Rough Computing Models for Acute Dengue Patients Gene Expression Data Analysis

[1]Dr. E. N. Sathishkumar, [2]Dr. K. Thangavel, [3]P. S. Raja
[1]Guest Lecturer, [2]Professor and Head, [3]Research Scholar
[1]Department of Computer Science,
[1]Periyar University, Salem – 636 011, India

Abstract: The analysis of gene expression datasets discover local structures composed by set of genes which lead to the development of sophisticated algorithms capable of extracting novel and useful knowledge from biomedical point of view. In the medical domain, these patterns are useful for understanding various diseases, and aid in more accurate diagnosis, prognosis, treatment planning, as well as drug discovery. Statistical measures evaluate a gene set theoretically, but the biological significance proves the real quality of the extracted genes. The Gene Ontology (GO) tool renders the biological significance in terms of function of the genes in the gene set. In this paper Rough Computing models such as Rough K-Mean Quick Reduct (RKMQR) and Weighted Rough Neural Network (WRNN) proposed and implemented for dengue gene expression dataset to find the marker genes. The biological significance of selected genes such as Biological Process, Cellular Component and Molecular Function are discussed and analyzed**.**

IndexTerms **- Rough Set, Gene Selection, Clustering, Classification, Dengue, Gene Expression Data.**

I. INTRODUCTION

Bioinformatics is an interesting area where many knowledge discovery tasks can be applied, among them, supervised and unsupervised classification (clustering) [1]. It is an inter-disciplinary field with roots in biology, statistics, mathematics and computer science. Simply, it can be described as the processing of biological information to solve problems within biological systems. From a computational perspective, bioinformatics carries its focus directly on the informatics activities on persistent data [2, 3, 4, 5]. The aim of bioinformatics is in achieving understanding of biological systems to help further the lives and possibilities of organisms. Richon [4] describes these aims as explaining:

- The processes of biological systems;
- The malfunctions of these processes that lead to diseases;
- Approaches to aid the discovery and development of drugs.

A problem with gene expression analysis is often the selection of significant genes within the data set that would enable accurate classification of the data to some output classes. These genes may be potential diagnostic markers too. A feature selection method based on rough set theory is studied for reducing genes from large gene expression database. The following are the good reasons for reducing the large number of genes:

- It will improve the quality of data for predictive accuracy and algorithm time performance
- Reducing the number of redundant and unnecessary genes can improve inference and classification
- More interpretable genes that can help identify and monitor the target diseases or function types
- To improve the comprehensibility of the output results and reduce the computational cost
- There is an opportunity to examine individual genes for further medical treatment and drug development

Feature selection (FS) is used synonymously with subset selection [6] and attribute selection [7], and in the field of biology, is termed discriminative gene selection [8, 9, 10, 11]. It is intimately related to dimension reduction. The intention of feature selection is to discover significant and informative features in the data set and discard any other feature as irrelevant, noisy and redundant information. In some cases, too many redundant or irrelevant features are overheads of main features for classification. In this study, Rough Set based feature selection is applied to select the genes which are highly expressed. Pawlak's Rough Set Theory [12, 13] has been used in bioinformatics to approach the task of feature selection by many researchers. Using three discrete degrees of correctness (vagueness), essentially a three-valued logic and robustness to imperfect data, it can give approximations to the statistical belief in the features (patterns). Depending on the measure used, we can then identify the usefulness of the pattern or rule and then return this information to the biologists for high level analysis on the data. The major interest of this research is investigating the application of rough set for feature selection to dengue gene expression data. Throughout this paper, the terms attribute, feature, variable and gene are used interchangeably.

The organization of this paper is as follows. Section 2 describes the proposed models. The dengue gene expression dataset, the experimental analysis and the results of rough set models are discussed in section 3. Section 4 details the biological significance and gene ontology for selected genes. Finally, the conclusion is drawn in Section 5.

II. ROUGH COMPUTING MODELS

Proposed Rough Computing models have been employed for dengue gene expression dataset to find the marker genes. Figure 1 represents the data flow diagram of proposed model.

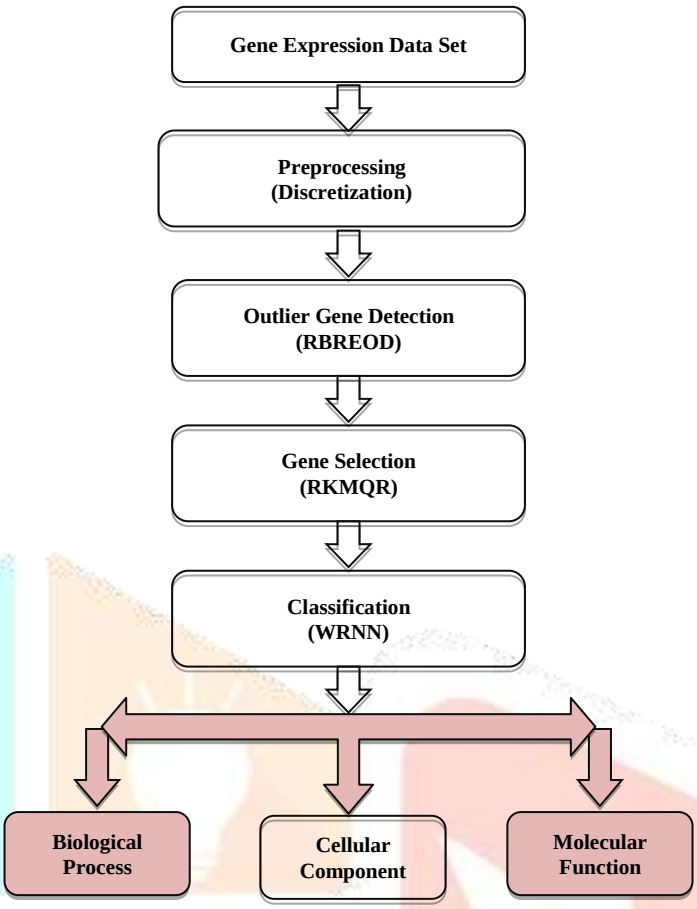

Fig. 1. Flow Diagram of Experimental Analysis

2.1Hybrid Rough K-Mean-Quickreduct (RKMQR) Algorithm

This section details the proposed RKMQR algorithm. The proposed RKMQR algorithm logically consists of three steps: First, to group the similar genes by applying Rough K-Means, because gene expression data may contain uncertain and inconsistent information. In such cases, a gene should belong to more than one cluster and as a result, cluster boundaries necessarily overlap. Second, the representative genes have been selected from resultant clusters using Quick Reduct. Finally, significant gene obtains based on their ACV measures. The RKMQR is described as in Algorithm.

Algorithm 1:RKMQR (C, D)

Inputs: Gene expression data contains *n* genes and a *m* class variable, G = {g_1, g_2, …, g_n} and D= {d_1, d_2, …, d_m}

Output: G_{best} – Selected gene

Step 1: Set $K = 5$ and $G_{best} \leftarrow \{\}$
Step 2: Do
Compute gene wise cluster, Rough K-Means(G, *K*) to select $GC_i = \{g_1, g_2, \ldots, g_q\}$,
End
Step 3: Discretize each GC_i using K-Mean Discretization
Step 4: for i = 1 to *K*
Compute reducts $RC_i = \{R_1, R_2, \ldots, R_r\}$ using Quick Reduct Algorithm
End
Step 5: Compute ACV for all refined RC_i
Step 6: Collect all the genes from each cluster, where ACV = 1
$R_k = \{RC_i / ACV(RC_i) = 1\} = \{RG_1, RG_2, \ldots, RG_r\}$
Step 7: Repeat step 2 to 6, for $K = 7$ and 10.
Step 8: $G_{best} = \bigcap_{k \epsilon Rk} R_k$
Step 9: Return G_{best}

2.2Weighted Rough Neural Network Classifier

In this section, a novel method Weighted Rough Neural Network (WRNN) is proposed to handle inconsistent, uncertain and class imbalance dataset. Imbalanced data means that one of the classes has more samples than the other classes. The class with more samples is called the majority class while the other is the minority class. Most classification techniques perform poorly with the

minority class. There are three suggested techniques to overcome imbalanced data problems. The first is to create or modify the existing classification algorithms to deal with class imbalance problems. Data re-sampling is the second technique which includes over sampling or under sampling the data set to adjust the size of data set. The last approach is the feature selection, which is used to select a subset of features that allow the classifier to reach optimal performance [14].

2.2.1Weighted Rough Neural Network Algorithm:

Rough set and Neural Networks can solve the complex and high dimensional problems, which are called RNNs [15, 16, 17]. A rough neuron can be viewed as a pair of neurons, in which one neuron corresponds to the upper boundary and the other corresponds to the lower boundary. Upper and lower neurons exchange information with each other during the calculation of their outputs. In WRNN algorithm, the rough set theory is integrated with backpropagation network to classify gene expression data. In the traditional rough set, all samples have equal weight without considering the distribution of samples. Here, we apply Class equal Sample Weighting (CSW) scheme [18] to build a weighted decision system WIS = (U, Aw, Cw, D). By using this method, samples belonging to majority class have smaller weight while samples in the minority class have larger weight. Based on the weighted information system we perform the 'inference decision making'. The weighted values are to be discretized since rough set based classification is proposed. The boundary values wBNDC(x), of weighted information system are considered as uncertain values, and inference decision making is done based on the similarity measure. The similarity measure is evaluated for elements of boundary with the centroid of each class lower approximation, and the decision value is updated according to the closest centroid. The algorithm of WRNN classifier method is described in Algorithm 2.

Algorithm 2 : WRNN (C, D)
Input:IS = (U, A, C, D) be a decision system data,C - Conditional attributes, D - Decision attribute,
Output: Predicted Decision attribute

Step 1: Set, *WIS* ← [ ], *ND* ← [ ]
Step 2: Do
Step 3: For every $a_i \in A$
Step 4: For every $d_j \in D$
Step 5: If $a_i \in d_j$then
Step 6: $$w_i \leftarrow \frac{1}{(n(D) \times n(A_j))}$$
where, n(D) – No. of decision classesn(A_j) – No. of samples classified as d_j
Step 7: Form Weighted Information System,
$WIS \leftarrow a_i \times w_i$ here, $WIS = (U, A_w, C_w, D)$
end
end
Step 8: Discretize *WIS* using K-Means clustering
Step 9: Compute the Upper Approximation
$$\bar{R}X_w \leftarrow \{x \in U \mid [x]_{Cw} \cap X \neq \Phi\}$$
Step 10: Compute the Lower Approximation
$$\underline{R}X_w \leftarrow \{x \in U \mid [x]_{Cw} \subseteq X\}$$
Step 11: Compute the Boundary Region
$$wBND_C(x) \leftarrow \cup\bar{R}X_w - \cup\underline{R}X_w$$
Step 12: $wCEN_d(x) \leftarrow Mean(\underline{R}X_w)\ d = 1,2, \ldots, |D|$
Step 13: For every $wBND_C(x)$
Step 14: For every $wCEN_D(x)$
Step 15: $D_{ij} \leftarrow Dist(wCEN_D(x), wBND_C(x))$ here, Dist – Euclidian distance
End
Step 16: $T \leftarrow index(min(D_{ij}, wCEN_D(x)))$
End
Step 17: *Update* $D \leftarrow T$
Step 18: *Call BPN(C, D)*

III. Experimental Analysis

3.1Data Set

In this paper, Acute Dengue Patients: Whole Blood Gene expression data is used for rough computing model experiments, which is obtained from the Gene Expression Omnibus [19], added recently to NCBI in platform GPL13158 with series GSE51808. It is an analysis of blood from patients with Acute Dengue Virus (DENV) infection and during convalescence. Results provide insight into

molecular mechanisms underlying host response to DENV infection. The dataset contains 54,715 total numbers of genes and 56 samples / patients. Out of 56 patients 47 patients are infected with dengue virus and 9 patients are in healthy control. Dengue infected patients who have different disease state such as Convalescent (19), dengue hemorrhagic fever (10), dengue fever (18).

3.2Experimental Results

This experimental analysis involves five major steps such as discretization, outlier gene detection, gene selection, classification and finally gene ontology of selected genes (see fig. 1). Preprocessing of dengue gene expression data set is performed in the first step. It contains continuous variables between the ranges 2.0190 to 14.6196, which represents as low gene expression value is 2.0190 and high expression value is 14.6196. Before applying rough computing models it needs to be discretized continuous features of a dataset. All the genes are discretized using entropy based discretization method. In second step, outlier genes are detected using RBREOD method [20] which discovers outliers from boundary objects using rough entropy measure. We obtained 914 genes as outlier out of 54,715 genes by applying RBREOD. First two steps are common to further experimental analysis. We obtained marker genes from 53,801 genes by applying proposed RKMQR gene selection method. The entire gene expression dataset is divided in to lower and upper approximation by applying Rough K-Means algorithm. The genes in the lower approximation belong certainly to the cluster; therefore they cannot belong to any other cluster. The genes in an upper approximation may belong to other cluster. For this experiment the number of clusters is chosen to be five, seven and ten (K = 5, 7, 10), then dengue data set will divide K number of groups using Rough K-Means clustering techniques. After that, representative genes were selected from a resultant clusters by applying Quick Reduct algorithm.

3.3 Result I

RKMQR divides the whole dengue gene expression dataset into five different clusters. When K = 5, Rough K-Means algorithm chooses five genes as random cluster center, one for each cluster. Figure 2 shows a scatter plot of five co-expressed gene clusters; every point in the plot shows the gene in the clusters. In this experiment 9162 genes are placed in cluster 1 and it is symbolized by green colored point symbol, 23363 genes are placed in cluster 2 and it is represented by magenta colored cross symbol, 3043 genes are placed in cluster 3 and it is represented by cyan colored square symbol, 7345 genes are placed in cluster 4 and it is represented by red colored plus sign symbol and 11802 genes are placed in cluster 5 and it is represented by blue colored diamond symbol.

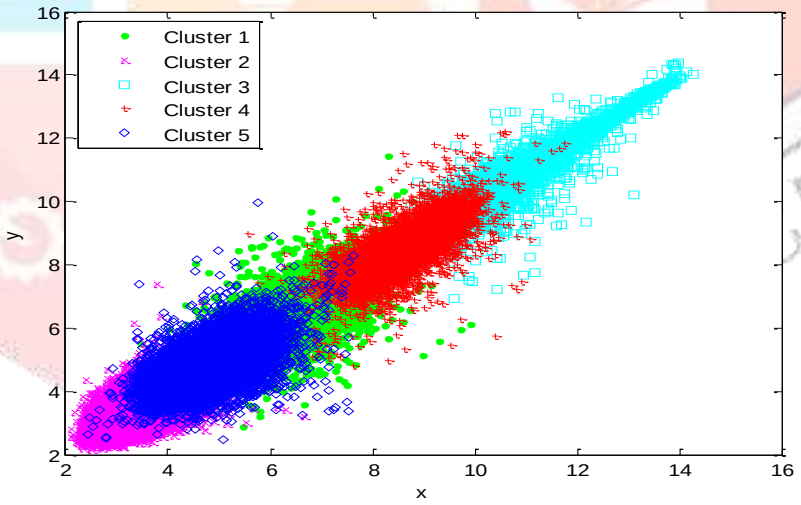

Fig. 2. Scatter plots of five clusters by applying RKMQR

Table 1: List of Selected Genes When K = 5 by using RKMQR

Cluster	No. of Genes (GCi)	Selected Genes
Cluster 1	9162	'1552632_PM_a_at', '221599_PM_at', '229891_PM_x_at'
Cluster 2	23363	1560864_PM_at', '227940_PM_at', 244668_PM_at'
Cluster 3	3043	'1405_PM_i_at', '208647_PM_at' '210951_PM_x_at' , '214665_PM_s_at'
Cluster 4	7345	'1554390_PM_s_at', '202944_PM_at' '203594_PM_at', '218696_PM_at'

Cluster 5	11802	'1552279_PM_a_at', '1556758_PM_at' '213954_PM_at', '218051_PM_s_at'

The number of genes grouped in each cluster and selected genes are tabulated in table 1 for the experiment performed when K = 5 by applying RKMQR. Three representative genes selected from Clusters 1 and 2 and four representative genes selected from each resultant clusters 3, 4 and 5. It is observed from cluster 2 that genes 1560864_PM_at, 227940_PM_at and 244668_PM_at are having low expressed values between 2.5 to 4.5 and cluster 3 genes 1405_PM_i_at, 208647_PM_at, 210951_PM_x_at and 214665_PM_s_at are having highly expressed values among 8.5 to 12.5. PC plot visualizes that gene '1405_PM_i_at' is highly expressed when compare with other representative genes.

3.4 Result II

Rough K-Means algorithm uses seven genes as random cluster centers, when K = 7. Figure 3 shows a scatter plot of seven co-expressed gene clusters. In this experiment 1657 genes are placed in cluster 1 and it is symbolized by green colored point symbol, 7519 genes are placed in cluster 2 and it is represented by magenta colored cross symbol, 11487 genes are placed in cluster 3 and it is represented by cyan colored square symbol, 6575 genes are placed in cluster 4 and it is represented by red colored plus sign symbol, 4322 genes are placed in cluster 5 and it is represented by blue colored diamond symbol, 5960 genes are placed in cluster 6 and it is represented by black colored asterisk symbol and 17195 genes are placed in cluster 7 and it is represented by green colored plus sign symbol.

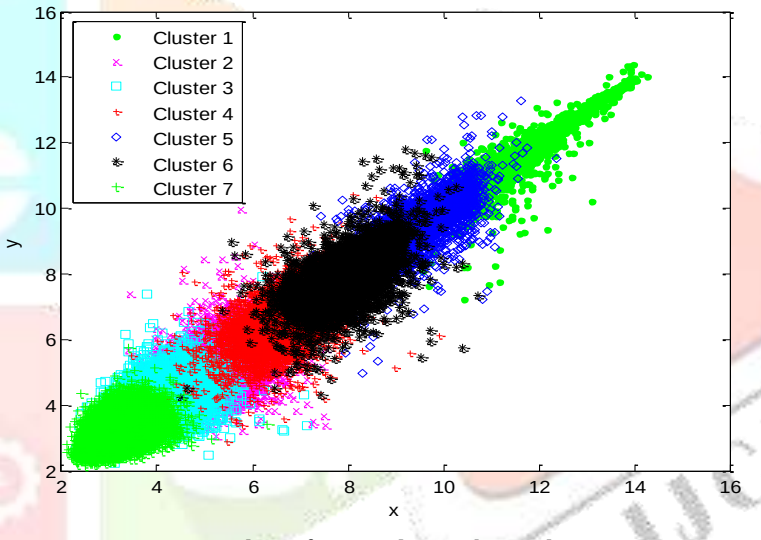

Fig. 3. Scatter plots of seven clusters by applying RKMQR

Table 2: List of Selected Genes When K = 7 by using RKMQR

Cluster	**No. of Genes (GCi)**	**Selected Genes (ID_Ref)**
Cluster 1	1657	'1405_PM_i_at','204588_PM_s_at', '207668_PM_x_at','216988_PM_s_at'
Cluster 2	7519	'1552256_PM_a_at','1560488_PM_at', '218051_PM_s_at','239666_PM_at',
Cluster 3	11487	'1316_PM_at','1566171_PM_at', '226499_PM_at','243045_PM_at'
Cluster 4	6575	'1487_PM_at','201831_PM_s_at', '212122_PM_at','231768_PM_at'
Cluster 5	4322	'200868_PM_s_at','212052_PM_s_at', '221637_PM_s_at'
Cluster 6	5960	'1552277_PM_a_at','216652_PM_s_at '225451_PM_at','235339_PM_at'
Cluster 7	17195	'1552359_PM_at','1555988_PM_a_at', '1561817_PM_at','233202_PM_at'

The genes grouped in each cluster and selected genes are tabulated in table 2 for the experiment performed when K = 7 by applying RKMQR. Four representative genes selected from Clusters 1, 2, 3, 4, 6 and 7 and three representative genes selected from cluster 5. It is observed from cluster 7 that genes '1552359_PM_at', '1555988_PM_a_at', '1561817_PM_at', and '233202_PM_at' are having low expressed values between 2.5 to 6 and cluster 1 genes '1405_PM_i_at', '204588_PM_s_at', '207668_PM_x_at' and '216988_PM_s_at' are having highly expressed values among 8.5 to 12.5. PC plot shows that gene '1405_PM_i_at' is highly expressed when compare with other representative genes..

3.5 Result III

This section elucidates the clustering solution obtained from one of the runs (when K=10) of RKMQR algorithm on dengue gene expression data. RKM algorithm chooses randomly ten genes as cluster centers, one for each cluster. The results are illustrated in figure 4 using scatter plot. Every point in the plot shows the gene in the respective clusters. In this experiment 4893 genes are placed in cluster 1 and it is symbolized by cyan colored point symbol, 4445 genes are placed in cluster 2 and it is represented by magenta colored cross symbol, 10234 genes are placed in cluster 3 and it is represented by cyan colored square symbol, 10932 genes are placed in cluster 4 and it is represented by red colored plus sign symbol, 2238 genes are placed in cluster 5 and it is represented by blue colored diamond symbol, 5496 genes are placed in cluster 6 and it is symbolized by blue colored asterisk symbol, 3625 genes are placed in cluster 7 and it is represented green colored downward-pointing triangles symbol, 7314 genes are placed in cluster 8 and it is represented by black colored pentagram symbol, 4738 genes are placed in cluster 9 and it is represented yellow colored asterisk symbol and 800 genes are placed in cluster 10 and it is represented by black colored circle symbol.

The genes grouped in each cluster and selected genes are tabulated in Table 3 for the experiment performed when K = 10 by applying RKMQR. Three representative genes selected from Clusters 2, 3 and 10 and four representative genes selected from each resultant clusters 1, 4, 5, 6, 7, 8 and 9. It is observed from cluster 4 that genes '1255_PM_g_at', '1562901_PM_at', '216516_PM_at' and '241079_PM_at' are having low expressed values between 2.2 to 4.2 and cluster 10 genes '203012_PM_x_at', '205922_PM_at' and '216988_PM_s_at' are having highly expressed values among 10 to 13.5.

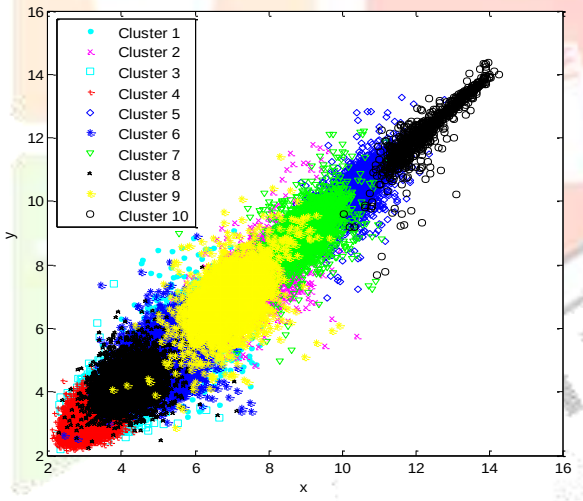

Table 3: List of Selected Genes When K = 10 by using RKMQR

Cluster	No. of Genes (GCi)	Selected Genes (ID_Ref)
Cluster 1	4893	'1007_PM_s_at','201831_PM_s_at', '222651_PM_s_at','241863_PM_x_at'
Cluster 2	4445	'1553856_PM_s_at','204023_PM_at', '221761_PM_at'
Cluster 3	10234	'1559623_PM_at','226983_PM_at', '244369_PM_at'
Cluster 4	10932	'1255_PM_g_at','1562901_PM_at', '216516_PM_at','241079_PM_at'
Cluster 5	2238	'1552611_PM_a_at','201078_PM_at', '201136_PM_at','214665_PM_s_at'
Cluster 6	5496	'1552283_PM_s_at','1557433_PM_at', '210690_PM_at','218051_PM_s_at'
Cluster 7	3625	'1294_PM_at','212604_PM_at', '217750_PM_s_at','226278_PM_at'

Cluster 8	7314	'1494_PM_f_at','212157_PM_at', '221991_PM_at','228060_PM_at'
Cluster 9	4738	'1552510_PM_at','202672_PM_s_at', '225712_PM_at','241991_PM_at'
Cluster 10	800	'203012_PM_x_at','205922_PM_at', '216988_PM_s_at'

Table 4: Selected Significant Genes

S.No	ID	Gene Title	Gene Symbol
1	'218051_PM_s_at'	'5"-nucleotidase domain containing 2'	'NT5DC2'
2	'1405_PM_i_at'	'chemokine (C-C motif) ligand 5'	'CCL5'
3	'214665_PM_s_at'	'calcium binding protein P22'	'CHP'
4	'216988_PM_s_at'	'protein tyrosine phosphatase type IVA, member 2'	'PTP4A2'
5	'201831_PM_s_at'	'USO1 vesicle docking protein homolog (yeast)'	'USO1'

Table 4 shows the five significant genes and their gene title, gene symbol. It is observed from the above experimental runs of RKMQR that five genes namely '218051_PM_s_at', '1405_PM_i_at', '214665_PM_s_at', '216988_PM_s_at' and '201831_PM_s_at' are selected as representatives in more than one experiment. Hence, it is ensured that these genes are most significant than other genes in dengue gene expression data set.

3.6 Classification Results

We obtained five genes as marker genes out of 54,715 genes by applying proposed rough computing models. The obtained marker genes are namely 'NT5DC2', 'CCL5', 'CHP', 'PTP4A2', 'USO1' and these genes are given as input to eight classifiers (WRNN, BPN, Naïve Bayes, JRip , J48, RF, K-Star, Decision Table) in this experiment. The performance of the classification are evaluated by using ten fold cross-validations and its measures such as Accuracy, Sensitivity, False Positive Rate, Precision, Recall, F-Measure, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

Table 5: Classification Accuracies Of Marker Genes

Classifiers	WRNN	BPN	Naïve Bayes	JRip	J48	RF	DT	K-Star
Accuracy	98.21	89.28	94.64	94.64	87.50	96.42	82.14	83.93
Sensitivity	98.20	89.30	94.60	94.60	87.50	96.40	82.10	83.90
FPrate	3.70	9.30	10.00	10.00	65.30	9.70	5.20	5.30
Precision	98.30	89.30	95.00	95.00	89.10	96.40	82.70	84.10
Recall	98.20	89.30	94.60	94.60	87.50	96.40	82.10	83.90
F-Measure	98.20	89.30	94.80	94.80	84.00	96.40	82.20	83.90
MAE	1.79	9.47	5.71	5.71	12.50	6.61	9.35	8.78
RMSE	13.36	24.93	18.55	18.55	35.36	15.41	25.20	27.49

The average classification accuracies obtained from the eight classification algorithms are tabulated in Table 5. The experimental results show that the best correctly classified instances are 55 (98.21%) out of 56 samples by WRNN and the lowest correctly classified instances is 46 (82.14%) by Decision Table algorithm. The BPN, Naïve Bayes, JRip, J48, Random Forest and K-Star give 89.28%, 94.64%, 94.64%, 87.50%, 96.42% and 83.93% accuracy respectively. Minimum false positive rate 3.7 is achieved with WRNN classifier and highest false positive rate 65.30 recognized from J48. The highest true positive rate or sensitivity (98.20) obtained from WRNN.

The classification accuracy of each classifier can be observed from Figure 5 which shows eight evaluation measures. Hence it is observed that the average classification accuracy as 90.84% for those five significant genes. The average of MAE and RMSE of all algorithms are ranging from 7.49 to 22.36. To compare with other methods proposed WRNN produces the lowest mean absolute error 1.79 and lowest root mean squared error 13.36. Finally we obtained average classification accuracy of selected genes is high when compared to the classification accuracy produced by entire genes.

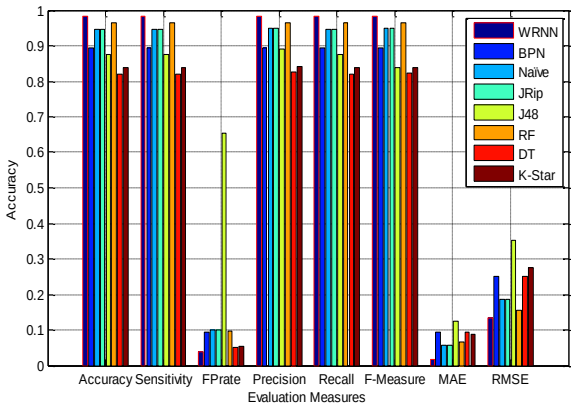

Fig. 5. Performance measures of each classifier using marker genes

IV. Biological Significance of Selected Genes

Gene expression profiles generated by microarrays can help us understand the cellular mechanism of biological process. Once a list of differentially expressed genes has been generated, the next task to determine the biological significance of the genes in that list. The interactions between biological processes are very complicated. One biological process may require involvement of hundreds of genes. One gene may also be involved in many biological processes. Many genes have been studied and their biological processes have been found; however there are still a lot of genes within biological processes whose involvement is unknown, even in well-studied organisms. Biological significance of selected genes such as Biological Process, Cellular Component and Molecular Function is originated using GeneMANIA and AmiGO [21, 22, 23].

4.1GeneMANIA

GeneMANIA helps to predict the function of selected genes and gene sets. It finds other genes that are related to a set of input genes, using a very large set of functional association data. Association data include protein and genetic interactions, pathways, co-expression, co-localization and protein domain similarity. If we enter a query gene list, such as 'NT5DC2', 'CCL5', 'CHP', 'PTP4A2' and 'USO1', GeneMANIA will output connections between query genes and three GO networks. In this network, five query genes are represented by the largest black circles. The graph shows the local neighborhood around the query genes as well as the top predictions. The combined network is constructed from co-expression, co-localization, pathways, genetic and physical interactions, and shared protein domains. Figure 6 illustrates biological process networks of query genes using GeneMANIA. According to this database, 50 genes were found as local neighborhood around the query genes. The following distribution characterized the types of interactions extracted from the BP network: consolidated-pathways 50.56%, co-expression 8.55%, genetic interactions 0.83%, co-localization 1.58%, Pathway 2.47%, physical interactions 32.05%, predicted 3.54%, shared protein domains 0.41%. CCL5 is one of the major participated genes in the top ranked consolidated pathway. The first four consolidated pathways are NOD-like receptor signaling pathway 19.22%, Chemokine receptors bind chemokines 10.83%, Syndecan1mediated signaling events 5.17% and CCR1 5.03%. We observed that gene CCL5 is genetically interacted with gene CCL2.

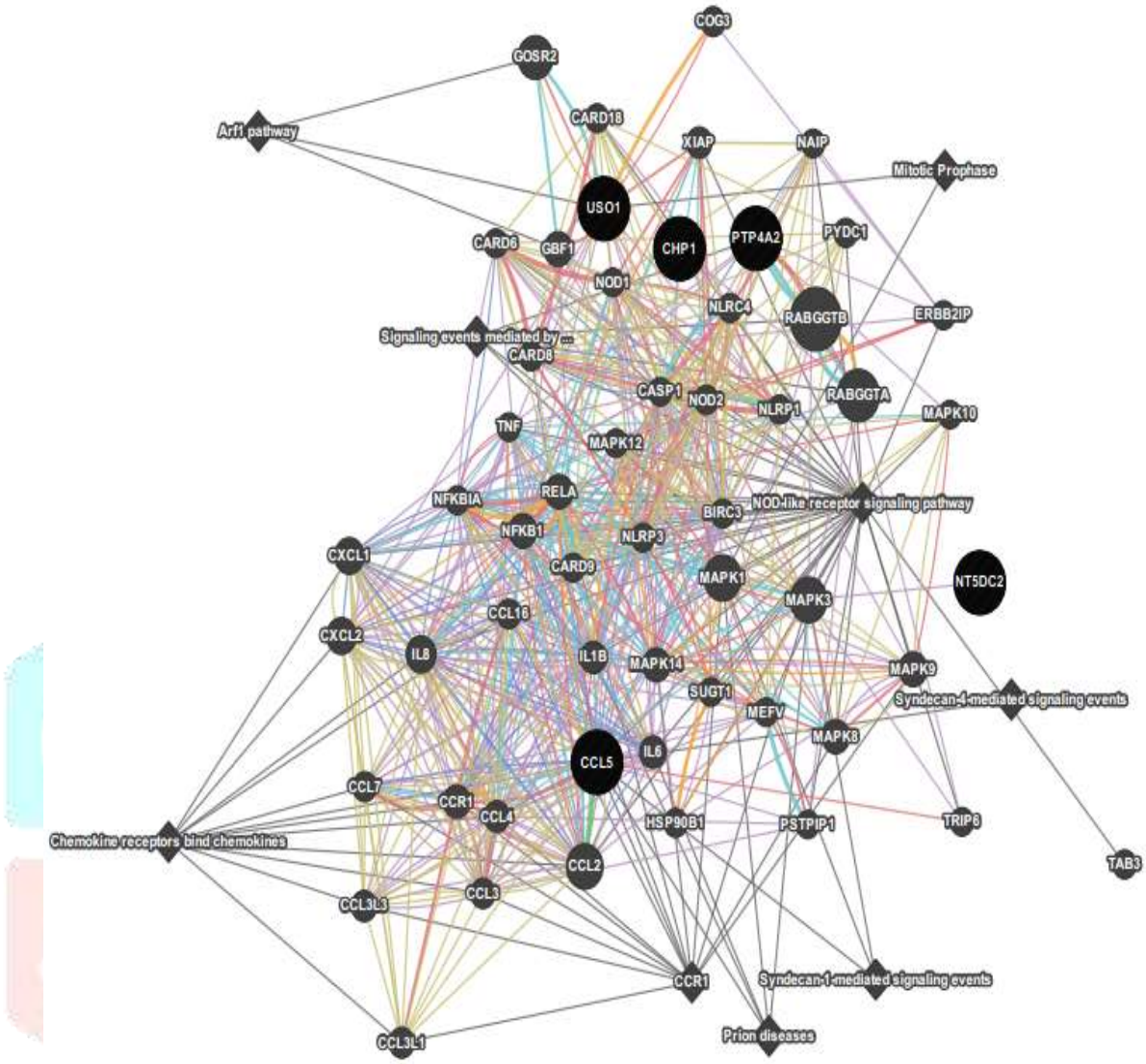

Fig. 6. Biological Process networks of query genes using GeneMANIA

Figure 7 illustrates the cellular component networks of query genes using GeneMANIA. In this network 20 genes were found as local neighborhood around the query genes. The following distribution characterized the types of interactions extracted from the CC network: consolidated-pathways 18.31%, co-expression 11.95%, genetic interactions 0.34%, co-localization 1.85%, Pathway 2.17%, physical interactions 60.13%, predicted 4.84%, shared protein domains 0.39%. CCL5 is one of the major participated genes in the top ranked consolidated pathway. The top most consolidated pathway is syndecan-1-mediated signaling events with 11.00% of distribution. We observed that the gene CCL5 is genetically and physically interacted with neighborhood genes. Compare with other query genes, CCL5 having more number of pathways between neighborhood genes. Those genes are participating in the same reaction within a pathway.

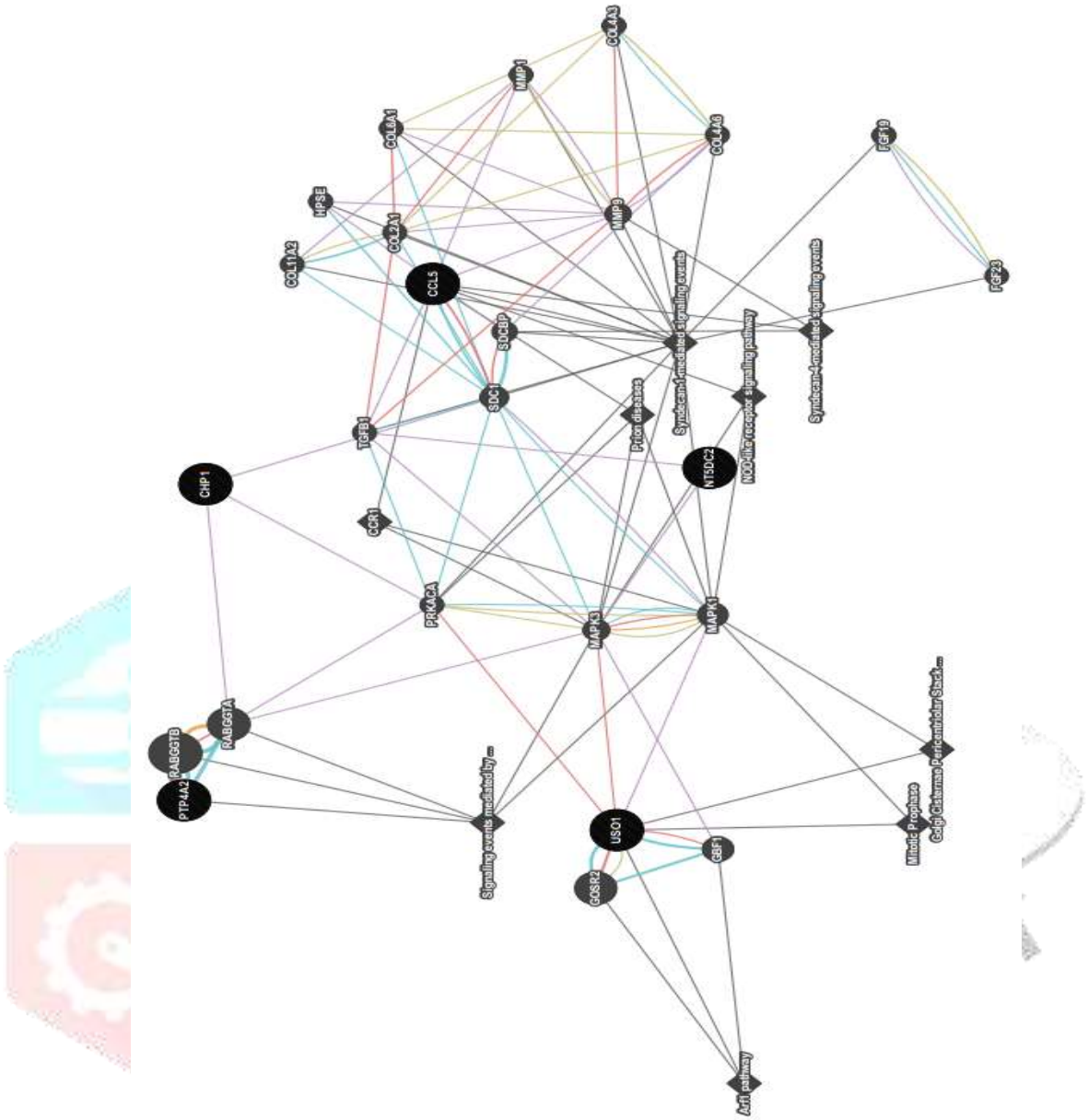

Fig. 7. Cellular Component networks of query genes using GeneMANIA

Figure 8 illustrates the molecular function network which describes the biochemical activity of a gene product. In this network 50 genes were found as local neighborhood around the query genes. The following distribution characterized the types of interactions extracted from the MF network: consolidated-pathways 41.08%, co-expression 6.52%, genetic interactions 1.47%, co-localization 2.45%, Pathway 1.88%, physical interactions 41.41%, predicted 3.41%, shared protein domains 1.79%. The top most consolidated pathway is chemokine receptors bind chemokines with 27.17% of distribution. The query gene CCL5 is one of the major participated genes in the top ranked consolidated pathway. It is observed that the gene CCL5 having number of shared protein domains with neighborhood genes.

We can ensure that from the above GeneMANIA enrichment results, gene CCL5 is biologically significant than other query genes. Further we extend the GO term enrichment analysis for CCL5 using AmiGO.

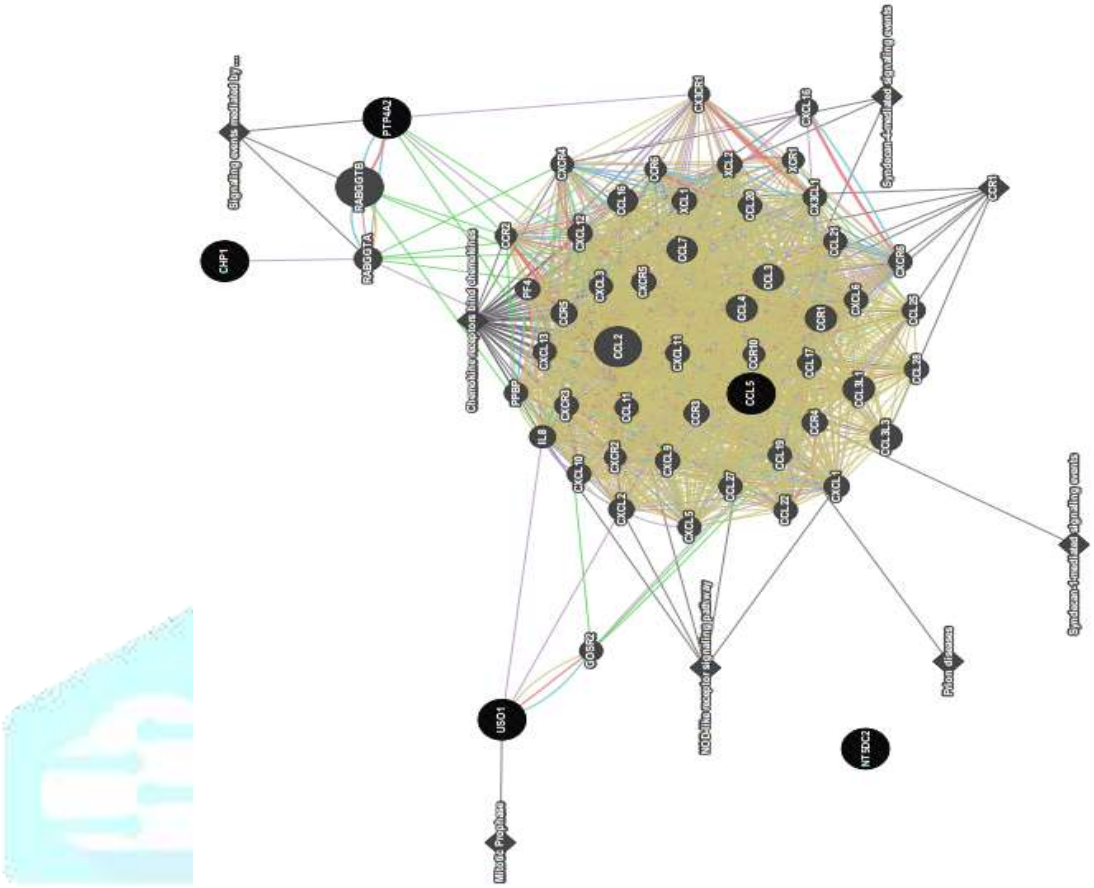

Fig. 8: Molecular Function networks of query genes using GeneMANIA

4.2AmiGO

AmiGO is a web application that allows users to query, browse and visualize ontologies and related gene product annotation (association) data. AmiGO can be used online at the Gene Ontology website to access the data provided by the GO Consortium [23]. The comparison of AmiGO biological significance for selected query genes namely 'NT5DC2', 'CCL5', 'CHP', 'PTP4A2' and 'USO1' is tabulated in Table 6. It is evident that CCL5 construct more biological significance than other query genes.

Table 6: Comparison Biological Significance of Selected Genes

S.No	Gene Symbol	BP	CC	MF	Total
1	'NT5DC2'	0	0	2	2
2	'CCL5'	28	4	7	39
3	'CHP'	4	2	4	10
4	'PTP4A2'	2	5	7	14
5	'USO1'	5	6	3	14

The gene CCL5 extracted using rough computing model and is responsible for various biological processes. The CCL5 query gene has 28 significant ontologies related to the biological process. A part of the GO annotations, GO: 0031622 is concerning the positive regulation of fever generation which is one of the biological process of acute dengue patients: whole blood gene expression data set. It contains 76 gene products and it defined as any process that activate or increase the frequency, rate, or extent of fever generation. CCL5 plays a role in up-regulation of fever, and it regulates osteoclast and chemotaxis.

Figure 9 provides graph view of term neighborhood for positive regulation of fever generation (GO:0031622). The ancestors of positive regulation of fever generation are tabulated in Table 7. Positive regulation of fever is one of the biological processes of query gene CCL5.

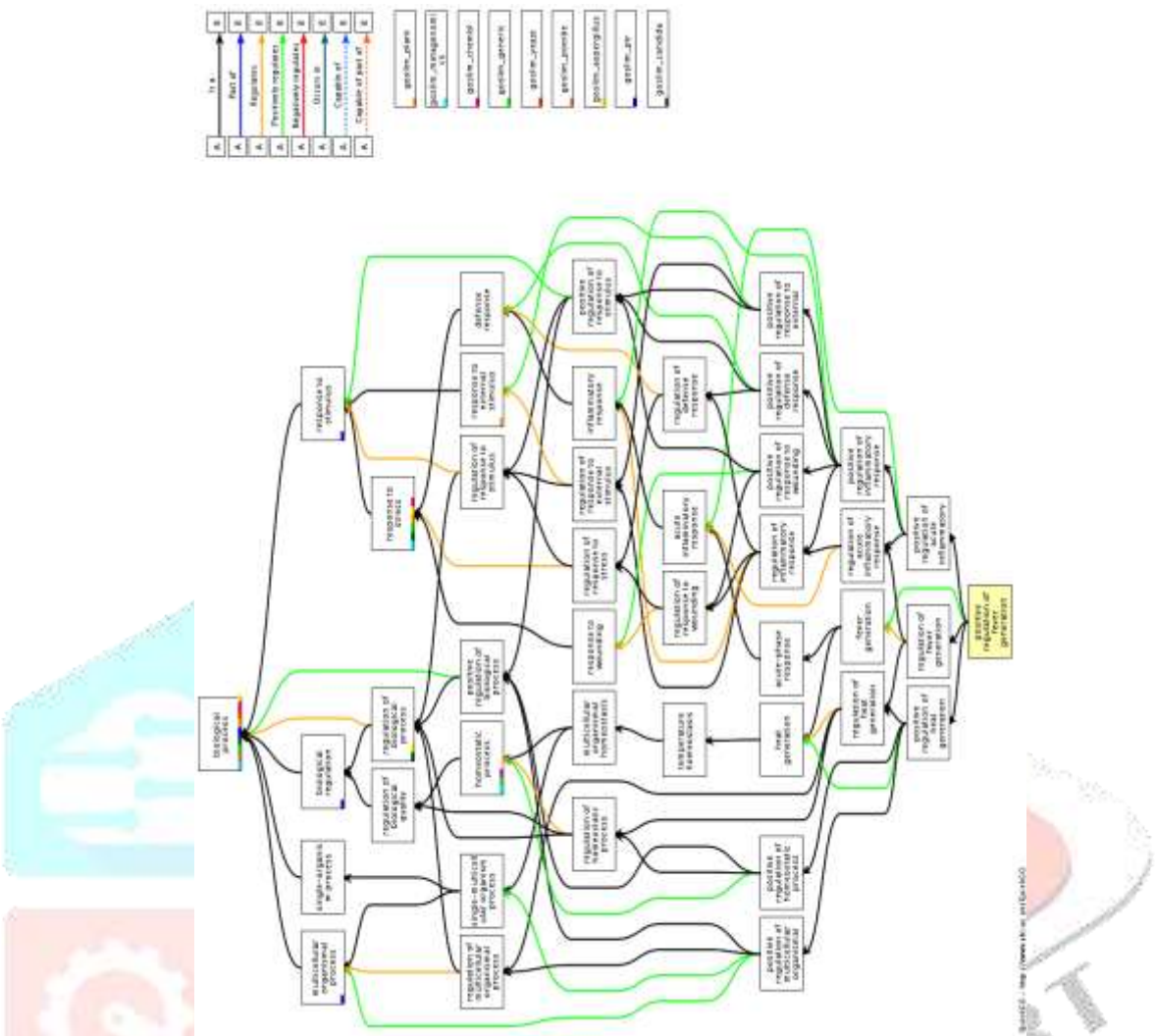

Fig. 9. Graph view of term neighborhood for positive regulation of fever generation (GO: 0031622)

Table 7: Ancestors Of Positive Regulation Of Fever Generation (Go: 0031622)

Relation	Object	Annotations
positively_regulates	fever generation (GO:0001660)	138
positively_regulates	heat generation (GO:0031649)	195
positively_regulates	Biological regulation (GO:0065007)	280420
positively_regulates	biological_process (GO:0008150)	992572
is_a	positive regulation of heat generation (GO:0031652)	93

V. CONCLUSION

The rough computing model has been proposed for improving the gene selection method in a simple and efficient way. In this paper, the informative genes (CCL5, NT5DC2, CHP, PTP4A2 and USO1) are selected from acute dengue patients: whole blood gene expression dataset by applying proposed rough set model. The experimental results showed that a proposed rough computing model provides highest accuracy of 98.21%, and minimum false positive rate of 3.7%. The biological significance of selected genes such as Biological Process, Cellular Component and Molecular Function is originated using GeneMANIA and AmiGO. It is observed that CCL5 plays an important role in positive-regulation of dengue fever and constructs more biological significance than other query genes. The proposed rough computing models gives sparse and interpretable classification accuracy and biological significance compared to the other gene selection methods.

VI. ACKNOWLEDGMENT

The present work is supported by Special Assistance Programme of University Grants Commission, New Delhi, India (Grant No. F.3-50/2011(SAP-II).

REFERENCES

[1] Jain A. K., Murthy M. N., Flynn P. J., "Data clustering: A Review", ACM Computing Surveys, Vol. 31(3), pp. 265-323, 1999.
[2] Jiang, F., Sui, Y., Cunge: "Outlier Detection Using Rough Set Theory", Springer, Heidelberg, pp. 39 – 57, 2005.
[3] Jinyan Li., Limson Wong., Qiang Yang., "Data Mining in Bioinformatics", IEEE Computer Society, pp. 16-18, 2005.
[4] Richon A. B., "A Short History of Bioinformatics", Network Science, http://www.netsci.org/ Science/ Bioinform/feature06.html, 2005.
[5] Ruchi Singh., Richa Sharma., "Bioinformatics: Basics, Algorithms and Applications", University Press, 2010.
[6] Sewell M, "Feature Selection", http://machine-learning.martinsewell.com/ feature-selection/feature-selection.pdf, pp. 1 – 5, 2007.
[7] Witten I. H. and Frank E., "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufman, Second Edition, 2005.
[8] Jianping Lu, "A novel feature selection method based on data normalization", International Conference on Computer Application and System Modeling (ICCASM), pp. Vol. 6, pp. 422-425, 2010.
[9] Kulkarni A., B. S. C. Naveen Kumar, V. Ravi et al., "Colon cancer prediction with genetics profiles using evolutionary techniques," Expert Systems with Applications, vol. 38, No. 3, pp. 2752-2757, 2011.
[10] Richard Jensen, "Combining rough and fuzzy sets for feature selection", University of Edinburgh, Vol. 149, Issue 1, pp. 5-20, 2005.
[11]Zhu Y., H. Li, D. Miller et al., "caBIGTM VISDA: Modeling, visualization, and discovery for cluster analysis of genomic data," BMC Bioinformatics, Vol. 9, No. 1, pp. 383 - 396, 2008.
[12]Pawlak. Z, "Rough Sets" International Journal of Computer and Information Sciences, pp. 341-356, 1982.
[13]Pawlak Z., "Rough set approach to knowledge-based decision support," European Journal of Operational Research, Vol. 99, No. 1, pp. 48–57, 1997.
[14]Rushi L., Snehlata S. D., Latesh M., "Class Imbalance Problem in Data Mining: Review", International Journal of Computer Science and Network (IJCSN), Vol. 2, pp. 226-230, 2013.
[15]Dongbo Zhang, "Integrated methods of rough sets and neural network and their applications in pattern recognition", Hunan University, pp. 256 – 272, 2007.
[16]Rajakeerthana K. T., Velayutham C. and Thangavel K., "Mammogram Image Classification Using Rough Neural Network", ICC3, Advances in Intelligent Systems and Computing, 246, pp. 133-138, Springer India, 2014.
[17]Weidong Zhao and Guohua Chen, "A survey for the integration of rough set theory with neural networks", Systems engineering and electronics, Vol. 24, No. 10, pp. 103-107, 2002.
[18]Hala S. Own, Ajith Abraham, "A Novel-weighted Rough Set-based Meta Learning for Ozone Day Prediction", Acta Polytechnica Hungarica, Vol. 11, No. 4, pp. 59 – 78, 2014.
[19]Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/
[20]E.N.Sathishkumar and K.Thangavel, "Rough Set Based RBREOD Algorithm for Outlier Detection", International Conference on Signal and Speech Processing (ICSSP) on 21, 22 and 23 August 2014.
[21]Ashburner, M., Ball, C., Blake, J., and et al., "Gene Ontology: tool for the unification of biology", Nature Genetics, pp. 25-29, 2000.
[22]David Warde-Farley et.al. "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function", Nucleic Acids Research, pp. 214 – 220, Vol. 38, 2010.
[23]Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, "AmiGO Hub; Web Presence Working Group", Amigo: online access to ontology and annotation data, Bioinformatics, Vol. 25, Issue 2, pp. 288–9, 2009.