Descriptive and visual modeling of expression data

P.Antal

antal@mit.bme.hu

BME

Overview

1. The types and nature of high-throughput molecular biological data
2. Descriptive, visual, statistical and causal data analysis
3. Distance, correlation, association, similarity
4. Dimensionality reduction
   (a) Distance preservation mappings: multidimensional scaling
   (b) Topology preservation mappings: self-organizing maps
   (c) Factor/variance preservation mappings
5. Clustering
   (a) Types: partitioning (parametric, non-parametric), hierarchical
   (b) Parametric partitioning (density-based) clustering
   (c) Non-parametric partitioning clustering
   (d) Hierarchical clustering: error functions, methods
   (e) Evaluation (external, internal) and fusion (early, mid, late)
   (f) Extensions: asymmetric, rejection,bi-clustering, proximity-based,etc

The knowledge- and data-rich biomedicine

High-throughput measuring methods at genomic, proteomic, metabolic level $\Rightarrow$

1. sequencing robots:genome sequences,
2. gene chips: **gene expression data** (mRNA),
3. protein chips: protein-protein interaction,

Types of data: (consensus) sequence, expression, interaction, (protein) structure, (protein) location, sequence variations (SNPs), . . . (recall: Collins: A vision for the future of the genomics research)
Nature of biomedical data (analysis):

1. high-dimensional,
2. noisy,
3. unexplored,
4. embedded in rich qualitative knowledge,
5. multilevel/structured, distributed,
6. for biologists (direct probabilistic, qualitive/graphical inferences) and clinicians (validity, ethical issues, practicality)

Types of data analysis

Assume a "preprocessed"[a] gene expression (GE) data set D_n about the objective expressed quantity of 10^3 genes containing 10^3 samples under 10 different conditions without and with knocking-out/silencing 9 genes separately. Data preprocessing and particularly GE preprocessing is an art, we focus on the more general
Questions and types (phases?) of data analysis (descriptive, visual, statistical and causal):

1. Descriptive statistics of the data set, e.g. means, variances, histograms, univariate/multivariate/joint, etc.
2. Visualization in lower dimensions with preserving distance, topology, factors, variance
3. Visualization by graphs (with weighted edges), by trees (with hierarchy), by partitions (clusters)
4. Statistical inference of (in)dependency models for each condition and intervention
5. Inference of causal model

[a] Instead of the specific GE preprocessing, here we focus on the general issues of data analysis

Bayesian data analysis

Bayes statistical framework for statistical and causal data analysis, because biomedical data (analysis) is

1. high-dimensional $\Rightarrow$ relative scarcity of data w.r.t. model complexity,
2. noisy $\Rightarrow$ robustness by model averaging,
3. unexplored $\Rightarrow$ large model classes without overfitting,
4. embedded in rich qualitative knowledge $\Rightarrow$ priors,
5. multilevel/structured, distributed $\Rightarrow$ normative fusion,
6. for biologists $\Rightarrow$ direct probabilistic statements
7. for clinicians $\Rightarrow$ own subjective priors, credibility, Bayes factor

Distances, associations, similarities I.

Dot/scalar product of vectors $\mathbf{a}, \mathbf{b}$ is

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

It is the standard inner product for Euclidean space, denoted by $\langle \mathbf{a}, \mathbf{a} \rangle$). The norm $\|\mathbf{a}\|$ in such space is defined as $\sqrt{\langle \mathbf{a}, \mathbf{a} \rangle}$. It defines the Euclidean metric (distance) $d(\mathbf{a}, \mathbf{b})$ as

$$d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\| = \left(\sum_{i=1}^{n} |a_i - b_i|^p \right)^{\frac{1}{p}}$$

with $p = 2$.
Using the geometric interpretation a "similarity" $s(\mathbf{a}, \mathbf{b})$ called cosine metric can be defined also as

$$s(\mathbf{a}, \mathbf{b}) = \cos\theta = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \, \|\mathbf{b}\|}$$

Distances, associations, similarities II.

The correlation of random variables X, Y is defined as

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{var(X)}\sqrt{var(X)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\ \sqrt{E(Y^2) - E^2(Y)}}.$$

Its estimator from sample is the Pearson correlation coefficient

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \text{ where } s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

which is equivalent to the cosine metric if the data is mean-centered.
Note that uncorrelated variables can be dependent in general (but e.g. not in normal distribution).
The (passive observational) dependency of (discrete) random variables X, Y is better represented by the mutual information

$$MI(X,Y) = KL(p(X,Y||p(X)p(Y))) = \sum_{x,y} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$$

Note that dependency, or specifically "correlation does not imply causation".

Distances, associations, similarities III.

The Spearman's rank correlation coefficient can test any (not just linear) monotonic relation of scores X, Y using their generated rankings of N objects r_i^X, r_i^Y. It is defined as

$$\rho = 1 - \frac{6 \sum_i (r_i^X - r_i^Y)^2}{N(N^2 - 1)}$$

Distance preservation mappings

Multidimensional scaling: hidden dimensions

Idea: map data points $x_i \in \mathcal{R}^n$ to $y_i \in \mathcal{R}^m$ $m < n$ preserving the original distances $d(x_i, x_j = d_{ij})$, as much as possible. E.g. by minimizing this distortion/error function (Sammon's mapping)

$$E = \frac{1}{\sum_i \sum_{j>i} d_{ij}} \sum_i \sum_{j>i} \frac{(d_{ij} - \|y_i - y_j\|)^2}{d_{ij}};$$

(A symmetric, positive matrix D with 0 diagonal representing distances of n objects can be exactly represented in $\mathcal{R}^m$, iff HAH is positive semidefinite with rank less than m, where $H = \ldots, A = \ldots$. The exact solution can be found as ...)

Topology preservation mappings

Idea: crumpled thread, paper in 3D

Self-Organizing Maps (SOMs): The SOM defines a mapping from high dimensional input data space onto a regular two-dimensional array of neurons. Every neuron i of the map is associated with an n-dimensional reference vector, where n denotes the dimension of the input vectors. The reference vectors together form a codebook. The neurons of the map are connected to adjacent neurons by a neighbourhood relation, which dictates the topology, or the structure, of the map. Adjacent neurons belong to the neighbourhood Ni of the neuron i. The number of neurons determines the granularity of the mapping.

In the learning process of the SOM sample vectors are randomly drawn from the input data set and the closest codebook vector and its neighbours are drifted towards it (by using similarity or distance, e.g. the common Euclidean distance measure).

Standard application: mapping high dimensional irregular data to a low dimensional regular grid

A reversed application (eg finding layout for graphs): mapping an irregular topology to regular data

Factor/variance preservation mappings

Factor analysis is a statistical technique used to explain variability among observed random variables in terms of fewer unobserved random variables called factors. The observed variables are modeled as linear combinations of the factors, plus "error" terms.
Principal components analysis (PCA) is a technique for simplifying a dataset, by reducing multidimensional datasets to lower dimensions for analysis.PCA is a linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA has the distinction of being the optimal linear transformation for keeping the subspace that has largest variance.
Assuming that the empirical mean of the distribution has been subtracted from the data set X, the Karhunen-Loève transform is defined by the eigenvectors $\mathbf{W}$ of the matrix of observed covariances

$$\mathbf{C} = \mathbf{X}\mathbf{X}^T = \mathbf{W}\mathbf{\Sigma}^2\mathbf{W}^T,$$

where the matrix $\mathbf{\Sigma}$ is n-by-n with the nonnegative eigenvalues on the diagonal and zeros off the diagonal.

Clustering

Dividing objects into hierarchical vs non-hierarchical, non-overlapping sets by minimizing an error function expressing coherence and distance preservation.

1. Types: partitioning (parametric, non-parametric), hierarchical
   (a) Parametric partitioning (density-based) clustering
   (b) Non-parametric partitioning clustering
   (c) Hierarchical clustering
2. Evaluation (external, internal) and fusion (early, mid, late)
3. Extensions: asymmetric ”distance", sample rejection, overlapping clusters, bi-clusters, proximity-based,etc

Parametric partitioning clustering

Parametric (density-based) clustering (unsuperwised learning).
Assume a Naive Bayesian network model with

1. discrete random variable Y as root, features $X = \{X_1, \ldots, X_N\}$ X_1,
2. conditional distribution and prior from a given parametric family (i.e. for the number of values of Y!)
3. incomplete observation $D_n^X = \{X^{(1)}, \ldots, X^{(n)}, \}$ (i.e. with unknown/hidden/missing label Y).

Goal: reconstruction of the missing labels $D_n^Y = \{Y^{(1)}, \ldots, Y^{(n)}, \}$
Solution: posterior for the labels by Bayesian model averaging (AutoClass), MAP model reconstruction => labeling
Solution': MAP model reconstruction and Bayes decision for labeling
Solution" (unsuperwised learning): missing data management: imputation, iterative imputation, k-means, ML parameter estimation with EM or MCMC with Gibbs sampling
Note the possibility of a generative/causal interpretation.
Drawbacks: parametric, number of clusters

”Non-parametric" partitioning clustering

The K-means algorithm clusters objects into k partitions by minimizing the total intra-cluster variance, or, the squared error function.

$$E = \sum_{i=1}^{K} \sum_{j \in S_i} |x_j - \mu_i|^2$$

where μ_i are the means for the clusters. It is a variant of the EM algorithm in which the goal is to determine the k means of data generated from gaussian distributions. It

Require: feature data set D_n^X, number of clusters K
Ensure: K cluster means
Ini: random or prior based selection of means μ_k
repeat
Label feature samples by the index of the closest cluster
Reestimate means
until NoChange, or NoImprovement(E_{t+1},E_t,t)

Advantages: ”satisficing" in practice (fast&good enough)
Drawbacks: number of clusters, local optimum, superpolynomial time - $2^{\Omega(\sqrt{n})}$ - to converge in the worst case,
Improvement: k-medoids (using samples as cluster centers instead of means)

Criterion functions for clustering I.

The sum-of-squared-error criterion

$$E = \sum_{i=1}^{|C|} E_i \text{ where } E_i = \sum_{x \in C_i} \|x - \mu_i\|^2 \text{ and } \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

because of the Euclidean distance over $x \in R^d$ (using $x^{(i)}, x^{(j)}, x^{(k)} \in C_i$)

$$E_i = \sum_{x^{(k)} \in C_i} \sum_{l=1}^{d} [x_l^{(k)} - \frac{1}{|C_i|} \sum_{x^{(j)} \in C_i} x_l^{(j)}]^2 = \frac{1}{|C_i^2|} \sum_{x^{(k)} \in C_i} \sum_{l=1}^{d} [\sum_{x^{(j)} \in C_i} (x_l^{(k)} - x_l^{(j)})]^2$$

$$= \frac{1}{|C_i^2|} \sum_{x^{(k)}} \sum_{l=1}^{d} \sum_{\substack{x^{(j)} \\ j \neq k}} (x_l^{(k)} - x_l^{(j)})^2 + \frac{1}{|C_i^2|} \sum_{x^{(k)}} \sum_{l=1}^{d} \sum_{\substack{x^{(j)} \\ j \neq k}} \sum_{\substack{x^{(i)} \\ i \neq j,k}} (x_l^{(k)} - x_l^{(j)})(x_l^{(k)} - x_l^{(i)})$$

for each l, $|C_i^2| E_i = 2 \sum_{m \neq n} ({x^{(m)}}^2 - x^{(m)} x^{(n)}) + (|C_i| - 2) \sum_{m \neq n} ({x^{(m)}}^2 - x^{(m)} x^{(n)})$, i.e. we can express E_i using only the first term. By noting that this term is $\|x^{(k)} - x^{(j)}\|^2$, we can rewrite E as

Criterion functions for clustering II.

$$E = \sum_{i=1}^{|C|} \frac{1}{2|C_i|} \sum_{x^{(k)}, x^{(j)} \in C_i} \|x^{(k)} - x^{(j)}\|^2$$

(As can be expected as the first expression measures the spread of the cluster linearly in $|C_i|$s and counting each sample once, whereas the second is quadratic in $|C_i|$s and counts each pair twice.)

This suggests the generalization of the error for non-metric $s(x, x')$ as

$$E = \sum_{i=1}^{|C|} \frac{1}{2|C_i|} \sum_{x^{(k)}, x^{(j)} \in C_i} s(x^{(k)}, x^{(j)})$$

Criterion functions for clustering III.

This pairwise formulation of the error function allows the separation of the inter/within- and between/intracluster terms for a given clustering

$$\sum_{x^{(k)},x^{(j)}} s(x^{(k)},x^{(j)}) = \overbrace{\sum_{k,j} s(x^{(k)},x^{(j)})^{1(x^{(k)}\in C_k, x^{(j)}\in C_j, j\neq k)}}^{\text{intracluster}} + \overbrace{\sum_{k,j} s(x^{(k)},x^{(j)})^{1(x^{(k)},x^{(j)}\in C_j)}}^{\text{intercluster}}$$

These terms can be expressed also using the means as the former E and the $\sum_{i=1}^{|C|} \|\mu - \mu_i\|^2$, where μ is the overall mean and their complementarity can be proved (i.e. analogue roles of the minimization of the within-cluster and the maximization of the within-cluster terms).

Hierarchical clustering

Problem: $|C| =$?

Idea: construct clusterings of n samples for $|C| = 1, \ldots, n$ (called at level $n-1, \ldots, 0$)

Refinement: require embedded clusterings (i.e. $x, x' \in C_i$ at level k, then $x, x' \in C_j$ at level $k > k'$) $\Rightarrow$ hierarchical clustering (HC).

A representation for a HC is the HC tree or dendrogram, which is a binary tree with internal nodes at the height of the similarity of their joined clusters i,j.

FIGURE

Definitions of similarity/distance of pair of clusters:

$$\boldsymbol{d}_{min}(C_i, C_j) = \min_{\substack{x \in C_i \\ x' \in C_j}} \|x - x'\|, \quad \boldsymbol{d}_{max}(C_i, C_j) = \max_{\substack{x \in C_i \\ x' \in C_j}} \|x - x'\|$$

$$\boldsymbol{d}_{avg}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{x' \in C_j} \|x - x'\|, \quad \boldsymbol{d}_{mean}(C_i, C_j) = \|\mu_i - \mu_j\|$$

Hierarchical clustering: methods

Methods: agglomerative/divisive

Require: pairwise distances
Ensure: topology and labelling
Ini: Define a cluster C_i and leaf at height 0 in tree T for each sequence i
for i=1 to L-1 **do**
Select pair of clusters i,j with minimal d_{ij}
Define new cluster $C_k = C_i \cup C_j$
Define a new node in T to be the parent of i and j at height d_{ij}
Remove C_i, C_j from set of clusters and insert C_k
End: Insert root for the final two clusters i,j at height d_{ij}

Optimal cluster number: analyzing the decrease of similarity by the merge
The usage of $\boldsymbol{d}_{min}$ and the insertion of an edge between nodes/samples i,j in case of their merge in the AC method, gives a minimum length (weight) spanning tree. It is called the single-linkage algorithm, if it is stopped before the distance exceeds a threshold.
The usage of $\boldsymbol{d}_{max}$ and the insertion of all the edges between samples in the merged clusters in the AC method, gives cliques for the clusters. It is called the complete-linkage algorithm, if it is stopped before the distance exceeds a threshold.
Note that $d'(x, x') = \min_l x, x' \in C_i^l$ (the minimum level where they are co-members) induces a distance from a non-metric $d(x, x')$, which is also ultrametric.

Evaluation: internal measures

Evaluation may use internal or external measures (i.e. with/without references,gold standard).
Silhouette coefficient (external) characterize the ratio between cluster coherence (intracluster distance) and cluster separation (intercluster distance). The Silhouette value for each element i in cluster k is

$$sc_{ik} = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

where $a(i)$ is the average dissimilarity of member i to all other members of its cluster and b(i) the dissimilarity of member i to the nearest member of the nearest cluster. With n_k the size of cluster k and n the total number of objects, the Silhouette coefficient per cluster SC_k and the overall Silhouette coefficient SC (with values between -1 and 1) are defined as

$$SC_k = \frac{1}{n_k} \sum_{i=1}^{n_k} sc_{ik}, SC = \frac{1}{n} \sum_{k} SC_k$$

Evaluation: external measures

Given a data D_n and their clustering $\mathcal{C}$, let define an nxn adjacency matrix M as $M_ij = 1(\exists C_k \in \mathcal{C} : x^{(i)}, x^{(j)} \in C_k)$. For two clustering $calC, calC'$ let N_{00},N_{10},N_{01},N_1 denote the respective counts of elementary pairs in the induced M, M'.
The Jaccard coefficient is defined as

$$J(\mathcal{C}, \mathcal{C}') = \frac{N_{11}}{N_{11} + N_{10} + N_{01}},$$

which ignores the false negative errors. The more balanced rand index is given as

$$R(\mathcal{C}, \mathcal{C}') = \frac{N_{11} + N_{00}}{N_{11} + N_{10} + N_{01} + N_{00}}$$

Fusion

Phases for fusion from multiple sources: early, mid, late

1. Early: The combination of the raw data (i.e. distance/dissimilarity function)
2. Mid: The combination of the distance/dissimilarity matrices
3. Late: The combination of the clustering tree

Extensions

1. asymmetric ”distance"
2. sample rejection
3. overlapping clusters
4. bi-clusters,
5. proximity-based