Dear Editor,

Enclosed please find our substantially revised manuscript “MGRFE: multilayer recursive feature elimination based on embedded genetic algorithm for cancer classification”. In this revised manuscript, we have carefully addressed all the concerns by the two reviewers. We greatly appreciate the Referee’s comments on our manuscript. The following is our point-by-point response to each comment of the reviewers. Furthermore, I would like to take this opportunity to thank you for handling the review of our manuscript.

*Our responses to the review comments are in blue and italic.*

Sincerely yours,

Ying Li, Ph.D.

----------------------------------------------------------------------------------------------------------------------

\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Editor Comments

Associate Editor

Comments to the Author:

This manuscript was reviewed by two experts.

Both of them have concerns on comparison with other methods, ways of computational experiments, and statistical tests.

Furthermore, one reviewer recommends that the type of the paper should be changed to regular one.

And, I agree with this opinion.

(For page length/paper type issue, please do not ask me instead ask to the editorial staff or the editor in chief.)

Based on these points, I recommend the authors to revise the manuscript with taking all reviewers' comments into account.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Response**: Thanks for providing us the opportunity to revise the manuscript. The revised version considering all remarks of the reviewers has just been submitted. We have substantially revised the previous manuscript and made significant efforts in responding to the review comments. In addition, we have updated the type of the paper in the online system and changed it to a regular research one as you suggested. More details can be found below.

---------------------------------------------------------------------------

Reviewer: 1

Recommendation: Author Should Prepare A Minor Revision

Comments:

A multilayer recursive feature elimination technique based on embedded genetic algorithm for cancer classification has been presented. The authors have proposed a hybrid technique comprising both filter and wrapper methods for gene subset selection. The work is interesting and the manuscript is well organized.

1. In the introduction section author has mentioned the phrase "lack an explicit decline of the feature number". The particular phrase is not clear. Please elaborate and explain clearly the lacuna of swarm intelligence based gene selection approach.

**Response**: Thank you for this comment. In the revised manuscript, we have added more detailed explanations for this limitation of swarm intelligence based gene selection approaches. Without any mandatory controls other than iteration on feature number, it is difficult to achieve fine-turning in the reduction of feature number and enhance convergence speed. The higher the value of objective function, the better the individual obtained. In addition, the coding length of binary encoding is fixed as the gene range to represent all genes and the only way to change the number of features in swarm intelligence based gene selection approaches is cross-variation operation between individuals, making the actual number of genes or features among individuals unable to be precisely controlled. Furthermore, rather than SVM-RFE strategy which explicitly removes features with low weights at each step and in return has a precise control on feature number, none of swarm intelligence based feature selection methods mentioned in that section of the paper have used the recursive feature decline technique, leading to low computational speed and the lack of an explicit feature reduction mechanism.

* swarm intelligence based gene selection approach对基因特征的数量没有强制性的控制手段，只是靠迭代，目标函数越高个体越优秀。0-1编码，只是靠个体之间的交叉变异操作来改变特征数量，对每个个体中1的数量即特征数量没有显式的精确控制方式。
* 在SVM-RFE过程中，每一步都会显式地去除权重低的基因，这样可以对基因特征的数量有准确地控制，而且算法更快。

2. In algorithm 1, it has been mentioned to sort the optimal gene combination in GC and to preserve the top ranked genes. On what basis the top ranked gene would be sorted? For sorting what procedure is used?

**Response**: Thank you for this comment. For the first question, we sort the top ranked gene based on two metrics, fitness and gene number. The fitness of an individual is defined respectively according to different datasets. For imbalanced datasets, fitness is defined as α\*accuracy + (1- α)average accuracy, while we take α 0.6 in our experiment and the average accuracy is the average of sensitivity and specificity. For balanced datasets, fitness is simply defined as accuracy. The individual with higher fitness is superior. For two individuals with the same fitness values, the one with a smaller gene number is superior.

When it comes to the second question, the default sorting procedure in this algorithm is in Python, Timsort. Timsort, derived from merge sort and insertion sort, is a hybrid stable sorting algorithm which performs well on diverse real-world data. It has made good effects when we carried out the research on the sorting of optimal gene combination.

3. In the search space reduction stage, it has been mentioned that top 1000 genes have been selected by a threshold of 0.05 in t-test technique and thereafter MIC has been applied on the 1000 genes to re-rank them. Is there any particular reason of selection 0.05 value as threshold? Is there any mathematical reason for selecting particularly this value for threshold in t-test? Or it has been selected experimentally and any other value can also be chosen? Clarify in detail.

**Response**: Thank you for this comment. The selection 0.05 value as threshold is on the basis of mathematical and statistically theories. It is worth noting that the standard level of significance used to justify a claim of a statistically significant effect is 0.05 and there are many theories to account for the use of 0.05 in denotation statistical significance, which can trace back to the influence of *Statistical Methods for Research Workers* proposed by R.A. Fisher[1]. For better or worse, the term statistically significant has become synonymous with P≤ 0.05. Meanwhile, it is convenient to take the value of this threshold as a limit in judging whether a deviation ought to be considered significant or not. Overall, in the majority of analyses, an alpha of 0.05 is used as the cutoff for significance, guaranteeing the feasibility of most researches and studies .

4. What is the rationale for using particularly t-test first and then MIC? Can any combination of other two filter methods be used in search space reduction task? Clarify in detail.Use any other combination of two filter methods and compare it to the proposed combination of t-test and MIC-based search space reduction.

**Response**: Thank you for this comment.

* t-test（基于统计）广泛用于gene selection in microarray，MIC（基于信息）也在这个问题上有良好表现。
* （其它组合也可以）。前端的filter过程仅筛出差异性表达基因，用来缩小特征搜索空间,（理论上可换别的filter methods）。后端的多层迭代特征选择过程更重要!!！
* **更多实验。**

1. The study used *t*-test and MIC for their efficiency and convenience in gene filtering process. The *t*-test has been widely used and validated for detecting differentially expressed genes in microarray [2-4]. But t-test has limitation in dealing with multi-class dataset for multi-variate *t*-test can't be performed directly. The recently proposed MIC shows excellent performance in detecting a wide range of associations in large datasets including microarray[5, 6], and MIC can cope with multi-class dataset. Thus, we combined t-test and MIC to complete the feature screen task.
2. Why first use *t*-test and then MIC is because gene number can be quickly reduced by p-values from *t*-test. By following the commonly used p=0.05 significance threshold, we can quickly locate the statistically significant genes with p-values below 0.05. We also noted that the MIC calculation is kind of time-consuming compared with *t*-test, thus it is reasonable to perform *t*-test first to reduce the gene number.
3. Other combination of two filter methods can also be used in the search space reduction task, provided that: a) The employed two filter methods should be validated and qualified methods for finding informative genes in microarray; and b) at least one filter method could cope with multi-class dataset.
4. We use combination of first Anova then Fold change (FC) to perform comparison with combination of first *t*-test then MIC as shown in **Table 1**. The experiment was carried on 3 balanced datasets (Adeno, Gas1 and Pros) and 3 imbalanced datasets (DLBCL, Leuk, and CNS) using 5-flod cross validation. For Anova, the p-value threshold was also set as 0.05 as in *t*-test. From **Table 1**, it can notice that with the combination of Anova+FC, the sizes of finally selected genes are 2, 4, and 8 on datasets Adeno, Leuk and CNS, respectivily. But by the original *t*-test+MIC combination, simply 1, 2 and 7 genes are needed to achieve the same performance on the 3 datasets. On the rest of 3 datasets, the two filter method combinations have similar performance. Thus, the filter method combination of Anova+FC is little inferior to the original combination of *t*-test+MIC in finding the minimal discriminative gene subset.

**Table 1**. Compare the filter method combination of Anova and Fold change (FC) with combination of *t*-test and MIC on 3 balanced datasets (Adeno, Gas1 and Pros) and 3 imbalanced datasets (DLBCL, Leuk, and CNS) using 5-fold cross validation.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Filter Methods | Dataset | Genes | *Sn* | *Sp* | *Acc* | *Avc* | *MCC* | *AUC* |
| Anova+FC | Adeno | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| Gas1 | 3 | 0.987 | 0.973 | 0.98 | 0.98 | 0.96 | 0.979 |
| Pros | 4 | 0.98 | 0.982 | 0.981 | 0.981 | 0.963 | 0.982 |
| DLBCL | 3 | 1 | 1 | 1 | 1 | 1 | 1 |
| Leuk | 4 | 1 | 1 | 1 | 1 | 1 | 1 |
| CNS | 8 | 1 | 1 | 1 | 1 | 1 | 1 |
| *t*-test+MIC | Adeno | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Gas1 | 3 | 0.986 | 0.973 | 0.98 | 0.98 | 0.961 | 0.99 |
| Pros | 4 | 0.98 | 0.982 | 0.981 | 0.981 | 0.963 | 0.98 |
| DLBCL | 3 | 1 | 1 | 1 | 1 | 1 | 1 |
| Leuk | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| CNS | 7 | 1 | 1 | 1 | 1 | 1 | 1 |

5. The researchers have used t-test and then MIC. The gene selected after MIC is used in the proposed MGRFE algorithm. In table 5, why the *t*-test-based gene ranking has been compared? **MIC based ranking** should also be compared.

**Response**: Thanks for your suggestion. In the Table 1 in our previous manuscript, we listed the *t*-test ranking results for all selected genes on 17 binary datasets (is table 5 a mistake of Table 1 ?). In the revised manuscript, we have added the MIC-based gene rankings to the Table 1 and the result explain in the subsection 3.1 “Results on Dataset One*”*. For convenience to review, the t-test and MIC-based gene ranking results are listed in **Table 2**.

From the Table 2, it could be noted that:

1. The relative positions of selected genes in the two ranking methods are consistent on most datasets. For example, on the DLBCL dataset, the selected genes are ranked [13, 39, 54] in the *t*-test sorting. Meanwhile, the MIC-based gene rankings for same genes are [8, 24, 52], keeping the same ascending order as in *t*-test.
2. The top-ranked genes in the *t*-test are also top-ranked in the MIC ranking. For example, the selected gene on dataset ALL1 is the top one (with ranking 1) in both *t*-test and MIC sorting process.
3. The selected discriminatory genes are usually top-ranked by the *t*-test and MIC methods. For 5 of 17 datasets, the top one gene according to the *t*-test appeared in the final selected gene subsets. The *t*-test and MIC could find the informative genes that are important for the later feature wrapper search process.

Therefore, the employed filter techniques are qualified for the search space reduction stage.

**Table 2**. The sizes of selected gene subsets and *t*-test and MIC-based gene ranking results on the 17 binary classification datasets.

|  |  |  |
| --- | --- | --- |
| Datasets | Genes | *t*-test/MIC-based gene rankings |
| DLBCL | 3 | [13/8, 39/24, 54/52] |
| Pros | 4 | [1/1, 15/47, 74/49, 694/618] |
| Colon | 6 | [15/6, 58/21, 176/297, 225/80, 240/555, 495/482] |
| Leuk | 2 | [4/3, 7/5] |
| Mye | 7 | [3/3, 15/103, 83/142, 143/13, 378/217, 404/644, 569/707] |
| ALL1 | 1 | [1/1] |
| ALL2 | 8 | [1/80, 52/395, 78/3040, 80/1297, 522/2448, 687/2038, 737/920, 760/1449] |
| ALL3 | 8 | [4/500, 52/3437, 75/3010, 142/393, 488/443, 510/795, 715/1551, 770/1321] |
| ALL4 | 6 | [1/2, 6/45, 39/356, 282/226, 535/497, 754/1377] |
| CNS | 7 | [9/907, 53/542, 130/620, 131/519, 272/57, 273/454, 520/49] |
| Lym | 3 | [4/7, 5/4, 669/135] |
| Adeno | 1 | [468/27] |
| Gas | 3 | [22/1, 77/32, 306/36] |
| Gas1 | 3 | [132/74, 248/167, 717/500] |
| Gas2 | 2 | [38/6, 89/62] |
| T1D | 7 | [14/2229, 25/1579, 113/1287, 559/1282, 578/353, 680/426, 978/1728] |
| Stroke | 4 | [1/3, 23/115, 129/543, 276/539] |

6. Elaborate the significance of '0' ranked gene in t-test.

**Response**: Thank you for this comment. It is well-known that the index of the first element in a general array in programming is '0', so in our proposed method, the value '0' is just assigned to the first gene in the sequence. Accordingly, the t-test-based gene ranking begins with '0'.

* 程序中一般数组中第一个元素索引为0，所以这里对排在首位的基因标号为0.
* ->文章中统一改成1？

7. The comparative results of SRBCT, ALL-AML and ALL have been shown in table 7, 8 and 9. However, the tables are very similar to work in kar et al. [28]. **The similar type of comparison should be given for all the dataset used i.e. 19 dataset in the present work.**

**Response**: Thank you for this comment.

* 是的，我们在kar et al.的表格上增加了新的条目，kar et al.整理的三个表格很详细，我们可以与之前在这三个数据集上的工作有比较直观的效果比较。
* **19个都做好像不太现实**。再整理一两个？

8. The proposed work has also been compared with Kar et al. [28] in computational performance. Kar et al. have applied a swarm intelligence-based method to the space of all genes. They have not reduced the search space prior to the optimization task. In contrast, the proposed method have applied MGRFE technique on the reduced search space. The reduce search space have been constructed by t-test and then by MIC technique. In my opinion the search space reduction is fixed. It is done once before the application of MGRFE. **In that regard, the comparison of computational time would not significant because it has been computed in the reduced search space. The genes outside the reduced search space could carry valuable information towards classification accuracy.**

**Response**: Thank you for this comment.

* 首先，Kar et al. 的 KNN+PSO没有使用filter方法来降低其搜索空间是 他们程序运行时间非常高的一个因素。但是，MGRFE中的快速的RFE过程也确实使得程序运行速度大大提高。

在数据集SRBCT，ALL\_AML，MLL上，初始基因特征数分别是2308,7129,12582，KNN+PSO分别用时2.7956, 2.7906 and 7.1488 hours。MGRFE对应的在filter过程后剩余的特征数分别是700，500，700，filter+wrapper程序完整用时分别是10.8230, 9.0108 and 8.8739 minutes。处理SRBCT上700个基因MGRFE包含filter过程用了10分钟，而2308个基因KNN+PSO用了2.7956小时，差距还是比较明显的。

* 在reduced search space (规模500对于二分类，700对于多分类)中已经包含了足够多的差异性表达基因来有效地区分不同疾病种类，因为在多数microarray中最终选出的是10条以内的基因，500已经提供了较大的备选空间。
* 确实在reduced search space外也有有价值的基因，但此处我们不做考虑。

9. In the Conclusion section, the authors will need to clearly address the research contributions in theory. The research contributions in theory must be fully stated in at least one paragraph.

**Response**: Thank you for this comment. To better illustrate the theoretical contributions of this research, we have added more detailed statements in the section of discussion and conclusion in the revised manuscript. The chief research contribution in theory is providing a novel feature selection method which combines embedded genetic algorithm with recursive feature elimination process, working as a creative thought for future study. To the best of our knowledge, none previous studies have designed an evolutionary algorithm using variable length integer encoding approach in a recursive process along with the truncation selection operator to deal with the problem of minimal discriminatory feature selection in high-dimension datasets, which is described in this paper. Meanwhile, through theoretical and experimental comparisons, our proposed MGRFE could outperform mostly other state-of-the-art algorithms for gene selection on microarray data, so it can play an essential role in more high-dimensional data analyses in the future.

* 在启发式算法GA中引入RFE思路，是首例。而且结果很好。
* …

10. In the Conclusion section, the authors need to fully discuss insightful and practical implications.

**Response**: Thank you for this comment. We have supplemented sufficient practical significance and applications to make the paper more convincing and valuable. The proposed MGRFE approach can be regarded as a promising application in the analyses of high-throughput genomic, proteomic and metabolomic data, and it is also conducive to the potential cancer diagnosis by selecting the most related and minimal discriminatory genes. Moreover, the biological association between selected gene subsets with the related cancer phenotypes are validated by the literature mining on PubMed, which could provide novel cancer bio-marks candidates for related phenotype researches.

* 看最新的Gene selection in microarray文章最后结尾这里怎么处理的。仿照来写。

Reviewer: 2

Recommendation: Author Should Prepare A Major Revision For A Second Review

Comments:

First of all, the paper is described as "Survey/Tutorial," but it appears to describe a claimed original contribution by the authors, namely the MGRFE algorithm. The proposed new algorithm is compared against several existing algorithms. Therefore, if at all the paper is to be published, it should be as **a regular research paper, and not as a survey/tutorial paper.**

**Response**: Thank you for this comment. For the type of the paper, it is indeed just like your suggestion and you are insightful. We have already changed the type to a regular research one in the online system.

The paper is a mixture of techniques that are by now standard in the world of computational biology. Given a very large number of features, first use some pre-filtering to eliminate perhaps 90% to 95% of the features, and then use recursive feature elimination (RFE) on the remaining features. I could not find any **compelling evidence that the proposed approach is superior to the existing methods.**

**Response**: Thank you for this comment.

* 我们把GA和RFE过程结合起来了，效果比两类方法单独时都要好。RFE类的方法是找到的基因识别准确率不够高。群智能类的方法晒出的基因子集中往往还有多余的基因，对数量没有控制手段，而且收敛慢。

两者结合后可以优势互补，既有群智能算法强大的启发式搜索能力，又有RFE过程强制的显式基因数量下降。使得找到的基因特征子集数量区分能力高，而且数量小。[可以看文章结尾部分的说法]

The authors claim to compare their method on 17 data sets. But I did not see any evidence that the finally determined feature set is validated on an independent data set of the same form of cancer for example. **All that the authors have done is five-fold cross-validation within the same data set. Without this sort of validation on an independent data set, the claimed performance figures by themselves are not very persuasive.** This is because cross-validation within the same data set does not take into account factors such as batch effect, platform variation, and the like.

**Response**: Thank you for this comment.

* 修改文中表达，更严密。
* 对于交叉检验，别的同类文章目前也是这么做的。
* **可以的话找独立的另外数据集，用当前筛选出的基因在上面看分类效果。**

1. There are two main difficulties for validate the determined feature set on independent gene expression data set.
2. The very limited available benchmark datasets for one typical disease. It is difficult to acquire sufficient and appropriate bio-samples due to high expense of micro-array sample collection and other various factors{Dougherty, 2001 #463}, thus the available benchmark datasets are limited and the sample number in each data set is usually small. For many diseases, we just have one widely used microarray benchmark, like the colon cancer (Colon){Alon, 1999 #464} and small round blue cell tumors (SRBCT){Khan, 2001 #465}.
3. For microarray benchmark datasets about same disease, the features and sample classes are usually different. Different microarray datasets usually have different gene features for the gene probes vary among different microarray analysis platform. For example, on the leukemia related datasets of Leuk and MLL used in this study, the gene probes are very different for generating from different microarray platforms.

Thus, the currently published gene selection algorithms on microarrays are commonly validated within each microarray benchmark dataset.

1. We validated the selected gene subsets of Leuk and Gas1/Gas2 on independent datasets.
2. The leukemia datasets Leuk and MLL both have the sample data for ALL (acute lymphoblastic leukemias) and AML(acute myeloid). First, on Leuk the selected gene probes are [M23197, M31523]. Second, the genes related to these two probes are [CD33, TCF3]. Third, for these two genes, the gene probes in MLL are [32874\_at, 36802\_at, 1373\_at, 1374\_g\_at]. Finally, we test the classification performance of the obtained 4 gene probes for ALL and AML samples in MLL dataset.
3. Gas1 is about non-cardia gastric cancer, and Gas2 is about cardia gastric cancer. These two datasets are both from ref. {Wang, 2013 #466} and have the same gene probes as features. The gene probes selected by MGRFE on Gas1 are [*215380\_s\_at*, *221928\_at, 214746\_s\_at*], and the gene probes selected on Gas2 are [*210125\_s\_at*, *206361\_at*]. We validated the gene probe subset on the other dataset.

**Table 3. Indipendent validation of features selected on Leuk and Gas1/Gas2 by 10-time 10-fold cross validation.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Feature test on | Feature from | *Sn* | *Sp* | *Acc* | *Avc* | *MCC* | *AUC* |
| MLL | **Leuk** | 0.963 | 0.96 | 0.963 | 0.962 | 0.934 | 0.993 |
| MLL | **MLL** | 1 | 1 | 1 | 1 | 1 | 1 |
| Leuk | **Leuk** | 0.99 | 1 | 0.993 | 0.995 | 0.987 | 1 |
|  |  |  |  |  |  |  |  |
| Gas1 | **Gas1** | 0.984 | 0.965 | 0.974 | 0.974 | 0.952 | 0.989 |
| Gas1 | **Gas2** | 0.917 | 0.929 | 0.923 | 0.923 | 0.853 | 0.967 |
| Gas2 | **Gas1** | 0.933 | 0.827 | 0.88 | 0.88 | 0.774 | 0.973 |
| Gas2 | **Gas2** | 1 | 1 | 1 | 1 | 1 | 1 |

For dataset MLL, only two common sample kinds of ALL and AML as in Leuk are taken into consideration in this experiment.

From Table 3, it can be noted that the selected genes on Leuk achieved satisfying performance on MLL, with accuracy being 0.963. The gene sebset

The authors' **preferred method of genetic algorithms is known to lack theoretical foundations**, to be very sensitive to various parameters in the algorithm, and to be extremely time consuming. In contrast, the original paper where RFE was proposed, by Isabel Guyon, used the support vector machine (SVM) which is very fast and for which lots of theoretical results are available. This is another reason for my not being overly enthusiastic about the paper.

**Response**: Thank you for this comment. For selecting informative gene features in a microarray, the state-of-the art methods are commonly evolutionary-computation based.

* SVM-RFE理论好，但效果没有启发式群智能算法好，当前效果最领先的gene selection in microarray的方法都是第二类的。
* 时间方面，MGRFE相比于GA收敛速度有很大提升，RFE过程使得算法非常快.
* 时间在这个问题上重要性居于次位，因为对每个数据集执行一次筛选过程即够了，更重要的是筛选出的基因的区分能力。

There are several places where the authors do not appear to be **aware of simple statistical facts. For instance, the accuracy is a weighted average of the sensitivity and the specificity.**  But the authors talk as though they are independent parameters. Equation (1) in the right column of page 1 is too wide.

**Response**: Thank you for this comment. In the revised manuscript, we have made great effort to modify and optimize the expressions about statistical terms. Also, the format of our revised manuscript has been adjusted and we make the equations more organized by following the professional suggestions.

* 修改文中表达方式，调一下公式格式。

In Section 2.3.1 the authors use the T-test and MIC to achieve a first-cut reduction in the feature set. I have found that using the so-called "volcano plot," which combines the T-test with a fold-change criterion, works better than just the T-test alone.

**Response**: Thank you for this comment.

* **多一点实验？**

**Reference**

1. Fisher, R.A., *Statistical Methods for Research Workers*. 1958: Oliver and Boyd. 66-70.

2. Cui, X.Q. and G.A. Churchill, *Statistical tests for differential expression in cDNA microarray experiments.* Genome Biology, 2003. **4**(4).

3. Lazar, C., et al., *A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis.* Ieee-Acm Transactions on Computational Biology and Bioinformatics, 2012. **9**(4): p. 1106-1119.

4. Shen, Q., W.M. Shi, and W. Kong, *Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data.* Computational Biology and Chemistry, 2008. **32**(1): p. 53-60.

5. Ge, R.Q., et al., *McTwo: a two-step feature selection algorithm based on maximal information coefficient.* Bmc Bioinformatics, 2016. **17**.

6. Reshef, D.N., et al., *Detecting Novel Associations in Large Data Sets.* Science, 2011. **334**(6062): p. 1518.