A hybrid machine learning algorithm for unlabeled gene expression data classification

Gokmen Zararsiz[1,*], Ahmet Ozturk[1], Erdem Karabulut[2], Ferhan Elmalı[1]

[1] Department of Biostatistics and Medical Informatics, Erciyes University, Kayseri, Turkey

[2] Department of Biostatistics, Hacettepe University, Ankara, Turkey

*Contact author: gokmenzararsiz@hotmail.com

1.Introduction

In gene expression data analysis, classification algorithms are widely used to classify biological samples and to predict clinical or other outcomes. But, these algorithms can not be used directly, if the data is unlabeled. Ignoring unlabeled data leads information loss and labeling them manually is a very difficult and expensive process. There are semi-supervised algorithms which were produced to label the unlabeled data [1-3]. However, these algorithms are impractical in wholly unlabeled datasets. Thus, it is very significant to generate the class label for this kind of wholly unlabeled datasets and clustering algorithms can be used to obtain such labels [4-5]. We proposed a hybrid unlabeled gene expression data classification (UGEDC) algorithm to determine the best clustering with an optimum cluster number, then classify the dataset using new generated class labels.

"In our algorithm, class labels are being provided by clustering, so clustering is the vital part of the study. Choosing the right clustering algorithm is complicated and given the same data set, different clustering algorithms can potentially generate very different clusters. Determining the right cluster number is another complicated process. Thus, we used cluster validation measures to choose the best clustering algorithm with a convenient cluster number."

2.UGEDC Algorithm

2.1 UGEDC Algorithm

i. Calculate z-scores of selected genes before clustering

ii. For k=2 to $\sqrt{n}$ perform UPGMA, k-means and SOM clustering algorithms to transformed data

iii. Calculate internal (Connectivity, Dunn, Silhouette) and stability (APN, AD, ADM, FOM) measures for each models

iv. For each measures rank the models

v. Aggregate the ranks and determine the overall winner model with Cross-Entropi Monte Carlo algorithm

vi. Perform the winner model and obtain the class labels for classification

vii. Classify the data by using new determined class labels with RBF-kernel SVM (Radial Based Function - kernel Support Vector Machines)

2.2 Clustering Algorithms Used in UGEDC

2.2.1 UPGMA

Unweighted Pair Group Method with Arithmetic Mean is an agglomerative, hierarchical clustering algorithm that yields a dendogram which can be cut at a chosen height to produce the desired number of clusters. Each observation is initially placed in its own cluster, and the clusters are successively joined together in order of their closeness. The closeness of any two clusters is determined by a dissimilarity matrix, and can be based on a variety of agglomeration methods.

2.2.2 K-means

K-means is an iterative method which minimizes the within-class sum of squares for a given number of clusters. The algorithm starts with an initial guess for the cluster centers, and each observation is placed in the cluster to which it is closest. The cluster centers are then updated, and the entire process is repeated until the cluster centers no longer move.

2.2.3 SOM

Self-organizing maps is an unsupervised learning technique that is popular among computational biologists and machine learning researchers. SOM is based on neural networks, and is highly regarded for its ability to map and visualize high-dimensional data in two dimensions [7].

2.3 SVM

As a powerful and popular multivariate machine-learning method, SVMs have been widely used in biological classification problems. A SVM constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space for classification and the key idea of the SVM is to maximize the margin separating the two classes while minimizing the total classification errors.

2.4 Cluster Validation Measures and Rank Aggregation

Table 1 – Internal and Stability Validation Measures

Measure	Formula	Value interval	Should be
Connectivity	$Conn(C)=\sum_{i=1}^{N}\sum_{j=1}^{L}x_{i,nn_{i(j)}}$	$[0,\infty]$	Minimum
Silhouette width	$S(i)=\frac{b_i-a_i}{\max(b_i,a_i)}$; $b_i=min_{C_k\epsilon C/C(i)}\sum_{j\epsilon C_k}\frac{dist(i,j)}{n(C_k)}$	$[-1,1]$	Maximum
Dunn Index	$D(C)=\frac{min_{C_k,C_1\epsilon C,C_k\neq C_1}(min_{i\epsilon C_k,j\epsilon C_1}dist(i,j))}{max_{C_m\epsilon C}diam(C_m)}$	$[0,\infty]$	Maximum
APN	$APN(K)=\frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}(1-\frac{n(C^{i,l}\cap C^{i,0})}{n(C^{i,0})})$	$[0,1]$	Minimum
AD	$AD(K)=\frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\frac{1}{n(C^{i,0})n(C^{i,l})}[\sum_{i\epsilon(C^{i,0}),j\epsilon(C^{i,l})}dist(i,j)]$	$[0,\infty]$	Minimum
ADM	$ADM(K)=\frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}dist(\bar{x}_{C^{i,l}},\bar{x}_{C^{i,0}})$	$[0,\infty]$	Minimum
FOM	$FOM(l,K)=\sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{i\epsilon C_k(l)}dist(x_{i,l},\bar{x}_{C_k(l)})}$	$[0,\infty]$	Minimum

N: Total number of observations	$C=\{C_1,\ldots,C_k\}$: Clusters	$C^{i,0}$: Cluster containing observation i in original data
M: Total number of genes	a_i: Average distance between i and all other observations in same cluster	$C^{i,l}$: Cluster containing observation i in column l removed data
L: Total number of nearest neighbors	b_i: Average distance between i and all other observations in nearest neighbor cluster	$\bar{x}_{C^{i,0}}$: mean of $C^{i,0}$
$nn_{i(j)}$: j th nearest neighbor of observation i	$dist(i,j)$: Euclidean distance between observations i and j	$\bar{x}_{C^{i,l}}$: mean of $C^{i,l}$
$x_{i,nn_{i(j)}}$: 0 if i and j are in same cluster, $1/j$ otherwise	$n(C_k)$: Cardinality of cluster C	$x_{i,l}$: value of i th observation, l th gene
K: Cluster number	$diam(C_m)$: maximum distance between observations in cluster C_m	$\bar{x}_{C_k(l)}$: average of cluster $C_k(l)$

Rank aggregation is helpful in reconciling the ranks and producing a "super"-list, which determines the overall winner and also ranks all the clustering algorithms based on their performance as determined by all the validation measures simultaneously [8,9].

2.5 Application

We used a public dataset to demonstrate our UGEDC [6]. There were expression values of 3971 genes belong to 101 samples (41 control, 60 marfan syndrome) in this dataset. First we ranked all genes from most significant to less significant on Marfan syndrome disease using t test p values. Then, we seperated the dataset to 2 parts: the most 500 significant genes to dataset A, remaining genes to dataset B. After that, we took 100 gene samples randomly from these datasets as 50%-50%, 75%-25%, 100%-0% respectively. Then, we repeated it 4 more times and did the same process for 200, 300, 400 and 500 genes. Finally, we performed our hybrid algorithm for all these datasets. Clustering and classification errors of all models were calculated for 75% training set and 25% test set.

Table 2 – Clustering and classification errors of UGEDC for different number of genes and different percentages from dataset A and dataset B.

Percentages from dataset A and dataset B	Number of genes: 100		200		300		400		500	
	UGEDC Clustering Error (%)	UGEDC Classification Error (%)	UGEDC Clustering Error (%)	UGEDC Classification Error (%)	UGEDC Clustering Error (%)	UGEDC Classification Error (%)	UGEDC Clustering Error (%)	UGEDC Classification Error (%)	UGEDC Clustering Error (%)	UGEDC Classification Error (%)
50-50	44.15 ±2.67	0.89 ±1.79	44.95 ±2.28	5,00 ±5.42	45.54 ±1.71	0.71 ±1.60	44.35 ±3.46	4.46 ±1.79	44.35 ±1.63	0.71 ±1.60
75-25	42.77 ±4.05	3.57 ±6.19	45.94 ±3.74	3.57 ±2.91	$\boxed{37.03 \pm 6.28}$	$\boxed{0.89 \pm 1.79}$	42.37 ±2.26	3.57 ±5.05	40.99 ±9.72	0.89 ±1.79
100-0	24.55 ±10.23	0,00	18.81 ±3.64	2.14 ±4.79	22.37 ±8.09	4.28 ±4.65	20.59 ±5.75	5.36 ±6.84	17.82	32.14

3.Results and Discussion

Results revealed that gene selection is very important to determine the right class labels. Even, clustering error is between 17.82-24.55% if we select all significant genes to our model. It is 37.03-45.94%, if we select 75 percent of genes from the significant dataset and 44.15-45.54%, if we select 50 percent of genes from the significant dataset. Also SVM is very adaptable with clustering results and we had very good results for all models. UGEDC algorithm was proposed in this project, and it is very useful to classify those data whose class labels and the number of clusters can not be provided in advance. The experiment results and the efficiency of our algorithm will be checked on synthetically generated datasets in following studies.

DEMONSTRATION of UGEDC ALGORITHM

-calculate z-scores of gene expression data

Case	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8	Gene9	...	Gene300	Class Label
Case1	2.04	1.824	1.795	2.552	2.114	2.362	2.062	1.723	1.813	...	2.646	?
Case2	1.945	2.033	1.754	2.678	1.91	2.314	1.992	1.809	1.861	...	0.67	?
Case3	1.997	2.028	2.031	2.548	2.116	2.539	2.072	1.749	2.025	...	-0.011	?
Case4	1.995	1.831	1.978	2.749	2.078	2.46	2.326	1.769	1.915	...	1.168	?
Case5	2.114	1.836	1.981	2.65	2.122	2.557	2.19	1.703	1.878	...	-0.973	?
Case6	2.052	2.043	2.056	2.604	2.131	2.579	2.206	1.718	1.994	...	-1.953	?
Case7	2.066	2.15	1.981	2.406	2.111	2.427	2.145	1.74	1.952	...	0.089	?
Case8	1.984	2.13	2.014	2.398	2.145	2.502	2.174	1.736	1.881	...	-0.542	?
Case9	2.023	1.955	1.905	2.234	2.013	2.304	2.048	1.749	1.885	...	-1.106	?
Case10	2.044	1.978	1.927	2.151	2.006	2.448	1.97	1.632	1.868	...	-1.272	?
Case11	1.83	1.866	1.74	2.444	1.938	2.273	1.989	1.67	1.902	...	0.122	?
Case12	1.78	1.942	1.762	2.485	1.997	2.308	2.048	1.647	1.784	...	0.537	?
.										.	.	.
.										.	.	.
.										.	.	.
Case101	1.919	1.801	1.815	2.282	2.012	2.303	1.982	1.599	1.761	...	1.135	?

-perform UPGMA, k-means and SOM for 2 to 10 clusters
-calculate internal and stability measures for each model

Clustering Algorithm	Validation Measure	Cluster Number: 2	3	4	5	6	7	8	9	10
UPGMA	APN	0.0001	0.0001	0.0004	0.0022	0.0069	0.006	0.0061	0.0384	0.0573
	AD	12.023	11.8424	11.6722	11.5144	11.3506	10.9069	10.7135	10.5553	10.4039
	ADM	0.0016	0.003	0.0066	0.0296	0.0657	0.1905	0.1899	0.3091	0.4257
	FOM	0.3376	0.3364	0.3354	0.3355	0.3352	0.3282	0.3271	0.3214	0.3183
	Connectivity	2.929	5.8579	8.7869	11.7159	14.6448	24.7571	30.2901	30.5012	33.4302
	Dunn	0.7039	0.5642	0.4936	0.4936	0.4936	0.5603	0.5603	0.5603	0.5603
	Silhouette	0.2941	0.1816	0.1363	0.1073	0.0967	0.0788	0.0715	0.0597	0.0491
k-means	APN	0.0025	0.0018	0.0104	0.0094	0.0128	0.0296	0.0119	0.0098	0.0079
	AD	11.4477	10.6651	10.3096	10.1046	9.9407	9.829	9.5854	9.3304	9.1632
	ADM	0.0268	0.0214	0.0972	0.1103	0.138	0.2455	0.1227	0.1655	0.1179
	FOM	0.3175	0.2942	0.2878	0.2848	0.2834	0.2833	0.2817	0.2748	0.2731
	Connectivity	17.7202	20.5913	44.271	51.2397	50.723	48.7151	59.8353	60.254	62.8734
	Dunn	0.4499	0.4499	0.4362	0.4449	0.4449	0.4571	0.4747	0.5193	0.5193
	Silhouette	0.1237	0.1376	0.1091	0.1071	0.1076	0.1072	0.1016	0.1075	0.1052
SOM	APN	0.4324	0.1717	0.0547	0.1662	0.0561	0.1557	0.2276	0.236	0.2929
	AD	12.2306	11.3379	10.3665	10.3446	9.9195	9.9347	9.8933	9.707	9.6649
	ADM	3.0267	1.5023	0.5049	1.4297	0.5465	1.2381	1.7576	1.8316	2.224
	FOM	0.319	0.303	0.2869	0.2794	0.2768	0.2767	0.2754	0.2731	0.273
	Connectivity	2.929	24.2353	44.271	48.0302	54.1738	71.5702	71.6706	93.544	114.7286
	Dunn	0.7039	0.3916	0.4362	0.4449	0.4563	0.4563	0.4563	0.4563	0.4563
	Silhouette	0.2941	0.0694	0.1091	0.1099	0.1084	0.0931	0.0795	0.0659	0.0545

-rank the models for each measure

Clustering Validation Measure	Model Ranks: 1	2	3	4	5	6	7	8	9	10
APN	UP-2	UP-3	UP-4	KM-3	UP-5	KM-2	UP-7	UP-8	UP-6	KM-10
AD	KM-10	KM-9	KM-8	SM-10	SM-9	KM-7	SM-8	SM-6	SM-7	KM-6
ADM	UP-2	UP-3	UP-4	KM-3	KM-2	UP-5	UP-6	KM-4	KM-5	KM-10
FOM	SM-10	SM-9	KM-10	KM-9	SM-8	SM-7	SM-6	SM-5	KM-8	KM-7
Connectivity	UP-2	SM-2	UP-3	UP-4	UP-5	UP-6	KM-2	KM-3	SM-3	UP-7
Dunn	UP-2	SM-2	UP-3	UP-7	UP-8	UP-9	UP-10	KM-9	KM-10	UP-4
Silhouette	UP-2	SM-2	UP-3	KM-3	UP-4	KM-2	SM-5	KM-4	SM-4	SM-6

UP-UPGMA, KM-k means, SM-SOM

-determine the overall winner model with CE algorithm

UPGMA clustering with 2 clusters

-perform the winner model to obtain class labels

Case	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8	Gene9	...	Gene300	Class Label
Case1	2.04	1.824	1.795	2.552	2.114	2.362	2.062	1.723	1.813	...	2.646	1
Case2	1.945	2.033	1.754	2.678	1.91	2.314	1.992	1.809	1.861	...	0.67	1
Case3	1.997	2.028	2.031	2.548	2.116	2.539	2.072	1.749	2.025	...	-0.011	1
Case4	1.995	1.831	1.978	2.749	2.078	2.46	2.326	1.769	1.915	...	1.168	1
Case5	2.114	1.836	1.981	2.65	2.122	2.557	2.19	1.703	1.878	...	-0.973	1
Case6	2.052	2.043	2.056	2.604	2.131	2.579	2.206	1.718	1.994	...	-1.953	1
Case7	2.066	2.15	1.981	2.406	2.111	2.427	2.145	1.74	1.952	...	0.089	1
Case8	1.984	2.13	2.014	2.398	2.145	2.502	2.174	1.736	1.881	...	-0.542	1
Case9	2.023	1.955	1.905	2.234	2.013	2.304	2.048	1.749	1.885	...	-1.106	2
Case10	2.044	1.978	1.927	2.151	2.006	2.448	1.97	1.632	1.868	...	-1.272	1
Case11	1.83	1.866	1.74	2.444	1.938	2.273	1.989	1.67	1.902	...	0.122	1
Case12	1.78	1.942	1.762	2.485	1.997	2.308	2.048	1.647	1.784	...	0.537	1
.										.	.	.
.										.	.	.
.										.	.	.
Case101	1.919	1.801	1.815	2.282	2.012	2.303	1.982	1.599	1.761	...	1.135	1

-Classify the dataset by using new determined class labels with RBF-kernel SVM

R packages used in the study

cluster – to perform UPGMA algorithm

kohonen – to perform SOM algorithm

clValid – to calculate cluster validation measures for each model

RankAggreg – to rank the models and determine the overall winner model

e1071 – to perform RBF-kernel SVM algorithm [10]

References

[1] S. Goldman, Y. Zhou (2000). Enhancing supervised learning with unlabeled data. In Proceedings of the 17th International Conference on Machine Learning, 327-334.

[2] Z. H. Zhou, M. Li (2005). Tri-training: exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering, vol. 17, 1529-1541.

[3] I.H. Witten, E. Frank (2005). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers.

[4] J. Xie, C. Wang, Y. Zhang, S. Jiang (2009). Clustering Support Vector Machines for Unlabeled Data Classification. 2009 International Conference on Test and Measurement (Hong Kong, China), pp. 34-38.

[5] W. Han, M. Kamber (2001). Data Mining Concepts and Techniques, Morgan Kaufmann Publishers.

[6] Z. Yao, J. C. Jaeger, W. L. Ruzzo et. al (2007). A Marfan syndrome gene expression phenotype in cultured skin fibroblasts. BMC Genomics, 8:319.

[7] T. Kohonen (1997). Self Organizing Maps. Springer-Verlag, second edition.

[8] G. Brock, V. Pihur, S. Datta, and S. Datta (2008). clValid, an R package for cluster validation. Journal of Statistical Software, 25(4).

[9] V. Pihur, S. Datta, S. Datta (2009). RankAggreg, an R package for weighted rank aggregation. BMC Bioinformatics, 10:62

[10] D. Meyer (2001). Support Vector Machines The Interface to libsvm in package e1071. R News Volume 1/3, 23-26.

Correspondence:

Gökmen ZARARSIZ

Department of Biostatistics and Medical Informatics,Faculty of Medicine, Erciyes University,38039 Kayseri / TURKEY

Phone: +90 352 4374937 - 23480

Fax : +90 352 4375285