Learning in Networks of Similarity Processing Neurons

Lluís A. Belanche
belanche@cs.upc.edu

Soft Computing Research Group
Computer Science Department

Technical University of Catalonia, Barcelona, Catalonia, Spain

Advanced Topics in Computational Intelligence
(ATCI-MAI)

March 2018

What the lecture is about

Analysis of non-standard data in neural methods

1. similarities, dissimilarities and distances
2. non-standard or heterogeneous data sources
3. special situations (missing, imprecise or nought values, ...)
4. neuron models as similarity measures
5. standard transformations
6. Similarity Neural Networks (SNNs)

Preliminaries

- Human beings use the notion of **similarity** and **dissimilarity** for problem solving: inductive reasoning, analogical reasoning, ...

- Computer Science:
  - Artificial Intelligence
  - Case-Based Reasoning
  - Information Retrieval
  - Pattern Matching
  - **Machine Learning**: kNN, Neural Networks, SVMs, ...

How should these objects be compared?

Preliminaries

- **Metric** dissimilarities have been deeply studied but they are tied to a particular transitivity (triangle inequality)

- Particularly, **Euclidean** distances are often used due to our natural understanding of Euclidean spaces

- Not all metrics are Euclidean and many interesting dissimilarities are non-metric

- What about similarities? Where do similarities (and dissimilarities) come from?

Preliminaries

Neural Networks are very useful tools for the modelling of non-linear systems. **Shortcomings**:

1. How to deal with non-standard data sources (*e.g.*, mixed data types)

2. How to deal with special situations (*e.g.*, missing values)

3. Euclidean space $\mathbb{R}^n$ and its geometry may not capture the required relations among inputs and outputs

4. Optimization process: Low convergence velocity, multiple local minima, multiple restarts, high computational burden

5. Determination of the number of hidden neurons (and number of layers!) is still black magic

6. Interpretability is arduous (black-box model) and getting worse

7. Difficulty to inject prior knowledge (specially symbolic)

TABLE 2.2. Features of the Horse Colic data set.

1	Surgery?:	1 = Yes, it had surgery; 2 = It was treated without surgery
2	Age:	1 = Adult horse; 2 = Young horse (<6 months)
3	Hospital Number:	the case number assigned to the horse
4	Rectal Temperature:	in degrees Celsius, linear
5	Pulse:	the heart rate in beats per minute, linear
6	Respiratory Rate:	linear
7	Temperature of Extremities:	1 = normal; 2 = warm; 3 = cool; 4 = cold
8	Peripheral Pulse:	1 = normal; 2 = increased; 3 = reduced; 4 = absent
9	Mucous Membranes:	1 = normal pink; 2 = bright pink; 3 = pale pink; 4 = pale cyanotic; 5 = bright red / injected; 6 = dark cyanotic
10	Capillary Refill Time:	1 = <3 seconds; 2 = >= 3 seconds
11	Pain:	1 = alert; no pain; 2 = depressed; 3 = intermittent mild pain; 4 = intermittent severe pain; 5 = continuous severe pain
12	Peristalsis:	1 = hypermotile; 2 = normal; 3 = hypomotile; 4 = absent
13	Abdominal Distension:	1 = none; 2 = slight; 3 = moderate; 4 = severe
14	Nasogastric Tube:	1 = none; 2 = slight; 3 = significant
15	Nasogastric Reflux:	1 = none; 2 = >1 liter; 3 = <1 liter
16	Nasogastric Reflux pH:	scale is from 0 to 14 with 7 being neutral, linear
17	Rectal Examination—Feces:	1 = normal; 2 = increased; 3 = decreased; 4 = absent
18	Abdomen:	1 = normal; 2 = other; 3 = firm feces in the large intestine; 4 = distended small intestine; 5 = distended large intestine
19	Packed Cell Volume:	the # of red cells by volume in the blood, linear
20	Total Protein:	linear
21	Abdominocentesis Appearance:	1 = clear; 2 = cloudy; 3 = serosanguinous
22	Abdomcentesis Total Protein:	linear
23	Outcome:	1 = lived; 2 = died; 3 = was euthanized

TABLE 2.1. The first 25 records of the Horse Colic data set.

Surgery?	Age	Hospital Number	Rectal Temperature	Pulse	Respiratory Rate	Temperature of Extremities	Peripheral Pulse	Mucous Membranes	Capillary Refill Time	Pain	Peristalsis	Abdominal Distension	Nasogastric Tube
2	1	530101	38.50	66	28	3	3	?	2	5	4	4	?
1	1	534817	39.2	88	20	?	?	4	1	3	4	2	?
2	1	530334	38.30	40	24	1	1	3	1	3	3	1	?
1	9	5290409	39.10	164	84	4	1	6	2	2	4	4	1
2	1	530255	37.30	104	35	?	?	6	2	?	?	?	?
2	1	528355	?	?	?	2	1	3	1	2	3	2	2
1	1	526802	37.90	48	16	1	1	1	1	3	3	3	1
1	1	529607	?	60	?	3	?	?	1	?	4	2	2
2	1	530051	?	80	36	3	4	3	1	4	4	4	2
2	9	5299629	38.30	90	?	1	?	1	1	5	3	1	2
1	1	528548	38.10	66	12	3	3	5	1	3	3	1	2
2	1	527927	39.10	72	52	2	?	2	1	2	1	2	1
1	1	528031	37.20	42	12	2	1	1	1	3	3	3	3
2	9	5291329	38.00	92	28	1	1	2	1	1	3	2	3
1	1	534917	38.2	76	28	3	1	1	1	3	4	1	2
1	1	530233	37.60	96	48	3	1	4	1	5	3	3	2
1	9	5301219	?	128	36	3	3	4	2	4	4	3	3
2	1	526639	37.50	48	24	?	?	?	?	?	?	?	?
1	1	5290481	37.60	64	21	1	1	2	1	2	3	1	1
2	1	532110	39.4	110	35	4	3	6	?	?	3	3	?
1	1	530157	39.90	72	60	1	1	5	2	5	4	4	3
2	1	529340	38.40	48	16	1	?	1	1	1	3	1	2
1	1	521681	38.60	42	34	2	1	4	?	2	3	1	?
1	9	534998	38.3	130	60	?	3	?	1	2	4	?	?
1	1	533692	38.1	60	12	3	3	3	1	?	4	3	3

Horse Colic data

from the book *Clustering*, by Rui Xu and Don Wunsch, John Wiley & Sons, 2008

Characterization of learning systems

Commonly used methods to deal with mixed data are:

Ordinal variables are treated as real-valued or using a *thermometer* scale

Categorical variables with c modalities are coded using a binary expansion representation (a.k.a a 1-out-of-c or one-hot code)

Binary variables need (almost) no treatment, but asymmetries are ignored

Fuzzy variables are almost always avoided

Circular variables are specially bad treated ...

Characterization of learning systems

Commonly used methods to deal with special situations are:

Missing information is difficult to handle, specially when the lost parts are of significant size. Typical approaches remove the involved observations (or variables) or "fill in the holes" with the mean, ~~median~~ or nearest neighbor value. Statistical approaches need to model the input distribution itself, or are computationally very intensive

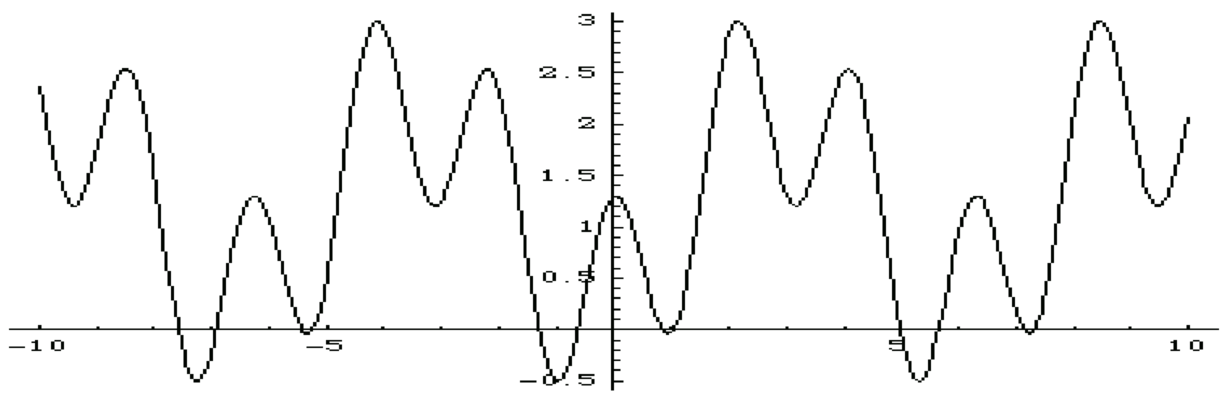
Multivalued variables may be considered as "set-valued"

Nought what to do in this case? (*e.g.*, number of pregnancies in males)

Characterization of learning systems

$\boxed{\text{similar observations} \Longrightarrow \text{similar targets}}$

(the converse is NOT true)

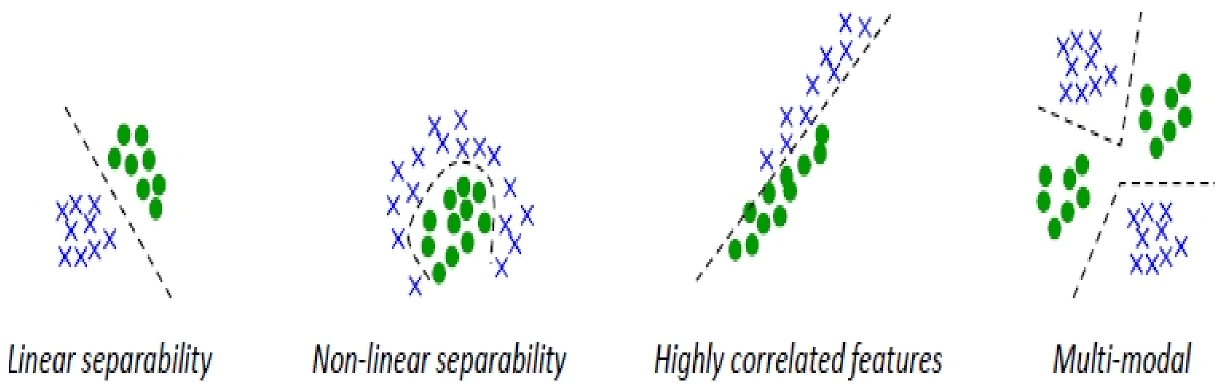


the **regression** case: truth of the principle is tantamount to continuity

Characterization of learning systems

$\boxed{\text{similar observations} \Longrightarrow \text{similar targets}}$

(the converse is NOT true)



the **classification** case: essentially true (otherwise generalization needs a huge number of observations, or perhaps the use of 1NN)

Characterization of learning systems

A **similarity measure** is a unique number expressing how "like" two observations are, given the available attributes (features, variables).

Ugly Duckling theorem[1]: learning is impossible without some sort of bias.

"Suppose that one is to list the attributes that plums and lawnmowers have in common in order to judge their similarity. It is easy to see that the list could be infinite: Both weigh less than 10,000 kg (and less than 10,001 kg), both did not exist 10,000,000 years ago (and 10,000,001 years ago), both cannot hear well, both can be dropped, both take up space, and so on. Likewise, the list of differences could be infinite ... any two entities can be arbitrarily similar or dissimilar by changing the criterion of what counts as a relevant attribute."

---

[1] Named after H.C. Andersen's story *"The Ugly Duckling"*.

Similarity measures

Can be defined as an upper bounded, exhaustive and total function

$$s : X \times X \to I_s \subset \mathbb{R}$$

where I_s is a non-empty interval (so $s_{max} := \sup I_s$ exists), such that:

Reflexivity: $s(\boldsymbol{x}, \boldsymbol{y}) = s_{max} \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$

Symmetry: $s(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{x})$

Lower boundedness: $\exists a \in \mathbb{R}$ such that $s(\boldsymbol{x}, \boldsymbol{y}) \geq a$, for all $\boldsymbol{x}, \boldsymbol{y} \in X$ (note this is equivalent to ask that ínf I_s exists)

Closedness: given a lower bounded s, $\exists \boldsymbol{x}, \boldsymbol{y} \in X$ such that $s(\boldsymbol{x}, \boldsymbol{y}) =$ ínf I_s (equivalent to ask that ínf $I_s \in I_s$)

Similarity measures

For every feature, an appropriate similary should be chosen. It is possible that variables of equal types require different similarity measures:

Example 1 *We have a numerical variable counting the number of members in a family, like* $1, 1, 2, 3, 4, 5, 7, 9, 11, 14, 17, 19, 21$.

We have $d(1, 3) = 2 = d(19, 21)$. *However, maybe we would like to regard those families numbering 19,21 members as more similar than those numbering 1,3 (notice that 3 is triple to 1)*

Is this possible? Yes, changing the metric to a non-Euclidean metric like the Clarke metric:*

$$d(x, y) = \frac{|x - y|}{x + y}$$

Now $d(1, 3) = 0{,}5 > d(19, 21) = 0{,}05$.

*This metric requires $x, y > 0$.

<u>Similarity measures</u>

Symmetry

The classical example for a possible lack of **symmetry** is found in the famous statements: "Butchers are like surgeons" as compared to "Surgeons are like butchers"

This is a tricky example!

$\longrightarrow$ the set of features is being changed for the two comparisons.

Similarity measures

Transitivity

Classical example against **transitivity** (or lack of!):

1. Jamaica is like Cuba

2. Cuba is like the Soviet Union

3. (therefore, obviously wrong) Jamaica is like the Soviet Union

This particular example is misleading in at least three ways:

1. Similarities between countries cannot be captured as equivalence relations (two countries not "equal or not equal")

2. The context is being changed from 1. to 2.: in 1., the features being used to establish similarity are geographical (being an island in the Caribbean), whereas in 2. are political (sharing the same political regime)

3. Transitivity need not be constrained to min-trans (the strongest one)

The Gower similarity measure

Choose codomain $I_s = [0, 1]$. For any two data vectors $\boldsymbol{x}_i, \boldsymbol{x}_j$ to be compared on the basis of feature k, a **score** $s_{ijk} := s_k(x_{ik}, x_{jk})$ and a **mask** δ_{ijk} are defined, detailed below:

1. Set $\delta_{ijk} = 0$ when the comparison of $\boldsymbol{x}_i, \boldsymbol{x}_j$ cannot be performed on the basis of feature k for some reason; for example, by the presence of missing values, by the feature semantics, etc

2. Set $\delta_{ijk} = 1$ when such comparison is meaningful

3. If $\delta_{ijk} = 0$ for all the features, then $s(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is undefined

The Gower similarity measure

Binary variables

Binary (dichotomous) variables indicate the presence/absence of a trait, marked by the symbols $+$ and $-$

	Values of feature k			
Observation $\boldsymbol{x}_i$	$+$	$+$	$-$	$-$
Observation $\boldsymbol{x}_j$	$+$	$-$	$+$	$-$
s_{ijk}	1	0	0	0
δ_{ijk}	1	1	1	0

Now this is an example of an *asymmetric* measure

The Gower similarity measure

Categorical variables can take a number of discrete values, which are commonly known as *modalities*. For these variables no order relation can be assumed. Their *overlap* is:

$$s_{ijk} := \begin{cases} 1, & \text{if } x_{ik} = x_{jk}; \\ 0, & \text{if } x_{ik} \neq x_{jk} \end{cases}$$

Real-valued variables are compared with the standard metric in $\mathbb{R}$:

$$s_{ijk} := 1 - \frac{|x_{ik} - x_{jk}|}{R_k},$$

where R_k is the *range* of feature k (the difference between the maximum and minimum values).

The Gower similarity measure

The overall **coefficient of similarity** is defined as the average score over all partial comparisons:

$$s_{ij} := s(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\sum_{k=1}^{d} w_k s_{ijk} \delta_{ijk}}{\sum_{k=1}^{d} w_k \delta_{ijk}}, \qquad \boldsymbol{x}_i, \boldsymbol{x}_j \in X$$

where $w_k \geq 0$ are optional weights.

The Gower similarity measure

Theorem 2 (Gower, 1971) *The matrix $S = (s_{ij})$ is positive semi-definite (PSD) if and only if there are no missing values in X.*

This property *may* be lost when there are missing values: consider three observations in $[1,5]^4$, that is, $R_k = 4$ as in:

Feature no.	#1	#2	#3	#4
observation $\boldsymbol{x}_i$	1.0	2.0	3.0	1.0
observation $\boldsymbol{x}_j$	1.0	3.0	3.0	?
observation $\boldsymbol{x}_l$	1.0	3.0	3.0	5.0

Example data. The symbol ? denotes a missing value.

The Gower similarity measure

$$S = \begin{pmatrix} 1 & \frac{11}{12} & \frac{11}{16} \\ \frac{11}{12} & 1 & 1 \\ \frac{11}{16} & 1 & 1 \end{pmatrix}, \qquad \det(S) = -\frac{121}{2304} < 0$$

and therefore S is not PSD. However, if we replace ? by *any* value in $[1, 5]$, then the matrix S is certainly PSD.

Some theoretical background

Definition 3 (Euclidean metric) *Call* $D = (d_{ij})$ *a dissimilarity matrix if* $d_{ii} = 0$ *and* $d_{ij} = d_{ji}$*. Let* $d : X \times X \to \mathbb{R}$ *be a metric (distance function); then* d *is Euclidean if for any positive* $N \in \mathbb{N}$ *and every choice of observations* $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ *forming its associated dissimilarity matrix* $D_{N\times N} = (d(\boldsymbol{x}_i, \boldsymbol{x}_j))$*, there exists a configuration of points* $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$ *in* $\mathbb{R}^M$*,* $M \leq N$*, such that* $d_{ij} = ||\boldsymbol{z}_i - \boldsymbol{z}_j||_2$*.*

Theorem 4 (Gower and Legendre, 1986) *Consider a similarity matrix* $S = (s_{ij})$*, with* $s_{ij} \in [0, 1]$ *and* $s_{ii} = 1$*, then* S *is PSD iff the matrix* $D = (d_{ij})$ *is Euclidean, with* $d_{ij} = (1 - s_{ij})^{\frac{1}{2}}$*.*

Similarity networks

Overview

data $X \xrightarrow{s_1}$ similarities between $X \xrightarrow{s_2}$ similarities between similarities $\xrightarrow{\mathcal{L}}$ targets Y

$\uparrow$ `original features`

$\uparrow$ `features are similarities`

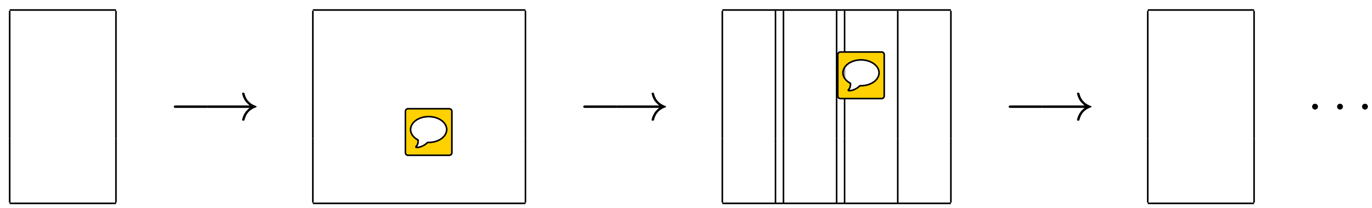
$\uparrow$ `features are similarities (higher-order)`

$\nwarrow \nearrow$

reduction process (along the way)

Similarity networks

Reduction process



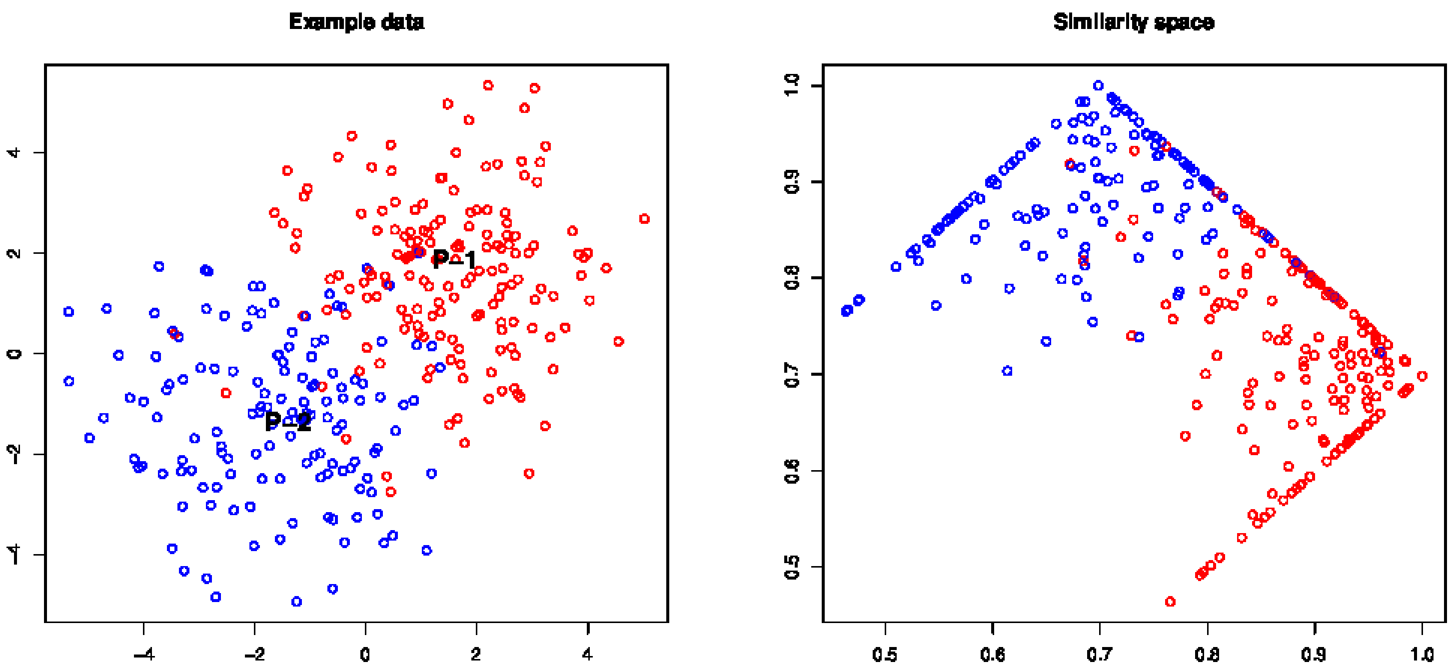
$X_{N\times d}$ $\qquad$ $S_{N\times N}$ $\qquad\qquad$ $S_{N\times d'}$

successive selection of prototypes of *decreasing* size (new features)

Obvious choice: clustering (hierarchical, probabilistic)

The result of this process is a layer of d' units, which we call S-neurons. Any learning method can now operate in the representation space spanned by the layer of d' S-neurons

Similarity measures



Left: simple 2-D problem with two classes in red and blue. The two chosen prototypes are marked as **P-1** and **P-2**. Right: similarity representation using the two prototypes.

Advantages

1. Similarities can handle non-standard data sources and special situations

2. Sparsity control (amount of reduction) can be learner-dependent

3. If the learner $\mathcal{L}$ is linear, training is fast, deterministic and optimal

4. Interpretability is greatly improved:

   a) similarities to a reduced set of known observations

   b) network weights are of the same type as corresponding inputs

   c) user can inject knowledge and play an active role

Clustering

Open questions:

1. How many clusters?

2. Which clustering method?

Clustering

Partitioning Around Medoids

- Data are clustered into clusters around "medoids" $\rightarrow$ PAM: a more robust version of K-means

- The goal is to find representative observations (medoids) which minimize the sum of the dissimilarities of all observations to their closest representatives (objective function J)

- The PAM algorithm finds a local minimum for J: a solution such that there is no single switch of an observation with a medoid that will decrease the objective

Experiments (I)

We study three approaches:

raw There is no effort in identifying variable types (all information is considered numerical, and scaled); missing values are either not identified or left as they come (for example, treated as zeros).

std All variable types are properly identified; non-numerical information is binarized with a standard dummy code. Missing values are identified and imputed with MICE.

sim Same as before with a first layer of S-neurons; then PAM selects $d' = \lfloor 0{,}05 \cdot N \rfloor$ prototypes in the learning part. Notice that, in this case, the model has the architecture of a neural network.

Experiments (I)

Name	#Obs	Def.	Missing	In→Out	Data types
PimaDiabetes	768 (500,268)	65.1 %	10.6 %	8 → 2	8R, 0N, 0D
HorseColic-23	363 (295,68)	61.4 %	25.6 %	22 → 3	7R, 7N,8D
HorseColic-24	364 (296,68)	63.5 %	25.6 %	22 → 2	7R, 7N,8D
Audiology	226 (200,26)	66.3 %	2.1 %	31 → 4	0R, 24N, 7D

- #Obs (learning, test)
- Def. (default accuracy)
- Missing (percentage of missing values)
- In→Out (no. of inputs and outputs)
- Data types: (R)eal, (N)ominal, or(D)inal

Experiments (II)

Generalization errors for the **raw** method.

	LogReg	Multinom	SVM	LDA
Pima	0.201	0.187	0.194	0.187
HorseColic-23	–	0.309	0.279	0.279
HorseColic-24	0.176	0.162	0.162	0.162
Audiology	–	0.231	0.154	0.269
AVERAGE	0.189	0.222	0.197	0.224

Generalization errors for the **std** method.

	LogReg	Multinom	SVM	LDA
Pima	0.190	0.198	0.205	0.201
HorseColic-23	–	0.265	0.279	0.353
HorseColic-24	0.147	0.191	0.147	0.147
Audiology	–	0.269	0.038	0.731
AVERAGE	0.169	0.231	0.168	0.358

Generalization errors for the **sim** method.

	LogReg	Multinom	SVM	LDA
Pima	0.183	0.190	0.194	0.175
HorseColic-23	–	0.324	0.294	0.309
HorseColic-24	0.162	0.176	0.176	0.191
Audiology	–	0.115	0.000	0.038
AVERAGE	0.172	0.201	0.166	0.178

Related work? Lots of ...

Pekalska and Duin's work

The Dissimilarity Representation for Pattern Recognition: Foundations and Applications

E. Pekalska and R. Duin (Delft University of Technology, The Netherlands)

"... a fundamentally new approach to pattern recognition in which objects are characterized by sets of dissimilarities to other objects instead of by using features"

Future Work

this is ongoing work ...

- Investigate more than one layer ("deep" architectures)
- Investigate other reduction techniques (supervised?)
- Incorporate weights w_k
- Extend Gower's similarity measure to other data types
- Make the S-neuron non-linear (free parameter)
- Use (more) challenging datasets for experimentation

"Representation of similar real-world objects must be close; there is no ground for any generalization on representations that do not obey this demand."

Computers and Pattern Recognition. Arkadev and Braverman, 1966