Meta-classification of Multi-type Cancer Gene Expression Data

Benny Y.M. Fung
Department of Computing,
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
csymfung@comp.polyu.edu.hk

Vincent T.Y. Ng
Department of Computing,
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
cstyng@comp.polyu.edu.hk

ABSTRACT

Massive publicly available gene expression data consisting of different experimental conditions and microarray platforms introduce new challenges in data mining when integrating multiple gene expression data. In this work, we proposed a meta-classification algorithm, which is called MIF algorithm, to perform multi-type cancer gene expression data classification. It uses regular histograms for gene expression levels of certain significant genes to represent sample profiles. Differences between profiles are then used to obtain dissimilarity measures and indicators of predictive classes. In order to demonstrate the robustness of the algorithm, 10 different data sets, which are individually published in 8 publications, are experimented. The results show that the MIF algorithm outperforms the simple majority-voting meta-classification algorithm and has a good meta-classification performance. In addition, we also compare our results with other researchers' works, and the comparisons are impressive. Finally, we have confirmed our findings with cancer/testis (CT) immunogenic gene families of heterogeneous samples.

Keywords

Gene expression, meta-classification, heterogeneous, multi-type

1. INTRODUCTION

Although DNA microarray techniques bring breakthroughs to cancer study, massive publicly available gene expression data, which are conducted by different laboratories with various experimental conditions and microarray platforms, introduce new challenges to conduct data mining with an integration of multiple and heterogeneous gene expression data. For gene expression data in cancer study, the advance of data mining leads to the discovery of global cancer profiling, patient classification, tumor classification, tumor-specific molecular marker identification and pathway exploration [15]. Different mining algorithms have been proposed, and significant findings are exploited corresponding to different algorithms. For most cases, validations of findings are done by a series of biological experiments or laboratorial works. However, in terms of efficiency and effectiveness of mining algorithms with respect to clinical applicability and robustness, the validations are mainly restricted by cross-validation or sub-sampling within a single data set [4], [11]. This validation scheme is not sufficiently to draw conclusions because of the problems of over-fitting and homogeneity within a single data set. To avoid these problems, there are two potential solutions: (1) it is required to validate mining algorithms with heterogeneous data sets consisting of different microarray platforms and experimental conditions, and (2) meta-analysis is performed with a number of heterogeneous data sets so that it can make meta-decisions with an integration of these data sets, rather than with individual data sets [5], [19].

To perform classification of heterogeneous data consisting of multi-type cancer, some common features (i.e. significant genes) must be founded in various cancer types. Subsets of genes, which are called cancer/testis (CT) immunogenic gene families, are recently proposed to have associations with one or more than one cancer type. Van der Bruggen et al. [23] suggested an approach to identify the molecular definition of tumor antigens recognized by T cells, and this approach leads to the discovery of various human tumor antigens, such as MEGEA1 and BAGE. Discovered tumor antigens are recently grouped into distinct subsets, and the subsets are named as cancer/testis (CT) immunogenic gene families. Currently, researchers have discovered 44 CT immunogenic genes families consisting of 89 individual genes in total [20].

In our previous works, we proposed a measure called "impact factors (IFs)" to improve the classification performance of heterogeneous gene expression data [7], [8]. In this paper, we extend the works and propose a meta-classification algorithm, which is called Majority-voting with Impact Factors (MIF) algorithm, to classify multi-type cancer gene expression data consisting of both different cancer types and microarray platforms. In order to validate the reliability and robustness of the MIF algorithm, 10 gene expression data sets, which are published in 8 different publications, are experimented, and the classification performance of the MIF algorithm is not only compared with the simple majority-voting meta-classification algorithm, but also with results of other researchers in [2].

2. RELATED WORKS

Recent progress in mining gene expression data is to discover knowledge from multiple and heterogeneous gene expression data. Some works are concerning theoretical flexibility to integrate gene expression data with various microarray platforms and technologies. Lee et al. [10] and Kuo et al. [9], respectively, described different approaches based on simultaneous mutual validation of large numbers of genes using two different microarray platforms. They used the NCI-60 data sets consisting of spotted cDNA arrays and Affymetrix oligonucleotide chips. Choi et al. [5] proposed a systematic integration of gene expression data based on normalizing data with an estimated means of other data sets.

For application level, classification is one of the common areas in data mining of gene expression data. Ng et al. [13] proposed a method to perform subtype classification with six different gene expression studies on Saccharomyces cervisiae. Recently, Bloom et al. [2] conducted a study of multi-platform, multi-type and multi-site classification on cancer gene expression data. In the study, 15 cancer types, published in 4 different publications, are experimented.

Meta-classification approaches are mainly divided into three categories [21]. The first category is to average individual

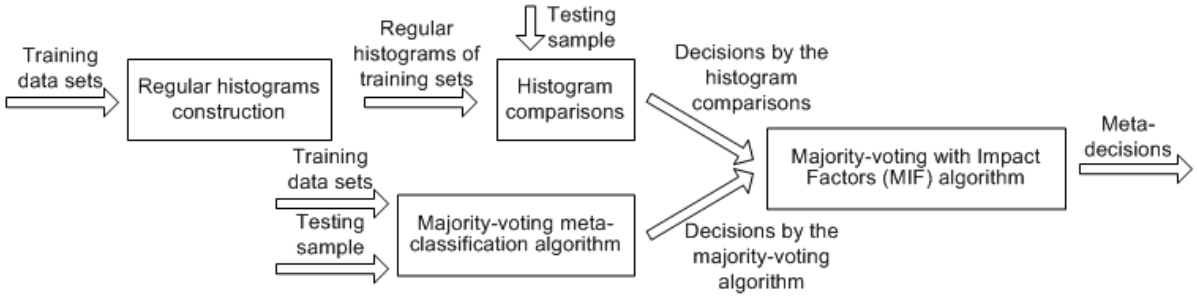

Figure 1. Process overview of the MIF algorithm.

decisions of different element classifiers without altering the original learning algorithms of the element classifiers. The second category is to predict the right learning algorithm or classifier for a particular problem from a set of element classifiers based on analyzing the fitness of the characteristics of testing data sets. The last category is to take a sub-sample of the entire data set and try each algorithm on this sub-sample. Among these three categories, the category of model averaging draws more attention in the literatures. For gene expression data, most works also belong to the category of model averaging. Some works include majority-voting [3], Bayesian combination [4], weighted-voting [4] and neural network ensembles [26].

3. MIF ALGORITHM

In this work, we proposed a meta-classification algorithm, called Majority-voting with Impact Factors (MIF) algorithm, to perform multi-type cancer gene expression data classification. It uses regular histograms for gene expression levels of certain significant genes to represent the profiles of samples. Differences between profiles are then used to obtain dissimilarity measures and indicators of predictive classes. The regular histograms are constructed by the uniform partitioning technique with maximum and minimum expression levels of the significant genes as upper and lower bounds. It aims at estimating densities of expression levels of significant genes in terms of relative positioning with respect to the upper and lower bounds. For a new sample, it compares its histograms with the histograms of individual classes in training sets. The classes with smaller dissimilarity measures are set as predictive classes for the new sample. As the same time, the majority-voting meta-classification algorithm is performed with the new sample too. If the decisions derived from the regular histogram comparisons and the majority-voting algorithm are the same, weighted scores corresponding to individual classes, which are based on the impact factors (IFs), are accumulatively adjusted the dissimilarity measures of the corresponding classes. On the other hands, if their decisions are different, there are no such weighted scores, and the dissimilarity scores are increased according to the results of the majority-voting algorithm. Figure 1 shows the process overview.

Here, we describe the MIF algorithm in details. First of all, individual regular histograms of every sample in each class in training sets are constructed [12]. Suppose that there are m training sets represented by the vector $X=(X_1, X_2, \ldots, X_m)$, and $X_i=(x_{i,1}, x_{i,2}, \ldots, x_{i,l}, x_{i,l+1}, \ldots, x_{i,n})$ be the training set i with l normal samples and $(n–l)$ cancer samples. The expression levels of gene g in X_i be represented by a vector $g=(e_{i,1}, e_{i,2}, \ldots, e_{i,n})$, where $e_{i,j}$ represents the expression level of g in sample j of set i (i.e. X_i), and $c=\{Normal,Cancer\}$ be the class vector such that $x_{i,j}.c$ representing the classes of sample j in set i. The algorithm for the regular histogram construction for training samples is shown in figure 2.

Inputs: aligned training samples sets X, number of bins n_b, number of significant genes n_g

Outputs: pairs of regular histograms for all training samples sets H_{Normal} and H_{Cancer}, sets of significant genes for all training sets G

1. **variables**:
2. $temp_{Normal}$ and $temp_{Cancer}$ be the temporary sets of regular histograms for each candidate of X_i, $temp_{Sig}$ be the temp set of significant for X_i, α be the percentage of bin candidates to be trimmed
3. *for i = 1 to size(X)*
4. $temp_{Normal} = \phi$;
5. $temp_{Cancer} = \phi$;
6. $temp_{Sig}$ = *find_sig_genes* (X_i);
7. $G = G + temp_{Sig}$;
8. *for j = 1 to size*(X_i)
9. *if* ($x_{i,j}.c$ = *Normal*)
10. $temp_{Normal} = temp_{Normal} + hist\_proc(x_{i,j}, n_b, temp_{Sig})$;
11. *else*
12. $temp_{Cancer} = temp_{Cancer} + hist\_proc(x_{i,j}, n_b, temp_{Sig})$;
13. *end if*
14. *end for*
15. $H_{Normal} = H_{Normal} + normalize\ (temp_{Normal},\alpha)$;
16. $H_{Cancer} = H_{Cancer} + normalize\ (temp_{Cancer},\alpha)$;
17. *end for*

Figure 2. Algorithm for calculating regular histograms for training samples sets.

In figure 2, for each training set X_i, where $X_i \in \{X\}$, significance of genes in X_i is calculated and ranked accordingly in the function "find_sig_genes" at code line 6. The common and widely used statistical method *t*-test is used to rank significance of the genes [6]. In the *t*-test, its sign is determined by the numerator. Therefore, the *t*-values are positive if the mean of normal class is larger than that of cancer class and negative if the mean of normal class is smaller than that of cancer class. Hence, taking genes from both tails from the sorted list, including positive and negative *t*-values, can assume that the same proportions of genes from both classes are considered. Extracted significant genes sets, $G=\{G_1, G_2, \ldots, G_m\}$, where G_i is the significant gene set in training X_i, are later used to construct and compare the histograms of testing samples.

At code lines 10 and 12 in figure 2, the function "hist_proc" is invoked to construct the regular histograms. The maximum and minimum expression levels among those extracted significant genes are set as the upper and lower bounds of the histograms.

Samples belong to the same classes of the same training sets may have different values for upper and lower bounds. However, we are only interested in the densities of expression levels with respect to sample-based maximum and minimum expression levels, which is in relative positioning. Therefore, if the absolute differences of a sample between two bounds are smaller than other samples, their global differences among significant genes will be smaller in a similar ratio as the bounds also. As a result, the effects of the absolute differences can be eliminated.

The uniform partitioning technique is used to evenly divide the distance between the upper and lower bounds into a required number of bins n_b. Each bin width is defined by *(upper-lower)/*n_b. Each data set should have *l* and *(n–l)* different regular histograms for normal and cancer samples, and all histograms should have n_b bins because of the uniform partitioning. For example, figure 3 shows an example. Assume that there are 100 significant genes, n_b is 10 bins, and the upper and lower bounds are 4917 and -652. By applying the uniform partitioning technique, each bin width is [4917-(-652)]/10=557 to nearest integer. Expression levels of identified significant genes are then mapped to different bins with respect to their expression levels, and the results are shown in figure 3. At the end, the regular histogram of the illustrated sample is represented by the vector of (0.11, 0.76, 0.07, 0.02, 0.01, 0, 0, 0.01, 0, 0.02).

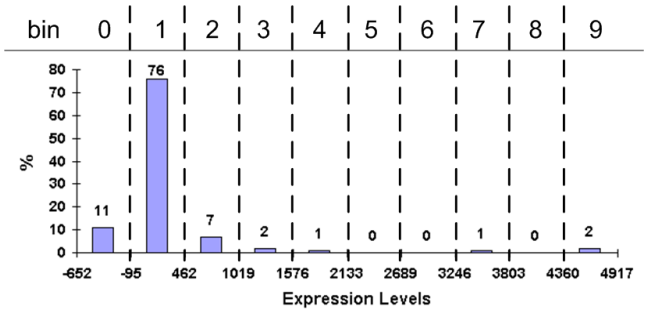

Figure 3. Example of regular histogram's construction for expression levels of significant genes.

After all the histograms corresponding to the same class of the same training sets (i.e. the *for* loop at code line 8) have been computed, $\alpha\%$ candidate bins with highest and lowest bin values are trimmed to eliminate the effects of outliners. Remaining bins are then accumulated to form a representative histogram of individual classes in the data sets. Since some entries are trimmed, the value of the sum of all bin values at the representative histograms can be unbounded. It causes inconsistent scaling when comparing with other histograms. In order to have consistent comparisons, normalization is done so that the sum of all bin values in a single representative histogram to have the sum equals to 1. Finally, all representative histograms for individual training sets are added to H_{Normal} and H_{Cancer}. To use the same example in figure 3, the resultant vector becomes (0.76, 0.07, 0.02, 0.01, 0, 0, 0.01, 0) after 5% of candidate bins with highest and lowest bin values are trimmed. In addition, the normalized vector becomes approximately (0.87, 0.09, 0.02, 0, 0, 0.02, 0) in order to have sum equals to 1.

With the computed H_{Normal} and H_{Cancer}, comparisons of the histograms between training and testing samples can be performed. Figure 4 shows the algorithm of the comparisons.

Inputs: pairs of regular histograms for all training sample sets H_{Normal} and H_{Cancer}, sets of significant genes for all training sets *G*, testing sample *s*, number of bins n_b

Outputs: predictive classes by the regular histogram comparisons C_{Hist}

1. **variables**:
2. H_s be the temporary variable of the regular histogram of the testing sample
3. *for i = 1 to size(*H_{Normal}*)*
4. H_s = *hist_proc(s,* n_b*,* G_i*);*
5. *if (dis(*H_s*,* $H_{Normal,\,i}$*) < dis(*H_s*,* $H_{Cancer,\,i}$*))*
6. C_{Hist} = C_{Hist} + *{Normal};*
7. *else*
8. C_{Hist} = C_{Hist} + *{Cancer};*
9. *end if*
10. *end for*

Figure 4. Algorithm for the comparisons of regular histograms between testing and training samples.

First of all, regular histogram of the testing sample *s* with respect to the significant genes set *G* of the training sets is computed. Then, dissimilarity measures between the testing sample and individual classes of training sets are computed, respectively. Assume that $H_s(b)$ be the regular histograms of the testing sample with bin *b*, and $H_c(b)$ is the regular histograms of the classes in the training sets with bin *b*, where *c={Normal,Cancer}*. Now, the dissimilarity measures, *dis*, between two histograms are calculated as:

$$dis(H_s,H_c|c\in\{Normal,Cancer\})=\sum_b\frac{|H_s(b)-H_c(b)|}{|H_s(b)+H_c(b)|} \qquad (1)$$

The second step is to compare the histogram of the testing sample to pairs of the histograms in each training set and determine predictive classes of the new sample with respective to individual training sets in the code segment from line 5 to line 9 in figure 4. For each training set, there are two histograms corresponding to it, one for each class. The dissimilarity measures of normal and cancer classes are compared, and the classes with smaller values of the measures are set as the predictive classes of the testing sample, and assigned as a new element in set C_{Hist}. Since there is a single prediction for each training set, so there are m elements in C_{Hist} for m different training sets.

At the same time, the majority-voting meta-classification algorithm is performed. In [8], we proposed an empirically-driven model averaging method to integrate individual classification decisions to form meta-decisions. Suppose that there is a data set *D*, and the data are arisen from *k* possible models (i.e. combinations of classifiers and data sets), $M=(M_1,\ldots,M_k)$. If Δ is the quantity of interest (i.e. classification performance), then its posterior distribution of Δ in data set *D* is:

$$pr(\Delta|D)=pr(\Delta|\beta_1,\ldots,\beta_k,D)=\sum_{i=1}^{K}(\beta_i\times pr(\Delta|M_k,D)) \qquad (2)$$

, where β_i is the quantity of pre-knowledge for model M_i, and it is defined as:

$$\beta_i = \frac{acc(D,M_i)\ x\ S_p(D,M_i)x\ S_n(D,M_i)}{\sum_{l=1}^{K} acc(D,M_l)\ x\ S_p(D,M_l)x\ S_n(D,M_l)} \quad (3)$$

, where $acc(D,M_i)$, $S_p(D,M_i)$ and $S_n(D,M_i)$ are the classification accuracy, specificity and sensitivity of model M_i with data set D.

To perform the majority-voting algorithm, K is set to 1 in equation 2. Therefore, we only consider a single model each time, and finally there are k individual decisions for k different models. Hence, the equation is rewritten as:

$$pr(\Delta \mid D) = pr(\Delta \mid \beta_i, D) = \sum_{i \in K} (\beta_i \times pr(\Delta \mid M_i, D)) \quad (4)$$

If there are m and k different training sets and classifiers, there will be $(m \times k)$ individual decisions for the testing sample (i.e. each model produce a decision). For each decision, it is determined by a pair of Δ. Since we are interested in predictive classes of testing sample s, represented as $s.c$, one way to make the prediction is to compare the values of $pr(s.c=Normal|D)$ and $pr(s.c=Cancer|D)$, where $c \in \{Normal, Cancer\}$. If $pr(s.c=Normal|D)$ is larger than $pr(s.c=Cancer|D)$, assigned predictive classes are normal. Otherwise, it is assigned as cancer. In order form meta-decisions among individual decisions, the majority-voting algorithm in equation 5 assigns predictive classes, C_{Vote}, which are the most often predictive classes of individual decisions $s.c_i$.

$$(s.c \mid s.c \in \{C_{Vote}\}) = \underset{c \in \{Normal, Cancer\}}{arg\,max} \sum_{i:\{s.c_i = c\}} 1 \quad (5)$$

Inputs: testing sample s, sets of significant genes for all training sets G, number of bins n_b, predictive classes by the regular histogram comparisons C_{Hist}, predictive classes by the majority-voting algorithm C_{Vote}, impact factors for normal and cancer classes IF_{Normal} and $IF_{Cancer,}$ pairs of regular histograms for all training sample sets H_{Normal} and H_{Cancer}, pre-knowledge measures corresponding to training sets β.

Outputs: meta-decisions C_{Pred}

1. **variables**:
2. d_{Normal} and d_{Cancer} be the dissimilarity values to normal and cancer classes, d_{Acc_normal} and d_{Acc_cancer} be the accumulative dissimilarity values to normal and cancer classes
3. *for i = 1 to size(C_{Hist})*
4. *if ($C_{Hist,\,i} = C_{Vote,\,i}$)*
5. *if ($C_{Hist,\,i}$ = Normal)*
6. $d_{Normal} = \beta_i \times IF_{Normal,\,i} / IF_{Cancer,\,i} \times dis(hist\_proc(s, n_b, G_i), H_{Normal,\,i}) / dis(hist\_proc(s, n_b, G_i), H_{Cancer,\,i})$;
7. *else*
8. $d_{Cancer} = \beta_i \times IF_{Cancer,\,i} / IF_{Normal,\,i} \times dis(hist\_proc(s, n_b, G_i), H_{Cancer,\,i}) / dis(hist\_proc(s, n_b, G_i), H_{Normal,\,i})$;
9. *end if*
10. *else*
11. *if ($C_{Hist,\,i}$ = Normal)*
12. $d_{Normal} = \beta_i \times IF_{Cancer,\,i} / IF_{Normal,\,i}$;
13. *else*
14. $d_{Cancer} = \beta_i \times IF_{Normal,\,i} / IF_{Cancer,\,i}$;
15. *end if*
16. *end if*
17. $d_{Acc\_normal} = d_{Acc\_normal} + log2\ (d_{Normal})$;
18. $d_{Acc\_cancer} = d_{Acc\_cancer} + log2\ (d_{Cancer})$;
19. *end for*
20. *if ($d_{Acc\_normal} < d_{Acc\_cancer}$)*
21. $C_{Pred} = C_{Pred} + \{Normal\}$;
22. *else*
23. $C_{Pred} = C_{Pred} + \{Cancer\}$;
24. *end if*

Figure 5. MIF algorithm.

Figure 5 shows the MIF (Majority-voting with Impact Factors) algorithm. It is an adoption of the decisions of the regular histogram comparisons, impact factors and majority-voting algorithm. In the figure, the combined meta-decisions are C_{Pred}. In the regular histogram comparisons, there are m individual decisions since there is a single decision corresponding to each training set. In contrast, there are $(m \times k)$ individual decisions from the majority-voting algorithm since there is a single decision corresponding to each training set together with a type of classifiers. Therefore, the decisions of the regular histogram comparisons are compared k times with that of the majority-voting algorithm of the same training set. IF_{Normal} and IF_{Cancer} are measures proposed in [7]. They define inter-experimental variations of a heterogeneous testing sample to normal and cancer classes of training samples, and they are expressed as IF_{Normal} and IF_{Cancer}.

Individual decisions of the regular histogram comparisons and the majority-voting algorithm are compared in the code segment from line 4 to line 16 in figure 5. If they are in the same decisions, equation 6 and 7 are applied for decisions of normal and cancer.

$$d_{Normal} = \beta_i \times IF_{Normal,i} / IF_{Cancer,i} \times dis(\alpha, H_{Normal,i}) / dis(\alpha, H_{Cancer,i}) \quad (6)$$

, where $\alpha = dis\ (hist\_proc(s, n_b, G_i)$

$$d_{Cancer} = \beta_i \times IF_{Cancer,i} / IF_{Normal,i} \times dis(\alpha, H_{Cancer,i}) / dis(\alpha, H_{Normal,i}) \quad (7)$$

, where $\alpha = dis\ (hist\_proc(s, n_b, G_i)$

For both equations, β_i is the magnitude of pre-knowledge for model M_i, which is calculated by equation 3. The factors of $(IF_{c1,i}/IF_{c2,i})$, given that $c1,c2 \in \{Normal, Cancer\}$ and $c1 \neq c2$, are linear scaling factors which minimize variations between two classes among different training sets. In fact, d_c's, where $c \in \{Normal, Cancer\}$, are measures with respect to overall gene expression levels in various training sets, but the ratio of gene expression levels between two classes in individual training sets are varied. Hence, d_c's should be rescaled accordingly in order to reduce the impacts of differential ratios between the two classes among various data sets. As a result, individual decisions are insensitive to bias of either class and variations of gene expression levels among training sets.

For the ratio of two different *dis*'s, it weights the results of the majority-voting algorithm by taking the similarity of shapes between two histograms. Remind that candidate i in the set $C_{Hist,i}$ is defined as:

$$C_{Hist,i} = (c1 \mid dis(c1,s) < dis(c2,s) \wedge c1,c2 \in \{Normal, Cancer\} \wedge c1 \neq c2) \quad (8)$$

Hence, the factor of $dis(c1,s)/dis(c2,s)$ makes β_i become smaller, and thus a higher degree of similarity is contributed to meta-decisions because of similarity of the regular histograms.

In contrast, if the two decisions are different, the factors, representing the similarity of the regular histogram comparisons, are excluded. The factors of $(IF_{c1,i}/IF_{c2,i})$ aim at minimizing variations between classes and bias of either class. Therefore, the factors are also used to adjust the values of β_i. However, the factors of $dis(c1,s)/dis(c2,s)$ are weighted factors which give higher ranks to decisions because of similarity of the regular histograms. For the case of different decisions between two

algorithms, the previous method is not appropriate. In fact, the histograms are constructed by a set of significant genes, which are selected and extracted after the accession numbers alignment. Also, the significant genes are ranked in terms of their differential gene expression levels between two classes, which is independent on variations of gene expression levels among different data sets. Therefore, it is possible that (1) some significant genes are omitted during the accession numbers alignment, and (2) selected and extracted significant genes, based on training sets, may cause misleading results. As a result, we use another method and have the following equations for the case of different decisions:

$$d_{Normal} = \beta_i \times IF_{Cancer,\,i} / IF_{Normal,\,i} \quad (9)$$

$$d_{Cancer} = \beta_i \times IF_{Normal,\,i} / IF_{Cancer,\,i} \quad (10)$$

Finally, calculated d_{Normal} and d_{Cancer} are adjusted on log2 scale, and individual results corresponding to their training sets are added together, expressed as d_{Acc_normal} and d_{Acc_cancer} for normal and cancer classes. Their magnitudes are compared, and the classes with smaller magnitudes become meta-decisions of the testing sample.

Table 1. Information of data sets.

Data set ID	*Cancer type*	*Authors*	*Accession annotation*	*Normal sample size*	*Cancer sample size*	*Training data*	*Testing data*
1	Bladder	Ramaswamy et al. [18]	Hu35K	7	11		√
2	Brain	Pomeroy et al. [16]	Hu35K	4	10		√
3	Colon	Notterman et al. [14]	GenBank	4	4		√
4	Lung	Bhattacharjee et al. [1]	U95A	17	126	√	√
5	Lung	Ramaswamy et al. [18]	Hu35K	7	8	√	√
6	Ovary	Welsh et al. [25]	Hu35K	3	30		√
7	Prostate	Singh et al. [22]	U95A	9	25	√	√
8	Prostate	Welsh et al. [24]	U95A	50	52	√	√
9	Prostate	Ramaswamy et al. [18]	Hu35K	9	10		√
10	Uterus	Ramaswamy et al. [18]	Hu35K	6	10		√

Table 2. Number of common genes between training and testing data sets.

		Testing data set ID									
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Training data set ID	*4*	7091	6153	6045	12599	7091	6153	12249	12599	12249	7091
	5	13774	8391	7840	7091	13774	8391	6808	7091	6808	13774
	7	6808	5949	5841	12249	6808	5949	12625	12249	12625	6808
	8	7091	6153	6045	12599	7091	6153	12249	12599	12249	7091

Table 3. Experimental results compared wit the majority-voting meta-classification.

Testing set ID	*Type*	*Approach*	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*Cost of learning savings*
1	Bladder	Majority-voting	73.61±9.49	39.29±31.68	95.45±5.25	5±3.92
		MIF algorithm	84.72±2.78	60.71±7.14	100.00±0	8.5±1
2	Brain	Majority-voting	75.00±7.14	25.00±35.36	95.00±5.77	1.5±2.38
		MIF algorithm	83.93±3.57	68.75±12.5	90.00±8.16	4.5±0.58
3	Colon	Majority-voting	87.50±0	75.00±0	100.00±0	6±0
		MIF algorithm	87.50±0	75.00±0	100.00±0	6±0
4	Lung	Majority-voting	96.50±0.81	94.12±0	96.83±0.92	28±1.15
		MIF algorithm	94.76±1.21	97.06±5.88	94.44±1.71	26±1.83
5	Lung	Majority-voting	75.00±3.21	42.86±20.2	95.45±9.09	5.5±1.91
		MIF algorithm	91.67±3.56	85.71±0	95.45±9.09	11.5±1
6	Ovary	Majority-voting	80.30±5.25	0.00±0	88.33±5.77	-3.5±1.73
		MIF algorithm	84.85±2.47	33.33±0	90.00±2.72	-1±0.82
7	Prostate	Majority-voting	100.00±0	100.00±0	100.00±0	18
		MIF algorithm	96.32±2.82	91.67±5.56	98.00±2.31	16±1.41
8	Prostate	Majority-voting	63.16±11.37	33.33±39.54	90.00±14.14	5±5.72
		MIF algorithm	57.11±5.85	15.50±12.58	97.12±1.11	14±12.25
9	Prostate	Majority-voting	75.00±3.21	42.86±20.2	95.45±9.09	5.5±1.91
		MIF algorithm	68.42±11.37	52.78±29.22	82.50±15	7.75±4.57
10	Uterus	Majority-voting	81.25±5.1	66.67±23.57	90.00±11.55	7±2
		MIF algorithm	81.25±0	75.00±9.62	85.00±5.78	7.5±0.58

4. EXPERIMETNS & DISCUSSIONS

To measure the classification performance, four measurements are used as performance indicators. Classification accuracy, sensitivity, specificity and learning cost savings are defined in terms of true positive (*TP*), true negative (*TN*), false positive (*FP*) and false negative (*FN*), and their definitions are [4], [13]:

- Accuracy (acc) – $acc=(TP+TN)/(TP+TN+FP+FN)$
- Sensitivity (S_n) – $S_n = TP/TP+FN)$
- Specificity (S_p) – $S_p=TN/(TN+FP)$
- Learning cost savings (sav) – $sav=[(FN+TP)*2]-(FP+2*FN)$

4.1. Data sets

In order to demonstrate the robustness of the MIF algorithm, 10 different data sets, which are individually published in 8 publications, are experimented. They are heterogeneous since they were conducted by different laboratories with different experimental objectives, microarray platforms and human genome arrays. Table 1 shows their information. Among all of them, two lung cancer (Bhattacharjee and Ramaswamy) and two prostate (Singh and Welsh) cancer data sets are arbitrarily selected as training data sets for extension and continuity of our previous works in [7], and all of them are used for testing.

As stated in table 1, there are three different accession numbers annotations, and therefore a process of standardization is required. We map the Hu35K and GenBank annotations into the U95A annotation according to the mapping table done by Ramaswamy et al. [17]. In fact, the mapping is not simply one-to-one mapping. There may be duplicated accession numbers in the mapped data set. Thus, an extra pre-processing step is performed to combine the expression levels by averaging all expression levels of the same accession numbers. After the standardization, it is required to find out those commonly existed genes for pairs of heterogeneous data sets and align their expression levels. In fact, the numbers of gene among different data sets are varied. Unavoidably, some expression levels are omitted because of missing data in either data set of pairs. Hence, the number of genes in aligned sets is either smaller or equals to the number of genes in the original data sets. Finally, we have table 2, which shows the number of commonly existed genes between training and testing data sets.

4.2. Results

In this section, we first compare the results of the MIF algorithm with that of the majority-voting algorithm, and then the results are compared with the works done by Bloom et al. [2]. Bloom's method is to perform multi-platform and multi-site microarray-based tumor meta-classification, and they used the measurement of classification accuracy as performance indicator. For parameters settings, the numbers of required bins n_b, and significant genes n_g, are set as 25 and 100. In addition, $\alpha\%$, which is the percentage of candidate bins to be trimmed, is set to 10% for achieving the optimal performance after some empirical studies. For classifiers training scheme, 70% of samples in each training data sets are selection for individual training at random, and all samples in testing data sets are used for performance measurements. In order to estimate the standard deviation of the performance, each training set is trained 100 times with different training candidates selected randomly.

In table 3, it shows that the MIF algorithm outperforms the majority-voting algorithm in terms of classification accuracy, sensitivity, specificity and cost of learning savings. Except for the cases of prostate cancer, the MIF algorithm achieves around 85% of accuracy, 65% of sensitivity, 90% of specificity and comparatively higher savings on learning cost.

For the classification accuracy, the data sets of lung cancer have the highest performance, but all cases of prostate cancer have little performance reduction. For lung cancer, the accuracy is higher than 90% for both cases (i.e. Bhattacharjee and Ramaswamy). Although there is 2% reduction for the data set Bhattacharjee, the accuracy for the data set Ramaswamy is increased from 75% to 91%. However, all data sets of prostate cancer have different degrees of performance degradation. There are reductions of 7%, 6% and 7% for the accuracy of the data set Singh, Welsh and Ramaswamy. In addition, it shows that two of them, which are Welsh and Ramaswamy, perform worse than other cases not only with the majority-voting algorithm, but also with the MIF algorithm. They only achieve around 60% for the accuracy, which is 20% lower than the average cases. For other cancer types, including bladder, brain, colon, and uterus, their average accuracy is around 85%. For the standard deviations of the accuracy, the MIF algorithm achieves smaller standard deviations for most cases. For the cancer types of bladder, brain, ovary and uterus cancers, the improvement is more than 50%. For the cancer types of lung and prostate cancers, the significance results are varied.

For the classification sensitivity and specificity, the MIF algorithm can have better balanced recall rates between normal and cancer samples, except for the cases of prostate cancer. Classification algorithms should have similar recall rates for samples in both classes so that the algorithms are unbiased to either class. Euclidean distance of sensitivity, S_n, and specificity, S_p, can be used to show the balance of recall rates between samples in two classes, and the distance is:

$$Euclidean(S_n, S_p) = \sqrt{S_n^2 + S_p^2} \qquad (11)$$

In table 4, it shows that the MIF algorithm outperforms the majority-voting algorithm for 6 cases (i.e. 1, 2, 5, 6, 8 and 10) and maintains the same performance for 2 cases (i.e. 2 and 3). Similar to the measurement of classification accuracy, the data sets of prostate cancer do not have impressive results. Testing set 7 and 9 show performance degradation (i.e. the majority-voting algorithm outperforms the MIF algorithm.).

Table 4. Balanced recall rates between normal and cancer sample.

Testing set ID	*Type*	*Majority-voting*	*MIF algorithm*
1	Bladder	1.03	**1.17**
2	Brain	0.98	**1.13**
3	Colon	**1.25**	**1.25**
4	Lung	**1.35**	**1.35**
5	Lung	1.05	**1.28**
6	Ovary	0.88	**0.96**
7	Prostate	**1.41**	1.34
8	Prostate	0.96	**0.98**
9	Prostate	**1.05**	0.98
10	Uterus	1.12	**1.13**

In addition, we have also compared our results with bloom's results in [2]. In table 5, it shows that the MIF algorithm outperforms Bloom's works for bladder and uterus cancers, and

maintains the same performance for lung cancer. However, there is performance reduction for prostate cancer.

Table 5. Comparison of results with other works.

Testing set ID	*Type*	*Classification accuracy (%)*	
		Bloom's results	*our results*
1	Bladder	77	**84**
5	Lung	**91**	**91**
9	Prostate	**94**	68
10	Uterus	74	**81**

4.3. Cancer/testis (CT) immunogenic gene families

Cancer/testis (CT) immunogenic gene families are subsets of genes, which are commonly existed in various cancer types. Some works show that most CT immunogenic gene families are expressed in more than one cancer types, but with various expression frequencies. In [20], Scanlan et al. have reviewed the expression frequencies of them in numerous cancer types consisting of bladder, brain, breast, colon, gastric, and etc. It shows that lung and melanoma cancers contain a higher percentage of CT genes examined at expression frequencies greater than 20%. In contrast, prostate and brain cancers have a relatively lower percentage of the CT genes examined at the same frequencies.

Table 6. Comparisons of the cancer/testis (CT) immunogenic gene families in various cancer types.

	Cancer type					
	Bla	Bra	Col	Lun	Ova	Pro
No. of included lowly-expressed CT genes with a low expression frequency, <= 20%	*17*	*5*	*12*	*29*	*11*	*11*
No of included CT highly-expressed genes with a high expression frequency, > 20%	*11*	*4*	*3*	*17*	*7*	*6*
Proportions of commonly existed highly-expressed genes to lung cancer	*7/11*	*3/4*	*2/3*	*29/29*	*5/7*	*2/6*
Proportions of commonly existed highly-expressed genes to prostate cancer	*4/11*	*1/4*	*1/3*	*2/29*	*2/7*	*6/6*

 # Abbreviations: Bla, bladder; Bra, brain; Col, colon; Lun, lung; Ova, ovary; Pro, prostate.

In our studies, we have analyzed how the proportions of shared highly-expressed CT genes between training and testing samples play a vital role in meta-classification performance of heterogeneous data. We investigated how the number of included lowly- and highly-expressed CT genes is varied with the classification performance. Table 6 shows the number of included lowly- and highly-expressed CT genes in various cancer types. Lung cancer has the highest proportions of both types of CT genes, and brain cancer has the lowest one. However, in [20], it has mentioned that the studies of brain cancer to the CT genes are insufficient in this moment. Therefore, brain cancer is exceptional and hence prostate and ovary cancers belong to the same family of having small proportions of both types of CT genes.

From our experiments, the data sets of prostate cancer only achieve classification accuracy of 75% in average, but the data set of ovary cancer can achieve 84% instead. Hence, it may be deduced that there is no direct and linear relationship between the number of included lowly- and highly-expressed CT genes and the classification performance.

Further, we have investigated how the number of shared highly-expressed CT genes between training and testing samples is in relation to the classification performance. In table 6, the last two rows show the proportions of the highly-expressed CT genes between the corresponding samples, and both lung and prostate cancers, respectively. If we consider the proportions together with the corresponding classification performance, we will have figure 6. In the figure, the classification performance has the same increasing and decreasing trends as the proportions of the CT genes to lung cancer, but reversed trends for the proportions to prostate cancer, except for the case of brain cancer.

Together with figure 6 and table 6, we can see that the proportions of shared highly-expressed CT genes between training and testing samples has impacts on classification performance, and the data sets of lung cancer have dominated roles at meta-decisions because of higher proportions of shared highly-expressed CT genes between training and testing samples. In the figure, lung cancer always has higher proportions of shared highly-expressed CT genes with other cancer types, except for the prostate cancer. The classification accuracy is higher than 80% in average. However, the classification accuracy for prostate cancer has been dropped significantly. It may be evidence to show that the decrease of the performance for prostate cancer is caused by lack of shared highly-expressed CT genes between training and testing

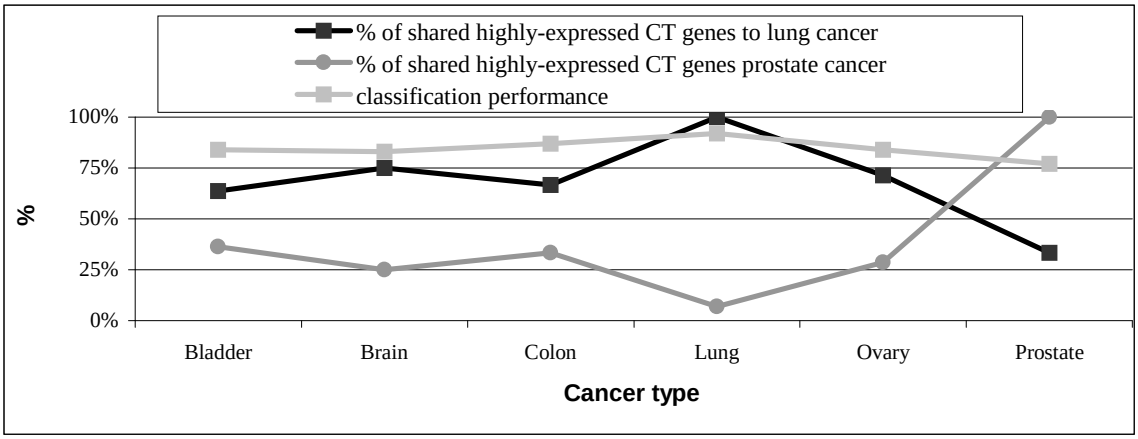

Figure 6. Relationship between shared highly-expressed CT genes and classification performance.

samples. Also, it is observed that the highly-expressed CT genes of prostate cancer in training samples are compromised with the lack of the genes between lung and prostate cancers. The possible explanation is that the number of included lowly- and highly-expressed CT genes in various cancer types. The fact is that the number of both included lowly- and highly-expressed CT genes to lung cancer is almost 3 times higher than that of prostate cancer, causing that data sets of lung cancer may have higher weights at meta-decisions. In addition, in figure 6, the decreasing rate of the performance for prostate cancer is less than that of the proportions of shared highly-expressed CT genes to lung cancer because of the increase of shared CT genes in prostate cancer (i.e. the ordinary type).

5. CONCLUSIONS

With the innovation of DNA microarray technologies, different mining algorithms have been proposed to discover knowledge in cancer gene expression data. Significant findings are recently exploited. However, most works are done with a single data set. In terms of efficiency and effectiveness of mining algorithms with respect to clinical applicability and robustness, it is too weak to draw conclusions because of the problems of over-fitting and homogeneity within a single data set.

In this work, we proposed the MIF algorithm to perform multi-type cancer gene expression data classification, which uses differences of regular histograms for gene expression levels of certain significant genes as parts of dissimilarity measures and indicators of predictive classes. In the experiments, we have intensively used 10 different data sets to show the reliability and robustness of the MIF algorithm. The results are impressive. The classification accuracy is around 85% in average for most cases, except for the data sets of prostate cancer.

To investigate the frustrated performance for prostate cancer, we have looked into the cancer/testis (CT) immunogenic gene families. We have discovered that the numbers of shared highly-expressed (i.e. expression frequencies > 20%) CT genes between training and testing samples have impacts on the classification performance of heterogeneous samples.

6. Acknowledgement

The work of the authors are supported in part by the Central Grant of The Hong Kong Polytechnics University, research project code HZJ89.

7. REFERENCES

[1] Bhattacharjee, A., Richards, W., Staunton, J., Li, C. and Monti, S. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. of the Natl. Acad. of Sci. USA, 98(24), pp. 13790-5.

[2] Bloom, G., Yang, I.V., Boulware, D. and Kwong, K.Y. (2004). Multi-platform, multi-site, microarray-based human tumor classification. Am. J. Pathol., 164(1), pp. 9-16.

[3] Cho, S.B. and Ryu, J.W. (2002). Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. Proc. of the IEEE, 90(11), pp. 1744-1753.

[4] Cho, S.B. and Won, H.H. (2003). Machine learning in DNA microarray analysis for cancer classification. Proc. of the First Asia Pacific Bioinformatics Conference, Adelaide, Australia, 19, pp. 189-198: Australian Computer Society.

[5] Choi, J.K., Yu, U., Kim, S. and Yoo, O.J. (2003). Combining multiple microarray studies and modeling interstudy variation. Bioinformatics, 19 Suppl., 1:i84-90.

[6] Cui, X. and Churchill G.A. (2003). Statistical tests for differential expression in cDNA microarray experiments, Genome Biol., 4(4):210.

[7] Fung, B.Y.M. and Ng, V.T.Y (2003). Classification of Heterogeneous Gene Expression Data. Special Issue on Microarray Data Mining, SIGKDD Explorations, 5(2), pp. 69-78.

[8] Fung, B.Y.M and Ng, V.T.Y. (2004). Selecting the Optimal Classification Results by an Empirically-driven Model Averaging. Proc. of the SIAM Bioinformatics Workshop 2004, Florida, USA, pp. 25-35: SIAM.

[9] Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L. and Kohane, I.S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics, 18(3), pp. 405-12.

[10] Lee, J.K., Bussey, K.J., Gwadry, F.G. and Reinhold, W. (2003). Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells. Genome Biology, 4:R82.

[11] Lu, Y. and Han, J. (2003). Cancer classification using gene expression data. Information Systems, 28(4), pp. 243-268.

[12] Nadimpally, V. and Zaki, M. (2003). A Novel Approach to Determine Normal Variation in Gene Expression Data. Special Issue on Microarray Data Mining, SIGKDD Explorations, 5(2), pp. 6-15.

[13] Ng, S.K., Tan, S.H. and Sundarajan, V.S. (2003). On Combining Multiple Microarray Studies for Improved Functional Classification by Whole-Dataset Feature Selection. Genome Informatics, 14, pp. 44-53.

[14] Notterman, D.A., Alon, U., Sierk, A.J. and Levine, A.J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res., 61(7), pp. 3124-30.

[15] Ochs, M.F. and Godwin, A.K. (2003). Microarrays in cancer: research and applications. Biotechniques, 34, Suppl., pp. 4-15.

[16] Pomeroy, S.L., Tamayo, P., Gaasenbeek, M. and Sturla, L.M. (2002). Gene Expression-Based Classification and Outcome Prediction of Central Nervous System Embryonal Tumors. Nature, 415, pp. 436-42.

[17] Ramaswamy, S., Ross, K.N., Lander, E.S. and Golub, T.R. (2003). Evidence for a molecular signature of metastasis in primary solid tumors. Nature Genetics, 33, pp. 49-54.

[18] Ramaswamy, S., Tamayo, P., Rifkin, R. and Mukherjee, S. (2001). Multi-class cancer diagnosis using tumor gene expression signatures. Proc. of the Natl. Acad. of Sci. USA, 98(26), pp. 15149-54.

[19] Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. and Chinnaiyan, A.M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res., 62(15), pp. 4427-33.

[20] Scanlan, M.J., Simpson, A.J.G. and Old, L.J. (2004). The cancer/testis genes: Review, standardization, and commentary. Cancer Immunity, 4(1), pp. 1-15.

[21] Seewald, A.K. and Frnkranz, J. (2001). An evaluation of grading classifiers. Proc. of the 4th Int. Symposium in Advances in Intelligent Data Analysis (IDA-01), Lisbon, Portugal, pp. 115-124: Springer-Verlag.

[22] Singh, D., Febbo, P.G., Ross, K. and Jackson, D.G. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 1(2), pp. 203-209.

[23] Van der Bruggen, P., Traversari, C., Chomez, P. and Lurquin, C. (1991). A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. Science, 254(5038), pp. 1643-7.

[24] Welsh, J.B., Sapinoso, L.M., Su, A.I. and Kern, S.G. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Research, 61, pp. 5974-78.

[25] Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M. and Kern, S.G. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc. Natl. Acad. Sci. USA, 98(3), pp. 1176-81.

[26] Yao, X. and Liu, Y. (1999). Neural networks for breast cancer diagnosis. Proc. of the 1999 Congress on Evolutionary Computation, New York, USA, pp. 1760-1767: IEEE Press.