT,CGCT,TGCT,GGCT,AATT,CATT,TAT T,GATT,ACTT,CCTT,TCTT,GCTT,ATTT,CTTT,TTTT,GT TT,AGTT,CGTT,TGTT,GGTT,AAGT,CAGT,TAGT,GAG T,ACGT,CCGT,TCGT,GCGT,ATGT,CTGT,TTGT,GTGT, AGGT,CGGT,TGGT,GGGT,AAAG,CAAG,TAAG,GAAG, ACAG,CCAG,TCAG,GCAG,ATAG,CTAG,TTAG,GTAG, AGAG,CGAG,TGAG,GGAG,AACG,CACG,TACG,GACG, ACCG,CCCG,TCCG,GCCG,ATCG,CTCG,TTCG,GTCG,A GCG,CGCG,TGCG,GGCG,AATG,CATG,TATG,GATG,A CTG,CCTG,TCTG,GCTG,ATTG,CTTG,TTTG,GTTG,AGT G,CGTG,TGTG,GGTG,AAGG,CAGG,TAGG,GAGG,ACG G,CCGG,TCGG,GCGG,ATGG,CTGG,TTGG,GTGG,AGG G,CGGG,TGGG,GGGG.

After creating the data file, those data files are loaded in the self-organising map toolbox.

3.1 Train the Data Using SOM

The self-organizing feature map (SOM) has been widely used as a tool for visualization of high dimensional data. The important features that includes the classification of data. The primary data items SOM is composed in two layers of neurons, input and output layers. A neighbouring relation with neurons defines the topology of the map.

1. The map node weight vectors are randomized.
2. Read all input datasets.
   1. Using Euclidean distance formula find the similarity between the input vector and the map nodes weight vectors.
   2. Track the node in smallest distance that is, BMU(Best Matching Units).
3. Update, the node and the near node of the BMU using the following formula,

$$w_v(s+1)=wv(s)\Theta(u,v,s)\alpha(s)\lambda()D(t) -w_v(s)) \quad s<\lambda$$

4. Increase s and do step 2.

Here, D(t)- input data vector, s-current iteration, D(s)-iteration progress, λ-iteration limit, w_v-current weight vector, Θ(u,v,s)-neighbourhood function.

The self-organizing map contains two types of training. One is batch train and another one is sequence train. After the training process the data can be visualized in u-matrix. The u-matrix is shown by 20-by-20 matric format and then the u-matrix contains hit ratio values of the data. Then the data are visualized in surface map through that we know data that can be grouped in similar types of data.
The data file also trained using the neural network toolbox [6].

3.2 SOM Architecture

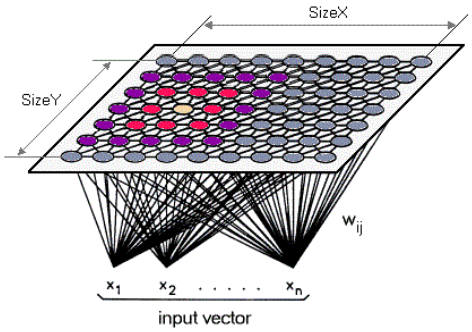

Fig-2:SOM Architecture

3.3 SOM Algorithm

1. Randomize the map's nodes' weight vectors
2. Grab an input vector $\mathbf{D(t)}$
3. Traverse each node in the map
   1. Use the Euclidean distance formula to find the similarity between the input vector and the map's node's weight vector
   2. Track the node that produces the smallest distance (this node is the best matching unit, BMU)
4. Update the nodes in the neighborhood of the BMU (including the BMU itself) by pulling them closer to the input vector
   1. **Wv**(s + 1) = **Wv**(s) + Θ(u, v, s) α(s)(**D**(t) - **Wv**(s))
5. Increase s and repeat from step 2 while $s < \lambda$

A variant algorithm

1. Randomize the map's nodes' weight vectors [7, 8]
2. Traverse each input vector in the input data set
   1. Traverse each node in the map
      1. Use the Euclidean distance formula to find the similarity between the input vector and the map's node's weight vector
      2. Track the node that produces the smallest distance (this node is the best matching unit, BMU)
      3. Update the nodes in the neighborhood of the BMU (including the BMU itself) by pulling them closer to the input vector
      4. **Wv**(s + 1) = **Wv**(s) + Θ(u, v, s) α(s)(**D**(t) - **Wv**(s))
3. Increase s and repeat from step 2 while $s < \lambda$.

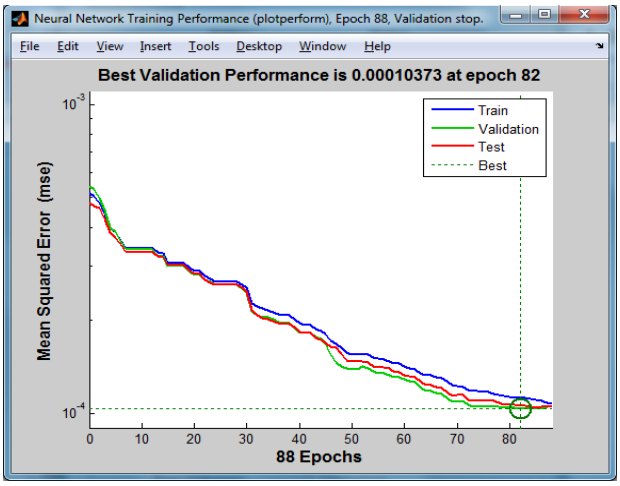

Fig-3: Best Validation Performance for Triplets.

Dimer Count:
Dimer Count is used to count dimers in nucleotide sequence. It represents the AGTC Count

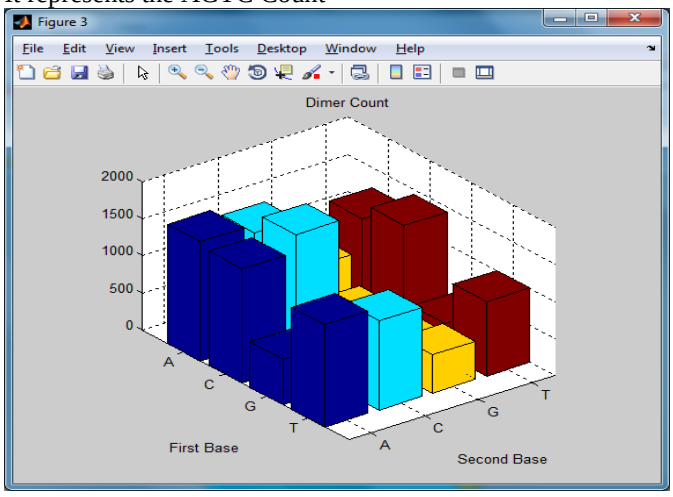

Fig-4: Dimer Count

3.4 Classification

Classification process is done using the self-organising map algorithm. In this the rotavirus types are classified that is, the rotavirus A,rotavirus B,rotavirus C,rotavirus D. Neural pattern recognition tool is also used to classify the virus types and through the NPR tool we find the confusion matrix [9, 10]. Neural network training performance also we find using the NPR tool. Through the neural network training error histogram shows the error rate in the virus sequence.
Bioinformatics tool is also used to know about the nucleotide density and the combination of AT and CG density. Through the bioinformatics toolbox we find the codonbias and also the codoncount. Rasmol tool is used to visualize the structure of the virus.

IV. RESULTS AND DISCUSSION

The experiments were carried out using two databases obtained from the UCI data repository and NCBI. A description of the methodology used for each experiment is presented in the following. Firstly, the dataset was analysed using the traditional SOM algorithm. Then, the dataset are classified using the SOM. Through the neural network pattern recognition tool, neural network clustering tool we find the variation between the data set. Rasmol tool is used to visualize the virus sequence and also it represent the structure of the particular virus.

Finally, SOM algorithm was applied over trained maps to obtain segmentation. Several comparative criteria were used, including the individual counting of errors obtained in the application of the algorithm over well-known databases, the use of quality measures present in toolbox and visual comparison of the trained maps, U-Matrix and segmented maps. The first criteria consists of labelling each of the neurons on the map with information about the class number it belongs to, in agreement with the number of input data instances that it represents.

For this to occur, each input data instance had to have a label to identify its class. Note that privileged information was not used during the training phase, only algorithm accuracy was verified. The second consists of using quality measures to evaluate and compare clustering algorithms. In this work, we used two included in the SOM Toolbox:

a) Data representation accuracy, measured using average quantization error between data vectors and their BMUs on the map.

b) Dataset topology representation accuracy, measured using topographic error, which is the percentage of data vectors for which the first and second BMUs are not adjacent units.

The third criterion is more subjective, because it is based on a visual comparison. As previously described, the U-matrix allows the visualization of representing neuron relations, thus SOM clusters.

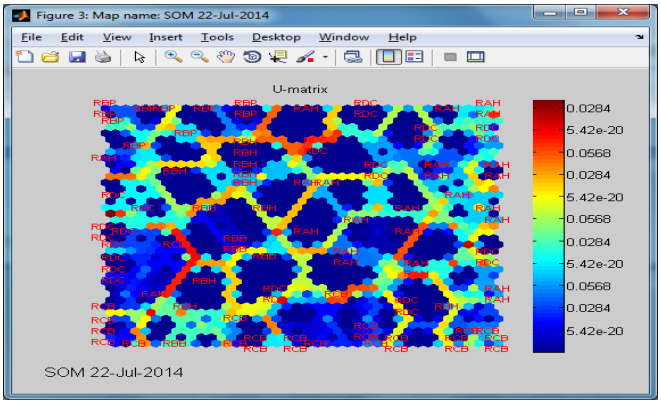

Fig-5: U-Matrix for Triplets

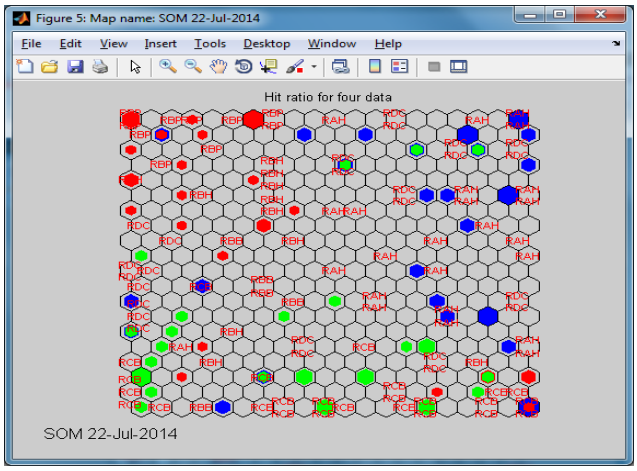

Fig-6: Hit Ratio for Four Data

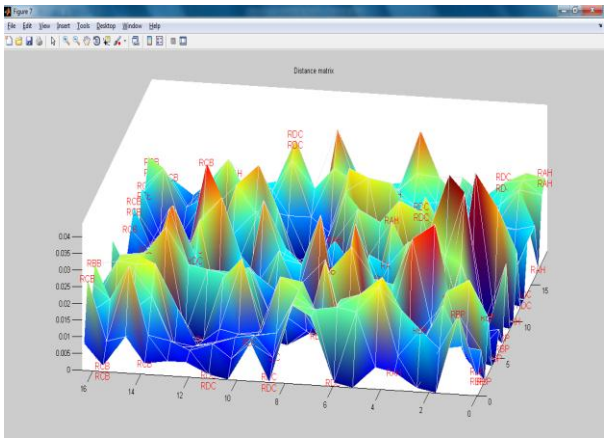

Fig-7:Distance Matrix

V. CONCLUSION AND FUTURE SCOPE

In this paper, we analyse the problem of approach to find the serotypes. In modern software implementations of artificial neural networks, the approach inspired by biology has been largely abandoned for a more practical approach based on statistics and signal processing. In some of these systems, neural networks or parts of neural networks (such as artificial neurons) are used as components in larger systems that combine both adaptive and non-adaptive elements. Self-organizing maps has been widely used in clustering applications. However, SOM approach is applied to single and local datasets. To solve problems to which neural network analysis has been applied successfully**.** In this analysis virus dataset have been used, that is rotavirus. Final results were compared with the trained dataset. In that four types of virus group are clustered and also some of the sub-clusters are found in each type of virus [11, 12]. Finally, the virus data will be used linear vector quantization or adaptive resonance theory to cluster the virus type to easily find serotype.

REFERENCES

[1] T. Kohonen, "The self-organizing map", In the Proceedings of the IEEE, 78, pp. 1464–1480, 1990.

[2] A. Ultsch, "Self-Organizing Neural Networks for Visualization and Classification", In: O. Opitzet al. (Eds). Information and Classification, Springer Berlin, pp. 301–306, 1993.

[3] J. Vesanto, "Using SOM in Data Mining", In Licentiate's Paper, Helsinki University of Technology, Espoo, Finland, 2000.

[4] J. A. F. Costa and M. L. de Andrade Netto, "Clustering of complex shaped data sets via Kohonen maps and mathematical morphology", In: B. Dasarathy (Ed.), Proceedings of the SPIE, Data Mining and Knowledge Discovery, 4384, pp. 16-27, 2001.

[5] F. L. Gorgônio and J. A. F. Costa, "Parallel Self-Organizing Maps with Applications in Clustering Distributed Data", In the International Joint Conference on Neural Networks (IJCNN'2008), 2008.

[6] S. Haykin, "Neural networks: A comprehensive foundation", 2nd ed., Macmillan College Publishing Company, New York, 1999.

[7] T. Kohonen, "Emergence of invariant-feature de-tectors in the adaptive-subspace self-organizing map", Biological Cybernetics, Vol.75, pp.281-291, 1996.

[8] T. Kohonen, "Self-Organizing Map", Springer-Verlag, 1995.

[9] R. Xu and D. Wunsch II, "Survey of Clustering Algorithms", IEEE Trans. on Neural Networks, Vol.16, Issue. 3, pp. 645-678, 2005.

[10] F. L. Gorgônio and J. A. F. Costa, "Parallel Self-Organizing Maps with Applications in Clustering Distributed Data", In International Joint Conference on Neural Networks (IJCNN'2008), 2008.

[11] P. Berkhin, "A Survey of Clustering Data Mining Techniques", In: J. Kogan et al., Grouping Multi-dimensional Data: Recent Advances in Clustering. Springer-Verlag, New York, 2006.

[12] T. Francis Thamburaj, "Analysis of Genome Signature Strength of SARS Coronavirus som neural network", IEEE, 2010.

Authors Profile

R. Vijayalakshmi M.Sc., MPhil., SET, She is currently working as Assistant Professor in Department of Computer Science, Sengamala Thayaar Educational Trust Women's College, Mannargudi. Her subjects of interests are Software Engineering, Programming Languages, Data Structures and Object Oriented Programming. She has qualified SET exam. She is a member of IACSIT, ISQEM.

S. Isabella M.Sc., MPhil., B.Ed, She is currently working as Assistant Professor in Department of Computer Science, Sengamala Thayaar Educational Trust Women's College, Mannargudi. She has 5 years experience in teaching.

