

A Novel Composite Method using the Simplified Cuckoo Optimization Algorithm and Harmony Search for Cancer Classification

M. Sofia, N. Tajunisha



Abstract: For each malignant development type, simply slight characteristics are using. The quality perseverance work remains a difficult 1. To conquer this issue, all of us propose the double degree quality dedication Technique known as MRMR-SCOA-HS. Within the principal stage, the base repeating and max-imam pertinence (MRMR) highlight willpower is employed to pick a subsection, subdivision, subgroup, subcategory, subclass of substantial qualities. The actual favored features are after that nourished right into a covering set up that combine another computation, SCOA-HS, making use of the help vector machine like a classifier. The particular strategy had been implemented in order to four microarray datasets, and also the exhibition has been broke down through forget about one particular cross-acknowledgment method. Temporary performance investigation from the expert introduced strategy to developmental computations suggested that this proposed calculations amazing is better than other program in selecting a less amount of qualities whilst safeguarding the greatest order accuracy. The methods in the pre-owned attributes were furthermore explored, also it was accepted that they select qualities tend to be organically vital that you every malignancy type.

Keywords: Gene selection, Minimum redundancy and maximum relevance (MRMR), Evolutionary algorithms, Simplified Cuckoo optimization algorithm-Harmony search algorithm (SCOA-HS)

I. INTRODUCTION

In the period of the most recent two decades, the start of DNA microarray method has offered open doors intended for customized pill by surveyed the connection phase's [1]. Microarray innovation possesses as of late also been utilized to slow down mine subtypes of unique malignancies relying on specific from the articulation stages of development of essential qualities [2-4]. This investment strategy gives issue by factor information on typically the hereditary makeup of a special malignant expansion tolerant, coupled these traces possibly bettering the excellence of cure opportunities manufactured by health facility [5].

Through microarray some, the quantity of components is important and It is above the quantity of testing [6, 7] and collection to an upcoming degree of perfection is examining a direct result of the particular marvel about dimensionality [8, 9]. To stay faraway from these troubles, quality widely used strategies are actually proposed that has just the most vital qualities will be chosen and also utilized for set up purposes [3]. There are actually number of points of interest to this technique of diminishing the quality of qualities along with choosing only qualities which might be increasingly prescient during request. By having a lot fewer qualities, not necessarily exclusively could be the preparing a chance to grouping totally diminished, the possibility of misclassification is moreover reduced. Additionally, utilizing a grand number of characteristics as side of the bargain to the Trier can make typically the classifier always be over-fitted [4].

Good quality choice approaches can be partitioned into a few fundamental strategies dependent on their whole communication while using classifier, especially, channel practices, wrapper systems and injected strategies [5]. Funnel strategies dissect the identical no associated with qualities searching at just the normal highlights on the information and remaining away from the effects of decided on qualities about the order setup [6]. Wrapper superior choice slope inclines toward the hunt process in the place of you can imagine quality subsets. They decided to go with qualities are actually then applied dependent on most of their capability to boost arrangement finely-detailed [9]. In the encase quality preference strategy, incorporate determination is usually connected to the agreement level; in fact, this connection is a lot regarding ground-breaking compared to the wrapper method. It is on the grounds that high quality choice throughout joined method is included often the classifier growth and the grouper is surveyed to give any foundation pertaining to highlight solution [1] (see Fig. 1). All the more currently, transformative measurements improved just for quality resolve have been applied inside the composition of wrapper systems [3]. Ultimately, another top quality choice system by strategies for blended relying on information group was planned, prescribing in which streamlining structured bunching could possibly choose slowly enlightening features to create arrangement exactness [4].

Every single quality persistence approach possesses benefits in addition to bad grades [4]. For instance, notwithstanding the fact that the exact channel method is basic plus computationally fruitful, its convention lingers right behind other.

Manuscript published on January 30, 2020.

* Correspondence Author

M. Sofia*, Ph.D Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts & Science for women, Coimbatore, India.

Dr.N. Tajunisha, Associate Professor, Department of Computer Science, Sri Ramakrishna College of Arts & Science for women, Coimbatore, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

A Novel Composite Method using the Simplified Cuckoo Optimization Algorithm and Harmony Search for Cancer Classification

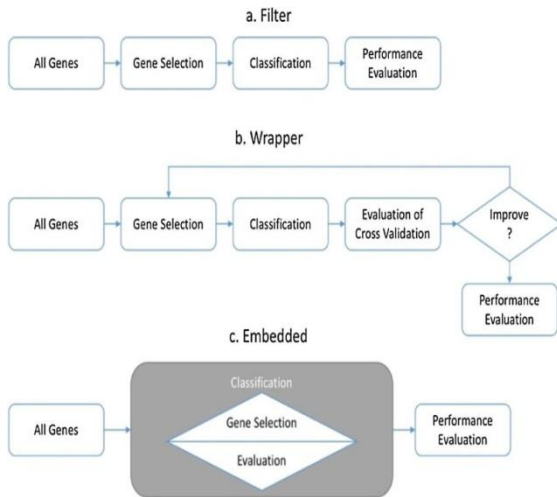


Fig. 1 Featured selection methods.

Strategies, since the arranger works autonomously and basically related throughout quality preference [5]. Alternately, typically the wrapper together with implanted technique, which combine the quality resolve process in the arrangement process, can boost higher group precision still experience the harmful effects of elasticity issues due to their high computational expenses and are also not a good choice for enormous datasets [6, 7].

Excessive order excellence is, naturally, of the relevance for custom made drug. In fact, biomarker is as well a region involving continuous study, where it is recommended analyze a number of number of components to spot patterns (e. grams., choosing very little qualities which might be on the whole on their own communicated amid various examples) [8, 9]. For this reason, in this research, the fundamental enterprises were to opt for the ideal variety of the best characteristics that can ideal recognize a pair of disease varieties. Least duplication and biggest significance (MRMR) include solution [3] was used to opt for a subset of the extremely important as well as least surplus qualities. The exact chose features were subsequently bolstered in a wrapper option that connects to the planned SCOA-HS refurbishment calculation which has a help vector machine (SVM) as a classer. The SVM was implemented as the classifier in this job, as it's obtain exhibitions has become demonstrated and also built up by simply various next to appraisals with assorted calculations. Two-level quality persistence joins the key benefits of both the funnel and wrapper techniques for good quality choice. The very strategies were being applied to a number of microarray datasets and the progress was examined by the just forget about one cross-approval (LOOCV) method.

II. MICROARRAY DATA

Microarray data associated with four varieties of cancer stages (leukemia, prostate, lymphoma, and colon) were especially analysis. Gene investigates data just for leukemia [1] along with prostatic most cancers [4] appeared to be seen with the Wide-ranging Corporation (www.broadinstitute.org). Gene types of records for lymphoma [5] ended upheld been assessed from the Lymphoma/Leukemia Molecular Profiling Job (lmpp. nih. gov). Some sort of gene phrase dataset intended for intestines melanoma [6] was initially applied on the

Princeton University or college Gene Phrase Work (<http://genomics-pubs.prince-ton.edu>). Most information set up to the datasets used in these kinds of study, for instance number of anatomy's genes, selections plus the a pair of instructional classes for each dataset, is shown in family table 1.

III. METHODOLOGY

3. Pre-Processing level

The conventional methodology is often illustrated during Fig. installment payments on your The data are actually divided into several states. An outstanding pre-processing levels, the top number of genes ended were been selected making use of MRMR. Typically the chosen passed genes ended up raised on in a wrapper build up including the main COA-HS criteria plus the SVM classifier to obtain the minimum amount variety of body's genes that offers completely accuracy along with reliability. And finally, the actual categorization regarding determine genes became measured with regards to accuracy when using the LOOCV procedure. To assess features of the COA-HS, the results happen to be checked for you to prospects founded from the evolutionary rules; for example it is partially algorithm (GA), the element swarm search engine marketing technique, the actual harmony seek out (HS) method, and the cuckoo optimization tactic (COT). A codes are anticipated in this examination were printed using Matlab 2014a.

3.1. Discretization

A gene key phrase data ended were been first categorized to a reduced amount of sounds also to produce the exact accuracy and reliability on the category final results [7]. The word associated with every single gene ended up being separated in a nine-state changing good indicate price (1) along with normal change (r) for your gene. For every one gene, much better protect nine recommends proposed maybe there gene has not been expressed (state zero) or perhaps expressed also much it had become over-expressed (states +1 that will +4) plus under-expressed (states 1 towards 4). Family table 2 essentials the different point out governments included in the info discretization.

3.2. Initial-phase gene collection using minimum redundancy maximum relevance (MRMR)

Gene expression truth is typically within a matrix format (see Fig. 3), where every single row mentioned a gene and each vertebrae pointed an illustration. The last range generally offers the class small name for each example of this. The class indicates (Ci) for a two-class classification venture is specified by often 1 or 2.

The actual attribute selection in a classification task is usually to discover subsection, subdivision, subgroup, subcategory, subclass involving popular features of which ideal define the info significance about the classification attempt [8]. MRMR, some form of filter a technique of gene suggestion, identifies individual's genes the fact produce a great deal more data dependent upon class label of the models.

In this examination, common specifics were employed along with redundancy involving body's genes and qualified intended for MRMR to express typically the pertinence classes.

Table 1
Basic information related to the microarray datasets used in this study.

Microarray dataset	Number of genes	Number of samples	Class 1	Class 2
Leukemia	7129	73	48(ALL)	25(AML)
Prostate	12,600	102	50 (Normal) (Germinal center B-	52(Cancerous) (Active B-
Lymphoma	4026	47	24DLCL)	23 DLCL)
Colon	7457	62	22(Normal)	40 Cancerous)

For you to Arranged in addition to evaluate the final result from the search engine marketing job. Considering that, within this analysis, the issue while using class involving microarray exact information can be building larger precision making use of the minimal variety of chosen gene history, the price strategy had been thought to be comes after: Exactly where Precision is definitely the reliability on the SVM classifier part with the LOOCV system, NOG would be the amount of chose family passed dow genes and a partnership. Since powerful data has become more important in comparison with number of characteristics, was permanent at multitude of to supply the exact accuracy plus reliability more importance within the charge functionality. Consequently, through reducing the fee function, how much selected gene history appeared to be lessening while exactness was initially well made use of. In the LOOCV method, one particular sample is definitely treated such as a test smaller sample files as the outstanding examples utilized to coach the very SVM and also consistency is set. If there are typically N trial run samples, this kind of perform is certainly long D occasions, whenever having a several structure, and also the typical accuracy and reliability will be worked out for your preferred passed body's genes. SVM useful for the categorization of precise genes, since SVM classer is an effective category formula and contains already been exhibited of exhibiting brilliant expansion in a variety of nerve classification obligations [9]. The LOOCV technique became chosen, mainly because it could defeat info over-fitting [4]. The actual functionality in the grouper seemed to be implemented LOOCV regarding correctness, empathy as well as specificity. Reliability reveals the actual ring and pinion ratio of most the right way categorized selections. Concern stated the particular proportion associated with appropriately labeled valuable selections and even specificity implies typically the percentage regarding in the correct way grouped damaging trials.

The latest amalgamated enhancement protocol, SCOA-HS, appeared to be much better simply by subscribing to often the not too long ago developed SCOA [1] together with HS rules. The final results were being in comparison as well as the PSO, GEORGIA, HS, as well as COA measurements. PSO can be an enhancement working out which is a result of the exact excitement involving feathered animals working. The particular GA computation is another significant developmental calculation that was very first presented by simply John Netherlands in 75. To completely depict the lately proposed COA-HS calculation, the very subtleties

connected with both COA and HS procedures will certainly beginning becomes clarified concerning the accompanying two segments?

3.3. Simplified Cuckoo optimization algorithm (SCOA)

SCOA is a multitude related achievement calculation which has been proposed just by Rajabion in 2011 and was basically roused by way of the life to a cuckoo suspended creature. The main cuckoo's web meeting in sleeping eggs is normally individual that is to say a cuckoo never encourages its custom home anytime laying offspring and is located

other feathered creatures' family homes to build its ovum. In doing the result, if the cuckoo's eggs require after to host's as it, almost certainly, the very cuckoo's along with will incubate and become engineered cuckoos. If you find the cuckoo's eggs happen to be concocted via the host winged animal, the main outsider ova will be obliterated. In the SCOA calculation, any egg from a home programs a potential collection and each cuckoo speaks towards a fruitful unique arrangement. The focus of the SCOA is to selecting the home most abundant in noteworthy chances of an egg's endurance. That way, the more offspring that make callable subsequent to remaining worked inside of a host family home, the most renowned the period of great benefit allocated to which will home. Around the point when opportunity shows up for the new house purchase of the adverse reports about them developed cuckoos, they switch towards the top home with the most upgraded endurance cost and lie down eggs extremely sweep from; this is identified as the egg laying range (ELR) and they are determined by Equalizer. (8).

$$\frac{\text{number or current cuckoo's eggs}}{\text{Total number of eggs}} = \text{Velocity}$$

Because, in character, the improvement of each populace is actually controlled, within the SCOA computation, a pedoman, Nmax, The gives the stage of confinement to the majority of extreme amount of cuckoos that may live in every stage. The way in which toward provide forth and also the movement associated with cuckoos in the direction of a superior house is rehashed multiple times (100 emphases) to appear through the greatest arrangement. Since it was, the price capacity is restricted through one hundred emphases from the SCOA. The flowchart in the SCOA calculations is delineated in Fig. 2.



3.3.2. Harmony search (HS) algorithm

Typically the amicability look for (HS) equation is a music-enlivened improvement solution [4]. In brighten, artists extemporize their instruments' pitch to seek out an ideal commitment, which can be attained.

3.4. Proposed algorithm (SCOA-HS)

With this section, the information of the recommended algorithm, SCOA-HS, for obtaining and picking out the most educational genes for you to improving the exact classification accuracy and reliability are defined. A flowchart of SCOA-HS is outlined in some. As evaluated, in the SCOA calculation (Section 3. several. 1), every single egg within the home combine an answer and every cuckoo echoes to another agreement. So that, during the quality connection examination, a reply for a good quality. The SCOA-HS calculation will start with the arrival of the cuckoos. At the actual emphasis, following your underlying people laid ovum, the benefit rates of the as it are decided by playing the cost ability. These agreements (eggs) will be then nourished into the HS calculation. HS is used to formulate the opportunity with decide the stainless region of your arrangement room or space, which SCOA alone cannot investigate. On the arrangement space or room, the immaculate location can be surveyed by the impulsive creation course of action through HMCR and PAR, which are fixed at 0.8 and 0.3, singularly. In a final result, an improvement option can be obtained by deterrent untimely combined SCOA. In the primary levels methodology, as soon as the arrangements via SCOA are actually executed in the HS working out, the HS emphasizes several times, after which the power an incentive to the arrangements proposed by the HS are relying on means of the fee capacity. The powers estimations belonging to the arrangements planned by the SCOA and HS are subsequently thought about as well as arrangements (eggs) with the better benefit regard are picked out to hold up against. From that point onward, these along with develop and turn cuckoos. Typically the endurance tempo of each cuckoo is determined or any cuckoos transfer towards the property with the most remarkable endurance pace and put in eggs inside ELR of the finest home (best position). When it is all said and done, the space for arrangements can be refined in regards towards the best blend. This ends one concentration of the SCOA-HS calculation and then the procedure is certainly rehashed too many times. Each time the very cuckoos build eggs from a further improved upon position, all these yields to locate a decent placement related about finding the charge capacity.

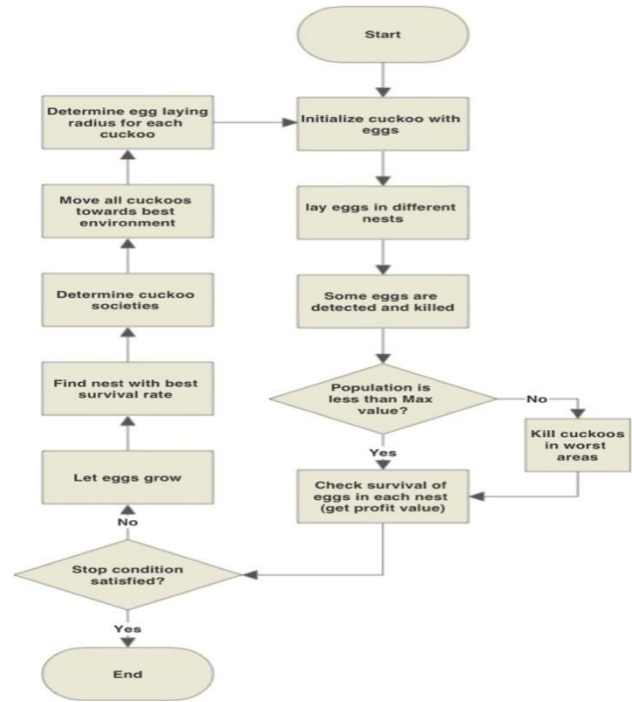


Fig 2: Flow chart

IV. FLOWCHART OF THE SIMPLIFIED CUCKOO OPTIMIZATION ALGORITHM

There are actually three quantities. The first grade is to engage in a review from recollection. The second point is to carry out a hit-or-miss pitch inside accept-able array of available pitch. Lastly, a final choice is that can be played a presentation adjacent to some sort of pitch into their recollection. From the HS mode of operation, these steps are, correspondingly, referred to as harmony memory (HM), the pitch adjustment rate (PAR) plus the harmony memory space consideration fee (HMCR). Typically the HS numbers presents a variety of steps, while marked listed below:

- Initialize the HM
- Initialize the HMCR
- Improvise a new harmony memory
- Update the HM
- Check the stopping criterion.

The actual HS formula has been efficiently used for various optimization problems, such as function selection, under the radar design adjustable and constant optimization issues.

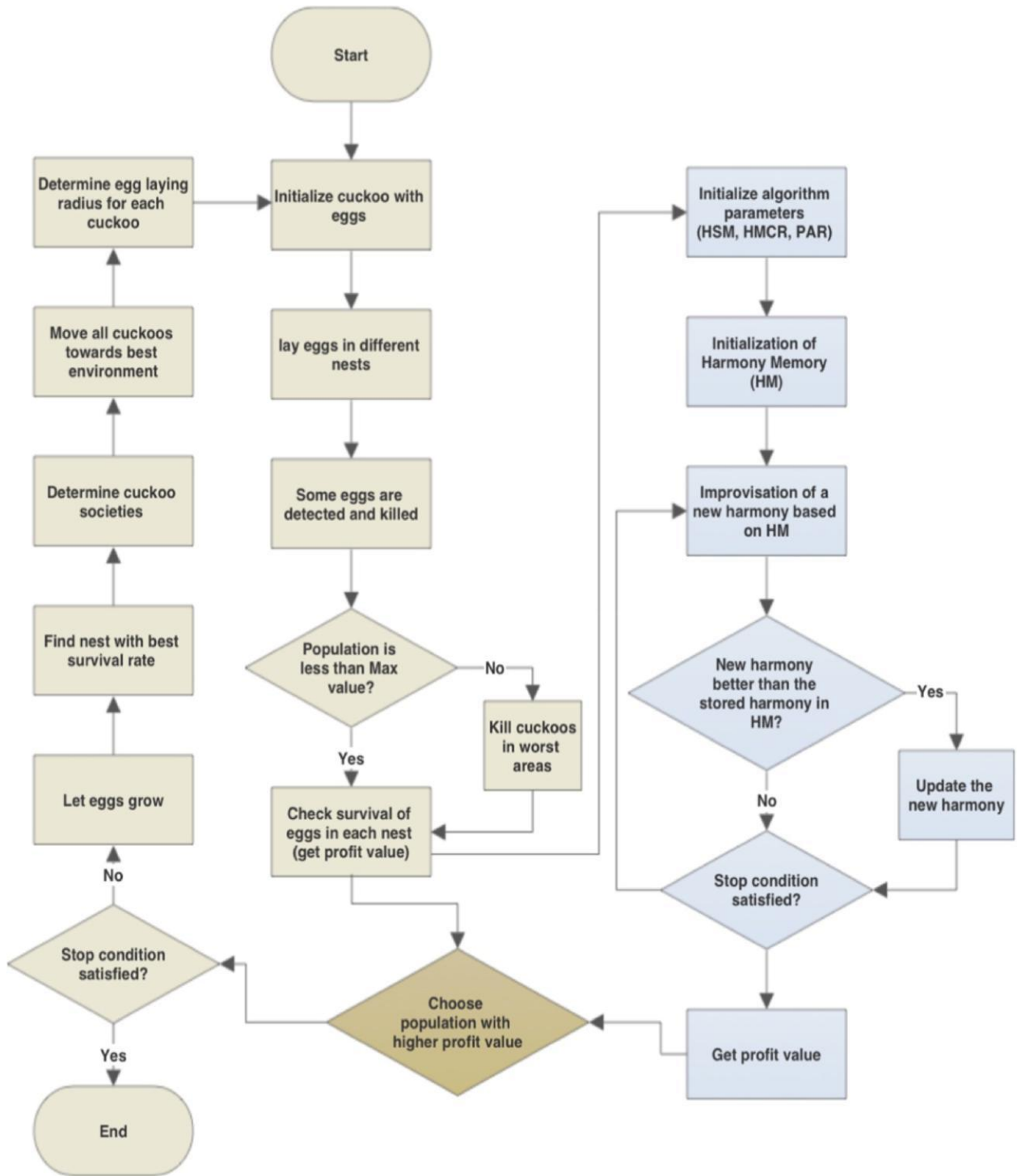


Fig. 3 Flowchart of the COA-HS algorithm

A Novel Composite Method using the Simplified Cuckoo Optimization Algorithm and Harmony Search for Cancer Classification

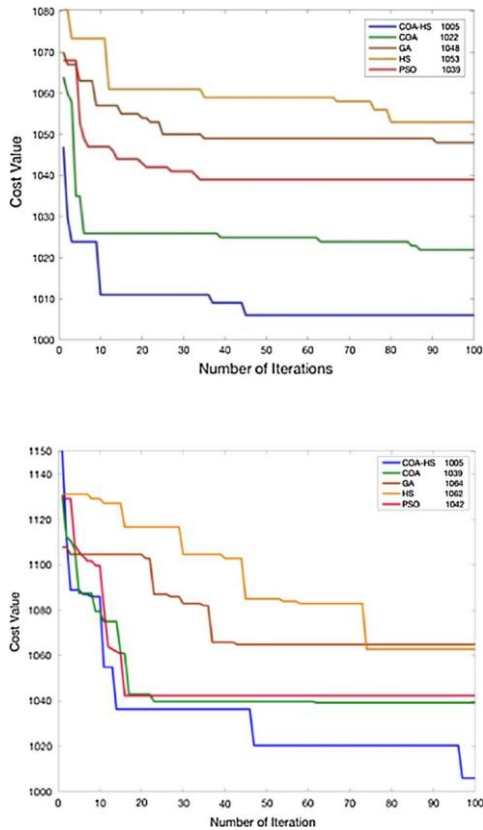


Table 2

Most informative genes selected using the COA-HS algorithm that achieve 100% Classification accuracy for all microarray datasets.

Dataset	Predictive gene
Leukemia	MPO myeloperoxidase
	Oncoprotein 18 (Op18) gene
	Proteasome iota chain
	ARHG ras homolog gene family, member G (rho G)
	LYN V-yes-1 yamaguchi sarcoma viral related oncogene
	OBF-1 mRNA for octamer binding factor 1
Lymphoma	UG Hs.120716 ESTs
	UG Ha. 1 69081 ets variant gene 6 (TEL oncogene)
	MCL1 myeloid cell differentiation protein
Prostate	(40508_at) Glutathione S-transferase alpha 4
	(37639_at) Hepsin
	(769_s_at) Annexin A2
	(1980_s_at) Non-metastatic cells 2, protein (NM23B)
Colon	(41661_at) Cadherin, EGF LAG seven-pass G-type receptor 1
	H. sapiens RON mRNA for tyrosine kinase
	H. sapiens mRNA for metallothionein (HUMAN)
	H. sapiens pterin-4a-carbinolamine dehydratase (PCBD)
	RETINOL-BINDING PROTEIN I, CELLULAR
	Human galactokinase (galK) mRNA, complete cds

SCOA-HS, to find the a great deal better genes with the highest reliability. The accuracy and reliability of the SVM classifier was basically measured making use of the LOOCV process.

V. RESULTS AND DISCUSSION

Throughout Fig. 2, several demonstrate the exact correctness on the SVM classifier to the top rated 60 gene record collected through the use of MRMR. Generally, as the level of genes boosts, the exactness increases. In some instances, much better protect classification persistence was declined as the selection of genes superior. For example , comes to the prostatic cancer dataset, the category exactness to the initial 6 gene history had been 97%, though the consistency has been diminished for the reason that volume of passed genes increased, having values regarding 91-93% at any time 90-100 anatomy's genes were definitely applied.

Accuracy and reliability, which was every time fewer than 100 percent. For the lym-phoma dataset, picking out 2 body's genes by the SCOA-HS algorithm bring about 100% correctness. However, various other algorithms essential a substantially higher wide variety of genes to have 100% finely-detailed.

Table a pair of lists the exact genes decided on by the SCOA-HS algorithm in which provided hundred percent classification precision of the SVM classifier per of the datasets.

To investigate the very relevancy of each one selected gene to the cancers type, their very own functions was investigated and are also described listed below.

MPO Myeloperoxidase is the feature enzyme on the myeloid family tree. The associated with acute myeloid leukemia (AML) is easy in the event that more than 3% of shot cells are generally confirmed being cytochemically

To assess the reduced number of genetics that can ideal distinguish involving two instructional classes of cancers, the number of choice genes was decrease for you to 50 applying MRMR. All these 50 family genes were subsequently fed straight into our planned algorithm,

Table 3

Accuracy (AC), sensitivity (SE), specificity (SP), and number of selected genes (#Genes) via 5 optimization algorithms when combined with the SVM as a classifier for four microarray datasets.

	Leukemia		Prostate		Lymphoma		Colon	
	#Genes	AC/SE/SP	#Genes	AC/SE/SP	#Genes	AC/SE/SP	#Genes	AC/SE/SP
GA	24	100/100/100	28	98.04/91.8/100	22	100/100/100	14	95.16/84.6/100
PSO	16	100/100/100	19	98.04/91.8/100	10	100/100/100	11	96.42/85.8/100
HS	25	100/100/100	33	98.04/91.8/100	21	100/100/100	12	95.16/84.6/100
SCOA	15	100/100/100	12	99.07/95.2/100	7	100/100/100	12	96.77/87.3/100
SCOA-HS6		100/100/100	5	100/100/100	3	100/100/100	5	100/100/100

VI. CONCLUSION

A number of dual-level gene selection processes that can uses MRMR and the SCOA-HS algorithm was first proposed to minimize the number of transferred genes that offers 99. 8% accuracy pertaining to cancer difference. To this very last part, MRMR appeared to be employed to reduce the variety of body's gene history to 70 so that the acceptable time for some sort of boosting criteria might quite possibly be diminished. The very 65 most of passed dow genes was obtained with UCI Archive subsequently implemented while advices to the subsequently level about gene line, that COA-HS ended up being enclosed while using Support Vector System classifier and even served in the form of wrapper gene variety way. The exact LOOCV strategy was initially executed to understand the improvement your particular proposed technique and the consequences were validated to individuals various other advertising throughout frisco tx algorithms that are included with PSO, ATLANTA, HS, as well as SCOA. The very associated cost minimization plots illustrate in which SCOA-HS several optimization rules in mainly because it reached an improved intercontinental minimum amount for all of anyone cancer datasets examined for the reason that study. Aside from, as seen in Table several, the SCOA-HS reached almost accuracy while using00 minimum a great deal of genes just for datasets com-pared to the other regulations. The options on the picked out body's genes happen to be further more reviewed, plus it was fundamentally confirmed the fact the selected genetic makeup are biologically relevant to each kind of cancers. Therefore, the principle dual-level gene selection recommends SCOA-HS is able to select very informative biomarker genes to achieve 100% finely-detailed in meticulous a second-class cancer category task.

REFERENCES

1. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science* 286 (1999) 531–537.
2. E. Blaveri, J.P. Simko, J.E. Korkola, J.L. Brewer, F. Baehner, K. Mehta, et al., "Bladder cancer outcome and subtype classification by gene expression", *Clin. Cancer Res.* 11 (2005) 4044–4055.
3. Z. Cai, R. Goebel, M.R. Salavatipour, G. Lin., "Selecting dissimilar genes for multi-class classification, an application in cancer subtyping, *BMC Bioinformatics*", 8 (2007) 206.
4. R. Wesolowski, B. Ramaswamy, "Gene expression profiling: changing face of breast cancer classification and management", *Gene Expr.* 15 (2011) 105–115.
5. H. Hijazi, C. Chan, "A classification framework applied to cancer gene expression profiles", *J. Healthc. Eng.* 4 (2013) 255–283.

6. K.Vengatesan, Mahajan S. B, Sanjeevikumar P, Sana Moin Submitted a paper on "The Performance Enhancement of Statistically Significant Bicluster Using Analysis of Variance" , *Advances in Systems, Control and Automation, Lecture Notes in Electrical Engineering* 442, DOI 10.1007/978-981-10-4762-6_64, Chapter No.: 64
7. A. Antoniadis, S. Lambert-Lacroix, F. Leblanc, "Effective dimension reduction methods for tumor classification using gene expression data", *Bioinformatics* 19 (2003) 563–570.
8. J. Cao, L. Zhang, B. Wang, F. Li, J. Yang, "A fast gene selection method for multi-cancer classification using multiple support vector data description", *J. Biomed. Inform.* 53 (2015) 381–389.
9. A. Jain, D. Zongker, "Feature selection: evaluation, application, and small sample performance", *IEEE Trans. Pattern Anal. Mach. Intell.* 19 (1997) 153–158.
10. K.Vengatesan, Mahajan S. B, Sanjeevikumar P., Mangrula Rupali A, Kala V, Pragadeeswaran S, Submitted a paper on "Performance analysis of Gene Expression data using Mann-Whitney U Test" , *Advances in Systems, Control and Automation, Lecture Notes in Electrical Engineering* 442, DOI 10.1007/978-981-10-4762-6_67 , Chapter No.: 67
11. C.E. Gillies, M.R. Siadat, N.V. Patel, G.D. Wilson, "A simulation to analyze feature selection methods utilizing gene ontology for gene expression classification", *J. Biomed. Inform.* 46 (2013) 1044–1059.

AUTHORS PROFILE



Mrs. M. Sofia ,MSc(CS),M.Phil.,(Ph.D), Assistant Professor, mmahalakshmins@gmail.com, Mrs.M.Sofia currently pursuing her PhD in Computer Science at Sri Ramakrishna College of Arts and Science for Women and received her M.Phil in Computer Sciece from Bharathiar University. She holds her M.Sc Computer Science degree from Bharathiar University and she obtained her Bachelor Degree in Information Technology from Bharathiyar University. She has 4 years of academic experience and 5 years of research experience. Her research areas include Data Mining, Big Data Analytics. She has about 10 technical and research publications to her credit in International Journals and presented 8 papers in National and International conferences.



Dr. N.TAJUNISHA MCA.,M.Phil., Ph.D., Associate Professor, tajcs@srcw.org, Dr.N.Tajunisha received her Ph.D in Computer Science from Mother Teresa Women University, Kodaikanal and M.Phil in Computer Sciece from Bharathiar University. She holds her MCA degree from Madurai Kamaraj University and she obtained her Bachelor Degree in Mathematics from Madurai Kamaraj University. She has 19 years of academic experience and 13 years of research experience. Her research areas include Data Mining, Big Data Analytics.