Eliciting Fuzzy Knowledge from the PIMA Dataset

Antonio d'Acierno
ISA – CNR
dacierno.a@isa.cnr.it

Giuseppe De Pietro, Massimo Esposito
ICAR – CNR
giuseppe.depietro@na.icar.cnr.it
massimo.esposito@na.icar.cnr.it

Our work

- Recently, we proposed a six-steps a data driven methodology to automatically build fuzzy inference systems [6].
  - The methodology produces FIS with an user defined number of rules.
  - Each step can be approached using several strategies
- In this paper, we use an implementation of our methodology to elicit knowledge from the PIMA dataset
- We obtain:
  - an interesting performance in terms of correct classification rate
  - linguistic variables are likely to be easily understood from human beings.

[6] A. d'Acierno, G. De Pietro, M. Esposito. Data Driven Generation of Fuzzy Systems: An Application to Breast Cancer Detection, 7th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2010), Palermo (Italy), 16-18 September 2010.

Introduction

- To increase the change of successful treatments, early detection of almost any disease is a key factor.
- The detection can be often formulated as a binary decision making problem:
  - uncertainty in form of information incompleteness, impreciseness, fragmentariness, not fully reliability, vagueness and contradictoriness often affects these problems.
- Computerized diagnostic tools to support physicians in interpreting data have been thus developed
  - Diagnostic Decision support Systems (DDSS)

DDSS

- A diagnostic tool should possess [1] three (often in contrast) characteristics:
  - it must attain the best possible performance in terms of correct classification rate.
  - It would be desirable the system not only provides a diagnosis but also a numerical value representing the degree to which the system is confident in the solution.
  - It would be also useful if the physician is not faced with a black box that simply outputs answers but the system should provide some insight into how the solution has been derived (interpretability).

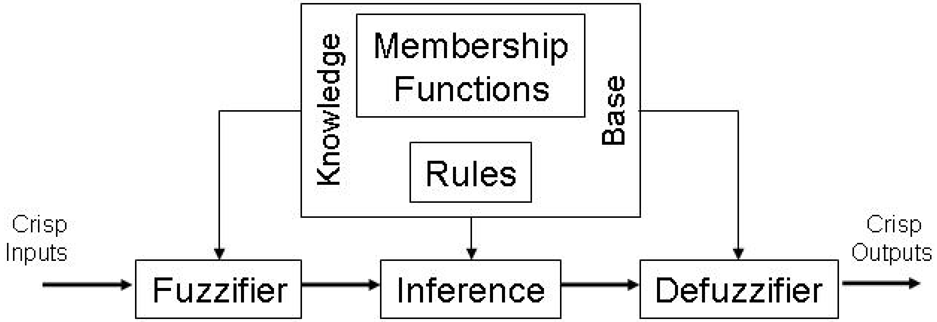
[1] C. A. Pena-Reyes and M. Sipper. A fuzzy-genetic approach to breast cancer diagnosis. Artificial Intelligence in Medicine, 17(2):131–155, 1999.

- Diagnostic tools, however, typically have unequal classification error costs so that straight CR cannot be assumed as a careful measure of the goodness of the classifier.
- A Receiver Operating Characteristic (ROC) graph has been showed to be a more accurate technique for selecting classifiers based on their performance.
- We guess that also the confidence χ can be used for selecting classifier since a good classifier should be highly confident with correctly classified examples while it should be "doubtful" with misclassified data points.

FIS

- A Fuzzy Inference System (FIS) is a system that (tries to) solve a (typically complex and nonlinear) problem by utilizing fuzzy logic methodologies and it is composed of

1. a fuzzifier (translates real-valued inputs into fuzzy values)
2. an inference engine (applies a fuzzy reasoning mechanism to obtain a fuzzy output),
3. a defuzzifier (translates this latter output into a crisp value),
4. of a knowledge base (containing both rules and membership functions).



FIS

- The inference process is performed by the engine using the rules contained in the rule base

  if antecedent then consequent

- The antecedent is a fuzzy-logic expression composed of one or more simple fuzzy expressions connected by fuzzy operators (the fuzzy equivalent of the classical and, or and not),
- In Mamdani systems, the consequent is an expression that assigns fuzzy values to the output:

  if service is good then tip is average

- In Takagi-Sugeno(TS) systems, the consequent expresses output variables as a function that maps the input space into the output space:

  if service is good then tip = f(service)

  where f is (typically) a first order linear function that becomes a constant in zero-order TS systems.

Fuzzy modeling

- Fuzzy modeling is the task of identifying the parameters of a FIS so that a desired behavior is attained.
- Knowledge driven approach:
  - When the available knowledge is complete and the problem space is not very large the system can be constructed directly using knowledge elicited from human experts.
- Alternatively, data driven fuzzy modeling can be used:
  - Available data and AI technologies are used to build the rules and/or membership functions.
- A problem: the knowledge base generated automatically from data may not be fully interpretable.

Interpretability

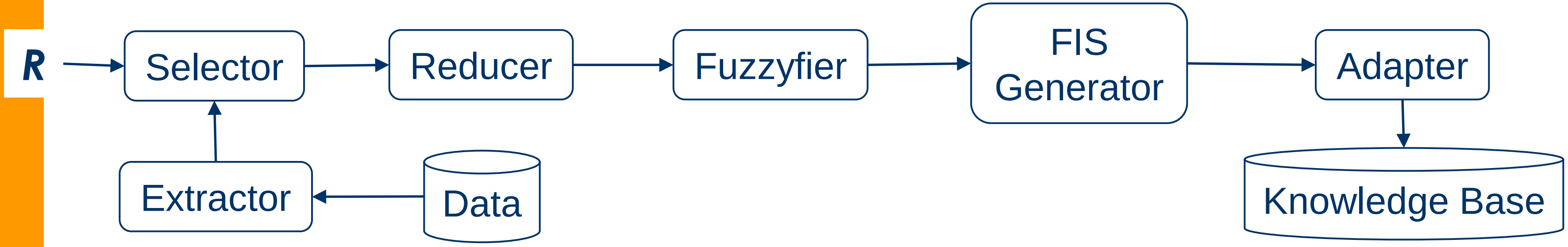
- Three conditions can be defined to obtain an interpretable fuzzy model [5]:

1. the fuzzy sets can be interpreted as linguistic labels (low, medium, high, medium-low, etc);
2. the set of rules must be as small as possible;
3. the if-part of the rules should be derived from a subset of independent variables rather than from the full set.

- Interpretability is a key feature in a DDSS.

5. Serge Guillaume. Designing fuzzy inference systems from data: An interpretability-oriented review. IEEE Transactions on Fuzzy Systems, 9(3):426–443, 2001.

The 6-steps methodology



- We extract crisp rules.
- If the case, we select R useful rules.
- If the case, we reduce the rules.
- Using the fuzzyfier, we build fuzzy rules.
- We generate the FIS (TS systems are used).
- We adapt membership functions.

The implemented methodology

- To extract crisp rules we use a full decision tree (without pruning) with a Gini's diversity index as split criterion and a split minimum factor equal to 1.
- Given a user defined number of rules R (assumed to be even), we select, for each class, the R/2 most covering leaf nodes.
- Each rule is simplified so that its antecedent contains each feature at most one time; three operator (atmosT9 are considered:
  - greater than a threshold (→ high)
  - less than a threshold (→ low)
  - between two thresholds (→ medium)
- ANFIS [7] is used to adapt membership functions

[7] J. S. R. Jang. Anfis: Adaptive network based fuzzy inference system. Systems, Man and Cybernetics, IEEE Transactions on, 23(3):665-685, May/June 1993.

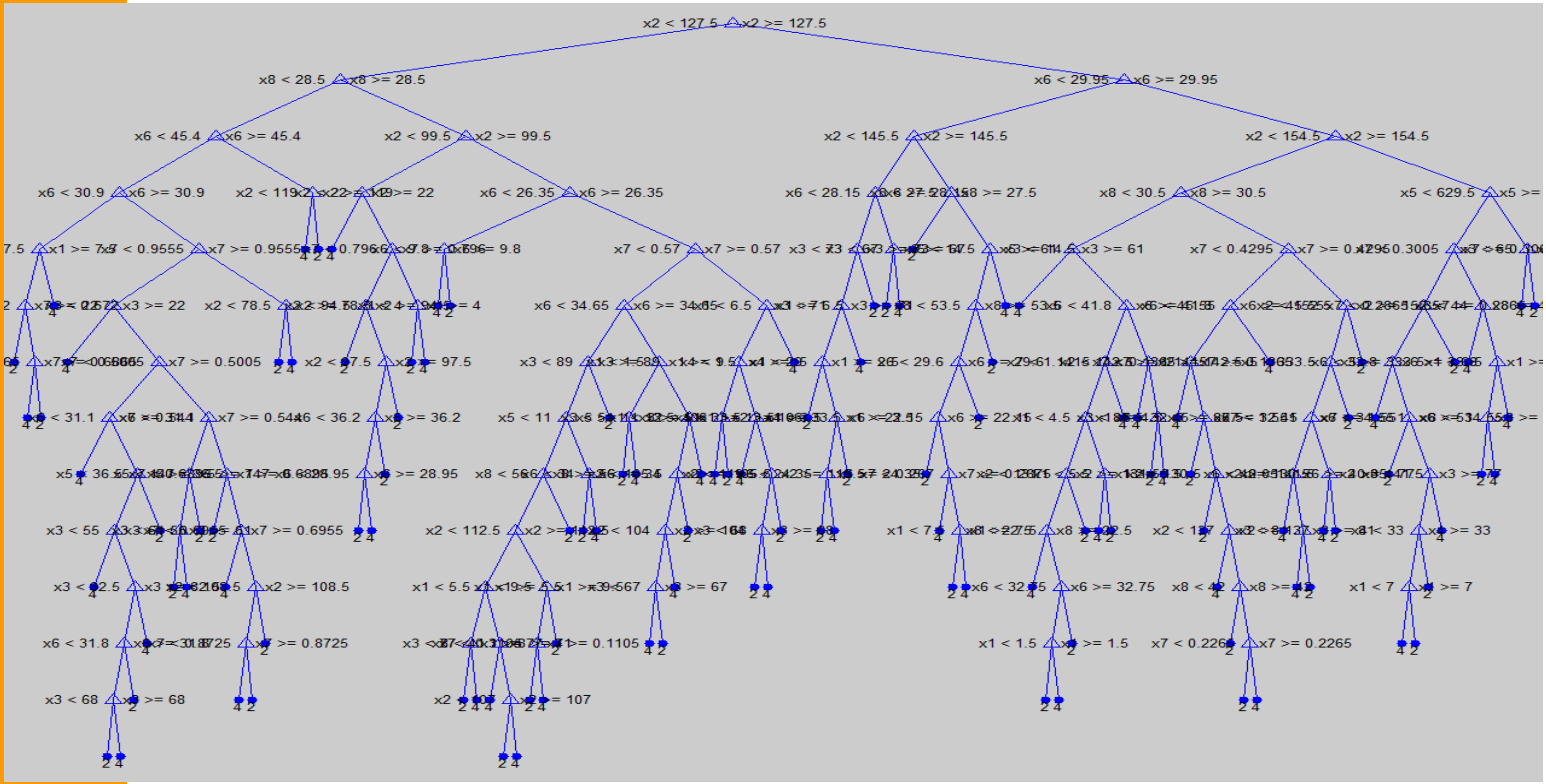
The PIMA Dataset

- A (complex) collection of medical diagnosis reports of 768 examples from a population living near Phoenix, Arizona, USA
- Patients here are females at least 21 years old of Pima Indian heritage.
- Binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria;
- There are 500 negative examples and 268 positive ones and for each patient there are 8 independent variables reported:
  1. number of times pregnant;
  2. plasma glucose concentration a 2 hours in an oral glucose tolerance test;
  3. diastolic blood pressure;
  4. triceps skin fold thickness;
  5. 2-Hour serum insulin;
  6. body mass index;
  7. diabetes pedigree function;
  8. age.

Experiments

- The proposed approach has been implemented using functions available in the standard version of the R2009a 64-bit version of MATLAB.
- We use a ten-fold cross validation that is repeated 10 times.
- We use TS systems.
- We use a threshold to classify the sample:
  - We choose the threshold that maximizes the classification rate on the learning set.
- We measure:
  - the average CR on the learning set (LS) and on the test set (TS) for both untrained FISs and adapted ones.

Preliminary results



100% on the learning set, 70% on the test set

Preliminary results

	UFIS		AFIS	
Rules	LS	TS	LS	TS
2	63,70%	63,50%	75,30%	73,80%
4	63,90%	63,90%	76,10%	74,20%
6	64,30%	64,30%	76,60%	74,10%
8	64,80%	64,80%	76,90%	74,40%
10	64,80%	64,70%	77,20%	74,50%
12	64,80%	64,60%	77,60%	74,90%
14	65,10%	65,00%	77,60%	75,00%
16	65,10%	65,10%	77,80%	75,00%
18	65,10%	65,00%	77,90%	74,90%
20	65,30%	65,10%	77,90%	74,90%

- An interesting performance is obtained using just two rules.
- The performance increases using more rules.
- In [6] it is obtained a 77.65% CR using 125 rules in a single run of a five fold cross validation.

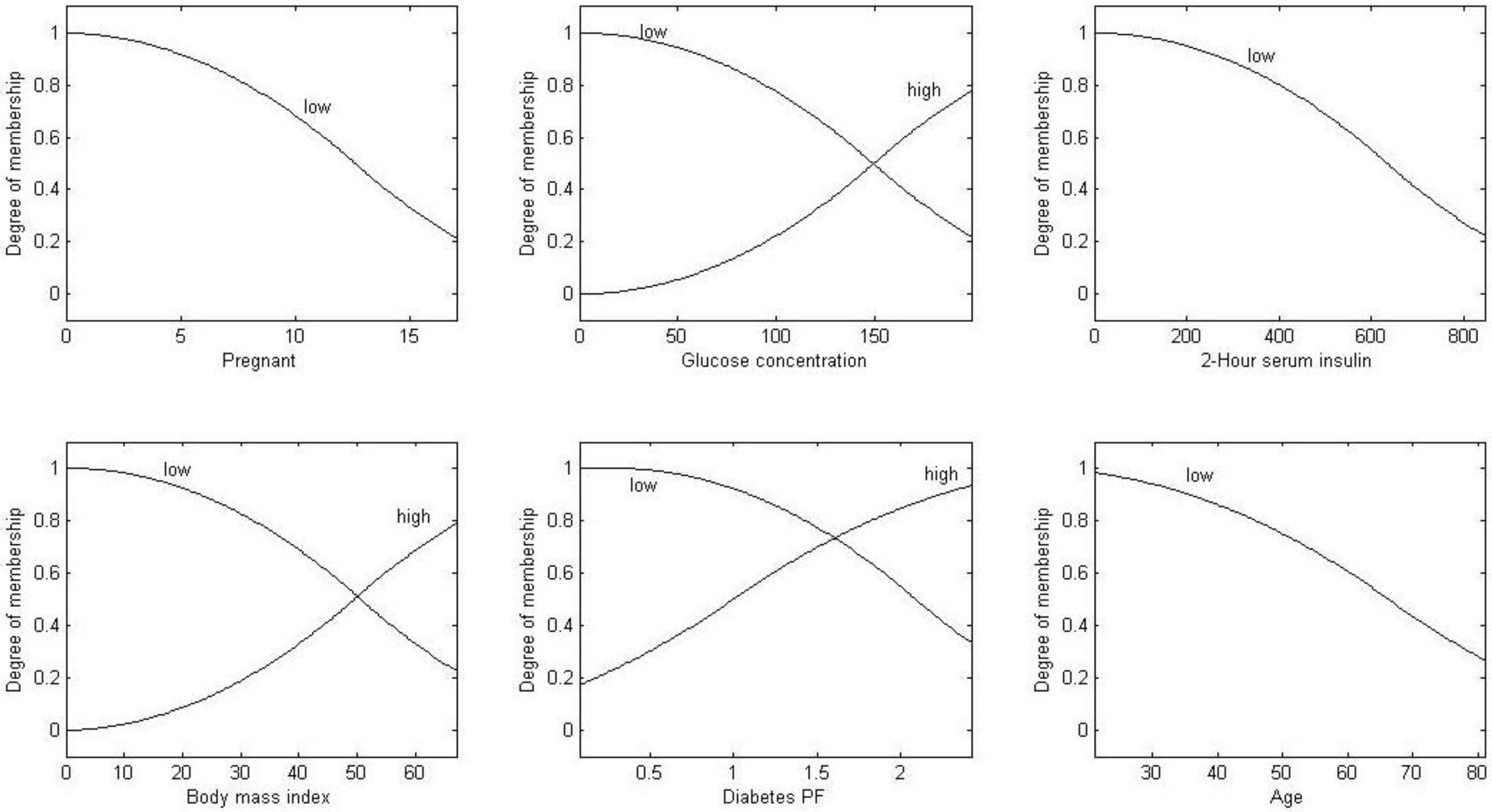
[8]S. N. Ghazavi and T. W. Liao. Medical data mining by fuzzy modeling with selected features. Artificial Intelligence in Medicine, 43(3):195-206, 2008.

2-Rules System

- Using the whole data set:

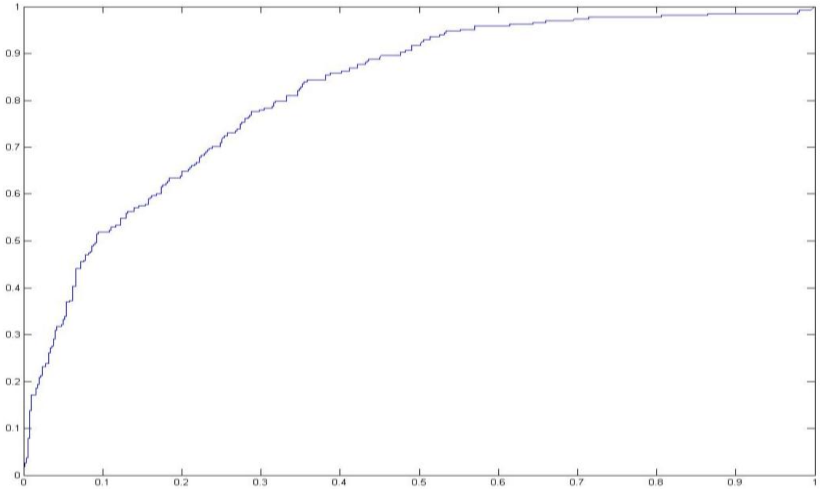
1. If (Pregnant is low) and
   (Glucose concentration is low) and
   (Body mass index is low) and
   (Diabetes PF is low) and
   (Age is low)
   then (Output is 2) (weight 0.7932)
2. If (Glucose concentration is high) and
   (2-Hour serum insulin is low) and
   (Body mass index is high) and
   (Diabetes PF is high) and
   (Age is low)
   then (Output is 4) (weight 1.5376)

2-Rules System



2-Rules System

- THR=2.97, AUC=0.819, CR=77.08



		Actual	
		P	N
Predicted	P	454	130
	N	46	138

14-Rules System

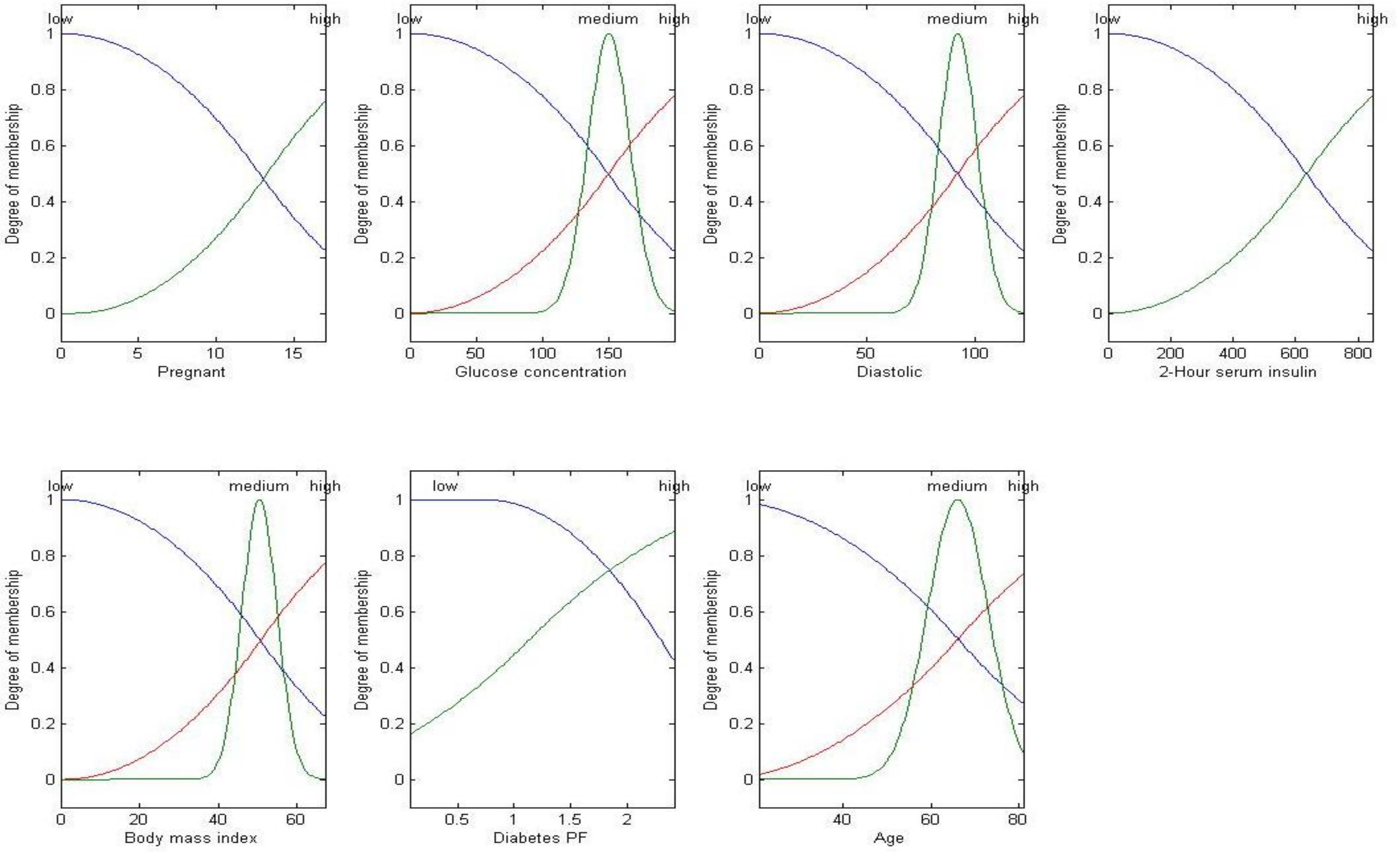
- Using the whole data set, is class1 if:
  1. If (Pregnant is low) and (Glucose concentration is low) and (Body mass index is low) and (Diabetes PF is low) and (Age is low)
  2. If (Glucose concentration is low) and (Diastolic is high) and (2-Hour serum insulin is high) and (Body mass index is medium) and (Diabetes PF is low) and (Age is low)
  3. If (Glucose concentration is low) and (Body mass index is medium) and (Age is high)
  4. If (Pregnant is low) and (Glucose concentration is medium) and (Body mass index is high) and (Diabetes PF is low) and (Age is high)
  5. If (Pregnant is low) and (Glucose concentration is low) and (Body mass index is low) and (Diabetes PF is high) and (Age is low)
  6. If (Glucose concentration is low) and (Diastolic is medium) and (2-Hour serum insulin is low) and (Body mass index is medium) and (Diabetes PF is low) and (Age is low)
  7. If (Pregnant is low) and (Glucose concentration is medium) and (Diastolic is high) and (Body mass index is medium) and (Diabetes PF is low) and (Age is low)

14-Rules System

- Using the whole data set, is class2 if:

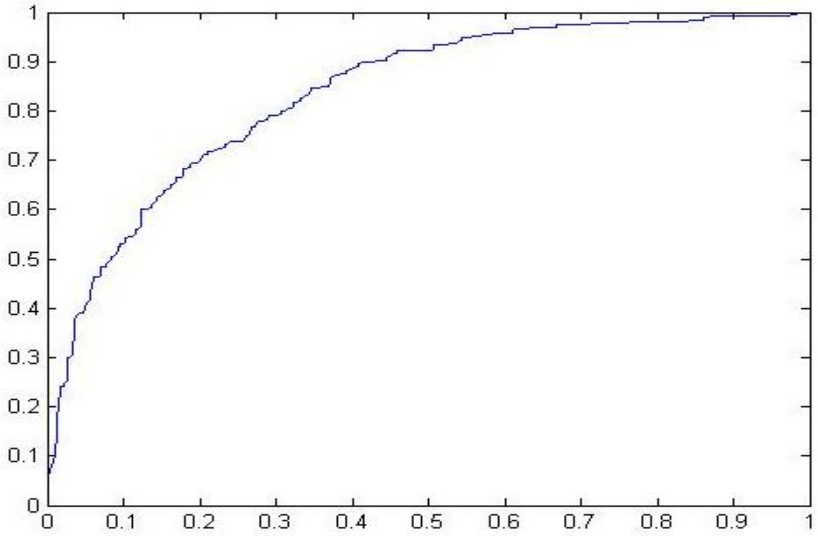
8. If (Glucose concentration is high) and (2-Hour serum insulin is low) and (Body mass index is high) and (Diabetes PF is high) and (Age is low)
9. If (Glucose concentration is medium) and (2-Hour serum insulin is low) and (Body mass index is medium) and (Diabetes PF is high) and (Age is high)
10. If (Pregnant is high) and (Glucose concentration is medium) and (Body mass index is high) and (Diabetes PF is high) and (Age is high)
11. If (Glucose concentration is high) and (2-Hour serum insulin is low) and (Body mass index is high) and (Diabetes PF is low) and (Age is medium)
12. If (Glucose concentration is high) and (Body mass index is low) and (Age is medium)
13. If (Glucose concentration is medium) and (Diastolic is low) and (Body mass index is high) and (Age is low)
14. If (Pregnant is high) and (Glucose concentration is high) and (2-Hour serum insulin is low) and (Body mass index is high) and (Diabetes PF is high) and (Age is high)

14-Rules System



14-Rules System

- THR=3.033, AUC=0.836, CR=77.99



		Actual	
		P	N
Predicted	P	439	108
	N	61	160

Preliminary results

- We define the confidence as

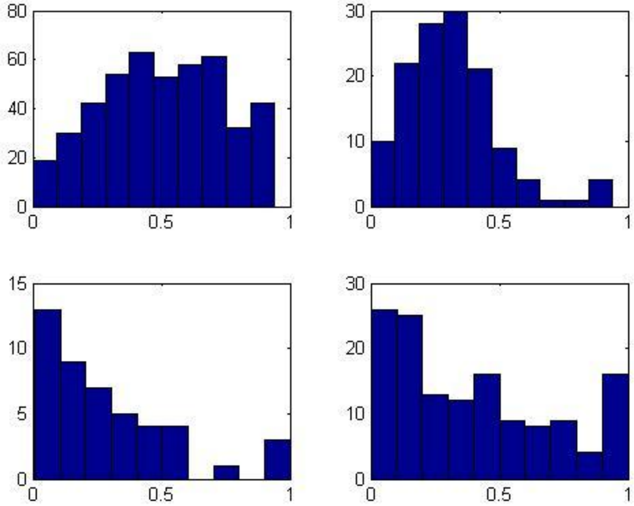
$$\chi=abs(O-\tau)/\alpha$$

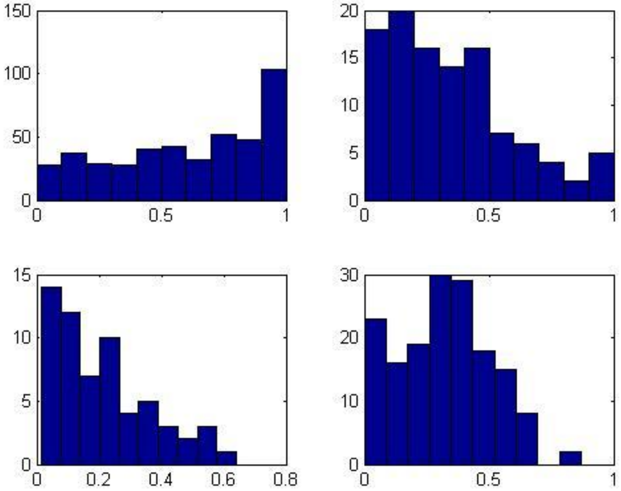
  - O is the output of the FIS
  - α is a normalizing factor so that $\chi \in [0,1]$
- A good classifier should be highly confident with correctly classified examples while it should be doubtful with misclassified data points.
- Starting from the confusion matrix on the test sets, we measure the cases correctly classified (NTP and NTN) with $\chi > 0.7$ and the number of instances incorrectly classified (NFP and NFN) with $\chi < 0.3$.

Preliminary results

2 RULES	
23,13%	54,62%
58,70%	21,01%

14 Rules	
46,24%	50,00%
72,13%	1,25%





Future directions

- Our methodology needs to be tested using:
  - other data sets;
  - different strategies for extracting rules.
- Different techniques to determine the correct thresholds for the FISs could improve the whole performance:
  - the one that minimizes the mean square error;
  - the one that minimizes a cost function that takes into account the unequal classification error costs.
- Weighted rules could be also used.
- It is in worth testing the possibility of using in parallel FISs with different numbers of rules:
  - their predictions could be combined using several strategies based on the confidence showed by each system.

Questions?