Mathematical Models for Feature Selection And their Application In Bioinformatics

Paola Bertolazzi, Giovanni Felici
Istituto di Analisi dei Sistemi ed Informatica IASI-CNR

Paola Festa
Dipartimento di Matematica e Applicazioni *R.M. Caccioppoli*, Università degli Studi di Napoli Federico II

Summary

Logic Data Mining System: online dmb.iasi.cnr.it

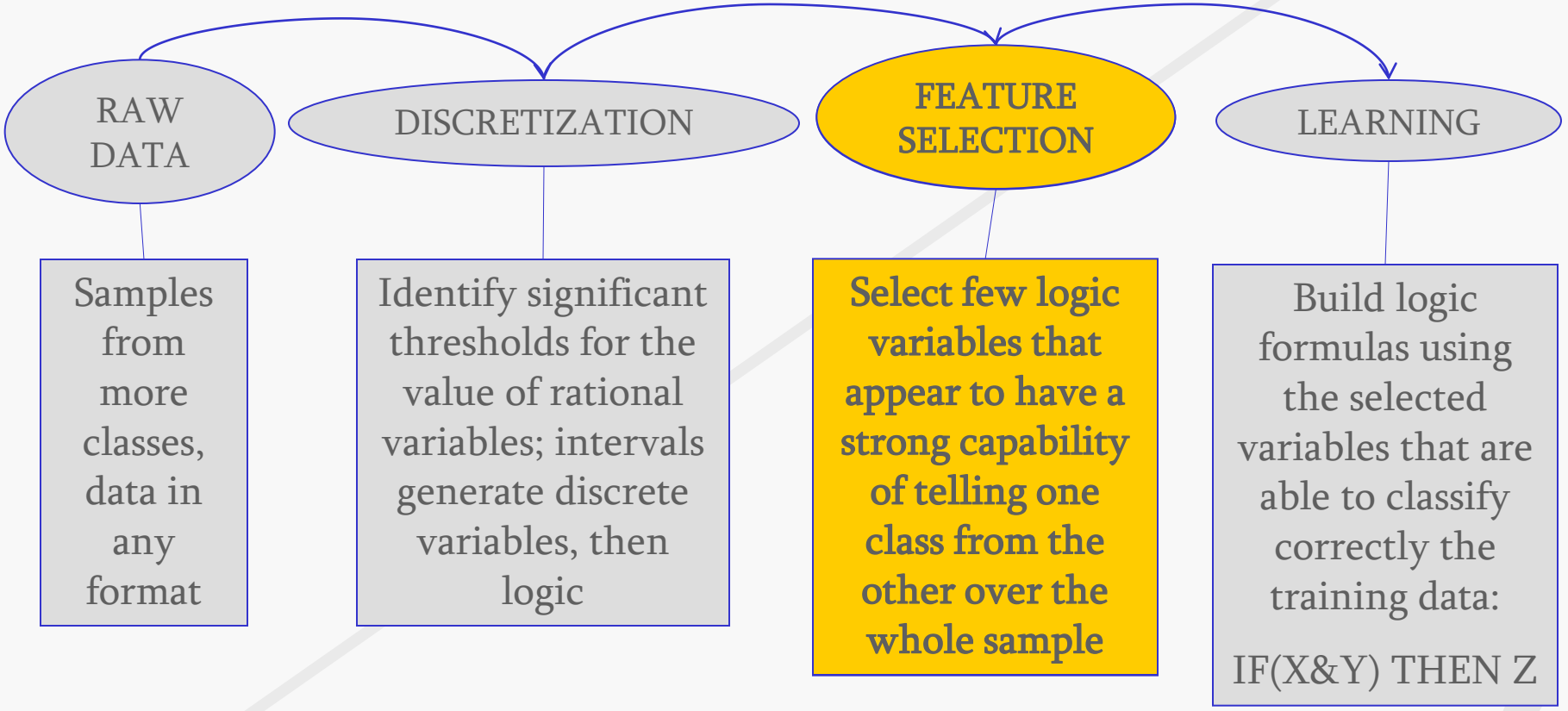
Focus on:

Formulation of the Feature Selection Problem

GRASP Methods

Applications

The Logic Data Mining flow



Feature Selection

- FS is a projection of a set of multidimensional points from their original space to a space of smaller dimension with little "loss of information" or large "reduction of noise".
- Information and noise must defined w.r.t. to the objective of the specific application: clustering, classification, synthesis...
- in supervised learning application, we want to preserve or enhance the relative distances between observations belonging to different groups.

FS as a Combinatorial Problem

When the projection of the points is simply a selection of a subset of the available dimensions, the FS problem has a combinatorial nature.

Such fact has been pointed out and exploited already in the literature:

- Garey M.R. and Johnson D.S. Computer and Intractability: a guide on the theory of NP-completeness. Freeman, San Francisco, 1979.
- E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, I. Muchnik, An implementation of logical analysis of data, IEEE Transactions on Knowledge and Data Engineering, 12 (2) 292-306 (2000).
- M. Charikar, V. Guruswami, R. Kumar, S. Rajagopalan and A. Sahai. Combinatorial Feature Selection Problems. In Proceedings of FOCS 2000.
- R. Beretta, A. Mendes, P. Moscato, Integer programming models and algorithms for molecular classification of cancer from Microarray Data, Proceedings of the Twenty-eighth Australasian Computer Science Conference, 38, 361 - 370 (2005)

Notations and Definitions

we assume that n m-dimensional points are the input data for the FS problems. The points are represented in the rational matrix A

M (resp. N) is the index set of the columns of A (resp. rows); then

$$n = |N|, m = |M|, A = n \times m, A \in R^{mxn}$$

An appropriate measure of the information contained in A is given by:

$$I(A) = \sum_i \sum_{j \neq i} \sum_k \left(a_{ij} - a_{jk}\right)^2$$

I(A) ≈ the average quadratic distance of the points in A, directly related to the **variance** expressed by A, a widely used measure in Statistics and Data Analysis.

A Simple Optimization Problem

Consider now the projection of A on a subset of its dimensions M', such that $|M'| = \beta < m$, and

$$x_k = \begin{cases} 1 & \text{if } k \in M' \\ 0 & \text{otherwise} \end{cases}$$

and therefore

$$I_x(A) = \sum_i \sum_{j \neq i} \sum_k \left(a_{ik} - a_{jk}\right)^2 x_k$$

represents the portion of information preserved by the projection of the points of A on their M' dimensions.

The simplest optimization problem that can be defined would be:

$$\max I_x(A) = \sum_i \sum_{j \neq i} \sum_k \left(a_{ik} - a_{jk}\right)^2 x_k$$

$$\sum_k x_k \leq \beta$$

$$x_k \in \{0,1\}$$

A (proper) extension: minimization of the infimum-norm

An alternative to the average approach consists in requiring a minimum level of distance between each pair, and requiring a projection that maximizes such level:

$$\max \alpha$$
$$\sum_k \left(a_{ik} - a_{jk}\right)^2 x_k \geq \alpha, \qquad \forall i, j, i \neq j$$
$$\sum_k x_k \leq \beta$$
$$x_k \in \{0,1\}$$

Relation between the two models

Let h = m x n and $\Omega \subset R^h$ be the Euclidean subspace where a point ω is defined as follows:

$$\omega = \left\{\omega_1, \ldots, \omega_h : \omega_l = \sum_k (a_{ik} - a_{jk})^2, l = i \times (n-1) + j, i \neq j\right\}$$

With a proper definition of the projection ω_x we have that the 2 models become:

$$\max \left\|\omega_x\right\|^1 \qquad \max \left\|\omega_x\right\|^{\inf}$$
$$\sum_k x_k \leq \beta \qquad \sum_k x_k \leq \beta$$
$$x \in \{0,1\}^m \qquad x \in \{0,1\}^m$$

1) Special Case: Binary Data

Let $d_{ij}^k = (a_{ik} - a_{jk})^2$

If data in A is binary, then $a_{ij} \in \{0,1\} \Rightarrow d_{ij}^k = \begin{cases} 1 & if\ (a_{ik} = a_{jk}) \\ 0 & \text{otherwise} \end{cases}$

and the FS problem can be rewritten as:

$$\max \alpha$$
$$\sum_k d_{ij}^k x_k \geq \alpha, \quad \forall i, j, i \neq j$$
$$\sum_k x_k \leq \beta$$
$$xk \in \{0,1\}$$

2) Special Case: Supervised Learning

The row vectors of A are partitioned into two different classes

$$A = \tilde{A} \cup \tilde{B}$$
$$a_i \in \tilde{A}, c(i) = \tilde{A}, \quad a_i \in \tilde{B}, c(i) = \tilde{B}$$
$$n_A = \left|\tilde{A}\right|, \ n_B = \left|\tilde{B}\right|$$

$$\max \alpha$$
$$\sum_k d_{ij}^k x_k \geq \alpha, \quad \forall i, j, c(i) \neq c(j)$$
$$\sum_k x_k \leq \beta$$
$$x_k \in \{0,1\}$$

Only the distance between points of different classes is taken into account; but the number of constraints is still very large, as it grows quadratically with n.

An example

features

samples

		1	2	3	4	5	6	7	8	9	10
1	A	1	1	1	0	1	0	0	1	0
2	B	1	1	0	0	0	1	0	1	1
3	B	0	1	1	0	1	1	1	1	0

constraint(1,2): x4 + x5+ X6 +X10 >=1

constraint(1,3): x1 + x6+ x7 >=1

solution with minimal size (X6 = 1, Xi = 0, i <>6)

 # constraints proportional to Na*Nb

With $\beta \leq 2$ the max value of α is still 1.

We need $\beta = 3$ for a solution with $\alpha = 2$

Variant 1) A Compact Model

Assume the case of supervised learning, and consider the subset of constraints related to row i belonging to class A, and add over the elements of class B:

$$\sum_k d_{ij}^k x_k \geq \alpha, \forall j, c(j) = \tilde{B} \qquad \sum_{j:c(j)=\tilde{B}} \left(\sum_k d_{ij}^k x_k \geq \alpha \right)$$

$$\sum_{j:c(j)=\tilde{B}} \sum_k d_{ij}^k x_k \geq \sum_{j:c(j)=\tilde{B}} \alpha \approx \sum_k \sum_{j:c(j)=\tilde{B}} d_{ij}^k x_k \geq \alpha \times n_B$$

$$\tilde{d}_i^k = \sum_{j:c(j)=\tilde{B}} d_{ij}^k, \qquad \tilde{d}_i^k = \begin{cases} 1 \\ 0 \end{cases}$$

k separates perfectly the 2 classes

k is useless for separation

A Compact Model (2)

The value $\tilde{d}_{ik}$ can be adopted as a direct measure of the importance of column k for row i:

$$f_{ik} = \begin{cases} \dfrac{\tilde{d}_{ik}}{n_B}, i : c(i) = \tilde{A} \\ \dfrac{\tilde{d}_{ik}}{n_A}, i : c(i) = \tilde{B} \end{cases} \qquad \tilde{f}_{ik} = \lfloor f_{ik} + \lambda \rfloor$$

$$\max \alpha$$
$$\sum_k \tilde{f}_{ik} x_k \geq \alpha, \qquad \forall i$$
$$\sum_k x_k \leq \beta$$
$$xk \in \{0,1\}$$

And λ controls the density of the constraint matrix of the IP problem

- if λ = 0.5, the coefficients of the constraints have value 1 only when the value of k for element i is different from the mode of the values of k over the element of the other class;
- If f is not rounded, then the constraints represent the maximization of the average hamming distance between the k^{th} coordinate of element i and the same coordinate of all the elements belonging to the other class.

How to solve those large (and hard) IPs ?

- At optimality: whit contained dimensions; else heuristics…

RELEVANT ISSUES

- The quality of solution depends on the chosen sample as well as on the solution algorithms
- There are many equivalent solutions for a given problem
- Cross validation approach: integrate the solutions obtained on different subset of the available data (re-sampling)

- It is required to solve many instances of the same problems over different input data …

Good heuristics seem to be the right approach…
Their weakness w.r.t. optimal methods are balanced by data sampling

Is it better to have MANY GOOD SOLUTIONS or FEW OPTIMAL ONES ?

H1) GRASP HEURISTICS

FS is NP-hard

- GRASP: Greedy Randomized Adaptive Search Procedure, successfully applied to find approx solutions to hard combinatorial problems (Festa and Resende, '02, '08).

- Each GRASP iteration consists of two phases:

1. an iterative greedy adaptive randomized construction phase that builds a feasible solution;
2. local search phase.

1. Define candidate elements C;
2. Apply a greedy function g(e), e in C;
3. Rank candidates in C according to greedy function values g(e);
4. Put well ranked candidates into a restricted candidate list RCL;
5. Randomly choose one element in RCL and add it to the solution under construction
6. Adaptive component: greedy function values depend on the partial solution under construction

GRASP for Feature Selection

The objective function is composed of three parts, with weights of decreasing importance:

- the value of α
- the number of rows covered at value larger than α,
- the total extra coverage spent on the rows.

swap local search procedure is applied to improve the solution, i.e. a new set of columns with lower cardinality (removal of redundant columns) and/or corresponding to a higher coverage;

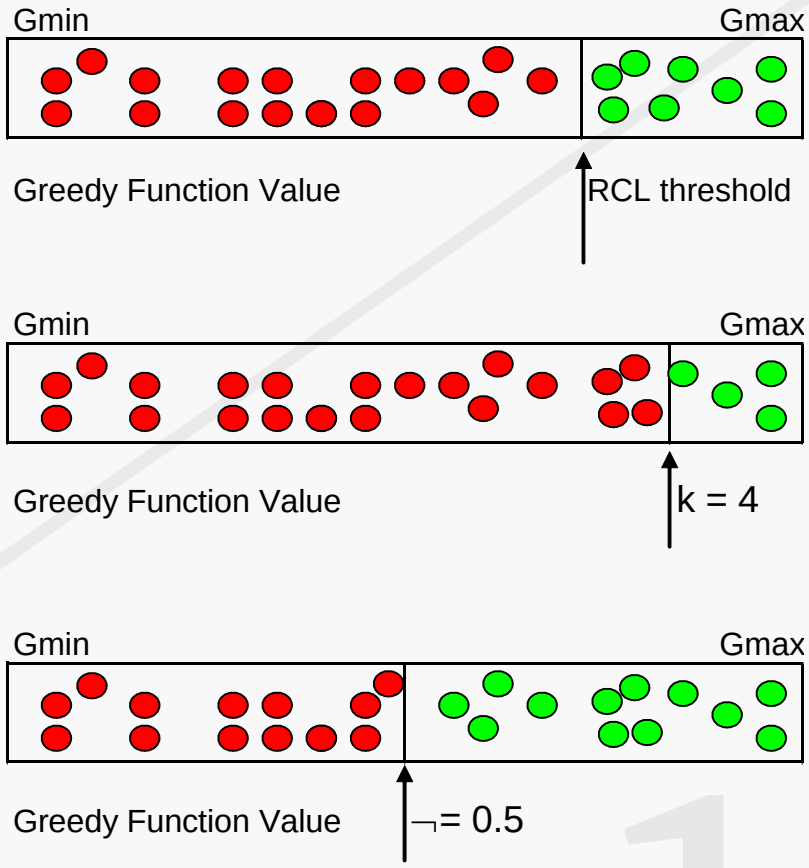
At each local search iterations, candidate sets of columns to be swapped are defined and all swaps are tested. Ad hoc data structures enable the construction and local search steps to be very fast.

RCL construction

$$g_{\min} = \min_{e \in C} g(e)$$

$$g_{\max} = \max_{e \in C} g(e)$$

- **Cardinality based**: RCL is made of the k elements with the best greedy value
- **Value based**: RCL is associated with a parameter δ in [0,1] and a threshold value $\mu = g_{min}+\delta(g_{max}-g_{min})$



- δ = 0: Pure greedy
- δ = 1: Pure random

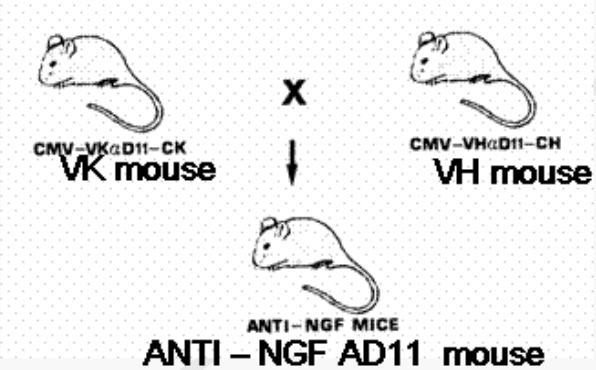
Results on Feature Selection

Name	dimension of Class1	Dimension of class 2	number of Clauses	Dimension of Clauses	Features	Rows of FS Problems	Support	Beta Value
t01	100	100	3	4	100	10.000	10	10
t02	200	200	3	4	100	40.000	10	10
t03	300	300	3	4	100	90.000	10	10
t04	100	100	3	4	1.000	10.000	20	30
t05	100	100	3	4	5.000	10.000	20	30
t06	200	500	4	8	50.000	10.000	20	30

Name	Number of Problems	maximum solution time (secs)	Best solution proved optimal	GRASP finds Best
t01	5	120	5/5	5/5
t02	5	120	5/5	5/5
t03	5	120	5/5	5/5
t04	5	900	3/5	4/5
t05	5	1.800	0/5	3/5
t06	5	3.600	0/5	5/5

Application:
Mining transcriptome data of the AD11 transgenic mouse model

Joint work with European Brain Research Institute Rita Levi-Montalcini, Roma, Italy (EBRI)



- The αD11 anti-NGF antibody is composed by the light (VK) and heavy (VH) chains. Crossing mice expressing the light chain (VK mice) with mice expressing the heavy chain (VH mice) yields double transgenic offspring, which expresses a functional αD11 antibody (anti-NGF AD11 mice).
- The AD11 anti-NGF mice represent a comprehensive transgenic model for an Alzheimer-like neurodegeneration, displaying in a progressive way a full complement of phenotypic hallmarks for the disease

For a total of 120 samples

Aims of the Project

1. To characterize the gene expression profile of the AD11 mice in different brain areas following temporal progression
2. To identify a limited set of genes able to discriminate between the neurodegeneration and the healthy state
3. Explain the onset of the Alzheimer disease and thus identify early biomarkers of the pathology

a) Discretization

The data is transformed from numerical to qualitative/binary
1) create many intervals for the expression of the genes
2) merge intervals based on explained entropy
Each gene receives its most appropriate number and type of intervals
genes that are not varying across the samples are discarded

b) Feature Clustering

Features with the same discretized profile over the samples are clustered

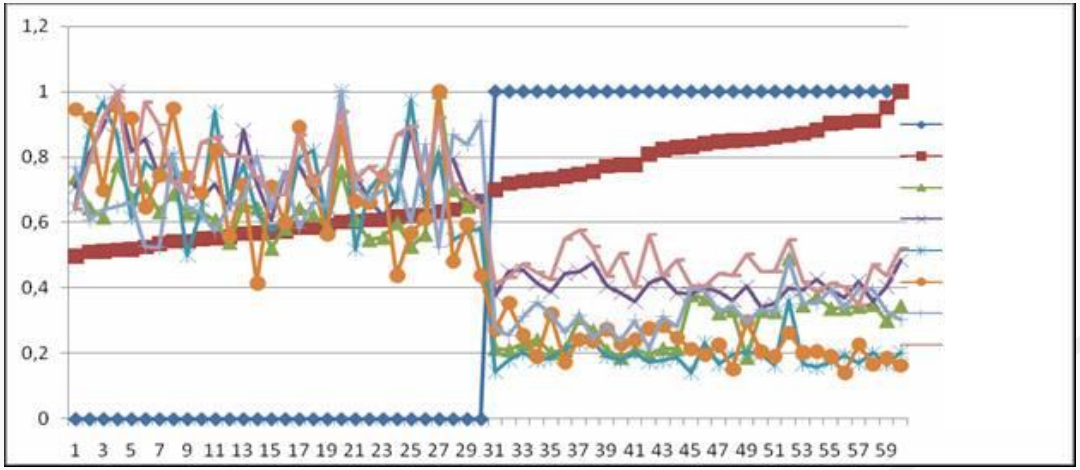
c) Feature Selection

The FS model is solved with GRASP with β values

d) Learning

Apply Lsquare as described on the reduced feature set

Some Results



Cluster size	Frequencies
1	3068
2	289
3	86
4	51
5	26
6	14
7	20
8	8
9	6
...	...
244	1
299	1
420	1
441	1
6650	1

- There are few genes (7) that are able, one by one, to separate exactly all the healthy from the sick mice (iterative application of the method) in leave-1-out cross validation
- The 7 genes are highly co-regulated or contro-regulated and identify a regulatory network that is presently under study at EBRI
- More genes are strongly co-regulated with the 7 genes network

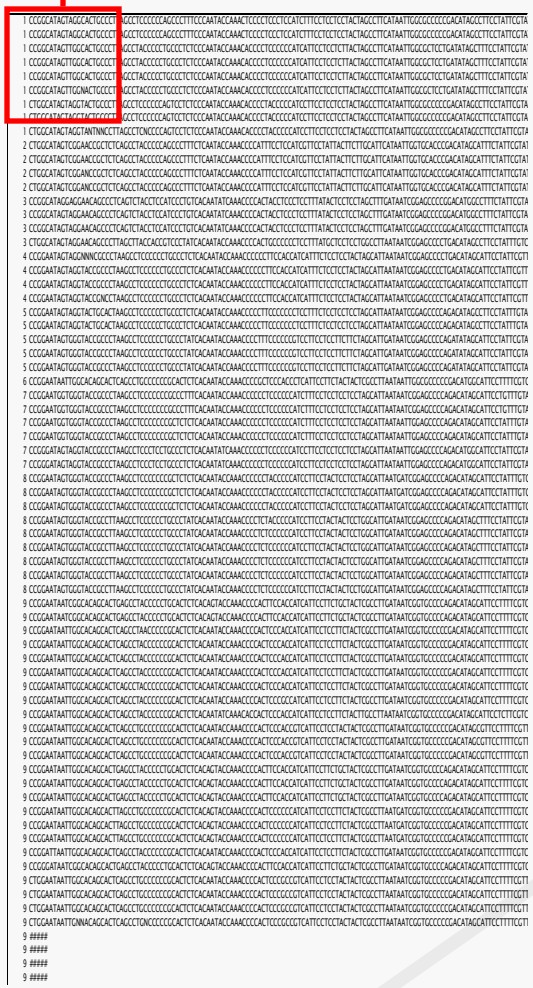
CLASS 1: A_52_P58XXXX>= 0.87
CLASS 2: A_52_P58XXXX < 0.87

Application:
Species Classification through Barcode

- A BARCODE is a small portion of mitochondrial DNA where the nucleotides change rapidly among specie
- Given samples from different species, the objective is to identify those combinations of muted nucleotides that have determined the differences from to in the evolution path
- BARCODING is a relatively new problem that is drawing attention of the bio-comp community

The international Consortium CBOL, funded by the Sloan Foundation, is investing money since 2005 in collecting barcodes of many species and in putting together a library of algorithms for its analysis. The CBOL website now makes available to researchers more than 2M barcodes

IASI is member of the Data Analysis Working Group of CBOL since 2006 and has developed a species classifier based on Logic Programming (BLOG) made available on the Consortium website



The whole picture...

SPECIE	BARCODE FRAGMENT
1	CTGGCATAGTAGGTACTGCCCTTAGCCTCCCCCCAGTCCTCTCCCAATACCAAACACCCCTACCCCCATCCT
1	CCGGCATAGTAGGCACTGCCCTTAGCCTCCCCCCAGCCCTTTCCCAATACCAAACTCCCCTCCCTCCATCTT
1	CCGGCATAGTTGGCACTGCCCTTAGCCTACCCCCTGCCCTCTCCCAATACCAAACACCCCTCCCCCCATCAT
1	CCGGCATAGTTGGNACTGCCCTTAGCCTACCCCCTGCCCTCTCCCAATACCAAACACCCCTCCCCCCATCAT
1	CCGGCATAGTTGGCACTGCCCTTAGCCTACCCCCTGCCCTCTCCCAATACCAAACACCCCTCCCCCCATCAT
1	CTGGCATAGTAGGTANTNNCCTTAGCCTCNCCCCAGTCCTCTCCCAATACCAAACACCCCTACCCCCATCCT
1	CCGGCATAGTAGGCACTGCCCTTAGCCTCCCCCCAGCCCTTTCCCAATACCAAACTCCCCTCCCTCCATCTT
1	CCGGCATAGTTGGCACTGCCCTTAGCCTACCCCCTGCCCTCTCCCAATACCAAACACCCCTCCCCCCATCAT
1	CTGGCATAGTAGGTACTGCCCTTAGCCTCCCCCCAGTCCTCTCCCAATACCAAACACCCCTACCCCCATCCT
2	CTGGCATAGTCGGAACCGCTCTCAGCCTACCCCCAGCCCTTTCTCAATACCAAACCCCATTTCCTCCATCGTT
2	CTGGCATAGTCGGAACCGCTCTCAGCCTACCCCCAGCCCTTTCTCAATACCAAACCCCATTTCCTCCATCGTT
2	CTGGCATAGTCGGANCCGCTCTCAGCCTACCCCCAGCCCTTTCTCAATACCAAACCCCATTTCCTCCATCGTT
2	CTGGCATAGTCGGANCCGCTCTCAGCCTACCCCCAGCCCTTTCTCAATACCAAACCCCATTTCCTCCATCGTT
3	CCGGCATAGGAGGAACAGCCCTCAGTCTACCTCCATCCCTGTCACAATATCAAACCCCACTACCTCCCTCCT
3	CCGGCATAGTAGGAACAGCCCTCAGTCTACCTCCATCCCTGTCACAATATCAAACCCCACTACCTCCCTCCT
3	CTGGCATAGTAGGAACAGCCCTTAGCTTACCACCGTCCCTATCACAATACCAAACCCCACTGCCCCCCTCCT
3	CCGGCATAGTAGGAACAGCCCTCAGTCTACCTCCATCCCTGTCACAATATCAAACCCCACTACCTCCCTCCT
4	CCGGAATAGTAGGTACCGCCCTAAGCCTCCCCCCTGCCCTCTCACAATACCAAACCCCCCTTCCACCATCAT
4	CCGGAATAGTAGGTACCGNCCTAAGCCTCCCCCCTGCCCTCTCACAATACCAAACCCCCCTTCCACCATCAT
4	CCGGAATAGTAGGNNNCGCCCTAAGCCTCCCCCCTGCCCTCTCACAATACCAAACCCCCCTTCCACCATCA
4	CCGGAATAGTAGGTACCGCCCTAAGCCTCCCCCCTGCCCTCTCACAATACCAAACCCCCCTTCCACCATCAT
5	CCGGAATAGTAGGTACTGCACTAAGCCTCCCCCCTGCCCTCTCACAATACCAAACCCCCTTCCCCCCCTCCT
5	CCGGAATAGTAGGTACTGCACTAAGCCTCCCCCCTGCCCTCTCACAATACCAAACCCCCTTCCCCCCCTCCT
5	CCGGAATAGTGGGTACCGCCCTAAGCCTCCCCCCCTGCCCTATCACAATACCAAACCCCTTTCCCCCCGTCCT
5	CCGGAATAGTGGGTACCGCCCTAAGCCTCCCCCCCTGCCCTATCACAATACCAAACCCCTTTCCCCCCGTCCT
5	CCGGAATAGTGGGTACCGCCCTAAGCCTCCCCCCCTGCCCTATCACAATACCAAACCCCTTTCCCCCCGTCCT
6	CCGGAATAATTGGCACAGCACTCAGCCTGCCCCCCCGCACTCTCACAATACCAAACCCCGCTCCCACCCTCA
7	CCGGGATAGTAGGTACCGCCCTAAGCCTCCCTCCTGCCCTCTCACAATATCAAACCCCCCTCCCCCCATCCT
7	CCGGAATGGTGGGTACCGCCCTAAGCCTCCCCCCCCGCCCTTTCACAATACCAAACCCCCCTCCCCCCATCT
7	CCGGAATGGTGGGTACCGCCCTAAGCCTCCCCCCCCGCTCTCTCACAATACCAAACCCCCCTCCCCCCATCT
7	CCGGAATGGTGGGTACCGCCCTAAGCCTCCCCCCCCGCTCTCTCACAATACCAAACCCCCCTCCCCCCATCT
7	CCGGGATAGTAGGTACCGCCCTAAGCCTCCCTCCTGCCCTCTCACAATATCAAACCCCCCTCCCCCCATCCT
7	CCGGAATGGTGGGTACCGCCCTAAGCCTCCCCCCCCGCCCTTTCACAATACCAAACCCCCCTCCCCCCATCT

Experiments

a) Discretization

The A,C,G,T values of the sites are associates to presence/absence logic variables

b) Feature Selection

Using the compact FS model solved optimally few sites are identified (10-30)

c) Learning

Apply Lsquare as described on the reduced feature set and obtain a formula that tells a species from the others

1. For each species k, we solve a 2-class learning problem:
   - ✓ class **A**: subset of samples in class k
   - ✓ class **B**: samples of class different from k
- We use a training subset (80%,90%) of the available data, and then test their classification capabilities on the remaining data.
- Training and testing samples are drawn at random maintaining the same proportion for each species

Experiments

1700 samples
150 different species
648 to 690 sites (or nucleotides).

			Error Rates	
β	α	test%	training	testing
10	4	10	8.02%	17.00%
10	4	10	10.14%	20.00%
10	4	20	11.90%	21.52%
10	4	20	13.50%	21.19%
average			10.89%	19.93%
15	6	10	0.87%	10.00%
15	6	10	1.93%	12.50%
15	6	20	1.50%	10.93%
15	6	20	2.04%	12.25%
average			1.58%	11.42%
20	8	15	0.20%	8.94%
20	8	15	0.61%	7.72%
average			0.40%	8.33%

Table 2: Optimal values and Error Rates (first data set)

826 samples
82 different species
660 sites (or nucleotides).

			Error Rates	
β	α	test%	training	testing
10	8	10	15.47%	15.06%
10	8	10	16.48%	15.90%
10	7	20	19.97%	21.03%
10	7	20	21.39%	22.56%
average			18.32%	18.64%
15	11	10	5.52%	6.67%
15	11	10	6.37%	8.33%
15	11	20	7.40%	10.42%
15	11	20	10.88%	10.42%
average			7.54%	8.96%
20	14	10	0%	1.38%
20	14	10	2.22%	3.08%
20	15	20	1.97%	5.50%
20	15	20	1.58%	5.13%
average			1.44%	3.77%

Table 4: Optimal values and Error Rates (second data set)

SPECIES	CC	WC	CLAUSE(S)
A1	1.00	0.00	(v100=c) and (v346=a) and (v499=t) and (v502=a)
A2	0.77	0.00	(v82=t) and (v238=t) and (v502=c)
A3	1.00	0.00	(v58=a) and not(v100=c) and not(v106=a)
A4	1.00	0.00	(v106=t) and (v139=g)
A5	1.00	0.00	not(v106=g) and not(v295=a) and not(v295=g)

Table 3: Logic Formulas for Species 1 to 5 (first data set)