

A hybrid machine learning algorithm for unlabeled gene expression data classification

Gokmen Zararsiz^{1,*}, Ahmet Ozturk¹, Erdem Karabulut², Ferhan Elmalı¹ 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

2.4 Cluster Validation Measures and Rank Aggregation

Table 1 – Interr	nal and Stability Validation Measures		-
Measure	Formula	Value interval	Should be
Connectivity	$Conn(\mathcal{C}) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}}$	[0,∞]	Minimum
Silhouette width	$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)}$ $b_i = \min_{C_k \in \mathcal{C}/C(i)} \sum_{j \in C_k} \frac{dist(i, j)}{n(C_k)}$	[-1,1]	Maximum
Dunn Index	$D(\mathcal{C}) = \frac{\min_{C_k, C_1 \in \mathcal{C}, C_k \neq C_1} (\min_{i \in C_k, j \in C_1} dist(i, j))}{\max_{C_m \in \mathcal{C}} diam(C_m)}$	[0,∞]	Maximum
ΔΡΝ	N M (cills cills)	[01]	Minimum

DEMONSTRATION of UGEDC ALGORITHM

						-calcula	te z-sco	ores of g	gene ex	pressio	nuala	
Case	Gene1	Gene2	Gene3 G	iene4 Ge	ne5 (Gene6 G	iene7 G	Gene8	Gene9		Gene300	Class Lak
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	2.6/8	2 116	2.314	1.992	1.809	1.861		0.67	ר ז
Cased	1.997	1 821	1 978	2.546	2.110	2.559	2.072	1.749	1 015		-0.011	: 2
Case4	2 114	1.836	1.978	2.745	2.078	2.40	2.520	1 703	1.913		-0 973	: ?
Case6	2.114	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	?
Case 10	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	?
										• • • • •		
												•
Case 101	1.919	1.801	1.815	2.282	2.012	2.303	1.982	1.599	1.761		1.135	?
						-perform	UPGN	1A, k-m	eans ai	nd SON	I for 2 to	o 10
								· ·				
						cilisters						
					(ciusters	to intorr	al and	stability	mode	uros for	oach
					-	ciusters -calculat	e intern	nal and	stability	/ measi	ures for	each
						ciusters -calculat model	te intern	nal and	stability	/ measi	ures for	each
	Clus	tering				ciusters -calculat model	te intern	nal and ^{uster Numl}	stability	/ measi	ures for	each
Algorithm	Clus	tering Validation	n Measure	2	3	ciusters -calculat model 4	te intern Clu 5 (nal and uster Numl	stability	/ measu	ures for	each
Algorithm UPGMA	Clus	tering Validation APN	n Measure	2 0.0001	3 0.0001	ciusters -calculat model	te intern Clu 5 0.0022	nal and uster Numl 6 0.0069	stability	/ measu 3 0.0061	ures for 	each
Algorithm UPGMA	Clus	tering Validation APN AD	n Measure	2 0.0001 12.023	3 0.0001 11.8424	ciusters -calculat model 4 0.0004 4 11.6722	te intern Clu 5 0.0022 11.5144	uster Numl 6 0.0069 11.3506	stability ber 7 8 0.006 10.9069	/ measu 3 0.0061 10.7135	ures for 0.0384 10.5553	each 10 10.09
Algorithm UPGMA	Clus	tering Validation APN AD ADM	n Measure	2 0.0001 12.023 0.0016	3 0.0001 11.8424 0.003	4 0.0004 11.6722 0.0066	te intern clu 5 0.0022 11.5144 0.0296	uster Numl 5 0.0069 11.3506 0.0657	stability ber 0.006 10.9069 0.1905	/ measu 0.0061 10.7135 0.1899	0.0384 0.0384 10.5553 0.3091	each 10.09 10.44 0.44
Algorithm UPGMA	Clus	tering Validation APN AD ADM FOM	n Measure	2 0.0001 12.023 0.0016 0.3376	3 0.0001 11.8424 0.003 0.3364	4 0.0004 11.6722 0.0066 0.3354	te intern Clu 5 0.0022 11.5144 0.0296 0.3355	uster Numl 0.0069 11.3506 0.0657 0.3352	stability ber 0.006 10.9069 0.1905 0.3282	/ measu 0.0061 10.7135 0.1899 0.3271	0.0384 0.0384 10.5553 0.3091 0.3214	each
Algorithm UPGMA	Clus	tering Validation APN AD ADM FOM Connectiv	n Measure /ity	2 0.0001 12.023 0.0016 0.3376 2.929	3 0.0001 11.8424 0.003 0.3364 5.8579	4 0.0004 11.6722 0.0066 0.3354 0.8.7869	te intern Clu 5 0.0022 11.5144 0.0296 0.3355 11.7159	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448	stability oer 0.006 10.9069 0.1905 0.3282 24.7571	/ measu 0.0061 10.7135 0.1899 0.3271 30.2901	0.0384 0.0384 10.5553 0.3091 0.3214 30.5012	each 0.02 10.44 0.42 0.33 33.43
Algorithm UPGMA	Clus	tering Validation APN AD ADM FOM Connectiv Dunn	n Measure /ity	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642	4 0.0004 11.6722 0.0066 0.3354 0.4936	te intern Clu 5 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936	stability oer 0.006 10.9069 0.1905 0.3282 24.7571 0.5603	/ measu 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603	0.0384 0.0384 10.5553 0.3091 0.3214 30.5012 0.5603	each 0.00 10.44 0.43 33.44 0.56
Algorithm UPGMA	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouetto	n Measure /ity e	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816	4 0.0004 11.6722 0.0066 0.3354 0.3354 0.4936 0.1363	te intern Clu 5 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788	/ measu 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715	0.0384 0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597	each 0.02 10.44 0.43 0.33 33.43 0.56 0.04
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouetto APN	n Measure /ity e	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018	4 0.0004 11.6722 0.0066 0.3354 0.3354 0.4936 0.1363 0.0104	te intern Clu 5 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788 0.0296	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119	0.0384 0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098	each 10 10.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouetto APN AD	n Measure vity e	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651	Clusters -calculat model 4 0.0004 11.6722 0.0066 0.3354 8.7869 0.4936 0.1363 0.0104 10.3096	te intern Clu 5 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788 0.0296 9.829	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854	0.0384 0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304	each 0.02 10.44 0.42 0.32 33.43 0.54 0.04 0.00 9.10
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouetto APN AD ADM	n Measure vity e	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214	4 0.0004 11.6722 0.0066 0.3354 0.4936 0.1363 0.0104 10.3096 0.0972	te intern Ch 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.1103	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788 0.0296 9.829 0.2455	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227	0.0384 0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304 0.1655	each 10 10.44 0.42 0.32 33.43 0.55 0.04 0.00 9.10 0.01
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouetto APN AD ADM FOM	n Measure vity e	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268 0.3175	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214 0.2942	4 0.0004 11.6722 0.0066 0.3354 0.3354 0.4936 0.1363 0.0104 10.3096 0.0972 0.2878	te intern Ch 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.1103 0.2848	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138 0.2834	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788 0.0296 9.829 0.2455 0.2833	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227 0.2817	0.0384 0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304 0.1655 0.2748	each 10 10.44 0.42 0.32 33.42 0.55 0.04 0.00 9.10 0.12 0.2
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouetto APN AD ADM FOM	n Measure vity e	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268 0.3175 17,702	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214 0.2942 20.5913	4 0.0004 11.6722 0.0066 0.3354 0.4936 0.4936 0.1363 0.0104 10.3096 0.0972 0.2878 44.271	te intern Ch 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.1103 0.2848 51 2397	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138 0.2834 50 723	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788 0.0296 9.829 0.2455 0.2833 48 7151	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227 0.2817	0.0384 0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304 0.1655 0.2748 60 254	each 10 10.44 0.42 0.32 33.42 0.56 0.00 9.16 0.12 0.22 62.82
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouetto APN AD ADM FOM Connectiv Dunr	n Measure vity e	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268 0.3175 17.7202 0.4499	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214 0.2942 20.5913	4 0.0004 11.6722 0.0066 0.3354 0.4936 0.4936 0.1363 0.0104 10.3096 0.0972 0.2878 44.271 0.4362	te intern 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.1103 0.2848 51.2397 0.4449	al and uster Numb 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138 0.2834 50.723 0.4449	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788 0.0296 9.829 0.2455 0.2833 48.7151 0.4571	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227 0.2817 59.8353 0.4747	0.0384 0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304 0.1655 0.2748 60.254 0.5192	each 10 10.44 0.42 0.32 33.43 0.56 0.04 0.04 0.32 0.56 0.04 0.04 0.12 0.22 62.83 0.55
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouetta APN AD ADM FOM Connectiv Dunn	n Measure vity e	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268 0.3175 17.7202 0.4499 0.1327	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214 0.2942 20.5913 0.4499 0.1275	4 0.0004 0.0004 11.6722 0.0066 0.3354 0.4936 0.1363 0.0104 10.3096 0.0972 0.2878 0.4362 0.4362 0.1091	te intern 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.1103 0.2848 51.2397 0.4449 0.1071	al and uster Number 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138 0.2834 50.723 0.4449 0.1076	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0296 9.829 0.2455 0.2833 48.7151 0.4571 0.1072	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227 0.2817 59.8353 0.4747 0.1016	0.0384 0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304 0.1655 0.2748 60.254 0.5193 0.1025	each 10 10.44 0.42 0.32 33.43 0.56 0.04 0.00 9.10 0.11 0.22 62.85 0.55
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouetta ADM FOM Connectiv Dunn Silhouetta Dunn	n Measure vity e vity	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268 0.3175 17.7202 0.4499 0.1237 0.4324	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214 0.2942 20.5913 0.4499 0.1376	4 -calculat model 4 0.0004 11.6722 3 0.0066 4 0.3354 9 8.7869 2 0.4936 3 0.0104 10.3096 0.0972 2 0.2878 3 44.271 9 0.4362 5 0.1091	te intern 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.1103 0.2848 51.2397 0.4449 0.1071	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138 0.2834 50.723 0.4449 0.1076	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0296 9.829 0.2455 0.2833 48.7151 0.4571 0.1072 0.1072	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227 0.2817 59.8353 0.4747 0.1016 0.2376	0.0384 0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304 0.1655 0.2748 60.254 0.5193 0.1075 0.326	each 10 10.44 0.42 0.33 33.43 0.56 0.04 0.00 9.10 0.11 0.22 62.88 0.51 0.11 0.21 62.88 0.51 0.11 0.21
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouette ADM FOM Connectiv Dunn Silhouette APN	n Measure vity e vity	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268 0.3175 17.7202 0.4499 0.1237 0.4324	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214 0.2942 20.5913 0.4499 0.1376 0.1717	4 -calculat model 4 0.0004 11.6722 3 0.0066 4 0.3354 9 8.7869 2 0.4936 3 0.0104 10.3096 0.0972 2 0.2878 3 44.271 9 0.4362 5 0.1091 7 0.0547	te intern 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.1103 0.2848 51.2397 0.4449 0.1071 0.1662	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138 0.2834 50.723 0.4449 0.1076 0.0561	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0296 9.829 0.2455 0.2833 48.7151 0.4571 0.1072 0.1557	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227 0.2817 59.8353 0.4747 0.1016 0.2276	0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304 0.1655 0.2748 60.254 0.5193 0.1075 0.236 0.236	each 10 10.44 0.42 0.33 33.42 0.54 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.05 0.04 0.04 0.05 0.05
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouette ADM FOM Connectiv Dunn Silhouette APN	n Measure vity e vity	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268 0.3175 17.7202 0.4499 0.1237 0.4324 12.2306	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214 0.2942 20.5913 0.4499 0.1376 0.1717 11.3379	4 -calculat model 4 0.0004 11.6722 3 0.0066 4 0.3354 9 8.7869 2 0.4936 3 0.0104 10.3096 0.0972 2 0.2878 3 44.271 9 0.4362 5 0.1091 7 0.0547	te intern 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.1103 0.2848 51.2397 0.4449 0.1071 0.1662 10.3446	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138 0.2834 50.723 0.4449 0.1076 0.0561 9.9195	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788 0.0296 9.829 0.2455 0.2833 48.7151 0.4571 0.1072 0.1557 9.9347	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227 0.2817 59.8353 0.4747 0.1016 0.2276 9.8933	0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304 0.1655 0.2748 60.254 0.5193 0.1075 0.236 9.707	each 0.02 10.44 0.42 0.33 33.42 0.54 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.05 0.04 0.04 0.04 0.05 0.04 0.04 0.05 0.04 0.05 0.04 0.05 0.04 0.05
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouette APN AD ADM FOM Connectiv Dunn Silhouette APN AD	n Measure vity e	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268 0.3175 17.7202 0.4499 0.1237 0.4324 12.2306 3.0267	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214 0.2942 20.5913 0.4499 0.1376 0.1717 11.3379 1.5023	4 -calculat model 4 0.0004 11.6722 3 0.0066 4 0.3354 9 8.7869 2 0.4936 3 0.0104 10.3096 0.0972 2 0.2878 3 44.271 9 0.4362 5 0.1091 7 0.0547 9 10.3665 9 0.5049	te intern 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.1103 0.2848 51.2397 0.4449 0.1071 0.1662 10.3446 1.4297	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138 0.2834 50.723 0.4449 0.1076 0.0561 9.9195 0.5465	stability ber 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788 0.0296 9.829 0.2455 0.2833 48.7151 0.4571 0.1072 0.1557 9.9347 1.2381	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227 0.2817 59.8353 0.4747 0.1016 0.2276 9.8933 1.7576	0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304 0.1655 0.2748 60.254 0.5193 0.1075 0.236 9.707 1.8316	each 10 10.44 0.44 0.33 33.44 0.54 0.00 9.11 0.12 62.88 0.55 0.11 0.22 9.66 2.1
Algorithm UPGMA k-means	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouette APN AD ADM FOM Connectiv Dunn Silhouette APN AD ADM FOM	n Measure vity e	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268 0.3175 17.7202 0.4499 0.1237 0.4324 12.2306 3.0267 0.319	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214 0.2942 20.5913 0.4499 0.1376 0.1717 11.3379 1.5023 0.303	4 -calculat model 4 0.0004 11.6722 0.0066 0.3354 8.7869 0.4936 0.1363 0.0104 10.3096 0.0972 0.2878 44.271 0.4362 0.1091 0.0547 10.3665 0.2869	te intern 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.103 0.2848 51.2397 0.4449 0.1071 0.1662 10.3446 1.4297 0.2794	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138 0.2834 50.723 0.4449 0.1076 0.0561 9.9195 0.5465 0.2768	stability der 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788 0.0296 9.829 0.2455 0.2833 48.7151 0.4571 0.1072 0.1557 9.9347 1.2381 0.2767	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227 0.2817 59.8353 0.4747 0.2817 59.8353 0.4747 0.1016 0.2276 9.8933 1.7576 0.2754	0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304 0.1655 0.2748 60.254 0.5193 0.1075 0.236 9.707 1.8316 0.2731	each 10 10.44 0.44 0.33 33.44 0.54 0.05 0.04
Algorithm UPGMA	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouette APN AD ADM FOM Connectiv Dunn Silhouette APN AD ADM FOM Connectiv Dunn	n Measure vity e vity e	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268 0.3175 17.7202 0.4499 0.1237 0.4324 12.2306 3.0267 0.319 2.929	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214 0.2942 20.5913 0.4499 0.1376 0.1717 11.3379 1.5023 0.303 24.2353	4 -calculat model 4 0.0004 11.6722 0.0066 0.3354 8.7869 0.4936 0.1363 0.0104 10.3096 0.0972 0.2878 44.271 0.0547 0.0547 0.2869 0.2869 44.271	te intern 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.103 0.2848 51.2397 0.4449 0.1071 0.2848 51.2397 0.4449 0.1071 0.1662 10.3446 1.4297 0.2794 48.0302	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138 0.2834 50.723 0.4449 0.1076 0.0561 9.9195 0.5465 0.2768 54.1738	stability der 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788 0.0296 9.829 0.2455 0.2833 48.7151 0.4571 0.1072 0.1557 9.9347 1.2381 0.2767 71.5702	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227 0.2817 59.8353 0.4747 0.2817 59.8353 0.4747 0.1016 0.2276 9.8933 1.7576 0.2754 71.6706	0.0384 10.5553 0.3091 0.3214 30.5012 0.5603 0.0597 0.0098 9.3304 0.1655 0.2748 60.254 0.5193 0.1075 0.236 9.707 1.8316 0.2731 93.544	each 10 10.44 0.44 0.33 33.44 0.54 0.04 0.04 0.04 0.11 0.22 62.88 0.55 0.11 0.22 9.66 2.1 0.21 9.66 2.1 0.12 9.66 2.1 0.12 9.66 2.1 0.14 0.22 9.66 2.1 0.14 0.22 0.15 0.14 0.22 0.15 0.14 0.22 0.15 0.14 0.22 0.15 0.14 0.22 0.15 0.14 0.22 0.15 0.14 0.22 0.15 0.11 0.22 0.22 0.11 0.22
Algorithm UPGMA	Clus	tering Validation APN AD ADM FOM Connectiv Dunn Silhouetta APN AD ADM FOM Connectiv Dunn Silhouetta APN AD ADM FOM Connectiv Dunn Silhouetta APN AD Connectiv Dunn	n Measure /ity e /ity	2 0.0001 12.023 0.0016 0.3376 2.929 0.7039 0.2941 0.0025 11.4477 0.0268 0.3175 17.7202 0.4499 0.1237 0.4324 12.2306 3.0267 0.319 2.929 0.7039	3 0.0001 11.8424 0.003 0.3364 5.8579 0.5642 0.1816 0.0018 10.6651 0.0214 0.2942 20.5913 0.4499 0.1376 0.1717 11.3379 1.5023 0.303 24.2353 0.3916	4 -calculat model 4 0.0004 11.6722 0.0066 0.3354 8.7869 0.4936 0.1363 0.0104 10.3096 0.0104 10.3096 0.2878 44.271 0.0547 10.3665 0.2869 44.271 0.2869 44.271 0.2869 44.271 0.3665 0.2869 44.271	te intern 0.0022 11.5144 0.0296 0.3355 11.7159 0.4936 0.1073 0.0094 10.1046 0.1103 0.2848 51.2397 0.4449 0.1071 0.1662 10.3446 1.4297 0.2794 48.0302 0.4449	uster Numl 0.0069 11.3506 0.0657 0.3352 14.6448 0.4936 0.0967 0.0128 9.9407 0.138 0.2834 50.723 0.4449 0.1076 0.0561 9.9195 0.5465 0.2768 54.1738 0.4563	stability oer 0.006 10.9069 0.1905 0.3282 24.7571 0.5603 0.0788 0.0296 9.829 0.2455 0.2833 48.7151 0.4571 0.1072 0.1557 9.9347 1.2381 0.2767 71.5702 0.4563	/ mease 0.0061 10.7135 0.1899 0.3271 30.2901 0.5603 0.0715 0.0119 9.5854 0.1227 0.2817 59.8353 0.4747 0.2817 59.8353 0.4747 0.2817 59.8353 0.4747 0.1016 0.2276 9.8933 1.7576 0.2754 71.6706 0.4563	2000 2000 2000 2000 2000 2000 2000 200	each 0.02 10.44 0.42 0.32 33.42 0.52 0.00 9.16 0.12 0.22 0.52 0.10 0.22 9.66 2.2 0.12 0.22 114.72 0.45 0.54 0.55 0.12 0.55 0.12 0.52 0.12 0.54 0.5

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 withinclass 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 highdimensional 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.



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.

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

-rank the models for each r	measure
-----------------------------	---------

Clustering		Model Ranks									
Validation Measure	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	

-determine the overall winner model with CE algorithm

UPGMA clustering with 2 clusters

-perform	the	winner	model to	o obtain	class	labels
periorini	uic	winner	mouch	oblam	01000	Tabelo

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	51	
Case 2	1.945	2.033	1.754	2.678	1.91	2.314	1.992	1.809	1.861	0.67	7 1	
Case 3	1.997	2.028	2.031	2.548	2.116	2.539	2.072	1.749	2.025	-0.011	L 1	
Case4	1.995	1.831	1.978	2.749	2.078	2.46	2.326	1.769	1.915	1.168	3 1	
Case 5	2.114	1.836	1.981	2.65	2.122	2.557	2.19	1.703	1.878	-0.973	3 1	
Case6	2.052	2.043	2.056	2.604	2.131	2.579	2.206	1.718	1.994	-1.953	31	
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	2 1	
Case9	2.023	1.955	1.905	2.234	2.013	2.304	2.048	1.749	1.885	-1.106	5 2	
Case 10	2.044	1.978	1.927	2.151	2.006	2.448	1.97	1.632	1.868	-1.272	2 1	
Case 11	1.83	1.866	1.74	2.444	1.938	2.273	1.989	1.67	1.902	0.122	2 1	
Case 12	1.78	1.942	1.762	2.485	1.997	2.308	2.048	1.647	1.784	0.537	7 1	
											•	
Case 101	1.919	1.801	1.815	2.282	2.012	2.303	1.982	1.599	1.761	1.13	5 1	

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.

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.



-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]

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