International Journal of Engineering & Technology, 7 (2.27) (2018) 12-16

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Parallel framework based gene signature-hierarchical random forest cluster for predicting human diseases

N. K. Sakthivel [1] *, N. P. Gopalan [2], S. Subasree [3]

[1] Research Scholar, Bharath University, Chennai – 600 073, Tamil Nadu, India
[2] Professor, Department of Computer Applications, National Institute of Technology, Trichy, TN, India
[3] Professor & Head, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Pampady – 680 588, Kerala, India
**Corresponding author E-mail: nksakthivel@gmail.com*

Abstract

Gene is not responsible for many Human Diseases and instead, diseases occur by different or group of genomes interacting together and cause diseases. Hence it is need to analyse and associate the complete genome sequences to understand or predict various possible human diseases. This research work focused i. Hierarchical-Random Forest based Clustering (HRF-Cluster), ii. Genetic Algorithm-Gene Association Classifier (GA-GA) and iii. Weighted Common Neighbor Classifier (wCN). These Classifiers were implemented and studied thoroughly in terms of Prediction Accuracy, Memory Utilization, Memory Usage and Processing Time. To improve the performances of the Gene Classifiers / Predictors further, this research work was proposed and implemented Gene Signature based HRF Cluster, G-HR. Results show that that the performances of the proposed Classifier G-HR is outperforming as compared with the identified three Classifiers in terms of Disease Pattern Prediction, Processing Time, Memory Usage and Classification Accuracy. To improve the performance of the system further in term of Processing Time, the proposed model G-HR is implemented under Parallel Framework and evaluated. That is the model is tested with Two, Four, Eight and Sixteen Parallel Processors and from the results, it is established that the Processing Time decreases considerably which will improve the performance of the Proposed Model.

Keywords: *Gene Association; Genetic Algorithm; Hierarchical Clustering; Human Genome Prediction; Parallel Framework; Random*

1. Introduction

DNA Microarrays have designed for measuring the transcriptional levels of DNA and RNA transcripts. These RNA transcripts were derived from group of genes of a genome [1-5]. The signature of Gene expression in the biomedical field used to identify a few human disease patterns [1], [2], [10]. Associating genes with genotypes or phenotypes is demanding research topic in bioinformatics which is called as disease-gene association research. This might be called as identification or prediction of disease genes.

Even though there are numerous classification mechanisms like Support Vector Machine, Neural Networks proposed to study and predict the complex disease patterns, we needed further more Models to achieve higher prediction accuracy [12-17]. It is noted from the literature survey that a few types of diseased-Gene prediction methods (Genetic Disorder) have proposed to identify Genes associated various diseases [18-23], [26], [29-31]. To improve the classification accuracy, a few methods established a set of Known Disease Genes [1], [2], [5], [11], [24], [25], which is used to predict various diseases by Computational Disease Gene Prediction Methods. This work identified recently proposed three classifiers and studied thoroughly. Based on observations, this work proposed an efficient Gene Classifier/Predictor called Gene Signature based HRF Cluster G-HR. This was implemented with Uni-Processor [1], [22], [27], [28] and Parallel Processors as well. The detailed procedures are discussed in the following sections.

This Research paper is arranged and written as follows. The Section 2 briefly described the recently proposed DNA Sequence Classifiers namely i. Hierarchical-Random Forest based Clustering (HRF-Cluster), ii. GA-Gene Association Classifier (GA-GA) and iii. Weighted Common Neighbor (wCN). The proposed Gene Signature based HRF Cluster, G-HR Model implemented in Uni-Processing and Parallel Framework as well is described in Section 3.The results and strengths of the proposed model in Uni-Processing as well as Parallel Processing is discussed at Section 4 and Conclusion was given in Section 5.

2. Sequence classifiers

This work identified three DNA Classifiers[1,2,3,4,5]. namely i. Hierarchical-Random Forest based Clustering (HRF-Cluster), ii. Genetic Algorithm based Gene Association Classifier (GA-GA), iii. Weighted Common Neighbour (wCN) for thorough study. The following sections discussed all the above mentioned three Classifiers.

2.1. Hierarchical-random forest based clustering (HRF-cluster)

The clustering analysis is basically proposed for evaluating the clustering methods and from the literature survey, it was noticed that Multiple Clusters could be employed to improve classification accuracy when Data Sets are larger in size. The Multiple Clusters need to grouped together for cross validation. This was proposed by

Copyright © 2018 N. K. Sakthivel et. al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

the researcher[1,2]. In that model, all Clusters have been analysed through Training Models with the known Test Data Patterns. The Euclidean distance method was used to calculate the similar and closer Clusters. This is an appropriate approach to construct closest cluster.
The Euclidean distance was calculated with the following equation 1.

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (1)$$

Hierarchical Clustering method was facilitating to group smaller closet clusters to relatively bigger one to compare comfortably. Since a few clusters practically have less number of points, it is necessitate to group all these smaller groups together.
To avoid over-fitting, the Random Forest Classifier is chosen by the author for this clustering approach. This is enabling the model to improve classification accuracy. A few strengths such as higher classification accuracy, parallel clustering and minimizing errors and outliers of the Hierarchical Random Forest approach was established [1-3].

2.2. Genetic algorithm based gene association classifier (GA-GA)

In this Genetic Algorithm based Gene Association study, the authors Koosha Tahmasebipour and et. al. [1], [2], [4] proposed a Genetic algorithm based Computational Disease-Gene Association method. From the literature survey, we noticed that the Genetic algorithms were suitable proven model to solve NP problems [1-3]. It generally tries to develop a population of candidate solutions ie chromosomes for the problem. According to this model, each and every candidate solution is assigned a fitness value and these values can be measured by fitness function. In this Genetic Model, the Mutation is representing ie mimicking the possible mutations when new individuals have been reproduced.
This method is trying to develop a group of candidate genes with the entire set of already-known disease genes. As represented in the Equation 1 which is called as Modularity Function Q, that developed by the author Luo [1], [2], [9] the modularity of the community is optimized from created Gene communities and the same can be optimized ie communities can be optimized. The Equation for the Modularity is given below.

$$Q(C) = \sum_{i \in C} \frac{K_i^{in}(C)}{K_i(C) = K_i^{in}(C) + K_i^{out}(C)} \quad (2)$$

Where

$$K_i^{in}(C) = \sum_{j \notin C} E(i, j) \quad (3)$$

This is representing the number of edges connecting i to another remaining nodes in C and

$$K_i^{out}(C) = \sum_{j \notin C} E(i, j) \quad (4)$$

This equation 4 is representing the number of edges connecting i to the nodes which are not exit in C. The concept of the scoring system which is as shown in the Equation 1 proposed by the authors [1,9] is that the Genes which are frequently selected to be in the optimizing communities will have more association with disease genes and this will have more score also as discussed.

2.3. Weighted common neighbor (WCN)

The Common Neighbors is the Clustering Scheme that used for considered as one of the frequently using local measures. It is also noticed that this model is for improving prediction accuracy, we could introduce weighted links.
From local measures, the authors [1], [2], [4-5] considered Common Neighbors and its weighted links and variants as Weighted Common Neighbors.
In each and every pair of disconnected vertices v_i and vj the score of link prediction measure can be computed as $S(v_i,v_j)$. After calculating scores for all link pairs, ranking either in ascending or descending order is possible.
Let us consider two vertices, v_i and v_j, which are likely connected together and the measure can be calculated as

$$S_{V_i,V_j}^{C} = \left| \Gamma(v_i) \cap \Gamma(v_j) \right| \quad (5)$$

The Weighted Common Neighbor (wCN) can be measured [1], [4-5] and defined as

$$S_{V_i,V_j}^{w} = \sum_{v_k \in \Gamma(v_i) \cap \Gamma(v_j)} w(v_i,v_k) + w(v_k, v_j) \quad (6)$$

The weighted average of local links connecting the common Neighbors of x, y and z can be measured as

$$wCN = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(y, z)}{2} \quad (7)$$

3. G-HR gene signature based HRF cluster

Identifying Gene Signatures for predicting the various Gene Patterns with highest accuracy is most essential and this can be used to construct high accuracy Gene Classifier/Predictor for clinical tests and applications[1,6,7,8]. Thus this research work proposed an efficient Gene Signature based HRF Cluster called G-HR. The Procedure is discussed elaborately in the following section. This model was proposed and implemented as shown in the Figure 1 to achieve higher Pattern prediction and classification accuracy.

3.1. G-HR Procedure

The procedure of the proposed G-HR method is follows. This can identify gene sets that are associated with genes expression and its subset clusters. It will form Clusters based on the distances of points which can calculate with Euclidean Distance Model. The proposed model capable of merging clusters depends on its sizes.
It is capable of eliminating noises and outlies so that the misclassification can be reduced which will help to maximize the classification accuracy. The Closest Cluster built by Hierarchical Random Forest Model was further optimized through Genetic Algorithm based Hierarchical Random Forest Model. As a whole, this proposed model achieves higher classification accuracy.

1) Collect Genome Sequence Training Data
2) Create Multiple Clusters through Euclidean Distance
3) Find Similar Clusters based on distance Calculated
4) Find Clusters with less points and merge together through Hierarchical Cluster
5) Validate through Hierarchical Random Forest
6) Minimize Misclassification Rate through GA-HRF
7) Maximize Area Under Curve (AUC) Measurement
8) Select Most Closest Cluster through GA-HRF
9) Remove Redundant Clusters through Spearman Rank Correlation Model

$$\rho = 1 - \frac{6\sum d^2}{N(N^2 - 1)} \quad (8)$$

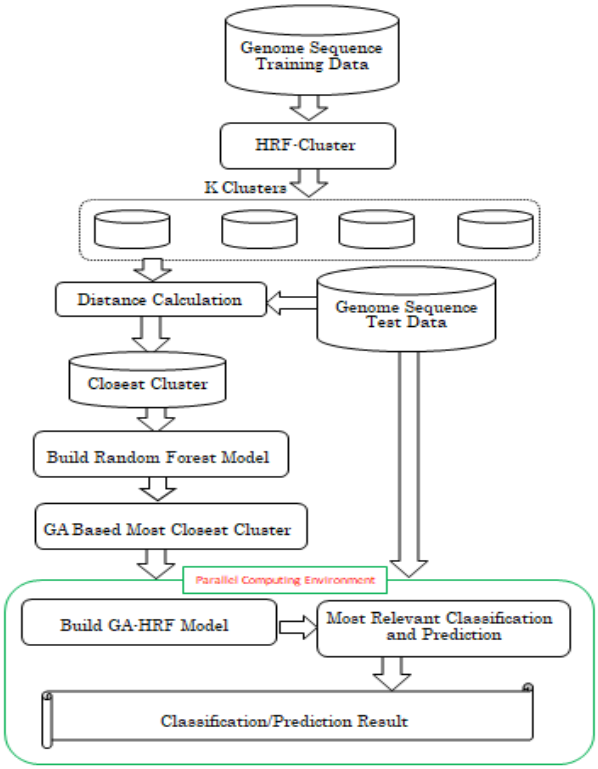


Fig. 1: Proposed Ghr Cluster.

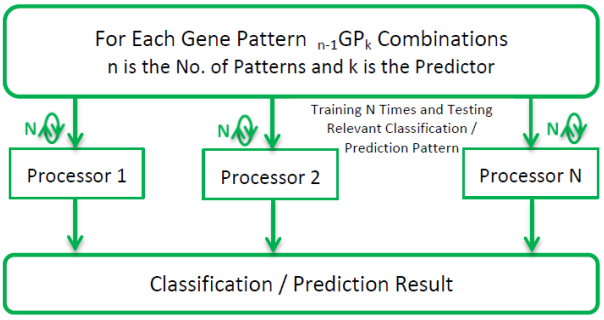


Fig. 2: Parallel Framework of the Proposed GHR Cluster.

3.2. Parallel computing framework

The proposed G-HR Method was implemented under Parallel Computing Framework to improve the performance of the proposed model in term of Execution Time. The Parallel Framework Architecture was shown in the Figure 2. The Model is designed to predict the Pattern in parallel by Processors. That is this work has evaluated the model with Two Processors, Four Processors, Eight Processors and 16 Processors.

4. Experimental study and analysis

The experimental procedures and simulations are carried out by this research work by using the Genome Sequence Data Sets, Master. MER [1],[2]. This was taken from NCBI for study which have described in this section.

Simulations are conducted to examine the performances and prediction abilities of the proposed Gene Signature based HRF Cluster G-HR along with the recently proposed classifiers i. Hierarchical-Random Forest based Clustering (HRF-Cluster), ii. GA-Gene Association Classifier (GA-GA) and iii. Weighted Common Neighbor (wCN) Classifiers.

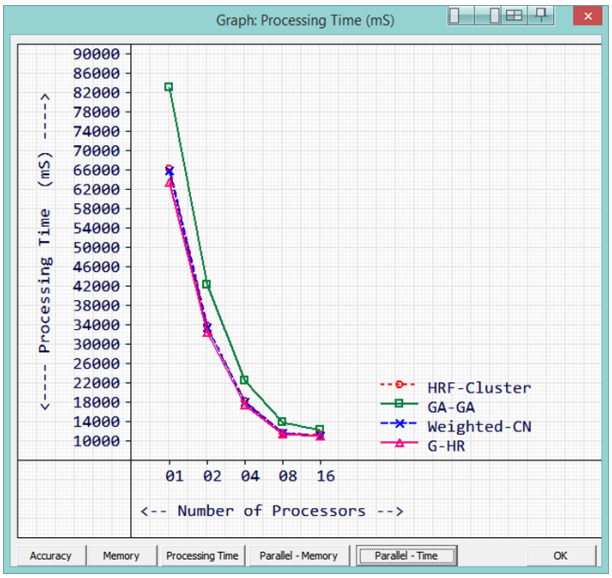


Fig. 3: Processing Time in Ms (Parallel Processing) vs. Classifiers

This paper considered 10 different Genome Genes Data Sets categories for predicting possible diseases and each category has 50,000 records and in total there are 500000 records used for performance analysis of the proposed model. The experiment with single processor was repeated number of times and average probabilities for predicting possible diseases were recorded.

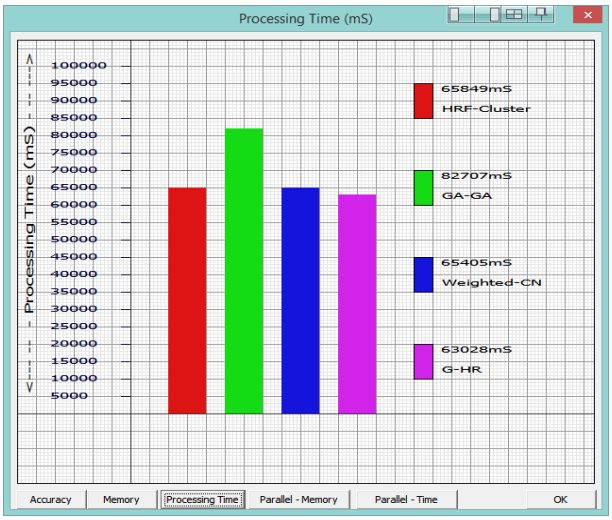


Fig. 4: Processing Time in Ms (Uni-Processing) vs. Classifiers.

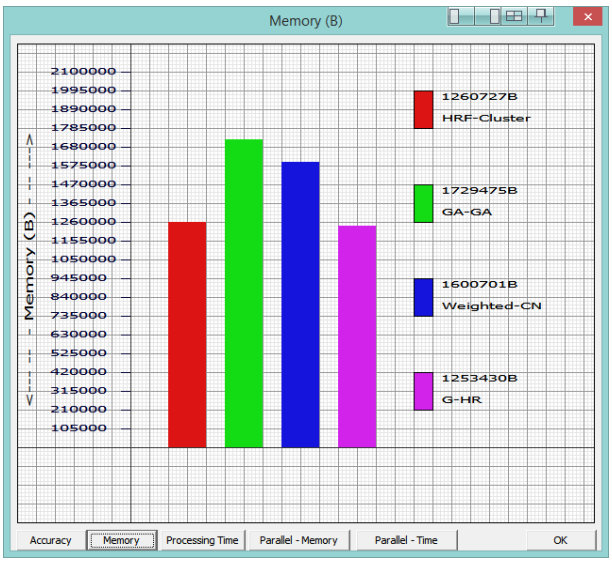


Fig. 5: Memory Usage in Bytes(Uni-Processing) vs. Classifiers.

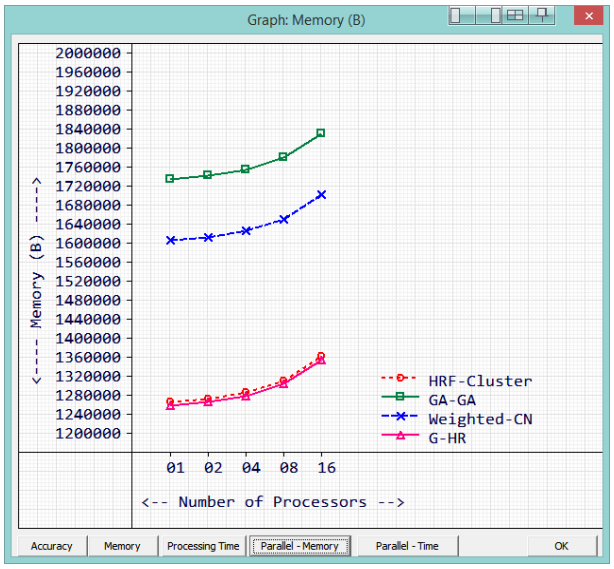


Fig. 6: Memory Usage in Bytes (Parallel Processing) vs. Classifiers.

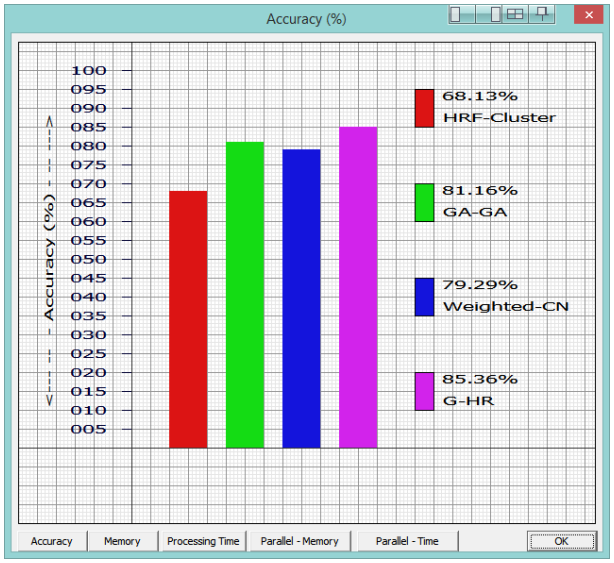


Fig. 7: Pattern Prediction Accuracy (Parallel Processing) vs. Classifiers.

As mentioned earlier, the Genome Data Sets of various human genome patterns were considered for simulation. The performances of the above discussed Genome Classifiers have been studied in terms of Execution / Processing Time, Memory Usage and Prediction Classification Accuracy.

The experiment with Multi-Processors say 2, 4, 8 and 16 Processors was repeated number of times and average probabilities for predicting possible diseases were recorded. It was noted that the Execution Time was reduced as number of processors involved were increased for Classification / Prediction.

This Research Work has developed the Interfacing Tool with the help of VC++ Programming Language to extract and validate the Gene Expressions which are downloaded from NCBI. The validated data is fed into BioWeka for analysing the proposed Genome Classifiers in terms of Execution / Processing Time, Memory Usage and Prediction Classification Accuracy.

Table 1: Performance Analysis of the Proposed Classifier G-HR

Classifiers			HRF-Cluster	GAGA	Weighted-CN	G-HR
Number of Records Identified	Bipolar Disorder		34101	40594	39939	43436
	Breast Cancer		34068	40664	39672	43000
	Cardiomyopathy		34836	40616	39848	42833
	Celiac Disease		34325	40711	39794	42838
	Cerebral Vascular Disease		34734	40654	39799	43195
	Diabetes		34661	41238	40036	43327
	Macular Degeneration		34633	41217	39911	43216
	Parkinson Disease		34617	40985	40236	42975
	Pericardial Disease		34208	40768	40010	43133
	Psoriasis		34837	40959	40376	43266
Accuracy (%)			68.13	81.16	79.29	85.36
Memory Usage (Bytes)	Processors	1	1267127	1735875	1607101	1259830
		2	1273527	1742275	1613501	1266230
		4	1286327	1755075	1626301	1279030
		8	1311927	1780675	1651901	1304630
		16	1363127	1831875	1703101	1355830
Execution Time (ms)	Processors	1	66273	83135	65837	63469
		2	33810	42236	33591	32400
		4	18236	22447	18121	17532
		8	11781	13888	11725	11430
		16	11218	12271	11191	11043

The proposed Gene Signature based HRF Cluster G-HR and the recently proposed classifiers namely i. Hierarchical-Random Forest based Clustering (HRF-Cluster), ii. GA-Gene Association Classifier (GA-GA) and iii. Weighted Common Neighbor (wCN) were implemented, executed with [1], [2], [4], [8], and 16 Processors and analysed thoroughly.

The experimental results of the comparative study of the proposed model executed by Uni Processor and Multi Processors interm of Processing Time, Memory Usage and Prediction Classification Accuracy were tabulated in the Table 1 and Figure 3 and Figure 7 as well. From the results, it was noticed that the proposed classifier predicts more diseases as compared with the existing Classifiers.

From the Figure 3 and Figure 4, it was noticed that the Execution Time by the Parallel Processors is lower than that of the Uniprocessor for prediction patterns.

From the Figure 5 and Figure 6, it was noticed that the Memory Usage by the Parallel Processors is higher than that of the Uniprocessor for pattern prediction. This is happened as each processor holds Datasets in Memory for Classification and Prediction.

It is also established that the Prediction Accuracy remains the same by the Uni-Processing Setup and Parallel Processing Framework, which are shown in the Figure 7 and Table 1.

5. Conclusion

This research work is implemented the three Classifiers, namely Hierarchical-Random Forest based Clustering (HRF-Cluster), Genetic Algorithm-Gene Association Classifier (GA-GA) and Weighted Common Neighbor Classifier (wCN) and studied thoroughly in terms of Prediction Accuracy, Memory Utilization, Memory Usage and Processing Time with around 500000 human genome patterns. From our experimental results, it is noted that the performances of these three classifiers are purely depend on the patterns of genomes. To improve the performances of the Gene Classifiers / Predictors further, this research work is proposed Gene Signature based HRF Cluster, G-HR. This was implemented and studied thoroughly. From the experimental results, it is noted that the performances of the proposed Classifier G-HR is outperforming as compared with that of the identified above mentioned three Classifiers in terms of Disease Pattern Prediction, Processing Time, Memory Usage and Classification Accuracy. It is also noticed that the execution time executed under Parallel Processing Framework is lesser than that of Uniprocessor.

References

[1] N. K. Sakthivel, N. P. Gopalan, S. Subasree, "G-HR: Gene Signature based HRF Cluster for Predicting Human Diseases", International Journal of Pure and Applied Mathematics, Volume 117 No. 9 (2017).

[2] N. K. Sakthivel, N. P. Gopalan, S. Subasree, "A Comparative Study and Analysis of DNA Sequence Classifiers for Predicting Human Diseases", ACM International Conference on Informatics and Analytics (ICIA-16), (2016). https://doi.org/10.1145/2980258.2982038.

[3] Thiptanawat Phongwattana, Worrawat Engchuan and Jonathan H. Chan, "Clustering-Based Multi-Class Classification of Complex Disease", seventh IEEE International Conference on Knowledge and Smart Technology (KST2015), (2015). https://doi.org/10.1109/KST.2015.7051475.

[4] Koosha Tahmasebipour and Sheridan Houghten, "Disease-Gene Association Using a Genetic Algorithm", 14th IEEE Computer Society conference on Bioinformatics and Bioengineering, Pp. 191-197, (2014). https://doi.org/10.1109/BIBE.2014.38.

[5] Gregorio Alanis-Lobato, "Exploring the Genetics Underlying Autoimmune Diseases with Network Analysis and Link Prediction" Middle East Conference on Biomedical Engineering (MECBME). (2014). https://doi.org/10.1109/MECBME.2014.6783232.

[6] Wei Hu, "High Accuracy Gene Signature for Chemosensitivity Prediction in Breast Cancer", Tsinghua Science and Technology. 530-536. Volume 20, Number 5, October, (2015).

[7] Conze, "Random Forests on Hierarchical Multi-Scale Supervoxels for Liver Tumor Segmentation in Dynamic Contrast-Enhanced CT Scans" IEEE 13th International Symposium on Biomedical Imaging (ISBI), April (2016).

[8] Desbordes Paul, "Feature selection for outcome prediction in esophageal cancer using genetic algorithm and Random Forest Classifier", Computerized Medical Imaging and Graphics, (2016).

[9] Feng Luo, James Z Wang, and Eric Promislow, "Exploring local community structures in large networks", Web Intelligence and Agent Systems. (2006).

[10] R. Ben-Hamo, S. Boue, F. Martin, M. Talikka, and S. Efroni, "Classification of lung adenocarcinoma and squamous cell carcinoma samples based on their gene expression profile Improved Diagnostic Signature Challenge", Systems Biomedicine, 1(4), 68-77. (2013). https://doi.org/10.4161/sysb.25983.

[11] Lilian Berton, "Link prediction in graph construction for supervised and semi-supervised learning", International Joint Conference on Neural Networks (IJCNN), Pp. 1-5. (2015).

[12] Witten, Ian H., and Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, (2005).

[13] Zaki, Mohammed J., and Wagner Meira Jr, "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, (2014).

[14] Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman, "Mining of Massive Datasets," Cambridge University Press, (2014). https://doi.org/10.1017/CBO9781139924801

[15] Nikam, Sagar S. "A Comparative Study of Classification Techniques in Data Mining Algorithms," Oriental Journal of Computer Science & Technology, Pp. 13-19, (2015).

[16] Han, Jiawei, Jian Pei, and Micheline Kamber, "Data Mining: Concepts and Techniques," Elsevier, (2011).

[17] Delveen Luqman Abd Al.Nabi, Shereen Shukri Ahmed, "Survey on Classification Algorithms for Data Mining (Comparison and Evaluation)," (ISSN 2222-2863) 4(8), (2013).

[18] W. Engchuan, J. H. Chan, "Pathway-Based Multi-Class Classification of Lung Cancer", Lecture Notes in Computer Science (LNCS), Vol. 7667 (Part V), pp. 697-702, (2012).

[19] X. Zhang and W. Xiao, "Clustering based Two-Stage Text Classification Requiring Minimal Training Data," International Conference on Systems and Informatics (ICSAI), (2012). https://doi.org/10.1109/ICSAI.2012.6223496.

[20] M. H. Chignell, B. G. Stacey, "The Classification of Patients into diagnostic groups using Cluster Analysis," Journal of Clinical Psycology, Vol. 37, pp. 151-153, (2006). https://doi.org/10.1002/1097-4679(198101)37:1<151::AID-JCLP2270370129>3.0.CO;2-4.

[21] Decap,D. et al., "Halvade: Scalable Sequence Analysis With Mapreduce," Bioinformatics, Pp. 2482–2488, (2015). https://doi.org/10.1093/bioinformatics/btv179.

[22] Gonzalez-Domı´nguez,J. et al., "Parallel and scalable short-read alignment on multi-core clusters using UPC++," Vol. 11, (2016).

[23] Jeffrey,D. and Sanjay,G, "Mapreduce: Simplified data processing on large clusters," Proceedings of the 6th Conference on Symposium on Opearting Systems Design and Implementation, pp. 10–10, (2014).

[24] Niemenmaa,M. et al., "Hadoop-bam: Directly Manipulating Next Generation Sequencing Data in the Cloud," Bioinformatics, Vol. 28, 876, (2012). https://doi.org/10.1093/bioinformatics/bts054.

[25] Pireddu,L. et al., "Seal: A Distributed Short Read Mapping and Duplicate Removal Tool," Bioinformatics, Vol. 27, Pp. 2159–2160, (2011). https://doi.org/10.1093/bioinformatics/btr325.

[26] Puckel wartz M.J. et al, "Supercomputing for the Parallelization Of Whole Genome Analysis," Bioinformatics, Vol. 30, Pp. 1508–1513, (2014).

[27] Wylie,K.M. et al., "Emerging view of the Human Virome," Translational Research, Vol. 160, Pp. 283–290, (2012). https://doi.org/10.1016/j.trsl.2012.03.006.

[28] Zaharia,M. et al., "Spark: Cluster Computing with working Sets," Proceedings of the Second USENIX Conference on Hot Topics in Cloud Computing, HotCloud'10, pp. 10–10, (2010).

[29] Puggini L, Doyle J, McLoone S, "Fault Detection Using Random Forest Similarity Distance," IFAC-Papers On Line, Vol. 48, Pp. 583–588, (2015). https://doi.org/10.1016/j.ifacol.2015.09.589.

[30] Kim E-Y, Kim S-Y, Ashlock D, Nam D, "MULTI-K: Accurate Classification of Microarray Subtypes using Ensemble KMeans Clustering," BMC Bioinformatics, (2009). https://doi.org/10.1186/1471-2105-10-260.

[31] Boongoen T, Garrett S and Price C, "New Cluster Ensemble Approach to Integrative Biological Data Analysis," International Journal of Data Mining and Bioinformatics, Vol. 8, Pp. 150-168, (2013). https://doi.org/10.1504/IJDMB.2013.055495.