Optimization and Data Mining in Biomedicine

Panos Pardalos

Center for Applied Optimization
Department of Industrial & Systems Engineering
Computer & Information Science & Engineering Department
Biomedical Engineering Program, McKnight Brain Institute
University of Florida
http://www.ise.ufl.edu/pardalos/

Mathematical Programming and Biomedicine

- In recent years, there has been a steady growth of interest in applications of optimization in biological and medical sciences.
- In many areas of biomedicine, optimization has become an indispensable tool (e.g. X-ray crystallograpgy, protein folding, etc.)
- Optimization is also frequently used for designing and modeling complex systems, which are abundant in biology.

Mathematical Programming and Biomedicine

- The types of optimization models utilized in biomedicine comprise a broad range of areas of mathematical programming, including linear programming, quadratic programming, general nonlinear programming, discrete optimization, etc.
- In particular, multi-quadratic 0–1 programming finds an important application in epilepsy research and fractional 0–1 programming is used to formulate the consistent biclustering problem in data mining.

Bibliography

- D.Z. Du, P.M. Pardalos and J. Wang (Eds.), Discrete Mathematical Problems with Medical Applications, DIMACS Series Vol. 55, AMS, 2000.
- C. Floudas and P.M. Pardalos (Eds.), Optimization in Computational Chemistry and Molecular Biology, Kluwer Academic Publishers, 2000.
- P.M. Pardalos and J. Principe (Eds.), Biocomputing, Kluwer Academic Publishers, 2002.

Bibliography

- P.M. Pardalos, J.C. Sackellares, P.R. Carney and L.D. Iasemidis (Eds.), Quantitative Neuroscience, Kluwer Academic Publishers, 2004.
- P.M. Pardalos, V.L. Boginski, A. Vazacopoulos (Eds.), Data Mining in Biomedicine, Series: Optimization and Its Applications, Vol. 7, Springer, 2007.

Epilepsy

- A symptom of a brain disorder distinguished by recurring seizures.
- Can begin at any age.
- Affects 1% of the population. World Health Organization estimates 50 million cases worldwide.
- Quality of Life: Affects self-esteem, career, social opportunities, restricted driving privileges.

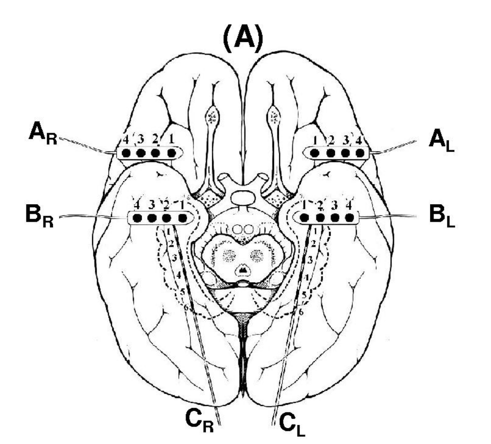
Epileptic Seizure Prediction

- Epilepsy consists of more than 40 clinical syndromes affecting 50 million people worldwide. At least 30% of patients with epilepsy continue to have seizures despite treatment with antiepileptic drugs.
- Epileptic seizure occurrences seem to be random and unpredictable. However, recent studies in epileptic patients suggest that seizures are deterministic rather than random. Subsequently, studies of the spatiotemporal dynamics in electroencephalograms (EEG's), from patients with temporal lobe epilepsy, demonstrated a preictal transition of approximately $\frac{1}{2}$ to 1 hour duration before the ictal onset

Epileptic Seizure Prediction

- The enormous number of neurons and dynamic nature of connections between them makes the analysis of brain function especially challenging.
- In order to perform a quantitative analysis of brain, one can treat certain groups of neurons (functional units of the brain) as vertices of a graph and investigate the connections between these functional units.

Epileptic Seizure Prediction



Lyapunov Exponents: what is it and why?

- Important measure that characterizes chaotic behavior of nonlinear system.
- Global Lyapunov Exponent: how fast nearby orbits of the system converge or diverge in infinitely large time interval.
- Local Lyapunov exponent characterize local predictability around a point x_0 in phase space
- Lyapunov Exponent has proven its efficiency in EEG analysis for predicting epileptic seizures

Formal Definition

Let a system be set by

$$\dot{X}(t) = F(X), \text{ where } X : \mathbb{R} \mapsto \mathbb{R}^n,\ \ F : \mathbb{R}^n \mapsto \mathbb{R}^n$$

The maximal Lyapunov Exponent λ can be defined as follows:

$$\lambda = \lim_{t \to \infty} \lim_{\delta X(0) \to 0} \frac{1}{t} \log_2 \frac{\delta X(t)}{\delta X(0)}$$

For short term maximal Lyapunov Exponent (STLmax) we can take "reasonable" t instead of external limit.

Estimation from Time Series

- In real life we often deal with one dimensional time series of noisy data (such as EEG signal) instead of explicit system of equations
- Wolf *et al* suggested algorithm for Lyapunov Exponent calculation from time series
  - A Wolf, J B Swift, H L Swinney, J A Vastano, Determining Lyapunov Exponenets from a Time Series, Phisica 16D (1985), pp. 285 - 317.
- We used Sackellares *et al* modification of Wolf's algorithm for STLmax calculation that handles noisy non-stationary data
  - L D Iasemidis, J C Principe, J C Sackellares, Measurement and Quantification of Spatiotemporal Dynamics of Human Epilepic Seizures. Nonlinear Signal Processing in Medicine, Ed. M. Akay, IEEE Press, 1999

Approach to Estimation

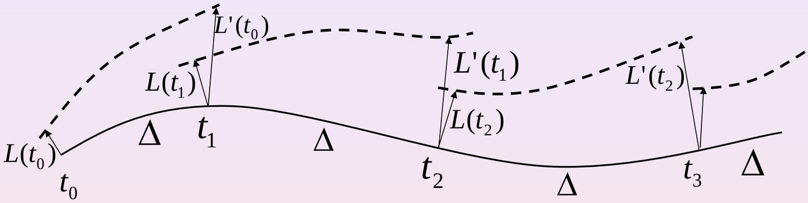


Figure: Evolution in phase space and replacement procedure used to estimate Lyapunov Exponents from experimental data

$$STL_{\max} = \frac{1}{t_M - t_0} \sum_{k=1}^{M} \log_2 \frac{L'(t_k)}{L(t_{k-1})}.$$

Short Term Largest Lyapunov Exponents

- Since **the brain is a nonstationary system**, algorithms used to estimate measures of the brain dynamics should be capable of automatically identifying and appropriately weighing existing transients in the data. In a chaotic system, orbits originating from similar initial conditions (nearby points in the state space) diverge exponentially (expansion process). The **rate of divergence** is an important aspect of the system dynamics and is reflected in the value of **Lyapunov exponents**.

Spatiotemporal Dynamical Analysis

- During the last decade, the advances in studying brain are associated with the extensive use of *electroencephalograms* (EEG) which can be treated as the quantitative representation of the brain function.
- EEG data essentially represent time series recorded from the electrodes located in different functional units of brain. We utilize the concept of *T-index* to measure the entrainment of two brain sites at a time moment.

Spatiotemporal Dynamical Analysis

- The T-index at time t between electrode sites i and j is defined as:

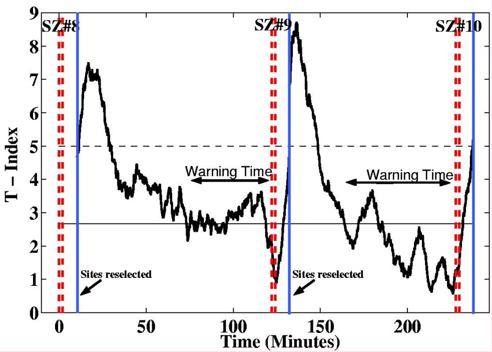
  $$T_{i,j}(t) = \sqrt{N} \times |E\{STL_{max,i} - STL_{max,j}\}|/\sigma_{i,j}(t)$$

  where $E\{\cdot\}$ is the sample average difference for the $STL_{max,i} - STL_{max,j}$ estimated over a moving window $w_t(\lambda)$.
- At the moment of a seizure some brain sites exhibit the convergence of their EEG signals, which is characterized by the drop of the corresponding T-index below $T_{critical}$.

Spatiotemporal Dynamical Analysis

- A natural graph representing the brain: each vertex corresponds to a functional unit/electrode, and there is a edge between two of them if T-index is below $T_{critical}$.
- The number of edges in this graph dramatically increases at seizure points, and it decreases immediately after seizures.

Epileptic Seizure Prediction



Epileptic Seizure Prediction

- One aspect of the analysis of the epileptic brain is finding a maximum clique in this graph. It provides us with the largest set of "critical" elecrode sites most entrained during the seizure.
- If the number of critical sites is set equal to k, we can formulate the problem of selecting the optimal group of critical site as a multi-quadratic 0–1 programming.

Electrode Selection Problem

- Let $x_i \in \{0,1\}$ denote if site i is selected:

$$\begin{array}{lll} \min & x^T A x & \\ \text{s.t.} & \sum_{i=1}^n x_i & = k \\ & x^T B x & \geq T_{critical} k(k-1) \\ & x \in \{0,1\}^n & \end{array}$$

- a_{ij} is the T-index between sites i and j during the seizure point.
- b_{ij} is the T-index between sites i and j 10 min after the onset of seizure.

Multi-Quadratic 0–1 Programming

$$\begin{aligned}
\min_{x\in\{0,1\}^n} \quad & f(x) = x^T A x, \\
\text{s.t.} \quad & \qquad Bx \qquad\quad \geq b, \\
& f_1(x) = x^T Q^1 x \quad \geq \alpha_1, \\
& f_2(x) = x^T Q^2 x \quad \geq \alpha_2, \\
& \qquad \dots \\
& f_k(x) = x^T Q^k x \quad \geq \alpha_k.
\end{aligned}$$

Applications

- Theoretical physics (spin glass models)
- Graph and network problems
- Engineering applications
- Medical applications
- Finance and economics
- Chemical applications

Conventional Linearization

- For each product $x_i x_j$, introduce a new variable $x_{ij} = x_i x_j$ (notice that $x_{ii} = x_i^2 = x_i$ for $x_i \in \{0, 1\}$.)
- The relation between new and old variables is defined by

$$x_{ij} \le x_i,$$

$$x_{ij} \le x_j,$$

$$x_{ij} \ge x_i + x_j - 1.$$

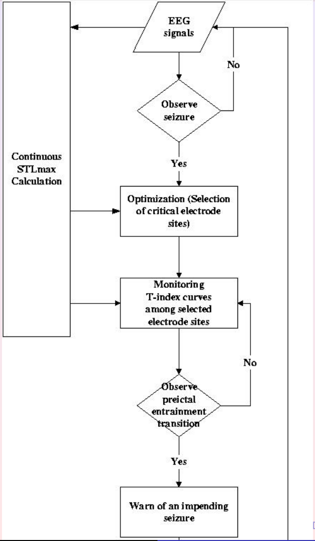
- The number of variables increases as $O(n^2)$. This is very inefficient from the computational viewpoint.

Improved Linearization

$$
\begin{aligned}
\min \quad & e^T s & & \\
\text{s.t.} \quad & Qx - y - s & &= 0, && (1)\\
& \textstyle\sum_{i=1}^{n} x_i & &= k, && (2)\\
& y & &\le M(1-x), && (3)\\
& Bx - z & &\ge 0, && (4)\\
& e^T z & &\ge T_\alpha k(k-1), && (5)\\
& z & &\le M' x, && (6)\\
& s, y, z & &\ge 0,\ x \in \{0,1\}^n, && (7)
\end{aligned}
$$

where $M = \max_i \sum_{j=1}^{n} q_{ij}$ and $M' = \max_i \sum_{j=1}^{n} b_{ij}$.

Seizure Warning Algorithm


EEG signals
Observe seizure
No
Yes
Continuous STLmax Calculation
Optimization (Selection of critical electrode sites)
Monitoring T-index curves among selected electrode sites
No
Observe preictal entrainment transition
Yes
Warn of an impending seizure

Epileptic Seizure Prediction

- P.M. Pardalos, W. Chaovalitwongse, L.D. Iasemidis, J.C. Sackellares, D.-S. Shiau, P.R. Carney, O.A. Prokopyev, V.A. Yatsenko, Seizure Warning Algorithm Based on Optimization and Nonlinear Dynamics, Mathematical Programming, Vol. 101/2 (2004), pp. 365–385.

- W. Chaovalitwongse, O.A. Prokopyev, P.M. Pardalos, Electroencephalogram (EEG) Time Series Classification: Applications in Epilepsy, Annals of OR, Vol. 148/1 (2006), pp. 227–250.

Massive Datasets

- The proliferation of **massive datasets** brings with it a series of special computational challenges. This **data avalanche** arises in a wide range of scientific and commercial applications.
- In particular, microarray technology allows one to grasp simultaneously thousands of gene expressions throughout the entire genome. To extract useful information from such datasets a sophisticated data mining algorithm is required.

Massive Datasets

Abello, J.; Pardalos, P.M.; Resende, M.G. (Eds.), Handbook of Massive Data Sets, Series: Massive Computing, Vol. 4, Kluwer, 2002.

Major Data Mining Problems

- **Clustering (Unsupervised):** Given a set of samples partition them into groups of similar samples according to some similarity criteria.
- **Classification (Supervised Clustering):** Determine classes of the test samples using known classification of training data set.
- **Feature Selection:** For each of the classes, select a subset of features responsible for creating the condition corresponding to the class (it's also a specific type of **dimensionality reduction**).
- **Outlier Detection:** Some of the samples are not good representative of any of the classes. Therefore, it is better to disregard them while preforming data mining.

Major challenges in Data Mining

- Typical noisiness of data arising in many data mining applications complicates solution of data mining problems.
- High-dimensionality of data makes complete search in most of data mining problems computationally infeasible.
- Some data values may be inaccurate or missing.
- The available data may be not sufficient to obtain statistically significant conclusions.

Biclustering

- Biclustering is a methodology allowing for feature set and test set clustering (supervised or unsupervised) simultaneously.
- It finds clusters of samples possessing similar characteristics together with features creating these similarities.
- The required consistency of sample and feature classification gives biclustering an advantage over other methodologies treating samples and features of a dataset separately of each other.

Biclustering

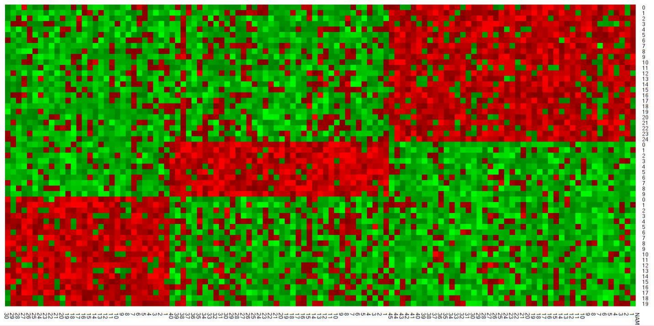


Figure: Partitioning of samples and features into 3 classes.

Survey on Biclustering Methodologies

- **“Direct Clustering” (Hartigan)**
- The algorithm begins with the entire data as a single block and then iteratively finds the row and column split of every block into two pieces. The splits are made so that the total variance in the blocks is minimized.
- The whole partitioning procedure can be represented in a hierarchical manner by trees.
- Drawback: this method does NOT optimize a global objective function.

Survey on Biclustering Methodologies

- **Cheng & Church's algorithm**
- The algorithm constructs one bicluster at a time using a statistical criterion – a low mean squared resedue (the variance of the set of all elements in the bicluster, plus the mean row variance and the mean column variance).
- Once a bicluster is created, its entries are replaced by random numbers, and the procedure is repeated iteratively.

Survey on Biclustering Methodologies

- **Graph Bipartitioning**
- Define a bipartite graph $G(F, S, E)$, where F is the set of data set features, S is the set of data set samples, and E are weighted edges such that the weight $E_{ij} = a_{ij}$ for the edge connecting $i \in F$ with $j \in S$. The biclustering corresponds to partitioning of the graph into bicliques.

Survey on Biclustering Methodologies

- Given vertex subsets V_1 and V_2, define

$$cut(V_1, V_2) = \sum_{i \in V_1} \sum_{j \in V_2} a_{ij}$$

and for k vertex subsets $V_1, V_2, \ldots, V_k$,

$$cut(V_1, V_2, \ldots, V_k) = \sum_{i<j} cut(V_i, V_j)$$

Survey on Biclustering Methodologies

- Biclustering may be performed as

$$\min_{V_1, V_2, \ldots, V_k} cut(V_1, V_2, \ldots, V_k),$$

  on G or with some modification of the definition of *cut* to favor balanced clusters.
- This problem is *NP*-hard, but spectral heuristics show good performance [**Dhillon**]

Biclustering: Applications

- **Biological and Medical:**
    - Microarray data analysis
    - Analysis of drug activity, Liu and Wang (2003)
    - Analysis of nutritional data, Lazzeroni et al. (2000)

Biclustering: Applications

- **Text Mining:** Dhillon (2001, 2003)
- **Marketing:** Gaul and Schader (1996)
- **Dimensionality Reduction in Databases:** Agrawal et al. (1998)
- **Others:**
  - electoral data - Hartigan (1972)
  - currency exchange - Lazzeroni et al. (2000)

Biclustering: Surveys

- S. Madeira, A.L. Oliveira, Biclustering Algorithms for Biological Data Analysis: A Survey, 2004.
- A. Tanay, R. Sharan, R. Shamir, Biclustering Algorithms: A Survey, 2004.
- S. Busygin, O.A. Prokopyev, and P.M. Pardalos, Biclustering in Data Mining, to appear in C&OR, 2007.

Data Representation

- A dataset (e.g., from microarray experiments) is normally given as a rectangular $m \times n$ matrix A, where each column represents a data sample (e.g., patient) and each row represents a feature (e.g., gene):

  $$A = (a_{ij})_{m \times n},$$

  where the value a_{ij} is the expression of i-th feature in j-th sample.

Definitions

- Data set of n samples and m features is a matrix

  $$A = (a_{ij})_{m \times n},$$

  where the value a_{ij} is the expression of i-th feature in j-th sample.

- We consider classification of the samples into classes

  $$\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_r,\ \mathcal{S}_k \subseteq \{1 \ldots n\},\ k = 1 \ldots r,$$

  $$\mathcal{S}_1 \cup \mathcal{S}_2 \cup \ldots \cup \mathcal{S}_r = \{1 \ldots n\},$$

  $$\mathcal{S}_k \cap \mathcal{S}_\ell = \emptyset,\ k, \ell = 1 \ldots r,\ k \neq \ell.$$

Definitions

- This classification should be done so that samples from the same class share certain common properties. Correpondingly, a feature i may be assigned to one of the feature classes

$$\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_r,\ \mathcal{F}_k \subseteq \{1 \ldots m\},\ k = 1 \ldots r,$$

$$\mathcal{F}_1 \cup \mathcal{F}_2 \cup \ldots \cup \mathcal{F}_r = \{1 \ldots m\},$$

$$\mathcal{F}_k \cap \mathcal{F}_\ell = \emptyset,\ k, \ell = 1 \ldots r,\ k \neq \ell,$$

in such a way that features of the class $\mathcal{F}_k$ are **“responsible”** for creating the class of samples $\mathcal{S}_k$.

Definitions

- This may mean for microarray data, for example, strong up-regulation of certain genes under a cancer condition of a particular type (whose samples constitute one class of the data set). Such a simultaneous classification of samples and features is called **biclustering** (or **co-clustering**).

Definitions

Definition

A *biclustering* of a data set is a collection of pairs of sample and feature subsets

$$\mathcal{B} = ((\mathcal{S}_1, \mathcal{F}_1), (\mathcal{S}_2, \mathcal{F}_2), \ldots, (\mathcal{S}_r, \mathcal{F}_r))$$

such that the collection $(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_r)$ forms a partition of the set of samples, and the collection $(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_r)$ forms a partition of the set of features.

Our Approach: Intuition

- Let us distribute features among the classes of training set such that each feature belongs to the class where its average expression among the training samples is highest.

$$\left(\mathcal{S}_1^0, \mathcal{S}_2^0, ..., \mathcal{S}_r^0\right) \rightarrow \left(\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_r\right)$$

- Now, if we transpose the matrix, take the feature classification as given, and re-classify the training samples according to highest average expression values in feature classes...

$$\left(\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_r\right) \rightarrow \left(\mathcal{S}_1^1, \mathcal{S}_2^1, ..., \mathcal{S}_r^1\right)$$

Intuition Behind Biclustering

- Will we obtain the same training set classification?

$$? \ \left(\mathcal{S}_1^0, \mathcal{S}_2^0, \ldots, \mathcal{S}_r^0\right) = \left(\mathcal{S}_1^1, \mathcal{S}_2^1, \ldots, \mathcal{S}_r^1\right) \ ?$$

- If yes, we will say that we obtained a **consistent biclustering**.

Consistent Biclustering

- Let each sample be already assigned somehow to one of the classes $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_r$. Introduce a 0–1 matrix $S = (s_{jk})_{n \times r}$ such that $s_{jk} = 1$ if $j \in \mathcal{S}_k$, and $s_{jk} = 0$ otherwise.
- The sample class *centroids* can be computed as the matrix $C = (c_{ik})_{m \times r}$:

$$C = AS(S^T S)^{-1}, \quad \left(c_{ik} = \frac{\sum_{j \in \mathcal{S}_k} a_{ij}}{|\mathcal{S}_k|} \right)$$

whose k-th column represents the centroid of the class $\mathcal{S}_k$.

Consistent Biclustering

- Consider a row i of the matrix C. Each value in it gives us the average expression of the i-th feature in one of the sample classes. As we want to identify the checkerboard pattern in the data, we have to assign the feature to the class where it is most expressed. So, let us classify the i-th feature to the class $\hat{k}$ with the maximal value $c_{i\hat{k}}$:

$$i \in \mathcal{F}_{\hat{k}} \;\Rightarrow\; \forall k = 1 \ldots r,\; k \neq \hat{k} :\; c_{i\hat{k}} > c_{ik}$$

Consistent Biclustering

- Using the classification of all features into classes $\mathcal{F}_1$, $\mathcal{F}_2$, ..., $\mathcal{F}_r$, let us construct a classification of samples using the same principle of maximal average expression. We construct a 0–1 matrix $F = (f_{ik})_{m\times r}$ such that $f_{ik} = 1$ if $i \in \mathcal{F}_k$ and $f_{ik} = 0$ otherwise. Then, the feature class centroids can be computed in form of matrix $D = (d_{jk})_{n\times r}$:

$$D = A^T F(F^T F)^{-1}, \quad \left(d_{jk} = \frac{\sum_{i\in\mathcal{F}_k} a_{ij}}{|\mathcal{F}_k|} \right)$$

whose k-th column represents the centroid of the class $\mathcal{F}_k$.

Consistent Biclustering

- The condition on sample classification we need to verify is

$$j \in \mathcal{S}_{\hat{k}} \;\Rightarrow\; \forall k = 1 \ldots r,\; k \neq \hat{k} :\; d_{j\hat{k}} > d_{jk}$$

Consistent Biclustering

Definition

A biclustering $\mathcal{B}$ will be called **consistent** if the following relations hold:

$$i \in \mathcal{F}_{\hat{k}} \;\Rightarrow\; \forall k = 1 \ldots r,\; k \neq \hat{k}:\; c_{i\hat{k}} > c_{ik}$$

$$j \in \mathcal{S}_{\hat{k}} \;\Rightarrow\; \forall k = 1 \ldots r,\; k \neq \hat{k}:\; d_{j\hat{k}} > d_{jk}$$

Consistent Biclustering

Definition

A data set is **biclustering-admitting** if some consistent biclustering for it exists.

Definition

The data set will be called **conditionally biclustering-admitting** with respect to a given (partial) classification of some samples and/or features if there exists a consistent biclustering preserving the given (partial) classification.

Consistent Biclustering

- **A consistent biclustering implies separability of the classes by convex cones.**

Theorem

Let $\mathcal{B}$ be a consistent biclustering. Then there exist convex cones $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_r \subseteq \mathbb{R}^m$ such that all samples from $\mathcal{S}_k$ belong to the cone $\mathcal{P}_k$ and no other sample belongs to it, $k = 1 \ldots r$. Similarly, there exist convex cones $\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_r \subseteq \mathbb{R}^n$ such that all features from $\mathcal{F}_k$ belong to the cone $\mathcal{Q}_k$ and no other feature belongs to it, $k = 1 \ldots r$.

Separation by Cones

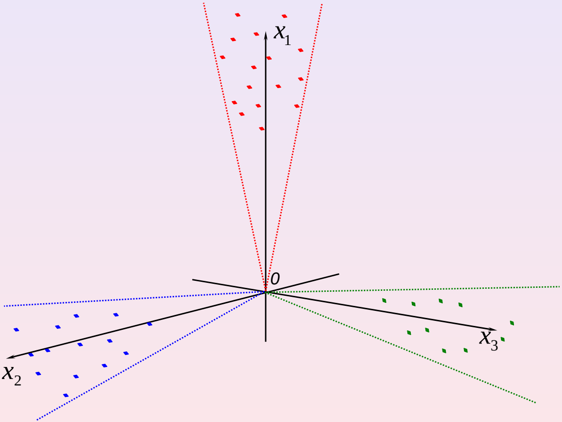


Figure: 3 classes are separated in 3D-space

Conic Separability

Proof

Let $\mathcal{P}_k$ be the conic hull of the samples of $\mathcal{S}_k$. Suppose $\hat{j} \in \mathcal{S}_\ell$, $\ell \neq k$, belongs to $\mathcal{P}_k$. Then

$$a_{\cdot\hat{j}} = \sum_{j \in \mathcal{S}_k} \gamma_j a_{\cdot\hat{j}},$$

where $\gamma_j \geq 0$. Biclustering consistency implies that $d_{\hat{j}\ell} > d_{\hat{j}k}$, that is

$$\frac{\sum_{i \in \mathcal{F}_\ell} a_{i\hat{j}}}{|\mathcal{F}_\ell|} > \frac{\sum_{i \in \mathcal{F}_k} a_{i\hat{j}}}{|\mathcal{F}_k|}$$

Conic Separability

Proof (cont'd)

Plugging the conic representation of $a_{i\hat{j}}$, we can obtain

$$\sum_{j \in \mathcal{S}_k} \gamma_j d_{j\ell} > \sum_{j \in \mathcal{S}_k} \gamma_j d_{jk},$$

that contradicts to $d_{j\ell} < d_{jk}$ (also implied by biclustering consistency).

Similarly, we can show that the formulated conic separability holds for feature classes.

α- and β-Consistent Biclustering

Definition

A biclustering $\mathcal{B}$ will be called α-**consistent** if the following relations hold:

$$i \in \mathcal{F}_{\hat{k}} \;\Rightarrow\; \forall k = 1 \dots r,\; k \neq \hat{k} : \; c_{i\hat{k}} > \alpha_i^F + c_{ik}$$

$$j \in \mathcal{S}_{\hat{k}} \;\Rightarrow\; \forall k = 1 \dots r,\; k \neq \hat{k} : \; d_{j\hat{k}} > \alpha_j^S + d_{jk}$$

where α is a vector of $\alpha_j^S \geq 0$ and $\alpha_i^F \geq 0$.

α- and β-Consistent Biclustering

Definition

A biclustering $\mathcal{B}$ will be called β-**consistent** if the following relations hold:

$$i \in \mathcal{F}_{\hat{k}} \ \Rightarrow \ \forall k = 1 \ldots r, \ k \neq \hat{k} : \ c_{i\hat{k}} > \beta_i^F c_{ik}$$

$$j \in \mathcal{S}_{\hat{k}} \ \Rightarrow \ \forall k = 1 \ldots r, \ k \neq \hat{k} : \ d_{j\hat{k}} > \beta_j^S d_{jk}$$

where β is a vector of $\beta_i^F \geq 1$ and $\beta_j^S \geq 1$.

α- and β-Consistent Biclustering

- If a biclustering $\mathcal{B}$ is α-consistent then it is consistent.

- If a biclustering $\mathcal{B}$ is β-consistent and $c_{ik} \geq 0$ and $d_{jk} \geq 0$, $\forall i, j, k$, then it is consistent.

- Both allow selecting the most representative subset of features and/or samples.

Supervised Biclustering

- One of the most important problems for real-life data mining applications is **supervised classification** of test samples on the basis of information provided by training data.
- A **supervised classification** method consists of two routines, first of which derives classification criteria while processing the training samples, and the second one applies these criteria to the test samples.

Supervised Biclustering

- In genomic and proteomic data analysis, as well as in other data mining applications, where only a small subset of features is expected to be relevant to the classification of interest, the classification criteria should involve dimensionality reduction and feature selection.

- We handle such a task utilizing the notion of consistent biclustering. Namely, we select a subset of features of the original data set in such a way that the obtained subset of data becomes conditionally biclustering-admitting with respect to the given classification of training samples.

Fractional 0–1 Programming Formulation

- Formally, let us introduce a vector of 0–1 variables $x = (x_i)_{i=1\dots m}$ and consider the i-th feature selected if $x_i = 1$.
- The condition of biclustering consistency, when only the selected features are used, becomes

$$\frac{\sum_{i=1}^{m} a_{ij} f_{i\hat{k}} x_i}{\sum_{i=1}^{m} f_{i\hat{k}} x_i} > \frac{\sum_{i=1}^{m} a_{ij} f_{ik} x_i}{\sum_{i=1}^{m} f_{ik} x_i}, \ \forall j \in \mathcal{S}_{\hat{k}}, \ \hat{k}, k = 1 \dots r, \ \hat{k} \neq k.$$

Fractional 0–1 Programming Formulation

- We will use the fractional relations as constraints of an optimization problem selecting the feature set. It may incorporate various objective functions over x, depending on the desirable properties of the selected features, but one general choice is **to select the maximal possible number of features in order to lose minimal amount of information provided by the training set**. In this case, the objective function is

$$\max \sum_{i=1}^{m} x_i$$

Fractional 0–1 Programming Formulation

- One of the possible fractional 0–1 formulations based on β-consistent biclustering criterion (suitable for microarrays):

$$\max_{\mathbf{x}\in\mathbb{B}^n} \sum_{i=1}^{m} x_i,$$

s.t.

$$\frac{\sum_{i=1}^{m} a_{ij} f_{i\hat{k}} x_i}{\sum_{i=1}^{m} f_{i\hat{k}} x_i} \geq \beta_j^S \frac{\sum_{i=1}^{m} a_{ij} f_{ik} x_i}{\sum_{i=1}^{m} f_{ik} x_i}, \ \forall j \in \mathcal{S}_{\hat{k}}, \ \hat{k}, k = 1 \ldots r, \ \hat{k} \neq k.$$

Fractional 0–1 Programming Formulation

- Generally, in the framework of fractional 0–1 programming we consider problems, where we optimize a multiple-ratio fractional 0–1 function subject to a set of linear constraints.

- We have **a new class** of fractional 0–1 programming problems, where fractional terms are not in the objective function, but in constraints, i.e. we optimize a linear objective function subject to fractional constraints.
- **How to solve fractionally constrained 0–1 programming problem**?

Linear Mixed 0–1 Formulation

- We can reduce our problem to a linear mixed 0–1 programming problem applying the approach similar to the one used to linearize problems with fractional 0–1 objective function.

T.-H. Wu, A note on a global approach for general 0–1 fractional programming, European J. Oper. Res. 101 (1997) 220–223.

Linear Mixed 0–1 Formulation

Theorem

A polynomial mixed 0–1 term $z = xy$, where x is a 0–1 variable, and y is a continuous variable, can be represented by the following linear inequalities:
(1) $z \le Ux$;
(2) $z \le y + L(x-1)$;
(3) $z \ge y + U(x-1)$;
(4) $z \ge Lx$,
where U and L are upper and lower bounds of variable y, i.e. $L \le y \le U$.

Linear Mixed 0–1 Formulation

- To linearize the fractional 0–1 program we need to introduce new variable y_k

$$y_k = \frac{1}{\sum_{\ell=1}^{m} f_{\ell k} x_\ell}, \;\; k = 1, \ldots, r.$$

Linear Mixed 0–1 Formulation

- In terms of the new variables fractional constraints are replaced by

$$\sum_{i=1}^{m} a_{ij} f_{i\hat{k}} x_i y_{\hat{k}} \geq \beta_j^S \sum_{i=1}^{m} a_{ij} f_{ik} x_i y_k$$

Linear Mixed 0–1 Formulation

- Next, observe that the term $x_i y_k$ is present if and only if $f_{ik} = 1$, i.e., $i \in \mathcal{F}_k$. So, there are totally only m of such products, and hence we can introduce m variables $z_i = x_i y_k$, $i \in \mathcal{F}_k$:

$$z_i = \frac{x_i}{\sum_{\ell=1}^{m} f_{\ell k} x_\ell}, \quad i \in \mathcal{F}_k.$$

Linear Mixed 0–1 Formulation

- In terms of z_i we have the following constraints:

$$\sum_{i=1}^{m} f_{ik} z_i = 1,\ k = 1 \dots r.$$

$$\sum_{i=1}^{m} a_{ij} f_{i\hat{k}} z_i \geq \beta_j^S \sum_{i=1}^{m} a_{ij} f_{ik} z_i \ \forall j \in \mathcal{S}_{\hat{k}},\ \hat{k}, k = 1 \dots r,\ \hat{k} \neq k.$$

$$y_k - z_i \leq 1 - x_i,\ z_i \leq y_k,\ z_i \leq x_i,\ z_i \geq 0,\ i \in \mathcal{F}_k.$$

Supervised Biclustering

- Unfortunately, while the linearization works nicely for small-size problems, it often creates instances, where the gap between the integer programming and the linear programming relaxation optimum solutions is very big for larger problems. As a consequence, the instance **can not be solved in a reasonable time** even with the best techniques implemented in modern integer programming solvers.
- **HuGE Index Data set: about 7000 features**
- **ALL vs. AML Data Set: about 7000 features**
- **GBM vs. AO data set: about 12000 features**

Heuristic

- If we know that no more than m_k features can be selected for class $\mathcal{F}_k$, then we can impose

$$x_i \leq m_k z_i, \; x_i \geq z_i, \; i \in \mathcal{F}_k.$$

Heuristic

Algorithm

1. Assign $m_k := |\mathcal{F}_k|$, $k = 1 \ldots r$.

2. Solve the mixed 0–1 programming formulation using the inequalities

$$x_i \le m_k z_i,\ x_i \ge z_i,\ i \in \mathcal{F}_k.$$

instead of

$$y_k - z_i \le 1 - x_i,\ z_i \le y_k,\ z_i \le x_i,\ z_i \ge 0,\ i \in \mathcal{F}_k.$$

3. If $m_k = \sum_{i=1}^{m} f_{ik} x_i$ *for all* $k = 1 \ldots r$*, go to 6.*

4. Assign $m_k := \sum_{i=1}^{m} f_{ik} x_i$ *for all* $k = 1 \ldots r$.

5. Go to 2.

6. STOP.

Supervised Biclustering

- After the feature selection is done, we perform classification of test samples according to the following procedure.
- If $b = (b_i)_{i=1\dots m}$ is a test sample, we assign it to the class $\mathcal{F}_{\hat{k}}$ satisfying

$$\frac{\sum_{i=1}^{m} b_i f_{i\hat{k}} x_i}{\sum_{i=1}^{m} f_{i\hat{k}} x_i} > \frac{\sum_{i=1}^{m} b_i f_{ik} x_i}{\sum_{i=1}^{m} f_{ik} x_i}, \; k = 1 \dots r, \; \hat{k} \neq k.$$

HuGE index data set: Feature Selection

- A computational experiment that we conducted was on **feature selection for consistent biclustering** of the Human Gene Expression (*HuGE*) Index data set. The purpose of the HuGE project is to provide a comprehensive database of gene expressions in normal tissues of different parts of human body and to highlight similarities and differences among the organ systems.
- The number of selected features (genes) is 6889 (out of 7070).

HuGE index data set: Feature Selection

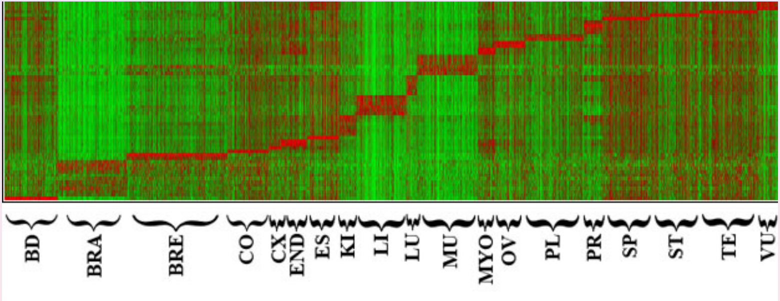


Figure: HuGE Index heatmap.

ALL vs. AML data set

- T. Golub at al. (1999) considered a dataset containing 47 samples from *ALL* patients and 25 samples from *AML* patients. The dataset was obtained with Affymetrix GeneChips.
- Our biclustering algorithm selected 3439 features for class *ALL* and 3242 features for class *AML*. The subsequent classification contained **only one error**: the *AML*-sample 66 was classified into the ALL class.
- The SVM approach delivers up to 5 classification errors depending on how the parameters of the method are tuned. The perfect classification was obtained only with one specific set of values of the parameters.

ALL vs. AML data set

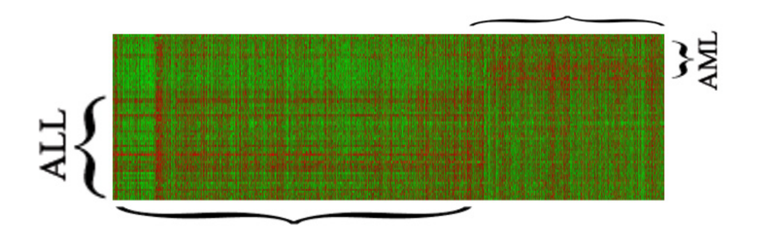


Figure: ALL vs. AML heatmap.

GBM vs. AO data set

- The algorithm selected 3875 features for the class GBM and 2398 features for the class AO. The obtained classification contained only 4 errors: two GBM samples (Brain_NG_1 and Brain_NG_2) were classified into the AO class and two AO samples (Brain_NO_14 and Brain_NO_8) were classified into the GBM class.

GBM vs. AO data set

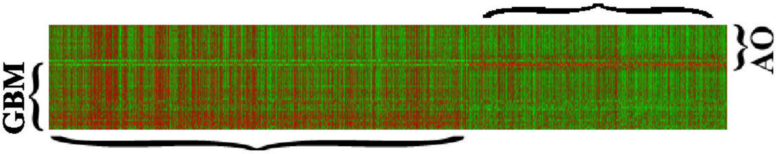


Figure: GBM vs. AO heatmap.

References

- S. Busygin, P. Pardalos, O. Prokopyev, Feature selection for consistent biclustering via fractional 0–1 programming, Journal of Combinatorial Optimization, Vol. 10/1 (2005), pp. 7–21.
- P.M. Pardalos, S. Busygin, O.A. Prokopyev, "On Biclustering with Feature Selection for Microarray Data Sets," BIOMAT 2005 – International Symposium on Mathematical and Computational Biology, R. Mondaini (ed.), World Scientific (2006), pp. 367–378.

Epilepsy Treatment

- Treatment: Drugs, Surgery, Electrical and Magnetic Stimulation.
- Vagus Nerve Stimulation
  - Electric stimulator implanted subcutaneously in the chest
  - Connected, via subcutaneous electrical wires, to the cervical left vagus nerve.

Vagus Nerve Stimulation parameters

- The VNS is programmed to deliver electrical stimulation at a set intensity, duration, pulse width, and frequency.
- Optimal parameters are presently determined on a case by case basis, depending on clinical efficacy (seizure frequency) and tolerability.
- Such parameter adjustment is time consuming and costly.
- There is a need to develop a reliable, objective and rapid method of determining the optimal stimulation parameters for each patient

EEG Data

- General Clinical Research Center in Shands Hospital at The University of Florida.
- Two patients A and B.
- 25 scalp-EEG channels
- Sampling rate 512 Hz

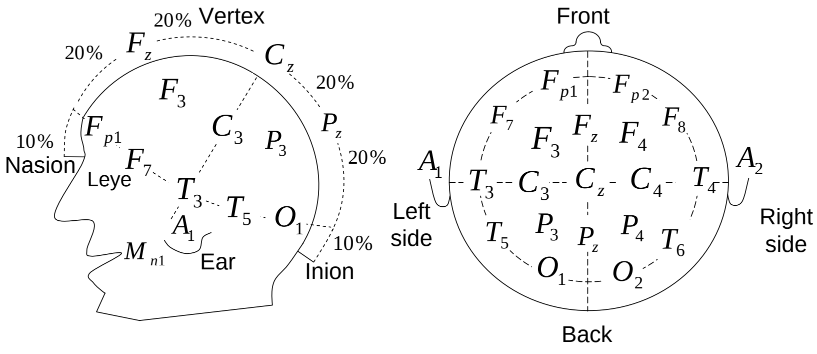


Figure: Montage for scalp electrode placement (10-20)

VNS Stimulation Settings

Parameter	Patient A	Patient B
Recorded time	$\approx$ 24 hours	
Signal duration	30 sec	
Rest duration	5 min	
Pulse width	500 μsec	250 μsec
Output current	1.75 mA	1.5 mA
Frequency	30 Hz	20 Hz
# of VNS cycles	255	237

Manually Activated Stimulation

- Stimulation may be activated manually, for example in case of seizure.
- During EEG recording session
  - Patient A did not undergo seizures.
  - Patient B experienced 14 seizures (manual stimulation was activated 14 times)
- Parameters for patient B's manual simulation
  - stimulation activated after 19-37 sec after seizure, output current 1.75 mA, signal frequency 20 Hz, pulse width 500 μsec, duration 60 sec.
- Manual stimulation is not included

Lyapunov Exponents Estimation: Parameters

- *Input:* EEG signal recorded with 512Hz ($\Delta t = 1.95 msec$)
- *Output:* STLmax series computed for every 4sec window of the source data
- *Algorithm parameters:* reconstructed dimension $p = 7$, lag step $\tau = 7$ (14 msec), evolution time $\Delta = 21$ (41 msec)
- 25 channels for patients A and B are processed

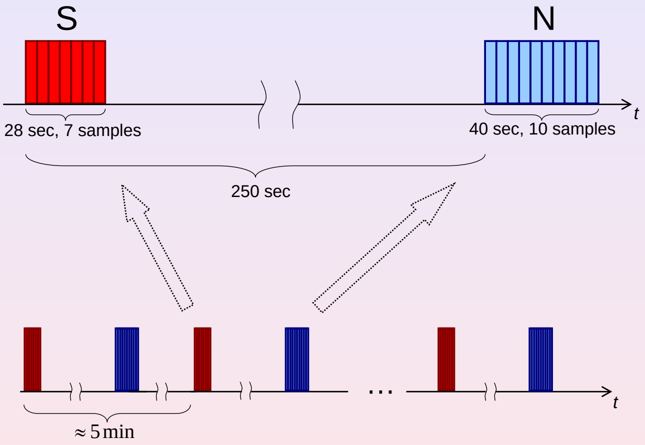


Figure: Building dataset for biclustering: STL_{max} points that are included for the analysis for each channel.

Biclustering Experiment

- Positive (stimulation) class: Each 30 sec stimulation provided 7 data points (STLmax that each correspond to a four seconds window)
- Negative (non-stimulation) class: 10 consecutive Lyapunov Exponents 250 sec after stimulation
- We averaged corresponding data points across all stimulation cases
- Thus, we obtained a 17×25 matrix. 17 samples (7 stimulation $+$ 10 non-stimulation) and 25 features (channels)

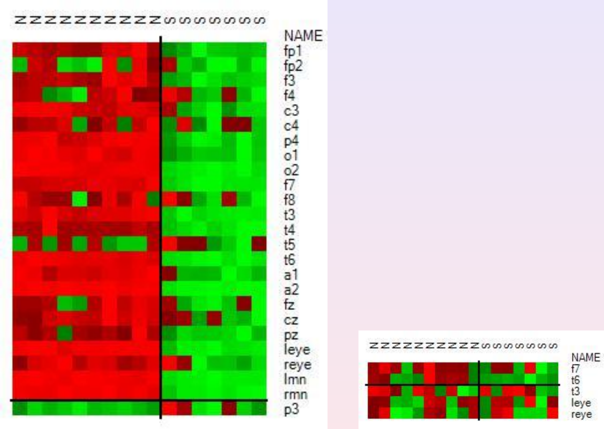


Figure: Heatmaps for patients A and B

Results

- Patient A
  - No features were excluded,i.e. patient's data were conditionally biclustering-admitting
  - STLmax data were consistently decreasing during the simulation except for channel $p3$
  - All samples and all classes are confirmed by cross-validation
- Patient B
  - five features selected
  - The leave-one-out cross-validation was passed for all but four samples

Thank You!

- Questions?