International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 7 (2019), pp. 1110-1116

Hybrid Feature Selection using Swarm and Genetic Optimization for DNA Copy Number Variation

Sanaa Fekry Abed Elsadek1[1], Mohamed Abd Allah Makhlouf [2], Ben Bella Sayed Tawfik El-Sayed[3], Hesham Nabeah El mahdy Mohamed[4]

[1]*Computer and Systems Engineering Department, Faculty of Engineering, Zagazig University, Zagazig, Egypt.*

[2,3]*Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt.*

[4]*Faculty of Computers and Informatics, Cairo University, Egypt.*

Abstract

Genomic variation can cause several types of human cancer so the selection of significant genes from the DNA copy number variations (CNVs) data carriages a hard challenge to researchers because of the high-dimensionality of features and multiclass categories being involved. Machine learning has important role in genomic variation it can help for predicting and analysis of DNA CNVs sequence. This paper presents models to classify a set of human malignancy types using machine learning to establish various models on set of cancers based on the CNV level values of 23,082 genes as a features for 2916 instances. Feature selection techniques used for dimensionally reduction of the CNVs data to find an optimal subset of genes which sufficiently represents the original set of features. A hybrid method of particle swarm optimization (PSO) and a genetic algorithm (GA) used to perform the genes selection. The proposed PSO-GA approach is compared with several learning algorithms.The performance of the proposed model is tested and compared with several machine learning algorithms with other benchmark using 10-fold cross-validation. The best performance achieved accuracy value 0.85 and ROC value 0.961 which are promising results not only effectively reduce the number of genes to 2051, but also attains a low classification error rate. The proposed model confirmed that genes from chromosome 3 have in evolving human cancers. It also predicted new genes not studied so far as important ones for the prediction of cancers.

Keywords: Practical Swarm Optimization, Genetic Algorithm, DNA Copy Number Variations, Machine learning, supervised classification.

1. INTRODUCTION

Optimization can be define as the process of how to adjusting the inputs to or characteristics in the mathematical process [1] , or in the experiment to find the minimum or maximum output or result as we wish. Trying to find feature selection for important gene expression analysis defining the type of cancer prediction often they use wrapper classification methods [22] to determine the type of tumor by reduction of the number of genes to investigate the case of a new patient , because in high dimensional data for CNV feature are stored can make negative affect on the performance of cancer classification in our works other words , when too many CNV features it sometimes make reduction on the classification accuracy and efficiency on the cancer classification system so if some of the features may be redundant it is the way of the optimization. It is important to say that no one can deny that Feature selection do an important role in trying to improve the performance in the machine learning methods trying to reduce the time for building the learning model and it try to increase the accuracy and decrease the time consume in calculation in the learning process. So the researchers take an important view on choosing the feature selection to do the enhancement required on the machine learning methods trying to identify the suitable optimal feature selection algorithms they do a very important for a given problem in machine learning task especially in high dimensional data. So it is important to know the study on the different feature selection algorithms for the research studies this is especial to dedicated for development the suitable.

Optimal feature selection algorithm [5] to make the good enhancement for improving performance of the machine learning jobs especially on high dimensional data to achieve the required objective.

Elsadek et.al [30] used machine learning to construct various models on the set of human cancer using the CNV level . The classification models developed in Elsadek research could provide a reasonable prediction of the cancer patients' stage based on their CNV level values.

Genetic search Algorithm(GA) optimization has become a famous heuristic technique in the optimization community[12] , with many researchers exploring the concepts, issue any type of Cancer malignant growth or tumor, caused by abnormal and uncontrolled cell division as we know. from CNV levels

in DNA so we will try In this paper to propose a new feature selection algorithm called Hybrid feature selection using swarm and genetic optimization for effective cancer genes based on DNA copy number variation to be used in six type of cancer detection system Feature selection is the process of choosing a subset of input variables by eliminating irrelevant features. The elimination of irrelevant features reduces the dimensionality of data. It may also allow learning algorithms to operate faster and more effectively. Feature selection is an active research area in machine learning, pattern recognition, and datamining genetic algorithm with a novel fitness function.

2. MATERIALS AND METHODS

The architecture of the proposed approach is represented in Figure 1. The details of the proposed approach are discussed in this section.

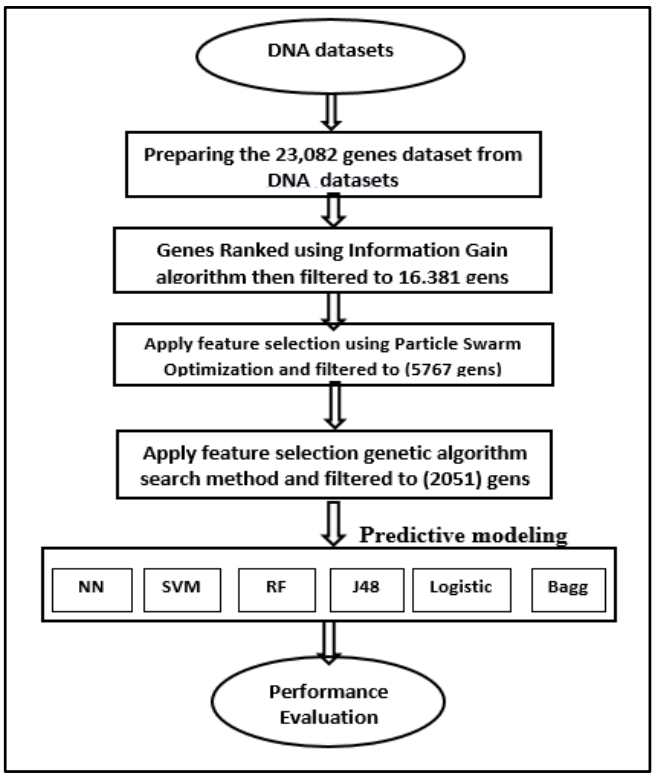

Figure 1: Steps of the proposed approach

2.1 Datasets

The dataset in this research used from [19] and [30] to establish the proposed model. It contains DNA CNVs level of 23082 genes, for 2916 instances .The data used the CNV in different types of cancers were downloaded from Cancer Genomics, [19–21]. We used a six cancer types called Breast invasive Carcinoma (BR) Bladder urothelial carcinoma (BL), Kidney renal clear-cell (KI), Colon and rectum (CO) .Glioblastoma multiform (GB) and .Head and neck squamous cell (HN) the following table(Table 1) is the description of the dataset.

Table 1: The number of instances for each cancer types in the dataset

Index	Cancer Type	Number of Cases
1 (BR)	BRCA (Breast invasive carcinoma)	847
2 (BL)	BL (Bladder urothelial carcinoma	134
3 (CO)	COAD/READ (Colon adenocarcinoma/Rectum adenocarcinoma)	575
4 (GB)	GBM (Glioblastoma multiforme)	563
5 (HN)	Head and neck squamous cell(HN)	305
6 (KI)	KIRC (Kidney renal clear cell carcinoma)	490
Total		2914

2.2 Features selection

Feature selection tools remove the irrelevant and make reduction for features this done from a dataset for improving the performance of the machine learning algorithms for increasing The accuracy and reduce the time for building the model. A DNA microarray can monitor the expression levels of thousands of genes simultaneously. A lot of research has demonstrated that this technology can be useful in the classification of cancers In recent year, researchers in the adversarial information retrieval community had moved towards machine learning approach to detect many researchers have expanded on.one of the most important algorithm is the particle swarm optimization (PSO) it is use before as a new method for global optimization [8], [9] Others have used PSO for comparison testing of other global optimization algorithms, including genetic algorithms and differential evolution [10], [11]. First Features Selection introduced by filter methods were the earliest techniques in the machine learning algorithms which Rank the features in descending order according to their importance [3] [4]. Hegazy, Ah et.al [31] introduced an improved binary version of the standard whale optimization algorithm for feature selection tasks in wrapper approach , the proposed whale model not only significantly enhances the basic whale optimization algorithm but also performs much superior to the other algorithms. Hegazy, Ah et.al [32] introduced improve the structure of basic Salp swarm algorithm to enhance accuracy, reliability and convergence speed. The integration of the inerita weight parameter in the improved Salp swarm algorithm achieve better performance than other algorithms. Hegazy, Ah et.al [33] presented a chaotic Salp swarm

algorithm to avoid these weaknesses, where chaotic maps are used to enhance the performance of Salp swarm algorithm. The produced superior results compared to other optimization algorithms.

2.3 Genetic Algorithms:

Genetic algorithms (Gas) are a type of optimization algorithm, meaning they are used to the optimal solution(s) to a given computational problem that maximizes or minimizes a particular function .Optimization for features required to improve performance, GA will do this by how changing GA parameters of mutation and cross over, meaning how they used to find the maximum or minimum of a function required fitness. GA one of the important techniques of the field of study in bioinformatics science called evolutionary computation [2]. In following the Figure 2 steps describe the main operations of the GA [12]:

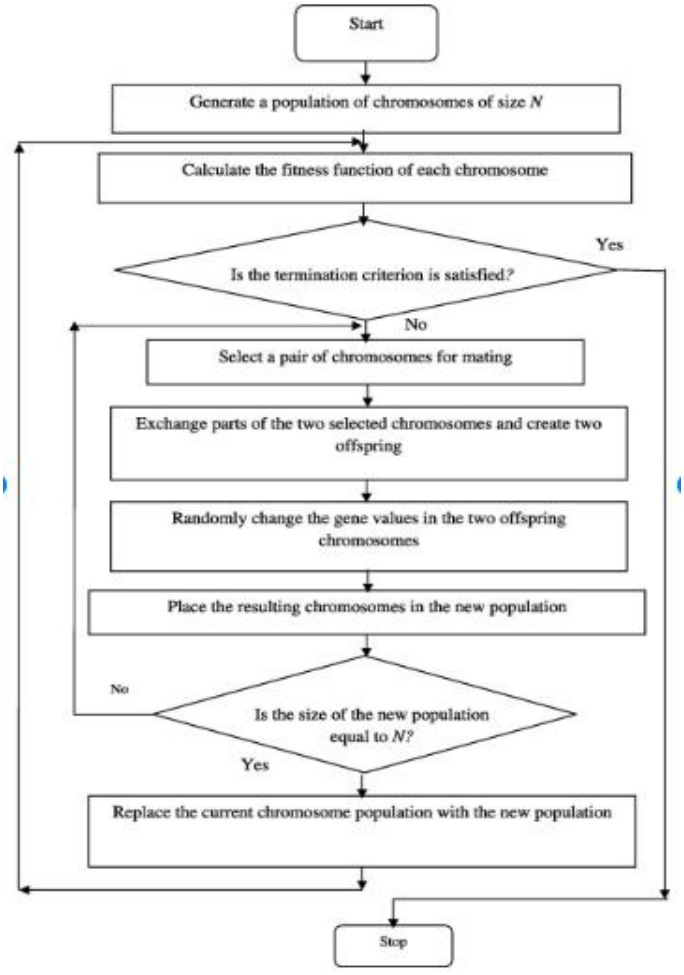

Figure 2: genetic algorithm flowchart.

1. A population of strings, states encoding candidate solutions (called individuals), is evolved toward better solutions. Evolution usually starts from a population of randomly generated individuals and happens in generation
2. For each generation evaluate the fitness of every individual in initial population.
3. Multiple individuals are stochastically selected from current population. New population is generated from selected individuals by: combining pairs of individuals
4. Crossover then, modifying some individuals randomly
5. Mutation new population is used in next iteration of algorithm. The termination of the algorithm when either: maximum number of generations has been produced. Satisfactory fitness level has been reached.

2.4 Particle Swarm Optimization (PSO)

Many of researchers applied the particle swarm optimization (PSO) as a new method for global optimization [13], [14] others have used PSO for comparison of other global optimization algorithms, including genetic algorithms and differential evolution [15], [17]. Defining a Standard PSO can be defined it as is apparently disorganized collection of moving individuals that tend to cluster together which each individual seems to be moving in a random direction it is simple in concept, easy to implement computationally efficient effective on a wide variety of problems. It is used for the optimization the population initialized by assigning random position and velocity, each particle keeps track of its best (highest fitness) the flowchart in Figure 3.

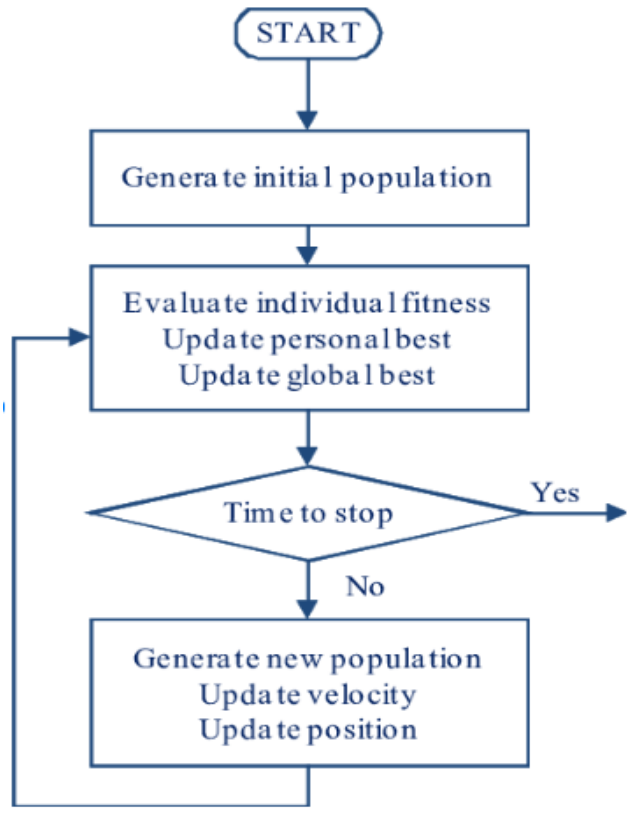

Figure 3: Particle Swarm algorithm flowchart.

3. EXPERIMENTAL RESULTS

Due to the high number of genes, many researchers studying how to select relevant genes effectively before using a classification method to decrease classification error rates and to determine the smallest subset of genes that can provide diagnostic information for disease prediction without trading off accuracy. The selected genes were ranked using Information Gain algorithm reduced the features to 16382 genes then feature selection reduced using PSO and filtered to (5767 gens) and feature selection reduced using GA and d filtered to (2051) gens.. The classification power of the different algorithms used in this study is evaluated by training / testing using 10-fold cross-validation method to examine the performances of the presented methods. The model performance was evaluated based on the actual and predicted classes [5-7]. Table2 represents the different models performance measures like accuracy, Receiver operating characteristic (ROC) area, Matthew Correlation Coefficient (MCC) and precision. The equations from 1-5 represent these measures.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{F-measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Recall}+\text{Precision}} \quad (2)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)\times(TP+FN)\times(FN+FP)\times(TN+FN)}} \quad (3)$$

$$Recall(I) = \frac{TP}{TP+FN} \quad (4)$$

$$\text{Precision}\ (I) = \frac{TP}{TP+FP} \quad (5)$$

Table 2: The accuracy and ROC area performance for algorithms.

Algorithms	Accuracy	MCC	ROC Area	Precision	Recall
Random forest	**80.59**	0.748	0.945	0.778	0.888
Support vector	80.38	0.752	0.910	0.818	0.831
J48	72.32	0.605	0.802	0.721	0.710
Logistic Reg.	**84.67**	0.787	0.961	0.844	0.855
Bagging	77.54	0.694	0.929	0.767	0.803

Table 3: The model performance of all algorithms for all of cancers.

Random forest	BR	BL	CO	GB	KI	HN
TP Rate	0.874	0.074	0.795	0.869	0.886	0.673
ROC Area	0 .944	0.838	0.954	0.979	0.980	0.938
Precision	0.771	0.526	0.762	0.889	0.861	0.728
F-Measure	0.819	0.130	0.778	0.879	0.873	0.699
Recall	0.874	0.074	0.795	0.869	0.886	0.673
FP Rate	0.106	0.003	0.061	0.026	0.029	0.030
Support vector	BR	BL	CO	GB	KI	HN
TP Rate	0.831	0.437	0.845	0.885	0.912	0.641
ROC Area	0.910	0.845	0.931	0.973	0.969	0.899
Precision	0.818	0.447	0.790	0.929	0.882	0.740
F-Measure	0.824	0.442	0.817	0.906	0.897	0.687
Recall	0.831	0.437	0.845	0.885	0.912	0.641
FP Rate	0.076	0.026	0.055	0.016	0.025	0.026
J48	BR	BL	CO	GB	KI	HN
TP Rate	0.718	0.178	0.757	0.815	0.853	0.539
ROC Area	0.802	0.577	0.846	0.889	0.919	0.736
Precision	0.721	0.270	0.679	0.818	0.831	0.591
F-Measure	0.720	0.214	0.715	0.817	0.842	0.564
Recall	0.718	0.178	0.757	0.815	0.853	0.539
FP Rate	0.114	0.023	0.088	0.043	0.035	0.044
Logistic Reg.	BR	BL	CO	GB	KI	HN
TP Rate	0.855	0.422	0.889	0.924	0.918	0.748
ROC Area	0.961	0.903	0.978	0.991	0.989	0.963
Precision	0.844	0.640	0.812	0.939	0.895	0.804
F-Measure	0.849	0.509	0.849	0.931	0.906	0.775
Recall	0.855	0.422	0.889	0.924	0.918	0.748
FP Rate	0.065	0.012	0.050	0.014	0.022	0.021
Bagging	BR	BL	CO	GB	KI	HN
TP Rate	0.803	0.096	0.805	0.840	0.890	0.641
ROC Area	0.929	0.798	0.938	0.966	0.975	0.920
Precision	0.767	0.406	0.723	0.878	0.821	0.681
F-Measure	0.785	0.156	0.762	0.858	0.854	0.660
Recall	0.803	0.096	0.805	0.840	0.890	0.641
FP Rate	0.100	0.007	0.076	0.028	0.039	0.035

The ROC curve represent the models performance measure the sensitivity vs. 1 – specificity, which is important to measure and evaluate the performance of models. Figure 4 shows the breast cancer ROC area for all the selected classifiers.

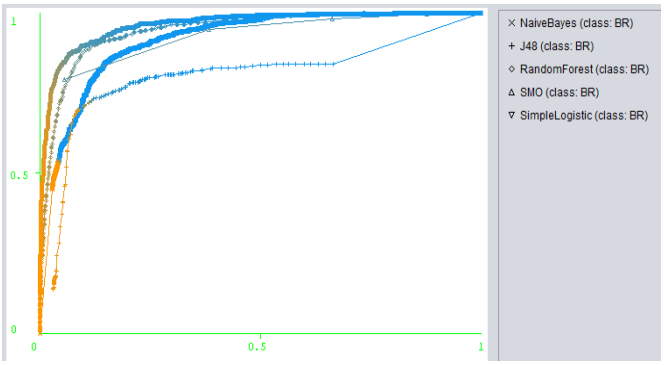

Figure 4: Models performance chart for breast cancer ROC area for all the selected classifiers.

4.1 Comparison with Other Studies

Ning Zhang et.al [23] implemented the Incremental Feature Selection methods and construct a classifications using Dagging classifier for CNVs of 23,082 genes, the model accuracy reached over 0.75 whereas our proposed model achieved the accuracy over 0.85 and ROC area 0.961, we analyzed the recognized genes using published database and compared the importance of selected genes with the other studies. The analysis proved that the identified genes are biologically relevant to contribute into molecular location on chromosome 3 genes. The research has found using the feature selection has discovered the high ranked genes belong to in the chromosome 3 as previous studies. The new is that in this research several genes were found biological significant but did not seem till now in cancer references.

Table 4. The top ranked genes description according.

Gene Symbol	Gene description	IGV rank	Reference in cancer	Chromosome
CRELD1	Encodes a member of a subfamily of epidermal growth factor-related proteins. [42].	0.3867	--	3
ARPC4-TTLL3	The read-through transcript encodes a fusion protein that shares sequence identity with each individual gene product[42].	0.3864	--	3
RPUSD3	protein that functions in the assembly of the mitochondrial ribosome by adding a pseudouridine group to 16S rRNA. [42].	0.3863	--	3
TTLL3	tubulin tyrosine ligase like 3	0.3861	--	3
PRRT3	proline rich transmembrane protein 3	0.3861	--	3
JAGN1	transmembrane protein. It functions in the early secretory pathway and is necessary for neutrophil differentiation and survival. [42].	0.386	--	3
CIDEC	Encodes a member of the cell death-inducing DNA fragmentation factor-like effector family. Members of this family play important roles in apoptosis. [42].	0.386	--	3
IL17RC	transmembrane protein coding gene	0.3871	[40,41]	3
CDC25A	Cell division cycle 25A	0.3075	[24,25]	3
ZMYND11	Zinc finger, MYND-type containing 11	0.2951	[26]	10
KBTBD6	Kelc repeat and BTB (POZ) domain containing 6	0.2813	[27]	13
EGFR	Epidermal growth factor receptor	0.2636	[28]	7
ZNF503-AS1	ZNF503 antisense RNA 1	0.2478	[29]	10
SEMA6A	Sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6A [42].	0.2155	[34]	5
CUL2	Cullin 2	0.2831	[36]	?
MIR4703	Microrna 4703	0.281	?	13
CDKN2A	Cyclin-dependent kinase inhibitor 2A	0.2831	[37]	9
CTBP2	C-terminal binding protein 2	0.2343	[38]	10
MMD2	encodes a member of the PAQR (progestin and adipoQ receptor) family	0.1347	[35]	7
RPS15	Encodes a member of the cell death-inducing DNA fragmentation factor-like effector family[42].	0.0746	--	19

IL17RC gene has the first ranked value in Table 4, is a transmembrane protein coding gene. In spite of its role in cancer is still dodgy, previous research presented that IL17RC is associated to prostate cancer. It isoforms are differently expressed in androgen-independent and androgen-dependent prostate cancers [39].

CDC25A gene is required for progression from G1 to the S phase of the cell cycle. It activates the cyclin-dependent kinase CDC2 by removing two phosphate groups [42].

CONCLUSION

Due to the high number of DNA CNV , researchers revising how to select the most relevant CNV genes before using a classification method to minimize the classification error rates and to decide the smallest subset of features that can help for disease prediction without trading off accuracy. We used a set of machine learning techniques to classify different cancer types (Breast, Balder, Colon, Glioblastoma, head and kidney) using the CNV level values. Using the a three hybrid levels of features selection information gain filter method then Swarm then genetic algorithm, to filter the most important genes that achieve 2051 subset of genes with accuracy of over 0.85 by using CNV levels of 23,082 genes. Examination of the top ranked genes may play major roles in discriminating human cancer types and help to understand tumorigenesis of different cancers. The proposed model offered that genes from the chromosome 3 have in developing human cancers. It also expected new genes not studied so far as important ones for the prediction of human cancers. As future work, the model can be used with deep learning techniques and apply wrapper feature selection techniques to improve the performance of the model and highlighting on the most effective genes.

REFERENCES

[1] Fraser A. : Simulation of genetic systems by automatic digital computers. I. Introduction.Aust. J. Biol. Sci. 10, 484–491 (1957)

[2] Holland J.H.: Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial in telligence. University of Michigan Press (1975)

[3] Mark A. Hall. Correlation-based feature selection for machine learning.Technical report, Department of Computer Science, University of Waikato, 1998.

[4] Barak Chizi and OdedMaimon. Dimension reduction and feature selection. In Data Mining and Knowledge Discovery Handbook, pages 83– 100.Springer New York Dordrecht Heidelberg London, 2010.

[5] Zhang, Ning, et al. "Classification of cancers based on copy number variation landscapes " *Biochimica et BiophysicaActa (BBA)-General Subjects* 1860.11 (2016): 2750-2755

[8] R. Eberhart and J. Kennedy (1995). A New Optimizer using Particle Swarm Theory.Proceedings of the Sixth International Symposium on

[9] J. Kennedy and R. Eberhart (1995).Particle Swarm Optimization. Proceedings of the 1995 IEEE International Conference on Neural Networks(Perth, Australia): IEEE Service Center, Piscataway, NJ, IV: pp 1942-1948.

[10] J. Vesterstrøm and R. Thomsen (2004).A Comparative Study of Differential Evolution, Particle Swarm Optimization, and Evolutionary Algorithms on Numerical Benchmark Problems. Proceedings of the 2004 IEEE Congress on Evolutionary Computation , Volume 2, pp. 1980 - 1987

[11] R. Eberhart and Y. Shi (1998).Comparison between Genetic Algorithms and Particle Swarm Optimization.

EP ’98: Proceedings of the 7th International Conference on Evolutionary Programming VII , pp. 611- 616.

[12] Kinnear, K. E. (1994). A Perspective on the Work in this Book. In K. E. Kinnear(Ed.),Advances in Genetic Programming (pp. 3-17). Cambridge: MIT Press

[13] R. Eberhart and J. Kennedy (1995). A New Optimizer using Particle Swarm Theory.Proceedings of the Sixth International Symposium on

[14] J. Kennedy and R. Eberhart (1995).Particle Swarm Optimization. Proceedings of the 1995 IEEE International Conference on Neural Networks(Perth, Australia): IEEE Service Center, Piscataway, NJ, IV:pp 1942-1948.

[15] J. Vesterstrøm and R. Thomsen (2004).A Comparative Study of Differential Evolution, Particle Swarm Optimization, and Evolutionary Algorithms on Numerical Benchmark Problems. Proceedings of the 2004 IEEE Congress on Evolutionary Computation , Volume 2, pp. 1980 – 1987

[17] R. Eberhart and Y. Shi (1998).Comparison between Genetic Algorithms and Particle Swarm Optimization. EP ’98: Proceedings of the 7th International Conference on Evolutionary Programming VII , pp. 611- 616.

[18] International Journal of Computer Applications (0975–8887)Volume 136–No.1, February 2016

[19] G. Ciriello, M.L. Miller, B.A. Aksoy, Y. Senbabaoglu, N. Schultz, C. Sander, Emerging landscape of oncogenic signatures across human cancers, Nat. Genet. 45 (2013) 1127–1133.

[20] E. Cerami, J. Gao, U. Dogrusoz, B.E. Gross, S.O. Sumer, B.A. Aksoy, A. Jacobsen, C.J. Byrne, M.L. Heuer, E. Larsson, The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, Cancer Discov. 2 (2012) 401–404.

[21] J. Gao, B.A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S.O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal,

Sci. Signal. 6 (2013) l1.

[22] Kohavi J, John GH (1998) The wrapper approach. In: Feature se-lection for knowledge discovery and data mining, pp 33–50

[23] Zhang, Ning, et al. "Classification of cancers based on copy number variation landscapes "*Biochimica et Biophysica Acta (BBA)-General Subjects* 1860.11 (2016): 2750-2755.

[24] D. Gasparotto, R. Maestro, S. Piccinin, T. Vukosavljevic, L. Barzan, S. Sulfaro, M. Boiocchi, Overexpression of CDC25A and CDC25B in head and neck cancers, Cancer Res. 57 (1997) 2366–2368.

[25] M.-Y. Huang, J.-Y. Wang, H.-J. Chang, C.-W. Kuo, T.-S. Tok, S.-R. Lin, CDC25A, VAV1, TP73, BRCA1 and ZAP70 gene overexpression correlates with radiation response in colorectal cancer, Oncol. Rep. 25 (2011) 1297–1309.

[26] H. Wen, Y. Li, Y. Xi, S. Jiang, S. Stratton, D. Peng, K. Tanaka, Y. Ren, Z. Xia, J. Wu, ZMYND11 links histone H3. 3K36me3 to transcription elongation and tumour suppression, Nature 508 (2014) 263–268.

[27] A.T. Lorincz, Cancer diagnostic classifiers based on quantitative DNA methylation, Expert. Rev. Mol. Diagn. 14 (2014) 293–305.

[28] J.G. Paez, P.A. Jänne, J.C. Lee, S. Tracy, H. Greulich, S. Gabriel, P. Herman, F.J. Kaye, N. Lindeman, T.J. Boggon, EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy, Science 304 ,2004, 1497–1500.

[29] N. Sengupta, C. Yau, A. Sakthianandeswaren, D. Mouradov, P. Gibbs, N. Suraweera, J.- B. Cazier, G. Polanco-Echeverry, A. Ghosh,M. Thaha, Analysis of colorectal cancers in British Bangladeshi identifies early onset, frequent mucinous histotype and a high prevalence of RBFOX1 deletion, Mol. Cancer 12 (2013).

[30] Elsadek, Sanaa Fekry Abed, Mohamed Abd Allah Makhlouf, and Mohamed Amal Aldeen. "Supervised Classification of Cancers Based on Copy Number Variation." *International Conference on Advanced Intelligent Systems and Informatics*. Springer, Cham, 2018.

[31] Hegazy, Ah E., M. A. Makhlouf, and Gh S. El-Tawel. "Dimensionality reduction using an improved whale optimization algorithm for data classification." *International Journal of Modern Education and Computer Science* 10.7 (2018): 37.

[32] Hegazy, Ah E., M. A. Makhlouf, and Gh S. El-Tawel. "Improved salp swarm algorithm for feature selection." *Journal of King Saud University-Computer and Information Sciences*(2018).

[33] Hegazy, Ah E., M. A. Makhlouf, and Gh S. El-Tawel. "Feature Selection Using Chaotic Salp Swarm Algorithm for Data Classification." *Arabian Journal for Science and Engineering*44.4 (2019): 3801-3816.

[34] M. Dhanabal, F. Wu, E. Alvarez, K.D. McQueeney, M. Jeffers, J. MacDougall, F.L. Boldog, C. Hackett, S. Shenoy, N. Khramtsov, Research paper recombinant semaphorin 6A-1 Ectodomain inhibits in vivo growth factor and tumor cell line-induced angiogenesis, Cancer Biol. Ther. 4 (2005) 659–668

[35] L.N. Cristiana, New insights into P53 signalling and cancer: implications for cancer therapy, J. Tumor 2 (2014).

[36] G.L. Semenza, Regulation of cancer cell metabolism by hypoxia-inducible factor 1, Seminars in Cancer Biology, vol. 19, Elsevier 2009, pp. 12–16.

[37] R. Bhati, C. Patterson, C.A. Livasy, C. Fan, D. Ketelsen, Z. Hu, E. Reynolds, C. Tanner, D.T. Moore, F. Gabrielli, C.M. Perou, N. Klauber-DeMore, Molecular characterization of human breast tumor vascular cells, Am. J. Pathol. 172 (2008) 1381–1390.

[38] G. Chinnadurai, The transcriptional corepressor CtBP: a foe of multiple tumor suppressors, Cancer Res. 69 (2009) 731–734.

[39] Z. You, Y. Dong, X. Kong, Y. Zhang, R.L. Vessella, J. Melamed, Differential expression of IL-17RC isoforms in androgen-dependent and androgen-independent prostate cancers, Neoplasia 9,2007 464–470.

[40] Zhang, Ning, et al. "Classification of cancers based on copy number variation landscapes " *Biochimica et Biophysica Acta (BBA)-General Subjects* 1860.11 (2016): 2750-2755.

[41] Z. You, Y. Dong, X. Kong, Y. Zhang, R.L. Vessella, J. Melamed, Differential expression of IL-17RC isoforms in androgen-dependent and androgen-independent prostate cancers, Neoplasia 9,2007 464–470.

[42] https://www.ncbi.nlm.nih.gov/gene/993/ortholog/?scope=7776 accessed 1/7/2013