Evaluation of Heuristic-Based MicroRNA Marker Selection Techniques for Classification of Cancer

Eliza Razak, Faridah Yusof, Raha Ahmad Raus
International Islamic University Malaysia, Kuala Lumpur, Malaysia
yfaridah@iium.edu.my

***Abstract*—Understanding and recognizing genetic sequences is one step towards the treatment of the genetic disorders. Cancer, which is a major leading genetic disorder and responsible for around 13% of all deaths world-wide. Since the discovery of DNA, there has been a growing interest in genetic sequence recognition and gene expression analysis, inspired by its promising potential to cure a broad range of genetic disorders. Conventional biopsy examinations are highly invasive since tissue samples are required to be extracted from patients. Blood-based biomarkers have given optimism about the future cancer management. There have indeed been a number of studies to identify novel miRNA-based cancer biomarkers. However, the existing diagnosis techniques using miRNA suffer from low diagnosis accuracy, sensitivity, and specificity. The low diagnosis accuracy and sensitivity of the existing techniques stems from the fact that there is extremely low miRNA count in body fluids. In this paper, we employed three marker selection algorithms to select relevant miRNAs that are directly responsible for cancer classification. Among the three methods, gain ratio (GR) results are quite encouraging. Despite much noise contaminated in the datasets, the predictive framework able to identify miRNA markers responsible for classification of cancer.**

***Index Terms*—miRNA, Cancers, Marker selection, Instance-Based k-nearest-Neighbor Classification (IBK)**

I. Introduction

Many people succumb to cancer every day. An early cancer detection and classification system is essential in order to save countless lives. Cancer is a family of diseases that involve uncontrolled cell growth wherein cells divide and grow exponentially, generating metastatic malignant tumors. Although currently available protein biomarkers and biopsy examination have low sensitivity, specificity and require invasive sampling procedures, they been widely used in cancer management cancer management [1]. Recently, there has been a tremendous increase in interest concerning circulating microRNAs (miRNAs) as a potent cancer biomarker to improve cancer management cancer management [2]. In fact, miRNAs are small non coding extracellular RNAs, approximately 18 to 22 nucleotides long, which are produced through a series of complex biogenesis pathway [3, 4]. Up to now, over two thousand human miRNA have been identified [5, 6]. miRNAs can be identified in body fluid such as blood, plasma, serum, urine and saliva [7, 8]. Ectopic microRNA expression have been associated with tumorigenesis, cancer metastasis [9]. In fact, miRNA expression data are suffering from high-level of noise (resulting from low miRNA sample count in body fluids and contamination in sample preparation). Furthermore, irrelevant features imply high dimensionality of input data. High dimensionality can cause over-fitting, reduce the generalization ability of the cancer classification system and can elevate the computational complexity of cancer classification systems. The robustness and the generalization ability of the classifier are directly proportional to its complexity [10]. Therefore, the marker selection process is an imperative stepping stone to accurate and reliable cancer classification.

Manuscript received June 3, 2016; revised July 11, 2016; accepted September 1, 2016.

II. Proposed Methods

A. Datasets

The proposed framework has been tested on two publicly available datasets which were gastric cancer (E-TABM-341) and ovarian cancer (E-TABM-343). Gastric cancer dataset contains total 353 samples and 315 miRNAs whereas ovarian cancer dataset contains total 84 samples and 1569 miRNAs.

B. Marker Selection

This study selected to employ three different marker selection algorithms which are gain ratio, correlation based filter and Relief. The main goal is to select out the most informative miRNAs subsets that can sufficiently discriminate classes of cancer prior to classification step. The aim of marker selection is to pick a subset of miRNA markers, $S \subset G$, that can sufficiently discriminate cancer type **Y**, given $|S| = s$ where $(s \ll g)$.

1) Gain ratio (GR)

Gain ratio (GR) is a modified version of the information gain algorithm as it offers normalized miRNA expression values. The GR of a miRNA is a number between 0 and 1 which approximately represents the degree of 'significance' of the miRNA in

doi: 10.18178/ijlbpr.5.1.18-21

discriminating different classes of cancer [11]. A GR of 0 roughly indicates that the corresponding miRNA has no significance in discriminating classes of cancer while a GR of 1 roughly indicates that the miRNA is significant in discriminating classes of cancer. The GR algorithm is visually depicted in Table I. The gain ratio (GR) between Y and s is defined to be:

$$GR(i;\mathbb{S}) \stackrel{\text{def}}{=} \frac{IG(i;\mathbb{S})}{H(\boldsymbol{Y})} \quad (1)$$

where IG denotes Information gain (IG) of a gene to the class label $\boldsymbol{Y}$ and $H(\boldsymbol{Y})$ indicates entropy of cancer class label $\boldsymbol{Y}$.

TABLE I. ITERATIVE MARKER SELECTION ALGORITHM BASED ON GAIN RATIO

Input: **X** (miRNA) and **Y** (cancer class)

Output: $\mathbb{S}$ (Selected miRNA markers)

Let $\mathbb{S} = \emptyset$;

Loop

$i = arg\,max_{i \epsilon G}\, GR\,(i;\mathbb{S})$;

$\mathbb{S} \leftarrow \mathbb{S} \cup \{i\}$;

up to $(|\mathbb{S}| \geq threshold)$ or $\left(\frac{d\,\mathrm{GR}(i;\mathbb{S})}{d|\mathbb{S}|} \approx 0\right)$

2) Correlation based filter algorithm (CFS)

Correlation based filter algorithm is a correlation based heuristic and arguably one of the fastest ways to classify cancer [12]. In the prediction phase, correlation based filter algorithm sequentially loops through all the samples and chooses the subsets of miRNA markers that are highly correlated with the cancer type **Y**. Correlation based filter algorithm is based on symmetrical uncertainty. The CFS algorithm is depicted in Table II. Symmetrical uncertainty (SU) measures the degree of correlation between markers and it is defined to be:

$$SU = 2.0 \times \frac{H(X) + H(Y) - H(X,Y)}{H(Y) + H(X)} \quad (2)$$

TABLE II. ITERATIVE MARKER SELECTION ALGORITHM BASED ON CFS

Input: **X** (miRNA) and **Y** (cancer class label)

Output: $\mathbb{S}$ (Selected miRNA markers)

1. Let $\mathbb{S} = \emptyset$;
2. **Loop**
3. $i = arg\,max_{i \epsilon G}\, SU\,(i;\mathbb{S})$;
4. $\mathbb{S} \leftarrow \mathbb{S} \cup \{i\}$;

until $(|\mathbb{S}| \geq threshold)$ or $\left(\frac{d\,\mathrm{SU}(i;\mathbb{S})}{d|\mathbb{S}|} \approx 0\right)$

3) Relief-F Algorithm

Relief-F algorithm is another type of marker subset selection algorithm with the principal of selecting markers randomly using attribute weighting method [13]. Furthermore, Relief-F normalizes expression values to a weight between −1 and 1. Additionally, Relief-F measures the conditional dependencies among attributes and it is applicable for both binary and continuous data. The Relief-F algorithm is depicted in Table III. The Relief-F is defined to be:

$$\begin{aligned} W_X &= P\left(different\ value\ of\ X \middle| \begin{matrix} nearest\ instances\ from \\ different\ class \end{matrix}\right) \\ &- P\left(different\ value\ of\ X \middle| \begin{matrix} nearest\ instances\ from \\ same\ class \end{matrix}\right) \end{aligned} \quad (3)$$

TABLE III. ITERATIVE MARKER SELECTION ALGORITHM BASED ON RELIEF-F

Input: **X** (miRNA) and **Y** (cancer class label)

Output: $\mathbb{S}$ (Selected miRNA markers)

1. Let $\mathbb{S} = \emptyset$;
2. **loop**
3. $i = arg\,max_{i \epsilon G}\, W_X\,(i;\mathbb{S})$;
5. $\mathbb{S} \leftarrow \mathbb{S} \cup \{i\}$;

until $(|\mathbb{S}| \geq threshold)$ or $\left(\frac{d\,W_X(i;\mathbb{S})}{d|\mathbb{S}|} \approx 0\right)$

C. Classification

Instance Based K-Nearest Neighbor (IBK) classification is arguably one of the easiest ways to classify cancer. It is also called memory-based learning or "lazy" learning because there is no training phase [14]. In the prediction phase, given an expression vector $\boldsymbol{x} \in \mathbb{R}^n$, an instance-based cancer classifier sequentially loops through all the samples and chooses the class label of the sample that is most similar to $\boldsymbol{x}$ in terms of some distance metric, which is traditionally Euclidean distance or edit distance, as output. An IBK, as the name suggests, simply chooses the majority of the class labels of the ƙ samples that are nearest to $\boldsymbol{x}$. IBK is calculated by the following equation.

$$d_i = \sqrt{(X_{N+1} - X_i)^2 (X_{N+1} - X_i)} \quad (4)$$

D. Validation

In order to assess the feasibility and validity of the proposed predictive framework, leave-one-out cross-validation (LOOCV) was applied [15]. The results will be then averaged to produce an estimate of the accuracy of the system. The following performance metrics were used to gauge the performance of the system: Accuracy and area under the curve (AUC) [16-18].

III. EXPERIMENTAL RESULTS

A. Gastric Cancer

The gastric cancer dataset contains 353 samples and expression values for 315 miRNAs. Figure 1 and 2 depict the accuracy and AUC of the predictive framework.

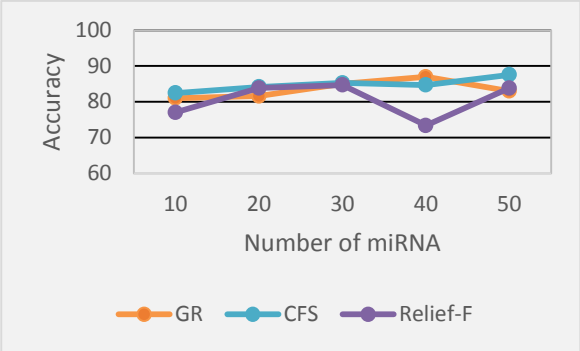

Figure 1. Performance benchmarking of accuracy for three search algorithms and IBK for gastric cancer dataset.

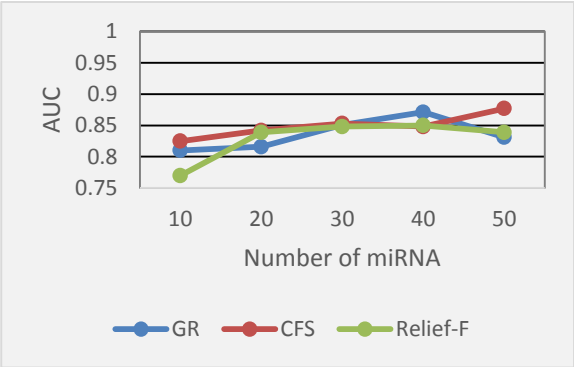

Figure 2. Performance benchmarking of AUC for three search algorithms and IBK for gastric cancer dataset.

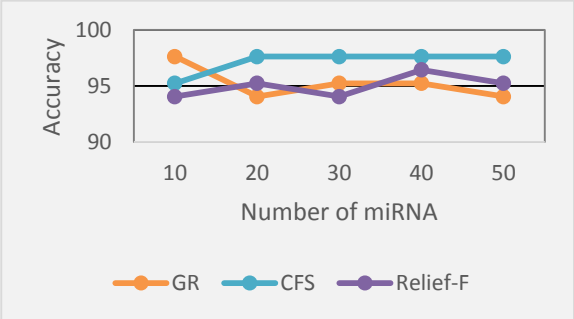

Figure 3. Performance benchmarking of accuracy for three search algorithms and IBK for ovarian cancer dataset.

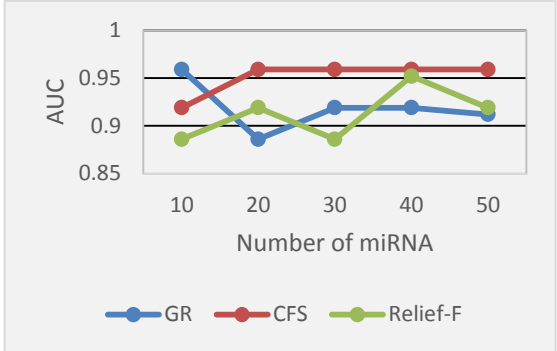

Figure 4. Performance benchmarking of AUC for three search algorithms and IBK for ovarian cancer dataset.

B. *Ovarian Cancer*

The ovarian cancer dataset contains 84 samples and 11,714 miRNAs. Figure 3 and 4 illustrate the accuracy and AUC of the predictive framework for ovarian cancer dataset.

C. *Performance Analysis*

Table IV summarizes the accuracies and number of selected miRNAs for IBK for gastric and ovarian datasets. For both datasets, CFS improves the accuracy of IBK. However, CFS takes longer evaluation time than GR and Relief-F. Similarly, GR can enhance the accuracy of IBK and at the same it is also a significantly fast search algorithm. The Relief-F takes lesser searching time than CFS but its performance is less robust than CFS and GR. The experimental results suggested that GR can provide most informative miRNAs subsets that can sufficiently discriminate classes of cancer in lesser amount of time with higher accuracy. In the training phase, the marker selection algorithms uncover markers based on a certain threshold. In the prediction phase, given an expression vector $\boldsymbol{x} \in \mathbb{R}^n$, all the miRNA are removed except the relevant miRNA markers. Finally, the output expression vector is $\boldsymbol{x} \in \mathbb{R}^s$ with reduced dimensionality. Therefore, the marker selection process is an imperative stepping stone to accurate and reliable cancer classification.

TABLE IV. ACCURACY AND NUMBER OF SELECTED MIRNAS FOR GR, CFS AND RELIEF-F ALGORITHMS

Title	Accuracy			Number of selected miRNAs		
Marker selection algorithms	GR	CFS	Relief-F	GR	CFS	Relief-F
Gastric cancer	86.9688%	87.5354 %	84.7025 %	40	50	30
Ovarian cancer	97.619 %,	97.619 %	96.4286 %	10	18	40
Average	**86.97%**	**92.58%**	**90.57%**	**25**	**34**	**35**

IV. CONCLUSION

This paper benchmarked the performance of three marker selection algorithms and IBK classifier to classify cancers from miRNA expression data. While comparing the accuracy and AUC of classification with respect to feature selection methods, Gain Ratio (GR) results are more encouraging. Despite much noise contaminated in

the datasets, the predictive framework able to identify miRNA markers responsible for classification of cancer.

REFERENCES

[1] K. Mäbert, et al., "Cancer biomarker discovery: current status and future perspectives," *International Journal of Radiation Biology*, vol. 90, no. 8, pp. 659-677, 2014.

[2] C. M. Croce, et al., "Diagnosis and treatment of cancers with microRNA located in or near cancer-associated chromosomal features," US Patent 20,150,368,647, 2015.

[3] D. M. Pereira, et al., "Delivering the promise of miRNA cancer therapeutics," *Drug Discovery Today*, vol. 18, no. 5, pp. 282-289, 2013.

[4] B. Xie, et al., "miRCancer: a microRNA–cancer association database constructed by text mining on literature," *Bioinformatics*, p. btt014, 2013.

[5] S. C. Eastlack and S. K. Alahari, "MicroRNA and breast cancer: understanding pathogenesis, improving management," *Non-Coding RNA*, vol. 1, no. 1, pp. 17-43, 2015.

[6] J. K. Rane, et al., "MicroRNA expression profile of primary prostate cancer stem cells as a source of biomarkers and therapeutic targets," *European Urology*, vol. 67, no. 1, pp. 7-10, 2015.

[7] X. Zhong, G. Coukos, and L. Zhang, "miRNAs in human cancer," *Next-Generation MicroRNA Expression Profiling Technology*, pp. 295-306, 2012.

[8] G. Cheng, "Circulating miRNAs: roles in cancer diagnosis, prognosis and therapy," *Advanced Drug Delivery Reviews*, vol. 81, pp. 75-93, 2015.

[9] T. A. Farazi, et al., "MicroRNAs in human cancer," *MicroRNA Cancer Regulation*, pp. 1-20, 2013.

[10] E. Alpaydin, *Introduction to Machine Learning*, MIT press, 2014.

[11] C. K. K. Reddy, C. Rupa, and B. V. Babu, "SLGAS: Supervised learning using gain ratio as attribute selection measure to nowcast snow/no-snow," *International Review on Computers and Software (IRECOS)*, vol. 10, no. 2, pp. 120-129, 2015.

[12] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 1, pp. 1-14, 2013.

[13] E. Mouelhi, W. Bouaguel, and G. B. Mufti, "A redundancy study for feature selection in biological data," *Mining Intelligence and Knowledge Exploration*, pp. 22-28, 2015.

[14] P. Manikandan, et al., "An improved instance based k-nearest neighbor (IIBK) classification of imbalanced datasets with enhanced preprocessing," *International Journal of Applied Engineering Research*, vol. 11, no. 1, pp. 642-649, 2016.

[15] M. G. Natrella, *Experimental Statistics*, Courier Corporation, 2013.

[16] N. J. M. Blackman and J. J. Koval, "Interval estimation for Cohen's kappa as a measure of agreement," *Statistics in Medicine*, vol. 19, no. 5, pp. 723-741, 2000.

[17] A. Ben-David, "About the relationship between ROC curves and Cohen's kappa," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 6, pp. 874-882, 2008.

[18] B. Rosner, *Fundamentals of Biostatistics*, Cengage Learning, 2010.

Eliza Razak is a Ph.D. student from Department of Biotechnology Engineering, Kulliyah of Engineering, International Islamic University Malaysia. Her research interests mainly lie in bioinformatics.