Use of PCA, hierarchical clustering, and k-means to group cancers by type using their gene expression profiles as features

Gene expression cancer RNA-seq data set

Asya Khleborodova

2018-12-17

Contents

0.1 Summary

Cancer types are characterized by their gene expression patterns. It follows that it may be possible to differentiate between cancer types or subtypes based on such patterns. The data set used here is 801 cancers samples, each cancer sample belongs to one of five cancer types. Each sample has 20,531 gene expression measurements as its features. In the folowing analysis I use unsupervised learning methods, PCA, hierarchical clustering, and k-means clustering, to subgroup each of 801 cancer samples based on its gene expression profile into one of 5 clusters. The best model result in 5 homogeneous clusters, each corresponding to a cancer type.

0.2 Data Gathering / Description / Wrangling

Gene expression cancer RNA-seq data set is located on the UCI Machine Learning Repository site. The data set has 20,531 features (genes) and 801 instances (one of 5 cancer types). The cancer types are Breast (BRCA), Colon (COAD), Kidney (KIRC), Lung (LUAD), and Prostate (PRAD). There are two files *data.csv* contains the RNA-seq gene expression levels as measured by illumina HiSeq platform and *lavels.csv* contains cancer type label for each of 801 samples.

0.2.1 Data Preparation and Wrangling

Gene expression data was imported and stored as *rna.data*, while cancer labels for 801 samples were stored as *rna.labs*.

```
library(dplyr)
rna.labs=read.csv(
"https://media.githubusercontent.com/media/asyakhl/cancerRNA_clustering/master/data/labels.csv",
  header=T)
rna.labs=as.character(rna.labs$Class)
rna.data=read.csv(
"https://media.githubusercontent.com/media/asyakhl/cancerRNA_clustering/master/data/data.csv",
  header = T)
```

First column (gene names) of the data set was deleted. Columns (genes) with zero expression were deleted from the data set, resulting in 20,264 features.

```
sample_names=rna.data$X
rna.data=rna.data[, -1]
rownames(rna.data)=sample_names
sums=colSums(rna.data)
sums0=(sums==0)
#267 columns (gene labels) were deleted from the data frame since all their values were 0
rna.data1=rna.data[,!sums0]
```

0.3 Data Analysis

0.3.1 Model Building

PCA, hierarchical clustering, and k-means are unsupervised techniques, meaning cancer type labels in *rna.labs* are not used with these techniques. The goal is to cluster the gene expression data into 5 separate groups. The resulting groups can then be checked against *rna.labs* to see that each cluster contains only/mostly one cancer type.

Considering the data and the techniques being used here, there are a few options for carrying out these methods: (1) the data can be scaled or unscaled, (2) hierarchical clustering and k-means can be done with or without inital principal component analysis, (3) hierarchical clustering linkage can be specified as *complete* (default), *single*, *average*, or *centroid*, (4) hierarchical clustering dissimilarity measure can be specified as Euclidean or correlation-based. Most variations of these options are tried here to find best unsupervised clustering model for the given data set.

Hierarchical Clustering

The code below shows that scaling the data has a negative effect on the accuracy of hierarchical clustering. When the data is scaled, hierarchical clustering assigns most observations to the same cluster #2. Well-performing model would separate 801 samples into one of 5 clusters based on sample gene expression profiles and comparison of samples within each cluster should reveal that their are nearly all of the same cancer type.

```
# unscaled hierarchical clustering: hclust(), complete linkage
hc.out=hclust(dist(rna.data1))
knitr::kable(cbind(cluster=c(1:5),
             as.data.frame.matrix(table(cutree (hc.out ,5), rna.labs))),
             caption = "Unscaled Hierarchical Clustering with Complete Linkage")
```

Table 1: Unscaled Hierarchical Clustering with Complete Linkage

cluster	BRCA	COAD	KIRC	LUAD	PRAD
1	0	0	0	0	135
2	22	23	0	139	0
3	271	0	0	2	1
4	0	0	146	0	0
5	7	55	0	0	0

```
# scale() scales variables to 0 mean and variance of 1
rna.data1.scaled=scale(rna.data1)
#scaled hierarchical clustering
hc.out.scaled=hclust(dist(rna.data1.scaled))
knitr::kable(cbind(cluster=c(1:5),
             as.data.frame.matrix(table(cutree (hc.out.scaled ,5), rna.labs))),
             caption = "Scaled Hierarchical Clustering with Complete Linkage")
```

Table 2: Scaled Hierarchical Clustering with Complete Linkage

cluster	BRCA	COAD	KIRC	LUAD	PRAD
1	10	0	0	1	10
2	244	77	145	134	125
3	40	1	0	5	0
4	6	0	1	1	0
5	0	0	0	0	1

Of the four linkage methods, complete linkage gives the best results. Compare the results below to that of unscaled *complete* linkage method above.

```
# average linkage (i.e. method="average") and Euclidean distance (i.e. dist())
hc.out.avg=hclust(dist(rna.data1), method="average")
knitr::kable(cbind(cluster=c(1:5),
             as.data.frame.matrix(table(cutree (hc.out.avg ,5), rna.labs))),
             caption = "Unscaled Hierarchical Clustering with Average Linkage")
```

Table 3: Unscaled Hierarchical Clustering with Average Linkage

cluster	BRCA	COAD	KIRC	LUAD	PRAD
1	289	78	145	140	135
2	0	0	0	0	1
3	10	0	0	0	0
4	1	0	0	1	0
5	0	0	1	0	0

```
#single linkage (i.e. method="single") and Euclidean distance (i.e. dist())
hc.out.single=hclust(dist(rna.data1), method="single")
knitr::kable(cbind(cluster=c(1:5),
             as.data.frame.matrix(table(cutree (hc.out.single ,5), rna.labs))),
```

```
            caption = "Unscaled Hierarchical Clustering with Single Linkage")
```

Table 4: Unscaled Hierarchical Clustering with Single Linkage

cluster	BRCA	COAD	KIRC	LUAD	PRAD
1	299	78	144	140	136
2	0	0	1	0	0
3	1	0	0	0	0
4	0	0	0	1	0
5	0	0	1	0	0

```
# centroid linkage (i.e. method="centroid") and Euclidean distance (i.e. dist())
hc.out.centr=hclust(dist(rna.data1), method="centroid")
knitr::kable(cbind(cluster=c(1:5),
            as.data.frame.matrix(table(cutree (hc.out.centr ,5), rna.labs))),
            caption = "Unscaled Hierarchical Clustering with Centroid Linkage")
```

Table 5: Unscaled Hierarchical Clustering with Centroid Linkage

cluster	BRCA	COAD	KIRC	LUAD	PRAD
1	299	77	144	141	136
2	0	1	0	0	0
3	0	0	1	0	0
4	1	0	0	0	0
5	0	0	1	0	0

Principal Component Analysis

So far the best model is hierarchical clustering with unscaled data and complete linkage method. Next, PCA analysis is performed to reduce dimentionality of the gene expression data set, *rna.data1* and capture most (~80%) between-cluster variance in fewer than 20,264 dimentions. The code below shows that once again unscaled data performs better than the scaled data.

```
# unscaled PCA
pc.out.unscaled=prcomp(rna.data1, scale=F)
# scaled PCA
pc.out.scaled=prcomp(rna.data1, scale=T)
# proportion of variance explained (pve) by each PC
pve.unscaled=summary(pc.out.unscaled)$importance[2,]
#cumulative pve
cumul.pve.unscaled=summary(pc.out.unscaled)$importance[3,]
head(cumul.pve.unscaled,20)
```

```
##     PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
## 0.15839 0.26343 0.35815 0.42316 0.45931 0.48904 0.51561 0.53124 0.54531
##    PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18
## 0.55758 0.56718 0.57608 0.58366 0.59089 0.59753 0.60391 0.60960 0.61478
##    PC19    PC20
## 0.61940 0.62386
```

```
# pve and cumulative pve for scaled data
pve.scaled=summary(pc.out.scaled)$importance[2,]
cumul.pve.scaled=summary(pc.out.scaled)$importance[3,]
head(cumul.pve.scaled,20)
```

```
##     PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
## 0.10540 0.19294 0.27104 0.32270 0.36299 0.39220 0.41574 0.43720 0.45352
##    PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18
## 0.46585 0.47640 0.48534 0.49385 0.50138 0.50804 0.51451 0.52060 0.52645
##    PC19    PC20
## 0.53174 0.53690
```

Result of Unscaled PCA as Input to Hierarchical Clustering

Unscaled PCA shows that 129 PCs account for 80% of variance between 5 clusters. Scores resulting from 129 PCs were used as unput to hierarchical clustering.

```
# complete linkage and Euclidean distance
hc.129out.unscaled=hclust(dist(pc.out.unscaled$x [,1:129]) )
knitr::kable(cbind(cluster=c(1:5),
as.data.frame.matrix(table(cutree (hc.129out.unscaled ,5), rna.labs))),
caption = "Hierarchical Clustering  with Scores Based
    on 129 PCs as Input from Unscaled Data ")
```

Table 6: Hierarchical Clustering with Scores Based on 129 PCs as Input from Unscaled Data

cluster	BRCA	COAD	KIRC	LUAD	PRAD
1	0	0	0	0	135
2	279	23	0	141	0
3	21	55	0	0	1
4	0	0	142	0	0
5	0	0	4	0	0

Unscaled hierarchical clustering with complete linkage outperforms the above model.

Unscaled PCA Result as Input to Hierarchical Cluster with Correlation-Based Dissimilarity

Here, correlation is used as dissimilarity measure instead of the Euclidean distance.

```
# correlations are transformed to distance via as.dist()
dd.unscaled=as.dist(1-cor(t(rna.data1)))
# correlation-based distance is used for hierarchical clustering
hc.corr.unscaled=hclust(dd.unscaled, method ="complete")
knitr::kable(cbind(cluster=c(1:5),
as.data.frame.matrix(table(cutree (hc.corr.unscaled, 5), rna.labs))),
caption ="Hierarchical Cluster with Correlation-Based
  Dissimilarity and Unscaled Data ")
```

Table 7: Hierarchical Cluster with Correlation-Based Dissimilarity and Unscaled Data

cluster	BRCA	COAD	KIRC	LUAD	PRAD
1	0	0	0	0	136
2	0	29	0	139	0
3	270	0	0	2	0
4	0	0	146	0	0
5	30	49	0	0	0

```
# correlations between PC scores are used as distances
dd.pc.unscaled=as.dist(1-cor(t(pc.out.unscaled$x [,1:129])))
hc.corr.pc.unscaled=hclust(dd.pc.unscaled, method ="complete")
knitr::kable(cbind(cluster=c(1:5),
as.data.frame.matrix(table(cutree (hc.corr.pc.unscaled ,5), rna.labs))),
caption ="Hierarchical Cluster with Correlation-Based Dissimilarity
    Based on 129 PCs as Input from Unscaled Data ")
```

Table 8: Hierarchical Cluster with Correlation-Based Dissimilarity Based on 129 PCs as Input from Unscaled Data

cluster	BRCA	COAD	KIRC	LUAD	PRAD
1	0	0	0	0	136
2	0	78	0	139	0
3	199	0	0	0	0
4	0	0	146	0	0
5	101	0	0	2	0

Unscaled hierarchical clustering with complete linkage outperforms models shown above.

K-MEANS

Unscaled K-means model outperforms the unscaled hierarchical clustering model.

```
# k-means method for K=5 (unscaled data)
km.out.unscaled=kmeans (rna.data1, 5, nstart =20)
# k-means results
head(km.out.unscaled$cluster)
```

```
## sample_0 sample_1 sample_2 sample_3 sample_4 sample_5
##        3        2        3        3        5        3
```

```
knitr::kable(cbind(cluster=c(1:5),
            as.data.frame.matrix(table(km.out.unscaled$cluster, rna.labs))),
            caption ="k-means method for K=5 (unscaled data)")
```

Table 9: k-means method for K=5 (unscaled data)

cluster	BRCA	COAD	KIRC	LUAD	PRAD
1	0	0	145	0	0

cluster	BRCA	COAD	KIRC	LUAD	PRAD
2	1	2	0	139	0
3	0	0	0	0	136
4	0	76	0	0	0
5	299	0	1	2	0

```
# k-means method for K=5 (scaled data)
km.out.scaled=kmeans(rna.data1.scaled, 5, nstart=20)
head(km.out.scaled$cluster)
```

```
## sample_0 sample_1 sample_2 sample_3 sample_4 sample_5
##        3        5        3        3        4        3
```

```
knitr::kable(cbind(cluster=c(1:5),
            as.data.frame.matrix(table(km.out.scaled$cluster, rna.labs))),
            caption ="k-means method for K=5 (scaled data)")
```

Table 10: k-means method for K=5 (scaled data)

cluster	BRCA	COAD	KIRC	LUAD	PRAD
1	0	73	0	0	0
2	0	0	145	0	0
3	0	0	0	0	134
4	247	0	0	2	1
5	53	5	1	139	1

0.3.2 Model Fit

Of the nubmer of models tried here, the best model for Gene expression cancer RNA-seq data set is Unscaled K-means model. Since there are 6 misclassifications, the unscaled k-means model has error rate of 100*(6/801)=0.75%. It appears that there are no significant outliers, since all 801 observations belong to one of five clusters in either k-means or hierarchical clustering.

0.3.3 Model Interpretation

Although, PCA was not very useful for model selection, PC loadings can be used to identify features (genes) that contribute most to differences in cancer clusters. The larger the loading of a feature in a PC loading vector, the more that feature contributes to differences in clusters. Also, features with similar size loadings are correlated.

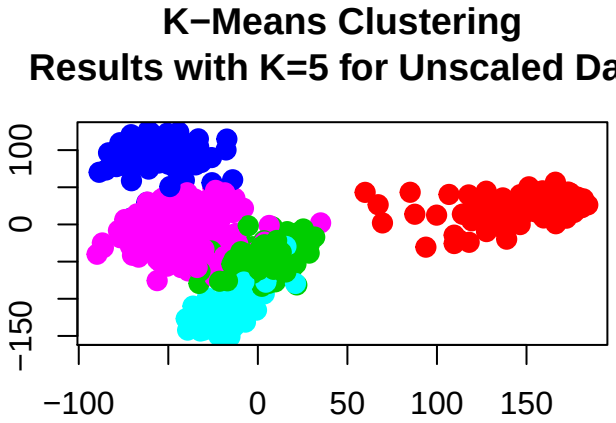
```
library(dplyr)
knitr::kable(head(data.frame(feature=names(pc.out.unscaled$rotation[,1]),
           PC1=pc.out.unscaled$rotation[,1])%>%arrange(desc(PC1))))
```

feature	PC1
gene_3439	0.0554663
gene_6733	0.0553963

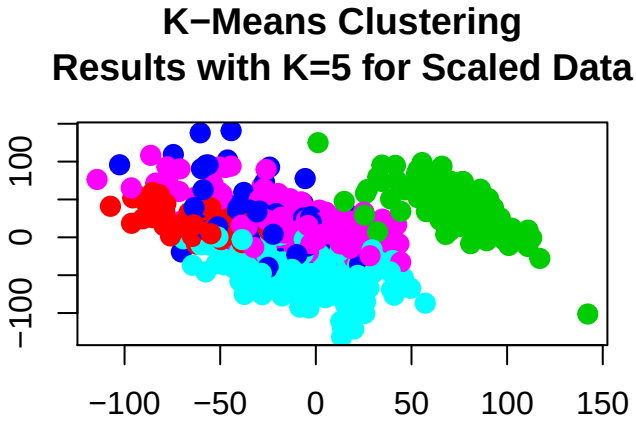
feature	PC1
gene_219	0.0535013
gene_1510	0.0524072
gene_16132	0.0522969
gene_16169	0.0518607

K-means results can be visualized using the first two PCs, although observations are not separated into 5 perfect clusters, because only two PCs or principal component scores are used to graph the 5 clusters.

```
plot(pc.out.unscaled$x[,1:2], col=(km.out.unscaled$cluster +1), main="K-Means Clustering
Results with K=5 for Unscaled Data", xlab="", ylab="", pch=20, cex=2)
```




```
plot(pc.out.scaled$x[,1:2], col=(km.out.scaled$cluster +1), main="K-Means Clustering
Results with K=5 for Scaled Data", xlab="", ylab="", pch=20, cex=2)
```



0.4 Conclusion

In conclusion, the best clustering model for Gene expression cancer RNA-seq data set, was unscaled k-means model with a very low error rate of 0.75%. PCA can be used for a closer look features (genes).