Meta-Classification of Multi-Gene Data with Alternative Feature Mapping

Victor C. LIANG Vincent T.Y. NG
Department of Computing, The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
{cscliang, cstyng}@comp.polyu.edu.hk

Abstract

In order to overcome the limitation on small size of gene datasets, many meta-classification methods which ensemble classifiers from different datasets have been developed. However, due to discrepancies of the characteristics among multiple heterogeneous datasets, the number of common and significant genes is usually small. Instead of matching common genes between heterogeneous datasets, we propose a novel solution, alternative feature mapping approach (AFM), to utilize related and discriminative gene expressions while not necessarily having exact matches. Genes in the training dataset are clustered and mapped to the test dataset as gene groups. Through analyzing the correlation within gene groups, significant genes can be matched and dataset dissimilarity factors can be used as weights for meta-classification. We conducted experiments consisting of 10 heterogeneous datasets with different cancer types and platforms. Our experiments show that classification performance is greatly improved using suitable significant genes selected by AFM, and weight voting method based on AFM provides more reliability for meta-classification.

Keywords: AFM, Gene Expression Data, Meta-classification, Heterogeneous and Feature Selection.

1. INTRODUCTION

For DNA microarray technology, most of available datasets are collected from different research institutes with particular objectives. Also, samples hybridized on different types of platforms like oligonucleotide-based or cDNA-based microarray may lead to large discrepancies [1]. The homogeneous sample resources that can be directly used for classification are limited. To overcome the insufficiency of homogeneous datasets, many research works turn to utilizing heterogeneous datasets to complement training samples. For single or multiple dataset(s), ensemble different classifiers trained from various irrelevant features, learning algorithms or datasets can be considered as effective strategies to improve the classification accuracy. Some works [2,3] have tried to partition high-dimensional features into several sets to train multiple classifiers for meta-classification. Nevertheless, the results were not totally satisfactory as the heterogeneous datasets had few common genes.

Several major compatibility problems may arise when handling heterogeneous gene expression datasets [4]. First, the number of probe sets configured for different types of microarray is not always consistent. For example, there are about 12600 probe sets contained in HG-U95v2 microarray, and approximately 22600 probe sets for HG-U133A. Second, in order to minimize the intra-microarray variation over several samples in the same dataset, normalization process needs to be performed to make raw signal comparable. As a result, the gene expression value may fall into different ranges by normalization controls. Third, experimenting on diverse cancer type samples can generate different gene expression profiles. These dissimilar profiles must own a part of distinctive functional genes different with other cancers, and result in making comparison impossible [5,6]. All methods mentioned above must initially extract the common probe sets shared by heterogeneous datasets before selecting significant genes and classifying cancer samples [7]. Obviously, analyzing on limited number of common genes may lose much additional information inside distinct features. Finding common significant genes among multiple gene datasets is a great challenge for meta-classification.

In this paper, we propose a mapping procedure to link between feature selection and classification and extend AFM for multiple gene datasets. The alternative feature mapping (AFM) is to match the significant features of training dataset with potential useful features in the test dataset instead of demanding exact feature matching. Genes in the training dataset are clustered and mapped to the test dataset as gene groups. After analyzing the correlation within each gene group, corresponding discriminative genes in test dataset can be identified by AFM for classification. Furthermore, the dataset comparison procedure in AFM can generate a dataset dissimilarity factor for each pair of datasets, which reflects the difference between training and test datasets. We take advantage of this dataset dissimilarity factor as weight for weight voting method in meta-decision. Meta-classification that assembles multiple classifiers associated with their weights improves classification accuracy and reliability.

2. ALTERNATIVE FEATURE MAPPING

2.1. Overview of AFM

Let $\boldsymbol{X_A} = \{\boldsymbol{x_{A_1}}, \boldsymbol{x_{A_2}}, \boldsymbol{x_{A_3}}, \cdots, \boldsymbol{x_{A_{m_A}}}\}$ be a training dataset A with m_A samples and n_A genes in each sample, and similarly $\boldsymbol{X_B}$ for test dataset B including m_B samples and n_B genes. A modified t-statistic feature selection method [8] is adopted as a hypothesis test to assign t-value for each gene. A large t-value means higher discrimination between two classes of normal and cancer sample. In AFM, initially, low informative genes need be filtered out with respect to small t-values in the training dataset, and only n'_A top order genes will be remained. After that a k-means cluster algorithm will be executed to partition n'_A candidates into k clusters to construct training gene groups G_A. Here, each gene group G_A contains p positive correlated genes and is represented as $G_A = \{g_1, g_2, g_3 ... g_p\}$. p may not be uniform and can be of different values for different groups. One significant gene g^*_A with the highest t-value will be extracted from each gene group to form the training feature vector $S_A = \{g^*_{A_1}, g^*_{A_2}, g^*_{A_3} ... g^*_{A_k}\}$. Next, we directly map each G_A to the test dataset $\boldsymbol{X_B}$ to construct k corresponding test gene groups G_B. Each group may contain q genes, and is represented as $G_B = \{g_1, g_2, g_3 ... g_q\}$. Comparing the difference between G_A and G_B, genes with low variation after mapping can be regarded as useful genes, and then be refined as a template to search more correlated genes across n_B genes in test dataset $\boldsymbol{X_B}$. Finally, similar to S_A, the corresponding significant features can be found out from these refined test groups, and generate a test feature vector $S_B = \{g^*_{B_1}, g^*_{B_2}, g^*_{B_3} ... g^*_{B_k}\}$. In this paper, we call genes in S_A and S_B discriminative genes. Two parameters k and n'_A can be set by users initially. A completed flow of AFM is illustrated in Fig. 1.

2.2. Gene Group Construction

The basic idea of AFM is to extend an individual significant gene to a group of related genes for mapping rather than the traditional direct mapping (TDM) that maps genes of same features between training and test samples. As these related genes in each group should be discriminative as well, features with high t-values in the filtered dataset are used to construct the training gene groups in Eq. (1).

$$\boldsymbol{X'_A} = FeatureFilter(\boldsymbol{X_A}, n'_A) \qquad (1)$$

After n'_A candidate genes are remained, a k-means cluster algorithm is applied to divide these genes into k gene groups, and each group may have different value of p. Garrett-Mayer E et al. [9] has demonstrated that positive correlation can best reflect the relationship of each gene in a group. Hence, the k-means algorithm in AFM uses the distance between each gene in terms of positive Pearson correlation coefficient, and minimizes the variation of genes within group as shown in Eq. (2). We believe that these independent groups built by k-means can cover different ranges of gene function pathways as many as possible. To generate a feature vector S_A containing k discriminative genes on the training dataset, gene with the highest t-value in each training group G_A will be selected respectively as shown in Eq. (3).

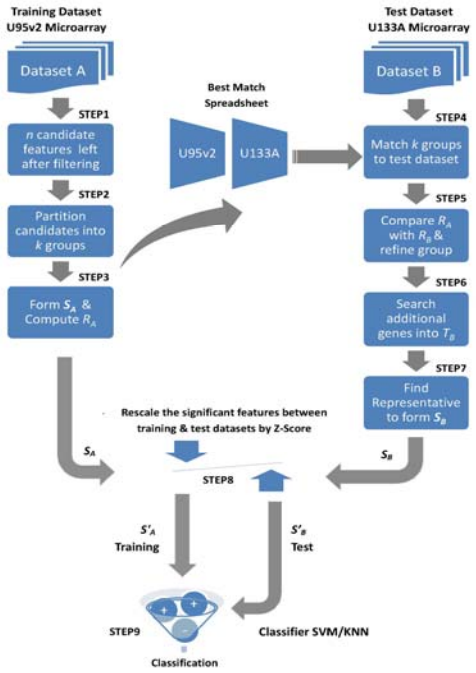

Training Dataset
U95v2 Microarray
Dataset A
STEP1
n candidate features left after filtering
STEP2
Partition candidates into k groups
STEP3
Form S_A & Compute R_A
Best Match Spreadsheet
U95v2
U133A
Test Dataset
U133A Microarray
Dataset B
STEP4
Match k groups to test dataset
STEP5
Compare R_A with R_B & refine group
STEP6
Search additional genes into T_B
STEP7
Find Representative to form S_B
Rescale the significant features between training & test datasets by Z-Score
S_A
S_B
STEP8
S'_A
Training
S'_B
Test
STEP9
Classifier SVM/KNN
Classification

Figure 1. Workflow of AFM.

$$G_{A_i} = k\_means(\boldsymbol{X'_A}, k) \quad i \in [1, k] \qquad (2)$$

$$S_A = Max_{t-value}(G_{A_1}) \cup Max_{t-value}(G_{A_2}) \cup \cdots \cup Max_{t-value}(G_{A_k}) \qquad (3)$$

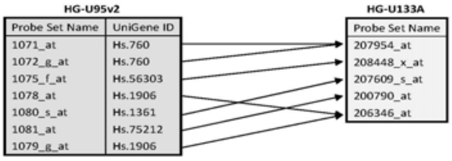
$$R_A = \begin{bmatrix} 1 & V_{1,2} & V_{1,3} & \cdots & V_{1,p-1} & V_{1,p} \\ V_{2,1} & 1 & V_{2,3} & \cdots & V_{2,p-1} & V_{2,p} \\ V_{3,1} & V_{3,2} & 1 & \cdots & V_{3,p-1} & V_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ V_{p-1,1} & V_{p-1,2} & V_{p-1,3} & \cdots & 1 & V_{p-1,p} \\ V_{p,1} & V_{p,2} & V_{p,3} & \cdots & V_{p,p-1} & 1 \end{bmatrix} \qquad (4)$$

The use of gene groups not only enables individual significant gene for group mapping, but also be able to measure the difference between two heterogeneous datasets. In order to compute the difference, AFM needs to calculate the correlation for each gene in a group to reflect the properties of G_A, which is termed as R_A. The detailed computation procedure is discussed in section 2.4.1. In this $p * p$ correlation matrix R_A shown in Eq. (4), V denotes the value of Pearson correlation coefficients of genes between each other in the same group.

2.3. Direct Dataset Mapping

For the mapping of different generations of Affymetrix microarrays, a preferable way is to use the spreadsheet provided by Affymetrix. According to the Best Match method shown in spreadsheet (HG-U95v2 vs HG-U133A) [10], one probe set in HG-U133A microarray may have multiple corresponding matched probe sets in HG-U95v2. It should be better to take average of same meaning probe sets in HG-U95v2 microarray so as to convert to one to one mapping as shown in Fig. 2. In AFM, because of the inconsistence of two microarrays, l unmatched genes may be lost and remain q genes in G_B, when G_A cannot be totally mapped to G_B.

$$G_B = Mapping(G_A)$$
$$q = \begin{cases} p & \text{full match of } G_A \text{ and } G_B \\ p-l & \text{partial match of } G_A \text{ and } G_B \end{cases} \qquad (5)$$



HG-U95v2

Probe Set Name	UniGene ID
1071_at	Hs.760
1072_g_at	Hs.760
1075_f_at	Hs.56303
1078_at	Hs.1906
1080_s_at	Hs.1361
1081_at	Hs.75212
1079_g_at	Hs.1906

HG-U133A

Probe Set Name
207954_at
208448_x_at
207609_s_at
200790_at
206346_at

Figure 2. Gene mapping between HG-95v2 and HG-U133A.

2.4. Gene Group Matching

2.4.1. Gene Group Comparison. Other than the l unmatched genes, differences of gene expressions in datasets can also influence the application of matched significant genes. For instance, some genes with strong discriminative power in the training dataset may no longer be significant for the test dataset. Therefore, it is necessary to introduce a strategy that can assess the discrepancies of two independent datasets and filter useless features which may exist in G_B. Since the comparison of individual gene is of low significant for heterogeneous datasets, the application of gene regulatory network to uncover the interaction information embedded in multiple correlated genes is obviously a suitable choice. If the correlation structure of a group has some changes after mapping, the dissimilarity between two groups can indicate that some genes may not perform well in the test gene profile.

Before assessing the difference between two datasets, AFM calculates the properties R_B of corresponding group G_B and then obtains the difference $|R_A - R_B|$ to measure the dissimilarity between two heterogeneous datasets. In difference matrix Eq. (6), NaN is the null value for l missing genes. Each gene in a group will be assigned a dissimilarity factor d_i derived from this difference matrix, which is the average of its correlation coefficients to other q non-null value genes to reflect its change from the training dataset. For each gene group, AFM introduces a group dissimilarity factor d_g that can indicate the discrepancies between G_A and G_B shown in Eq. (8). The group dissimilarity factor could also be extended for identifying the difference between two entire datasets, and be used for handling multiple training datasets in meta-classification as described in Section 2.5. For Eq. (7) and Eq. (8), we have $d_i, d_g \in [0,2]$.

$$|R_A - R_B| = \begin{bmatrix} 0 & NaN & V'_{1,3} & \cdots & V'_{1,p-1} & V'_{1,p} \\ NaN & 0 & NaN & \cdots & NaN & NaN \\ V'_{3,1} & NaN & 0 & \cdots & V'_{3,p-1} & V'_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ V'_{p-1,p} & NaN & V'_{p-1,3} & \cdots & 0 & V'_{p-1,p} \\ V'_{p,1} & NaN & V'_{p,3} & \cdots & V'_{p,p-1} & 0 \end{bmatrix} \qquad (6)$$

$$d_i = \sum_{j=1}^{q} V'_{i,j} / q \quad i, j = 1,2,3...q \qquad (7)$$

$$d_g = \sum_{i=1}^{q} d_i / q \quad i = 1,2,3...q \qquad (8)$$

2.4.2. Refining Matched Gene Group. Although there are numerous genes in a microarray, only a minority of genes showing distinction between normal and cancer samples are regarded as useful features. Therefore, in a gene group some genes with high variation after mapping to the test dataset would have a high possibility to be converted into the useless genes. Since the aim of AFM is to identify a set of potential significant genes representing all G_B, noisy genes contained in each gene group may bring interference to correctly obtain these expectative representative genes. Filtering these noisy genes with large d_i can further refine the matched gene groups G_B. In one test gene group, for the purpose of controlling the amount of genes which need to be filtered, the number of refined genes r is determined by d_g. If the difference between G_A and G_B is relatively large, AFM will only keep a few useful genes, and vice versa. Therefore, r is calculated by the inverse proportion of d_g to q genes in G_B, and β in Eq. (9) is a constant used to reduce the proportion of noisy genes in some cases. The remaining r genes are regarded as a template $T_B = \{g_1, g_2, g_3 ... g_r\}$, where $T_B \subset G_B$, to search additional correlated genes in $\boldsymbol{X_B}$. In AFM, it is necessary to refine all k test gene groups to improve the quality of templates.

$$r = (1 - d_g / 2 - \beta) * q \qquad (9)$$

2.4.3. Searching Representative Feature. Since the r remaining genes in the template T_B may not be sufficient to locate a representative gene in one group, additional $p - r$ genes which are correlated with T_B across the whole test dataset should be adopted to complement G_B. One advantage of gene complement is to make the number of genes in both training and corresponding test group equivalent. In AFM, we set a weight w_i to every gene in the whole test dataset by calculating their correlations with each other gene in template T_B as shown in Eq. (10). After that $p - r$ additional genes with highest w_i in $\boldsymbol{X_B}$ will be joined into the template T_B to reconstruct the gene group G_B. For k gene groups in the test dataset, this procedure will repeat k times until all test gene groups have been filled in. In order to speed up the process of searching, we can initially filter out low variation genes in $\boldsymbol{X_B}$ to make n_B smaller.

$$w_i = \sum_{j=1}^{r} Pearson(g_i, g_j)/r \quad g_i \in \boldsymbol{X_B}, g_j \in T_B \qquad (10)$$

With the steps above, all genes in a test group G_B would be related to a corresponding g_A^*. Therefore, in a group G_B we need to indentify the most representative gene from q potential useful genes to match significant gene g_A^* in G_A. Similar to Eq. (10), in Eq. (11) AFM computes the correlation for one gene with the rest of genes in G_B as a weighting factor to indicate its conjunction to G_B. It is obviously that the gene that shows the highest correlation w_i' among other genes in a group can be considered as a significant gene to represent G_B. Finally, all genes with highest w_i' in each G_B are collected to generate a test feature vector S_B corresponding to the training feature vector S_A for classification. Due to the different scaling among heterogeneous microarray datasets, it is also necessary to rescale S_A and S_B before inputting these discriminative genes into any kind of classifiers. Here, we simply convert S_A and S_B to standard normal distribution S_A' and S_B' with zero mean and unit standard deviation by z-score normalization, and complete all steps of alternative feature mapping.

$$w_i' = \sum_{i=1}^{p} Pearson(g_i, g_j)/p \quad i, j = 1,2,3...p \qquad (11)$$

$$S_B = Max_{w_i'}(G_{B_1}) \cup Max_{w_i'}(G_{B_2}) \cup \cdots \cup Max_{w_i'}(G_{B_k}) \qquad (12)$$

2.5. Meta-Classification

Usually, classification performance depends on the dataset that is used to train classifier. However, the size of training samples is usually small for an individual gene expression dataset, which may not be effective to train a high performance classifier. One possible way is to ensemble multiple classifiers that are generated from several heterogeneous datasets to make a meta-decision on test dataset. As AFM is to match significant genes between two datasets, problem of handling multiple training dataset can be solved by making pair of each training dataset with the particular test dataset, and performing classification individually. The meta-decision then can be determined by summarizing individual classification results of each classifier associated with their weights as shown in Fig 3.

For meta-classification we adopt a weight voting method to combine various classifiers that are trained by multiple gene datasets. Suppose there are h training datasets and 1 test dataset. Firstly, each training dataset need make a comparison with the test dataset to identify significant genes by AFM, and generate a dataset dissimilarity factor d_D for weight voting. In Section 2.4.1, we have introduced an indicator called group dissimilarity factor d_g calculated by making difference between R_A and R_B to reflect the discrepancies between two heterogeneous microarray datasets. We take average of d_g to derive d_D as weight for each pair of training and test datasets. In Eq. (13) d_g is the group dissimilarity factor, k is the number of gene groups forming in training dataset. Since $d_g \in [0,2]$, we normalize the dataset dissimilarity factor d_D into a range [0, 1] divided by 2.

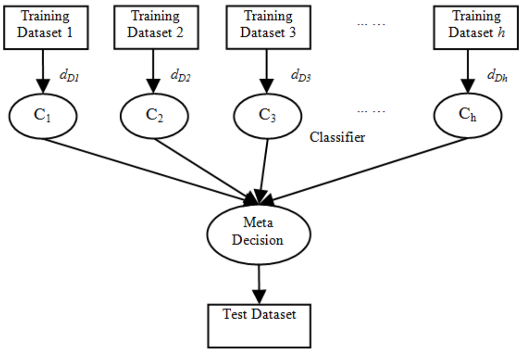


Figure 3. Meta-classification by combing h classifiers

$$d_D = (\sum_{i=1}^{k} d_{gi}/k)/2 \qquad (13)$$

Secondly, h classifiers trained by heterogeneous datasets are used to classify each test sample, and assign their classification labels, normal or tumor. The weight of classifier would then be voted to the corresponding label of test sample respectively. Finally, to make a meta-decision, we calculate the total weights on each classification label of test sample. By comparing the total weights of normal and tumor label, the final decision can be made by selecting the label that has smaller weights. In Eq. (14) if $classifier(x_i) = c$, $sign(classifier(x_i) = c) = 1$.

$$C_{ensemble} = \underset{c=\{Normal,Tumor\}}{\arg\min} \left(\sum_{i=1}^{h} d_{Di}\, sign(classifier(x_i) = c)\right) \qquad (14)$$

3. EXPERIMENTS

3.1. Experiment Settings

To demonstrate how AFM works well for heterogeneous gene expression data, we arranged two types of classifications, single-classification and meta-classification. For single-classification, only one classifier is used to classify the test dataset. In our experiments, 8 out of 10 gene expression datasets were separated into 3 kinds of heterogeneous datasets combination, which involved in datasets with same cancer type but different platforms, different cancer types with the same platform and different cancer types with different platforms. Detailed information about 10 microarray datasets is illustrated in Table I. Table II lists each combination of heterogeneous datasets. Two kinds of widely used classifiers, SVMs and KNN, were introduced to classify heterogeneous datasets using the discriminative genes selected by AFM. For comparison purpose, a set of comparison experiments with traditional direct mapping (TDM) was also performed to compare the effectiveness of direct matched genes and alternative genes. The procedure of TDM is much simpler than AFM, which first selects top *k* significant genes from the common genes between training and test dataset, matches corresponding genes to the same meaning genes in the test dataset, rescales significant genes by z-score normalization.

For meta-classification, we combined 6 classifiers that were trained by all six training dataset as shown in Table I to classify test dataset. The final meta-decision was made by using weight voting method according to the dataset dissimilarity factor of each classifier. In our experiment, 4 sets of experiments were conducted to verify whether AFM and work well on meta-classification scenario.

TABLE I. INFORMATION OF HETEROGENEOUS DATASETS

Dataset Name	Microarray	Cancer Type	Normal Samples	Tumor Samples	# of genes	Usage
GSE1987	U95v2	Lung	9	28	12599	Test
GSE2514	U95v2	Lung	19	20	12625	Test
GSE10072	U133A	Lung	49	58	22283	Training
GSE2443	U133A	Prostate	10	10	22283	Test
Singh	U95v2	Prostate	50	52	12600	Training
GSE6631	U95v2	HNSCC	22	22	12625	Training
GSE9476	U133A	Leukemia	38	26	22283	Training
GSE6012	U133A	Skin	10	10	22283	Test
GSE7670	U133A	Lung	31	35	22283	Training
GSE6919	U95v2	Prostate	81	25	12625	Training

TABLE II. COMBINATION OF HETEROGENEOUS DATASETS FOR CLASSIFICATION

Training Datasets	Test Datasets	Cancer Type	Platform
GSE10072 Singh	GSE2514 GSE2443	Lung Prostate	U133A -> U95v2 U95v2 -> U133A
Singh GSE10072	GSE1987 GSE6012	Prostate -> Lung Lung -> Skin	U95v2 U133A
GSE9476 GSE6631	GSE2514 GSE6012	Leukemia -> Lung HNSCC	U133A -> U95v2 U95v2 -> U133A

3.2. Experiment Results and Discussions

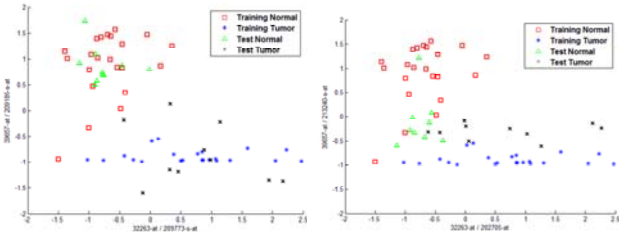
Table III demonstrates the single-classification results using TDM with 5 and 20 discriminative genes, and Table IV shows the corresponding classification performance by AFM. Firstly, when comparing the classification accuracies by using TDM and AFM, we find great accuracy improvement occurred for AFM in most experiments except the first case (GSE10072->GSE2514) with about 2.5% to 5% decrease. Furthermore, it is obviously that features selected by TDM do not successfully classify different cancer type datasets. Only 41%~75% accuracies for 5 significant genes case and 35%~70% for 20 significant genes case are reached by TDM. However, by contraries, AFM demonstrates its powerful performance for classification using appropriate genes not the direct matched genes. For example, TDM can only reach about 50% accuracy in experiment (GSE10072->GSE6012); while a high accuracy near to 100% has been achieved by AFM. Secondly, besides the comparison with TDM and AFM, for the different number of significant genes (5 and 20 genes) chosen in our experiments, classification accuracies for AFM using 5 significant genes have a slightly better than 20 genes in most cases. That may be the reason that small number of discriminative genes means fewer gene groups, and a relative larger size of genes in each test gene group, which could cover more potential useful genes. Thirdly, since two kinds of classifiers, SVMs and KNN are used to test the performance of AFM, when assessing the classification results of these two classifiers, both SVMs and KNN show satisfied classification accuracies in different experiments. Therefore, AFM has a good compatibility with different kinds of classifiers.

TABLE III. SINGLE-CLASSIFICATION RESULTS BASED ON DIRECT MAPPING METHOD (TDM)

Single-classification (TDM)	5 significant genes		20 significant genes	
	SVMs Acc (%)	KNN Acc (%)	SVMs Acc (%)	KNN Acc (%)
GSE10072->GSE2514	100	100	100	100
Singh->GSE2443	50	50	55	55
Singh->GSE1987	59.46	62.16	64.86	70.27
GSE10072->GSE6012	45	50	60	60
GSE9476->GSE2514	41.03	48.72	41.03	41.03
GSE6631->GSE6012	75	70	35	50

TABLE IV. SINGLE-CLASSIFICATION RESULTS BASED ON ALTERNATIVE FEATURE MAPPING (AFM)

Single-classification (AFM)	5 significant genes		20 significant genes	
	SVMs Acc (%)	KNN Acc (%)	SVMs Acc (%)	KNN Acc (%)
GSE10072->GSE2514	97.44	97.44	94.87	94.87
Singh->GSE2443	65	70	65	75
Singh->GSE1987	86.49	83.78	81.08	83.78
GSE10072->GSE6012	100	100	100	95
GSE9476->GSE2514	79.49	87.18	97.44	84.62
GSE6631->GSE6012	95	90	85	90



(a). Feature mapping using AFM. (b). Feature mapping using TDM.

Figure 4. Distribution of training and test samples.

We especially concentrate on an experiment (GSE6631->GSE6012) to investigate how AFM matches appropriate features. Fig. 4(a) illustrates the distribution of normal and tumor samples in the training and test datasets. AFM chose genes 32263_at and 39657_at from the training dataset and two corresponding discriminative genes 209773_s_at and 209185_s_at to observe the distribution of samples. Fig. 4(b) shows the distribution based on gene 32263_at, 39657_at and gene 202705_at, 213240_s_at matched by TDM. Through the comparison, we can notice that the test normal and tumor samples in the test dataset, as shown in Fig. 4(a) are totally separated according to the two feature genes selected by AFM; while in Fig. 4(b), normal samples are mixed with tumor samples, which could decrease the classification accuracy.

TABLE V. META-CLASSIFICATION RESULTS BASED ON ALTERNATIVE FEATURE MAPPING (AFM)

	SVMs (5 significant genes)						
	GSE 10072	Singh	GSE 9476	GSE 6631	GSE 7670	GSE 6919	Meta Decision
GSE1987							
Acc (%)	81.08	78.38	48.65	78.38	86.49	48.65	**86.49**
d_D	0.188	0.235	0.234	0.207	0.156	0.254	
GSE2514							
Acc (%)	97.44	94.87	89.74	74.36	94.87	74.36	**94.87**
d_D	0.069	0.196	0.234	0.201	0.083	0.252	
GSE2443							
Acc (%)	50	50	70	45	45	40	**55**
d_D	0.330	0.205	0.228	0.236	0.306	0.283	
GSE6012							
Acc (%)	90	85	75	90	95	70	**95**
d_D	0.334	0.230	0.256	0.222	0.299	0.264	

Table V shows the meta-classification results on four test datasets by combing six heterogeneous classifiers. For each test dataset, we listed classification accuracy of every used classifier and its dataset dissimilarity factor d_D as weight respectively. The accuracy of meta-decision is the final result after performing weight voting method by merging six classifiers. For one particular test dataset, we can find each classifier trained by heterogeneous training dataset may have totally different classification performance. Some classifiers works well and achieve high classification accuracy, and some are not. For test dataset GSE1987, 86.49% classification accuracy was reached by classifier GSE7670, while only 48.65% test samples were classified correctly by classifier GSE9476 and GSE 6919. Similar situations happened on other three test datasets as well. Apparently, it is not reliable to determine the class of test sample in terms of only one classifier. Furthermore, since classifiers are trained by heterogeneous datasets, we have no prior knowledge about the performance of classifier when it is applied on different types of datasets. Therefore, merging different classifiers could provide stable classifications and reach relative high accuracies. As we can see in Table V the accuracy of meta-decision is basically close to the highest accuracy among six classifiers. The reason is that for a high performance classifier, the dataset dissimilarity factor d_D is relatively low. When taking weight voting, classifiers with lower dataset dissimilarity factors would have more effect on final results, and ensure a reliable high classification performance by combing multiple classifiers.

4. CONCLUSIONS

In this paper, we propose a novel alternative feature mapping method (AFM), which can match two related but both highly discriminative genes from training and test dataset. Contrary to the traditional direct mapping method (TDM) that maps genes of exact features, AFM takes advantage of a group of correlated genes instead of individual significant genes for mapping. Based on the concept of gene regularity networks, a set of appropriate genes from test groups can be recognized by AFM, which may be more suitable for classifying heterogeneous cancer samples. Moreover, the dataset dissimilarity factor derived by comparison of heterogeneous datasets can be applied on meta-classification through weight voting method. To test the performance of AFM, we conducted 6 single-classification and 4 meta-classification experiments consisting of 10 heterogeneous datasets. In order to compare the classification results with AFM, we also classified the same gene expression data using the traditional direct mapping. Our experiments show that classification accuracies obtain great improvement based on the appropriate discriminative genes selected by AFM, and AFM can also provide reliable classification performance by combing multiple classifiers from heterogeneous datasets.

References

[1] Nancy Mah, Anders Thelin, et al., “A comparison of oligonucleotide and cDNA-based microarray systems”, **Physiol. Genomics**, Vol. 16, Feb. 2004, pp. 361-370.

[2] Kyung-Joong Kim and Sung-Bae Cho, “Ensemble classifiers based on correlation analysis for DNA microarray classification”, **Neurocomputing**, Vol. 70(1-3), Dec. 2006, pp. 187-199.

[3] Dragomir A, Maraziotis I and Bezerianos A, “An ensemble approach for phenotype classification based on fuzzy partitioning of gene expression data”, **Conf Proc IEEE Eng Med Biol Soc. (EMBS 06)**, Jan. 2006, pp. 5834-5837.

[4] Järvinen AK, Hautaniemi S, et al., “Are data from different gene expression microarray platforms comparable?”, **Genomics**, Vol. 83(6), Jun. 2004, pp. 1164-1168.

[5] Cahan P, Rovegno F, et al., “Meta-analysis of microarray result: challenges, opportunities, and recommendation for standardization”, **Gene**., Vol. 401(1-2), Oct. 2007, pp. 12-18.

[6] Griffith OL, Pleasance ED, et al., “Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses”, **Genomics**, Vol. 86(4), Oct. 2005, pp. 476-488.

[7] Liu HC, Chen CY, et al., “Cross-generation and cross-laboratory predictions of Affymetrix microarrays by rank-based methods”, **Biomedical Informatics**, Vol. 41(4), Aug. 2008, pp. 570-579.

[8] Singh D, Febbo PG, et al., “Gene expression correlates of clinical prostate cancer behavior”, **Cancer Cell**, Vol. 1(2), Mar. 2002, pp. 203-209.

[9] Garrett-Mayer E, Parmigiani G, et al., “Cross-study validation and combined analysis of gene expression microarray data”, **Biostatistics**, Vol. 9(2), Apr. 2008, pp. 333-354.

[10] Hwang KB, Kong SW, et al., “Combining gene expression data from different generations of oligonucleotide arrays”, **BMC Bioinformatics**, Vol. 5: 159, Oct. 2004.