Neural Network Classification with Prior Knowledge for the Analysis of Biological Data

Danilo Abbate
Department of Computer Science, University of Bari, Italy

Mario R. Guarracino
High Performance Computing and Networking Institute
National Research Council, Naples, Italy

Altannar Chinchuluun Panos M. Pardalos
Industrial and Systems Engineering Department,
University of Florida, USA

Clustering vs. Classification

- Clustering
  - objects are not labeled with any class information
  - density estimation is performed
  - clusters of objects are constructed based on similarities between their features
- Classification
  - capability of a system to learn from a set of input/output pairs
  - TRAINING SET $\Rightarrow$ input: *vector of features, class label*; output: *classification model*
  - TEST SET $\Rightarrow$ input: *vector of features, classification model*; output: *predicted class label*

Clustering

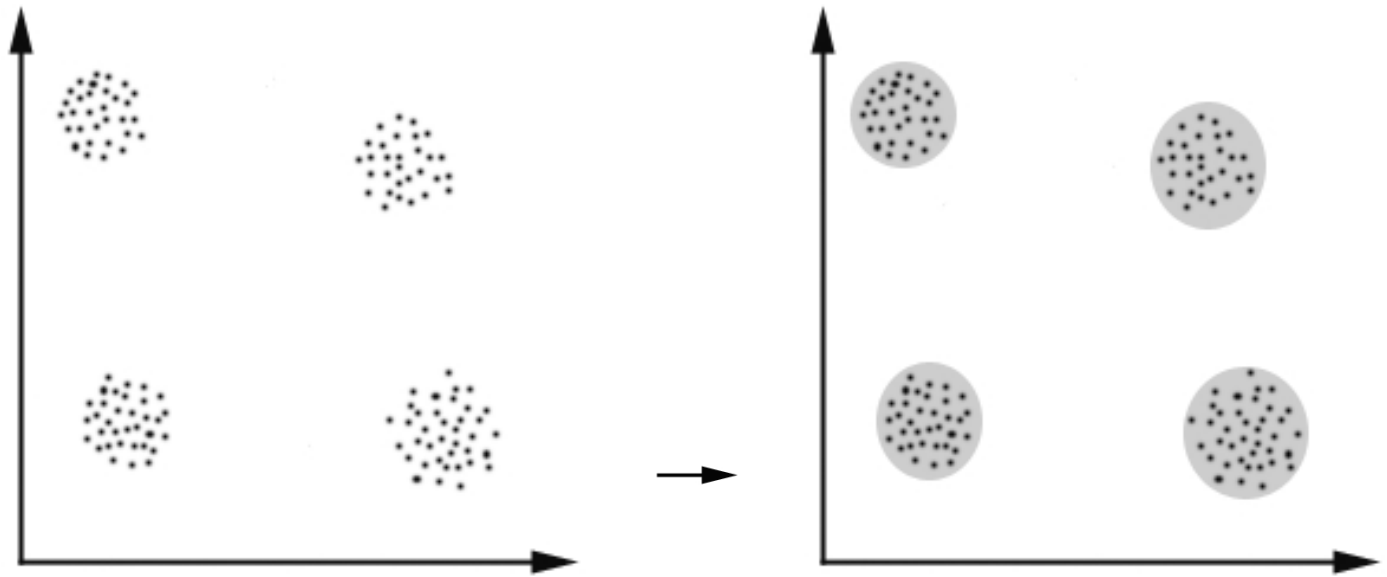
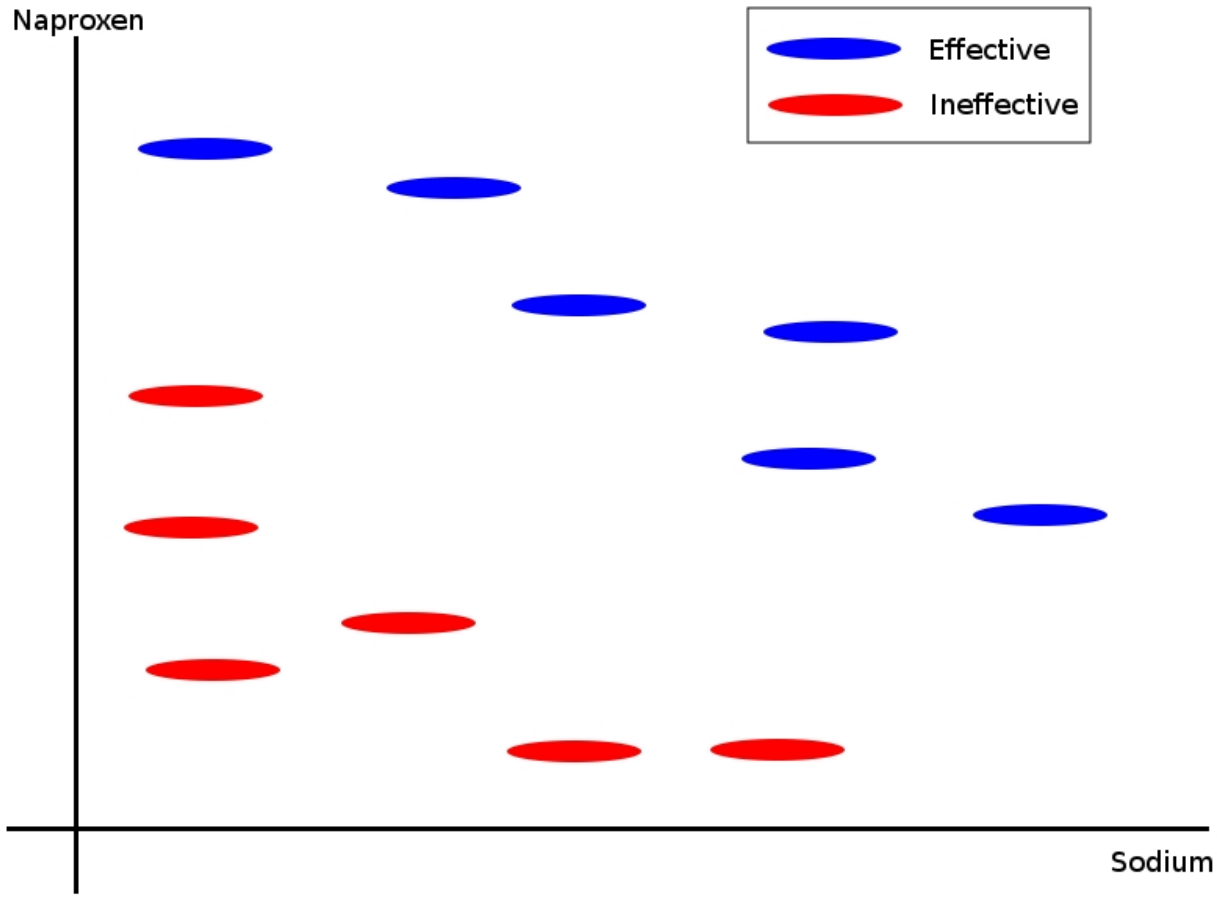
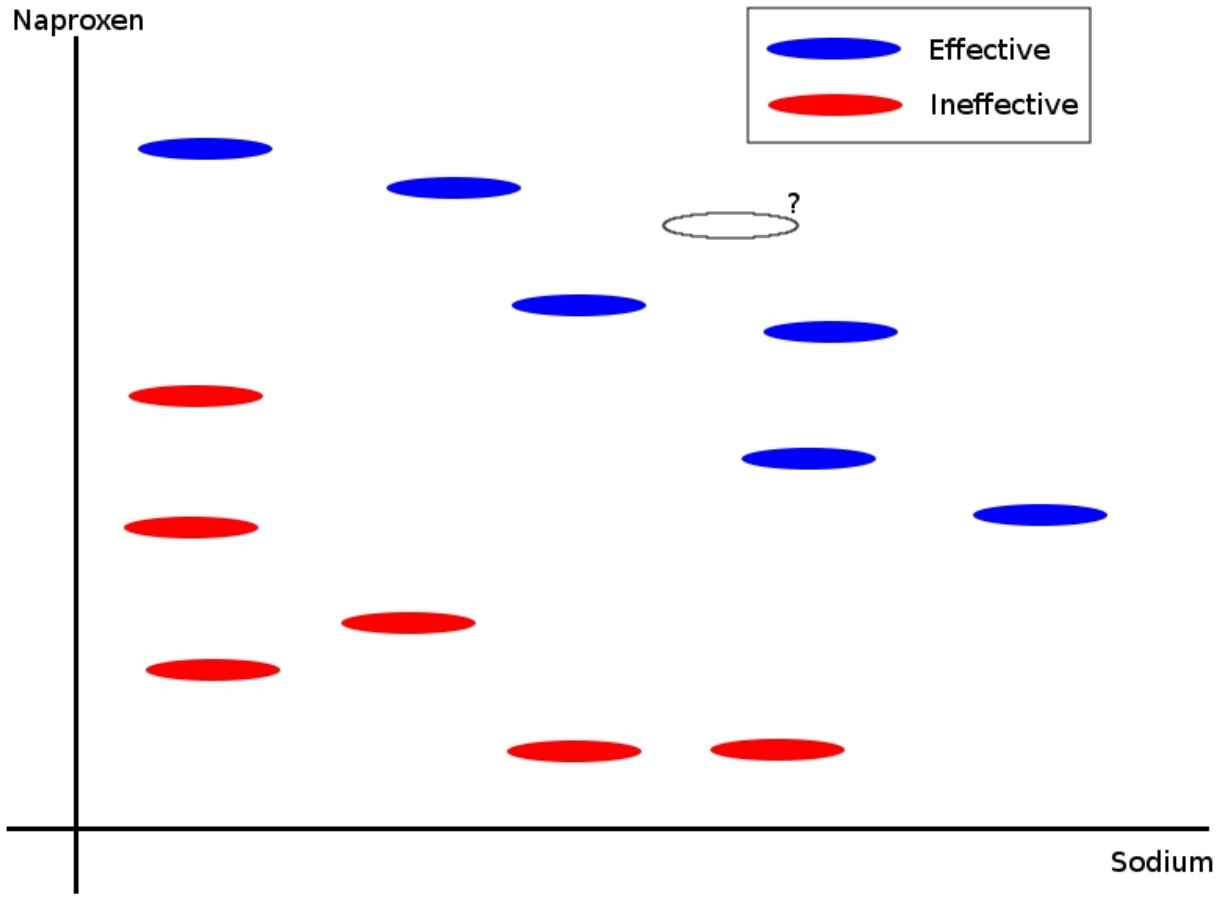


Figure: *k*-means Clustering

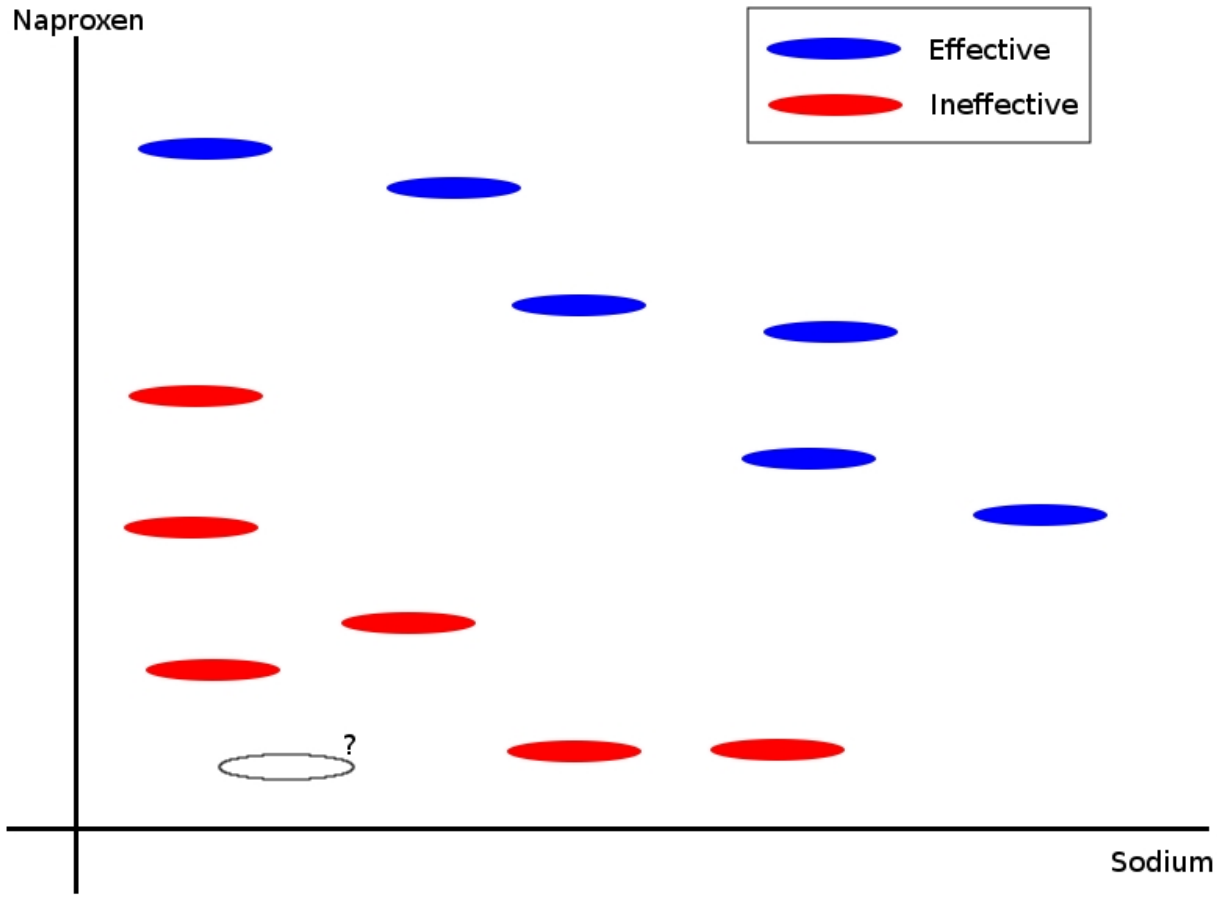
Binary classification



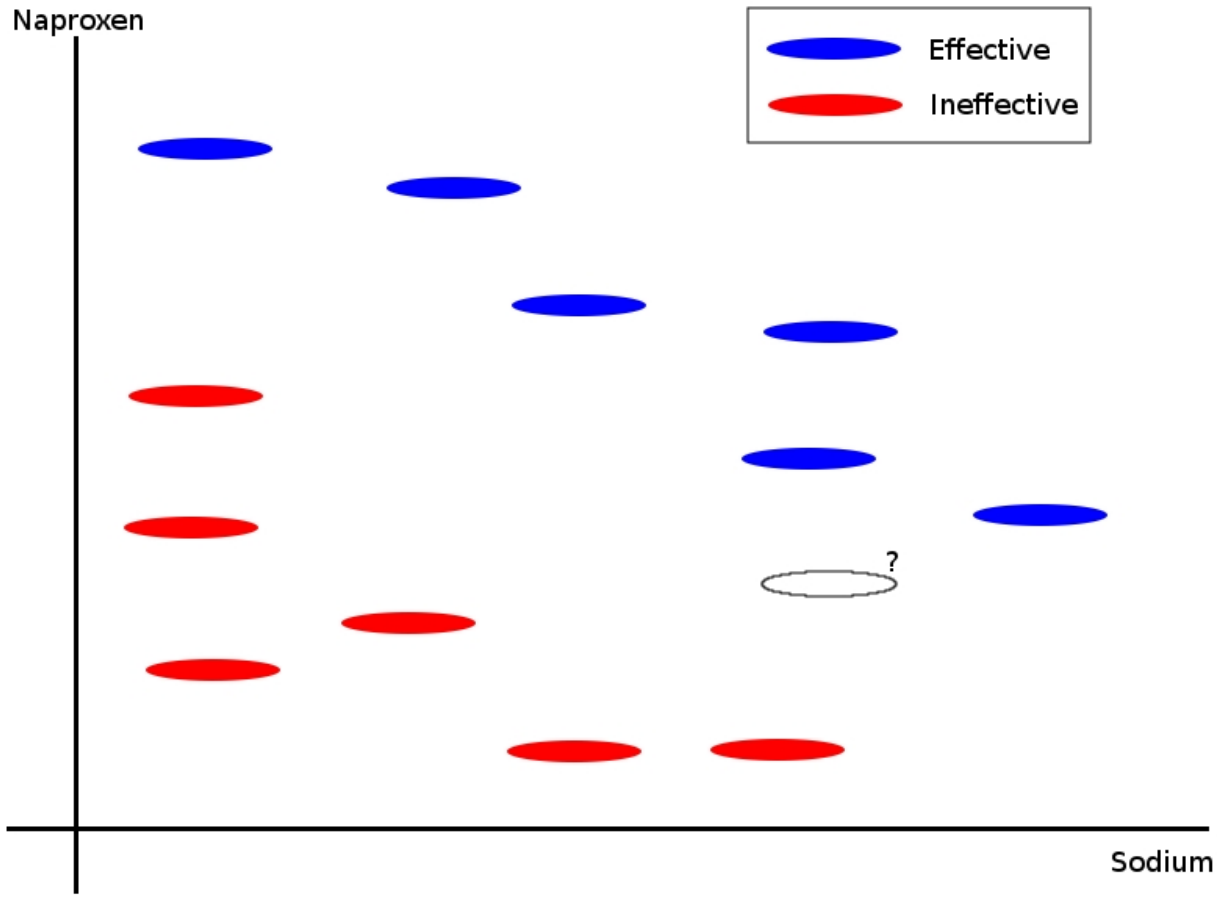
Binary classification



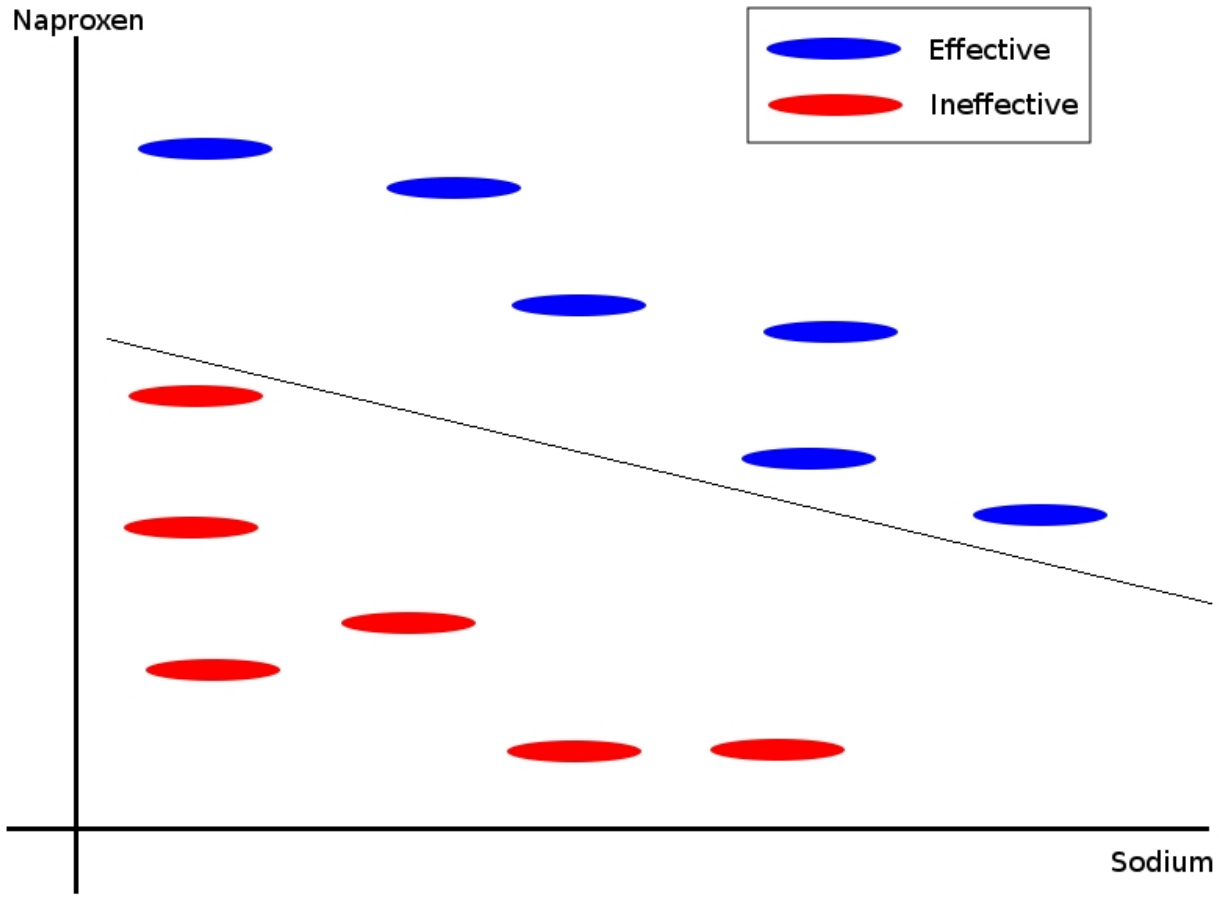
Binary classification



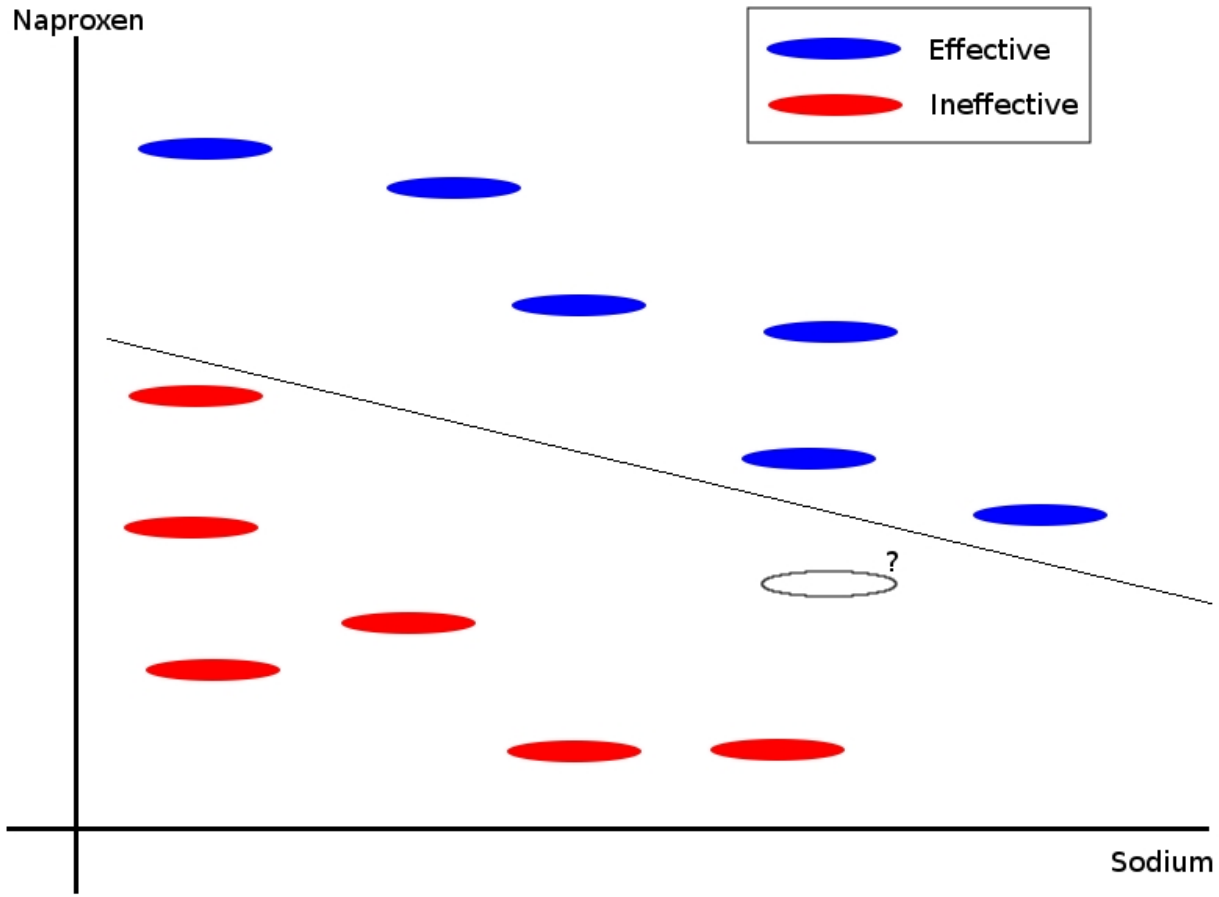
Binary classification



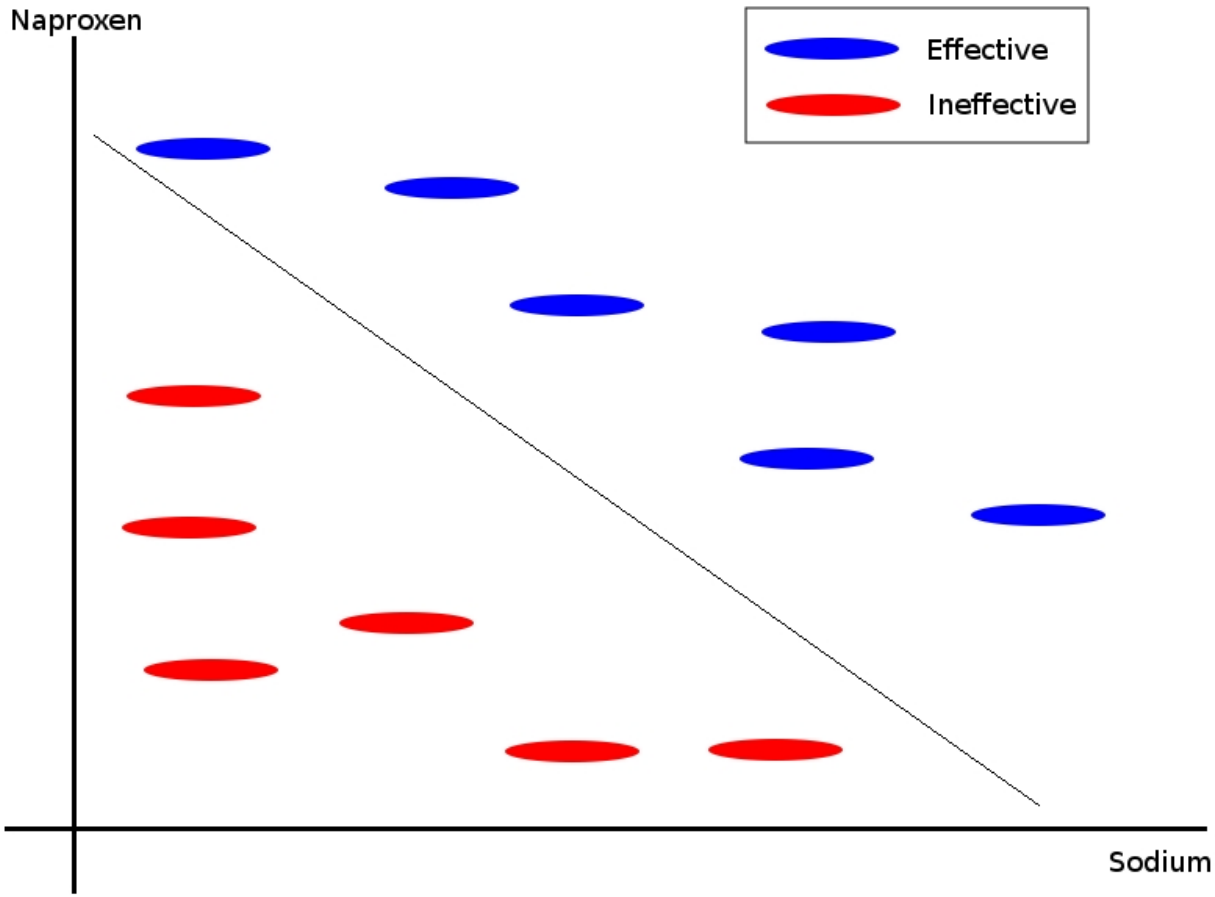
Binary classification



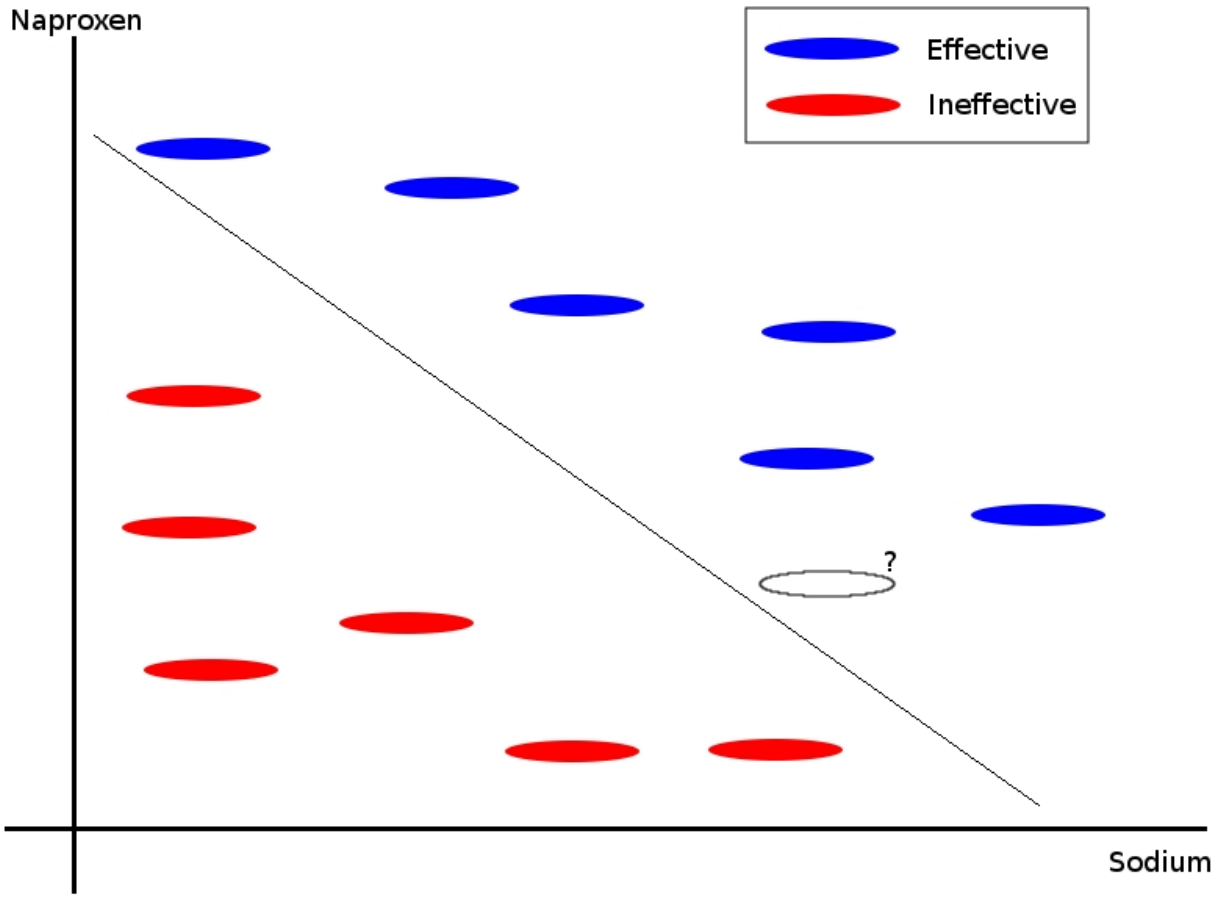
Binary classification



Binary classification



Binary classification



Introduction

- Given a set of data and two classes, -1 and $+1$, the purpose of binary classification is to build a model that divides the set into two disjoint classes, so that each data can be assigned to the correct class.
- Classification methods have been successfully applied for loan applications by banks, fiscal evasion by the Internal Revenue Service, face detection and optical character recognition in security applications.
- The most promising applications of those methods are in the field of biomedicine and bioinformatics.

Biomedical applications

- In biomedicine, binary classification is used for example in medical prognosis, to predict whether a patient will have a recurrence of the disease after a fixed time interval.
- From a mathematical point of view, given a set of points $\Gamma \subset \mathbb{R}^n$, a binary classifier is a function:

$$h(x) : \Gamma \subseteq \mathbb{R}^n \to \mathbb{R}, \quad x \in \Gamma,$$

whose sign represents the class of the point x.

Related work

- Examples of classifiers are neural networks, decision trees and support vector machines (SVM).
- The performance of a binary classifier can be evaluated through:
  - Misclassification error represents the percentage of misclassified samples.
  - Sensitivity is the percentage of true positives among all positives tested.
  - Specificity is the percentage of true negatives among all negatives tested.
- Most of the real world problems deal with irregular and noisy data for which optimal classification accuracy is hard to achieve.

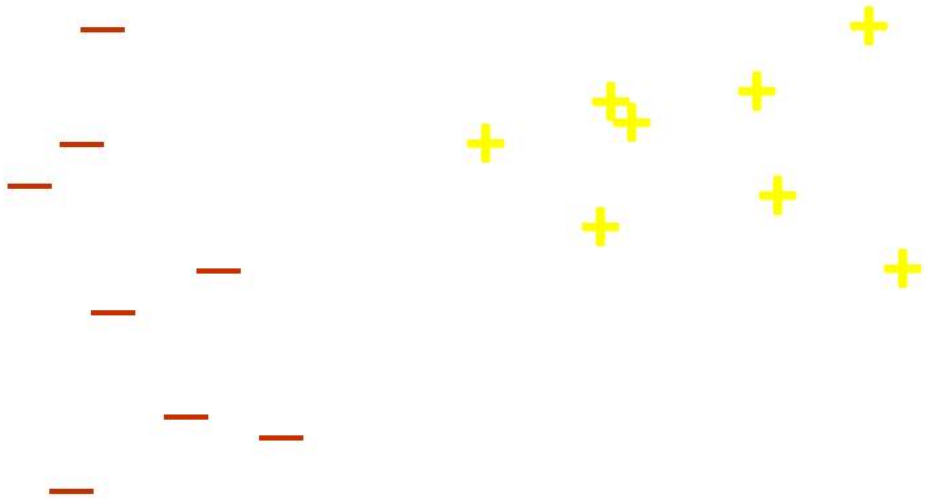
Related work

- A natural approach to use prior knowledge in a classifier is to add more points to the data set.
- This usually results in higher computational complexity and overfitting.
- On the other hand, Lee and Mangasarian propose to analytically express prior knowledge as additional constraints to the cost function of the optimization problem defining SVM.
- This solution has the advantage of not changing the dimension of the training set, and it avoids poor generalization of the classification model.

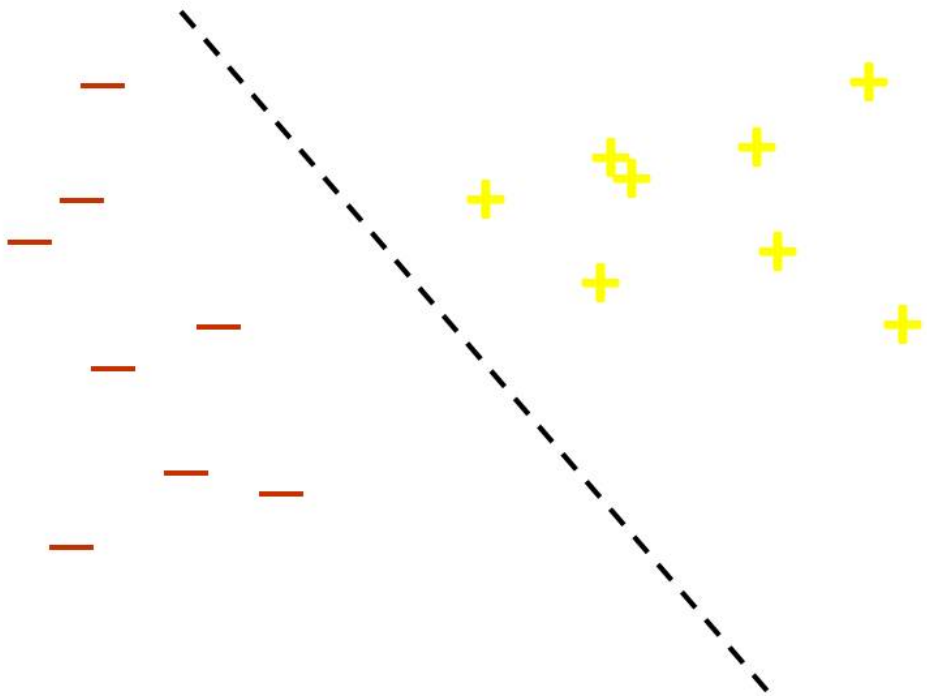
Contribution

- In this work, we show how prior knowledge can be applied to neural network classifiers.
- In particular, we will focus on Radial Basis Function Neural Networks (RBF-NN).
- Computational properties of RBF-NN make it a good candidate to understand how prior knowledge can be used to improve classification methods.
- We show that the proposed method with prior knowledge can substantially increase the original neural network accuracy.

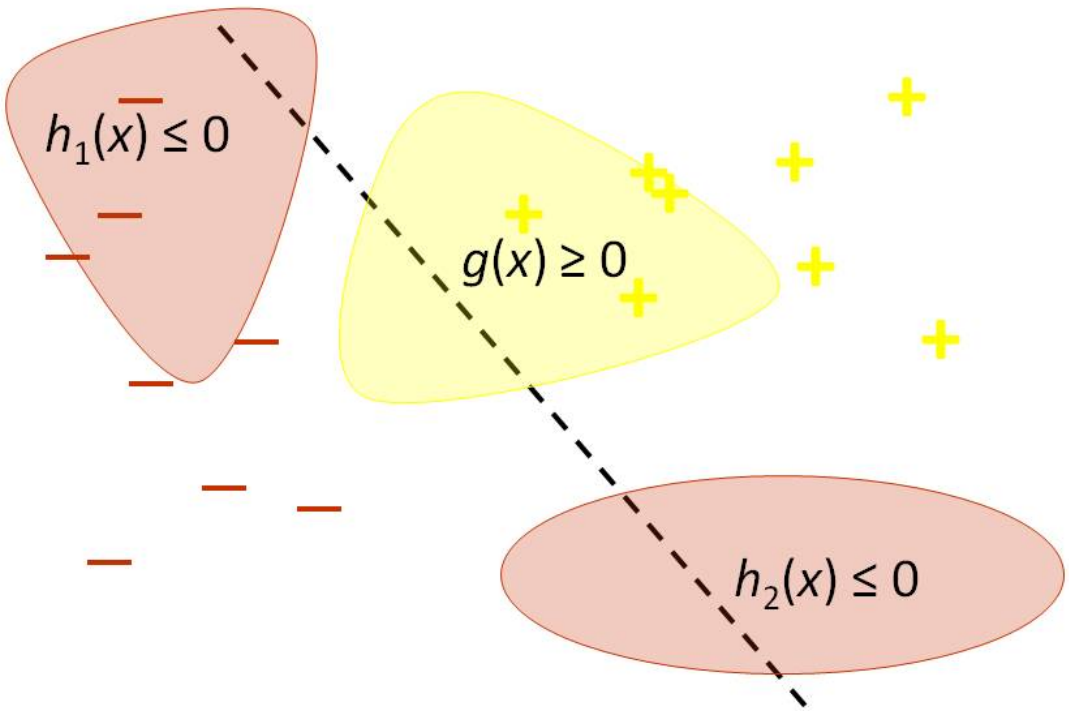
Prior knowledge as knowledge regions



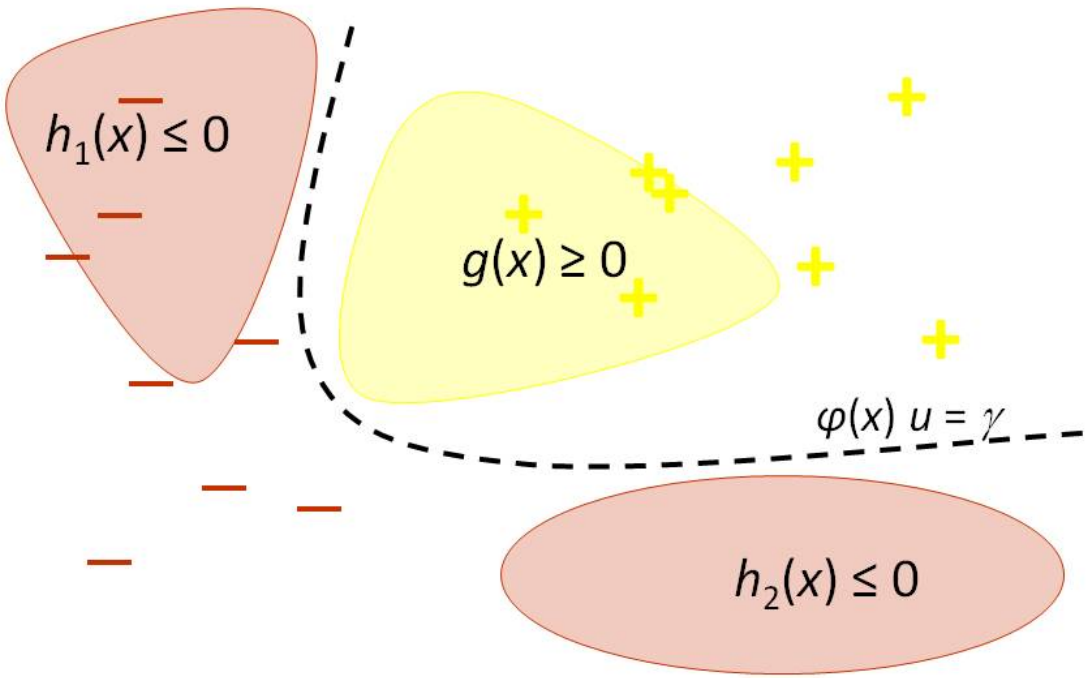
Prior knowledge as knowledge regions



Prior knowledge as knowledge regions



Prior knowledge as knowledge regions



RBF Neural Networks

- Radial Basis Function neural networks are a particular kind of neural network.
- They can approximate a continuous function defined on a compact set with a linear combination of radial basis functions

$$\varphi(x) = e^{-\frac{x^2}{2\sigma}}$$

- A RBF neural network is composed of three layers:
  1. an input layer
  2. an hidden layer of RBF nodes
  3. a layer of output nodes with linear activation function.

RBF Neural Networks

- Let y_i be the class label of input vector $x_i \in \mathbb{R}^n$, with $i = 1, \ldots, m$. The output of the network with input x_i will be represented by the function:

$$h(x_i) = \Sigma_{j=1}^{m} w_j \varphi(\|x_i - t_j\|),$$

  where t is the mean of the activation function evaluated on each point of the training set.
- Exact interpolation is obtained when $y_i = h(x_i), i = 1, \ldots, m$.

RBF Neural Networks

- The hidden layer, as it is based on neurons with a radial basis activation function, creates a response localized on the input vector x; the binary output will then be calculated as a weighted sum of these localized responses.
- Training a RBF network is a procedure divided into two phases:
  1. With an unsupervised learning technique, the parameters of the radial bases function σ and t are calculated.
  2. Values of the weights w, which determine the binary output y, are then computed.

RBF Neural Networks

- Traditionally there are two strategies for this first phase of unsupervised learning.
- The classic strategy calculates these parameters through different clustering techniques.
  - These aim to divide the training set into a fixed amount of homogeneous groups, organized according to the distance of the points in the training set.
- Besides clustering, it is possible to have an incremental approach.
  - In this way, one seeks to reduce the mean quadratic error under a threshold ϵ by adding nodes to the hidden layer.

RBF Neural Networks

- In the second part of the training, we search for values of the weights w which determine the binary output y.
- Such weights are calculated by minimizing the following error function:

$$E = \frac{1}{2} \sum_{i=1}^{m} (h(X_{i.}) - y_i)^2 \tag{1}$$

  which tells the distance of the actual solution from the desired one.
- Prior knowledge is added by a modification to this phase.

Prior knowledge in RBF Neural Networks

- Prior knowledge is then added as a set of constraints to obtain the following minimization problem:

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i=1}^{m}(h(X_{i.}) - c_i)^2 \\ s.t. \quad & Bw \geq 0. \end{aligned} \tag{2}$$

- The constraints of this problem force the hyperplane solution of the above equation to leave the p points represented by the matrix $B \in \mathbb{R}^{p\times n}$ in one halfspace.

A case study

- The method has been tested on the Wisconsin Prognostic Breast Cancer data set, from UCI repository.
- Results are compared with SVM using misclassification error, sensitivity and specificity.
- SVM results taken from Mangasarian and Wild (2006), results for RBF neural networks evaluated using a GNU/linux PC, kernel 2.6.9-42 with AMD Opteron 64 bits of the series 284 (2.2GHz), 4 Gigabyte RAM.
- Codes implemented with Matlab (R2006b).
- Accuracy, sensitivity and specificity were calculated upon Leave-One-Out (LOO) classification.
  - Parameters are computed through a ten-fold cross validation grid-search.

A case study

- The Wisconsin Prognostic Breast Cancer data set provides 30 cytological features plus tumor size and the number of metastasized lymph nodes.
- For each of the 198 patients, it provides the number of months before a new cancer has been diagnosed.
- If there has been no recurrence, the data set contains information on how long the patient has been under analysis.

A case study

- In our work, we want to identify those patients which had a recurrence in a period of 24 months, discriminating them from those who did not have any recurrence.
- This is a subset *Upsilon* of the data set:

$$\Upsilon = \{x \in WPBC | \text{ property p holds for x}\}$$

where the property p is defined as follows:

$$\text{p holds iff} \begin{cases} \text{the patient has a recurrence in the 24 month period,} \\ \text{the patient had not recurrence.} \end{cases}$$

- After this filtering, the remaining data set contains 155 patients.

A case study

To simulate the expertise of a surgeon, we used the same areas identified by Managasarian and Wild and described by the following formulas:

$$\left\| \begin{pmatrix} (5.5)\, x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} (5.5)\, 7 \\ 9 \end{pmatrix} \right\| + \left\| \begin{pmatrix} (5.5)\, x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} (5.5)\, 4.5 \\ 27 \end{pmatrix} \right\| - 23.0509 \leq 0$$

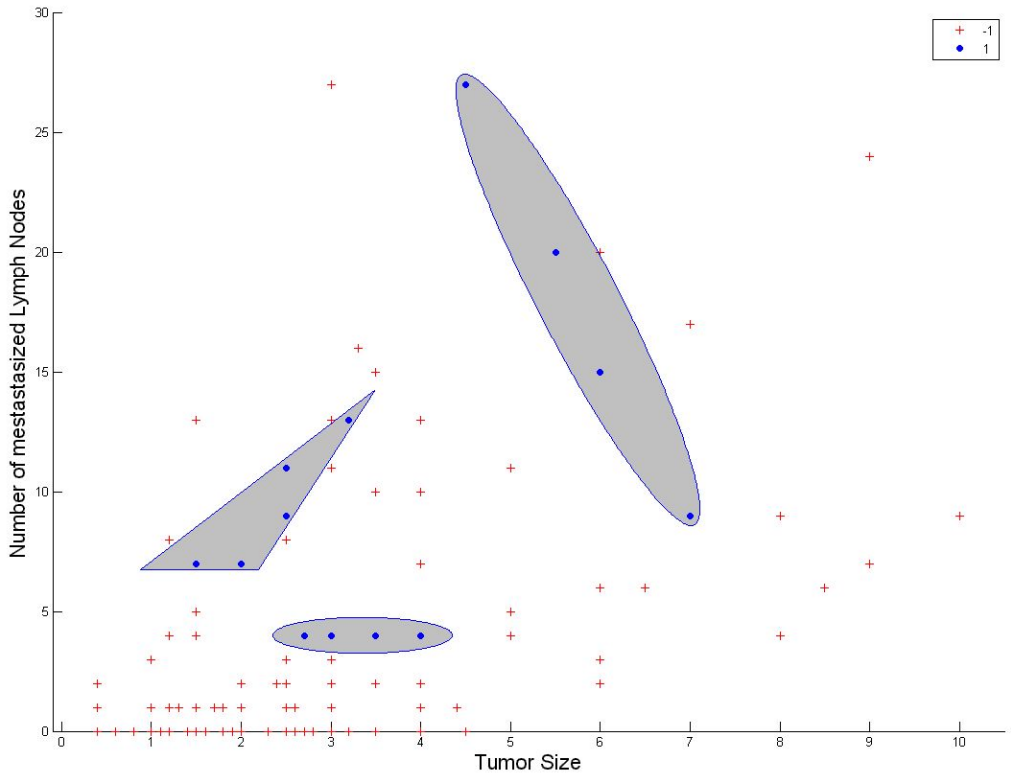
$$\Rightarrow f(x) \geq 1$$

$$\begin{pmatrix} -x_2 + 5.7143x_1 - 5.75 \\ x_2 - 2.8571x_1 - 4.25 \\ -x_2 + 6.75 \end{pmatrix} \leq 0 \Rightarrow f(x) \geq 1$$

$$\frac{1}{2}\left(x_1 - 3.35\right)^2 + \left(x_2 - 4\right)^2 - 1 \leq 0 \Rightarrow f(x) \geq 1.$$

These equations describe three areas in a two dimensional representation of the data set.

A case study


-1
1
Number of mestastasized Lymph Nodes
Tumor Size
0
5
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10

A case study

- The x-axis is the tumor size (the next to last feature of the data set) while the y-axis is the number of metastasized lymph nodes (the last feature of the data set).
- Following the work by Mangasarian and Wild, we decided to take those 14 points which belong to the three areas.
- We note that those 14 points are among the support vectors that have been misclassified by SVM in leave one out validation.

A case study

- So far, the lowest accuracy error was 13.7% (Bennett, 1992). Adding knowledge it decreases to $\approx 9.0\%$.

Classifier	Misclassification error	Sensitivity	Specificity
SVM	0.1806	0	1.000
SVM with knowledge	0.0903	0.5000	1.000
RBF-NN	0.1806	0	1.000
RBF-NN with knowledge	0.0968	0.4643	1.000
Improvement due to knowledge	50.0 %		

Leave One Out misclassification percentage, sensitivity and specificity on WPBC data set (24 months).

Discussion

- When knowledge is used, both methods have the same accuracy, and nearly same values of sensitivity and specificity.
- The slightly lower value of sensitivity is due to the fact that RBF-NN misses one point of class $+1$.
- The accuracy of the classifier, with respect to class -1, is measured in terms of specificity. Note that, with our approach, specificity is maximum.

Numerical experiments

- Thyroid: 215 patients along with 5 cytological features. 65 patients are affected by a thyroid disease (30.23%).
- Heart: 270 patients with 13 characteristics (age, sex, chest pain, blood pressure, ...). 120 patients, 44.44% are patients affected by heart disease.
- Pima Indians diabetes: 768 females, with or without the symptoms of diabetes. Among features, the number of pregnancies, glucose concentration in plasma, and diastolic blood pressure. 268 are positive (34.90%).
- Banana is an artificial data set of 2-dimensional points which are grouped together in a shape of a banana.

Numerical experiments

- We decided to choose points to add in prior knowledge, for each data set, executing a Leave One Out cross validation and chosing the misclassified points.
- In the next table, we report classification accuracy, sensitivity and specificity for RBF-NN with and without prior knowledge.

Numerical experiments

	Results without knowledge		
Data set	Misclassification error	Sensitivity	Specificity
Thyroid	0.1488	0.5538	0.9800
Heart	0.1926	0.7833	0.8267
Diabetes	0.3216	0.6493	0.6940
Banana	0.1399	0.8304	0.8829
	Results with knowledge		
Data set	Misclassification error	Sensitivity	Specificity
Thyroid	0.0977	0.7231	0.9800
Heart	0.1296	0.8500	0.8867
Diabetes	0.2227	0.8731	0.6940
Banana	0.1110	0.8565	0.8967

Leave One Out misclassification error percentage, sensitivity and specificity on different data sets.

Conclusions and future work

- We have proposed a method to incorporate prior knowledge in Radial Basis Function neural networks.
- The accuracy of the new algorithm well compares with other methods and the accuracy is improved with respect to neural networks without knowledge.
- Further investigation will be devoted to the identification of knowledge regions, in order to improve generalization of the classification model.
- We will investigate how the expression of a prior knowledge in terms of probability of a patient to belong to one class can affect classification models.


K. P. Bennett.
Decision tree construction via linear programming.
In *Procedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference (Utica, Illinois)(M.Evans,ed.)*, pages 82–90, 1992.


C. M. Bishop.
Neural networks for pattern recognition.
Oxford Press, 1995.


A. Bjorck.
Numerical Methods for Least Squares.
SIAM, Philadelphia, 1996.


G. H. Golub and C. F. van Loan.
Matrix Computation.
John Hopkins University Press, 1996.


M.R. Guarracino, D. Abbate, and R. Prevete.
Nonlinear knowledge in learning models.
In *Workshop on Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery, European Conference on Machine Learning*, pages 29–40, 2007.


Y. Lee and O. L. Mangasarian.
Ssvm: A smooth support vector machine for classification, 1999.


O. L. Mangasarian and E. W. Wild.
Nonlinear knowledge-based classification.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(1):68–74, 2006.


MATLAB.
User's guide.

The Mathworks, Inc., Natick, MA 01760, 1994–2006.
http://www.mathworks.com.

B. Scholkopf and A. J. Smola.
Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.
MIT Press, Cambridge, MA, USA, 2001.