Classifying Iris Data Based on Choquet Integral

Li Zhang
Graduate Section

Introduction

Singed Efficiency Measure: μ **:** $\mathscr{F} \rightarrow (-\infty, \infty]$ **is a** ***signed efficiency measure*** **iff** $\mu(\emptyset) = 0$

Interaction and Nonadditivity: The ***nonadditivity of μ*** **describes the interaction among the attributes.**

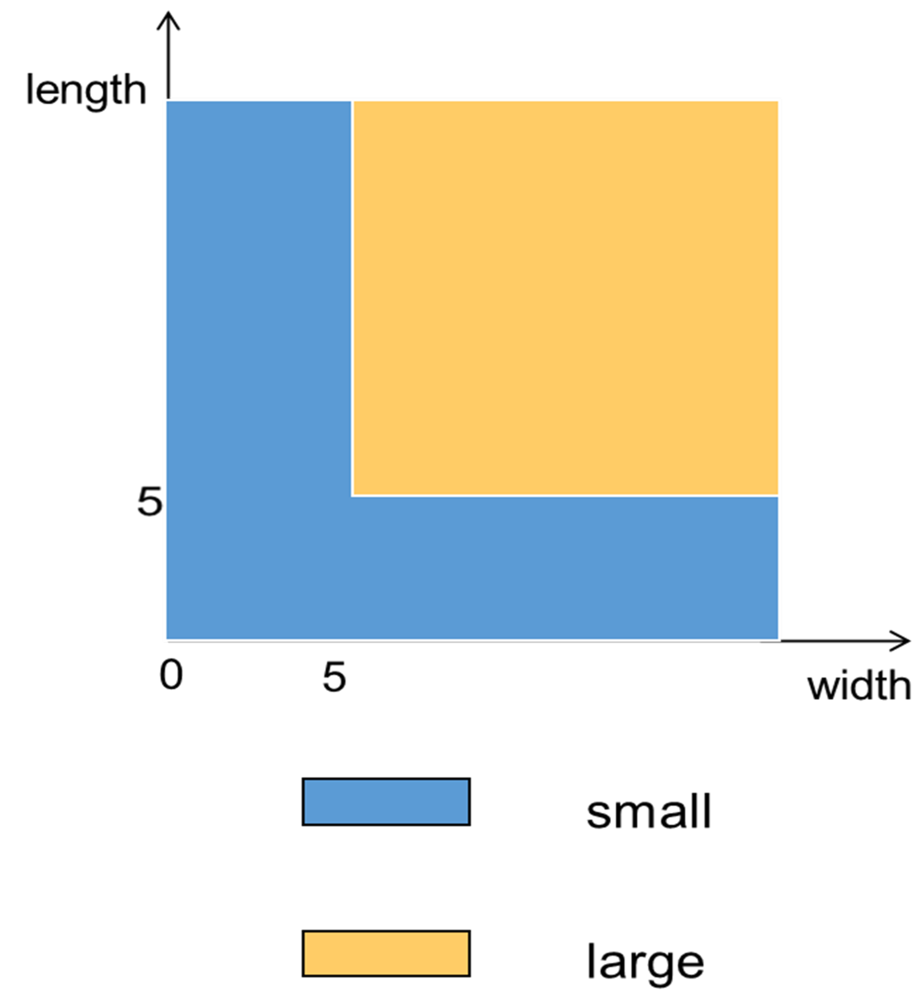
There are two special cases:

- **Subadditivity.** $\mu(E \cup F) \leq \mu(E)+\mu(F)$ **for any** $E \in C$ **and** $F \in C$ **.**
- **Superadditivity.** $\mu(E \cup F) \geq \mu(E)+\mu(F)$ **for any** $E \in C$ **and** $F \in C$ **with** $\cap F = \emptyset$**.**

The Choquet Integral: (C) $\int(a+bf)d\mu$ serves as an aggregation tool, optimally projecting the feature space onto a real axis. Regarding it as a functional of the integrand, its contour can be used as a nonlinear classifying boundary. Particularly, a, b balance various dimensions of attributes.

Take a look at this example: a mail box is large enough, but its slot is only 5 inches long. Thus, the envelops are classified into two classes:

(1) ***small*** that can be inserted into the mail box;
(2) ***large*** that cannot be inserted into the mail box.



Such a classification is not linear. There is a strong interaction between the length and width towards the classifying criterion. In fact, the classifying boundary can be expressed as a contour of the Choquet Integral shown later.

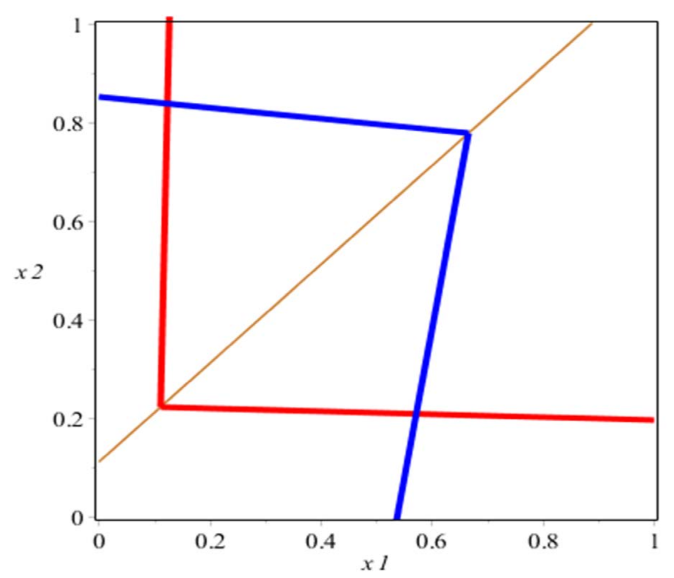
Classification

Contours of Two Choquet Integrals

(C) $\int(a_1 + b_1 f)d\mu_1 = d_1$, (C) $\int(a_2 + b_2 f)d\mu_2 = d_2$

where

- $a_1 = 0.2$
- $a_2 = 0.1$
- $b_1 = 0.9$
- $b_2 = 0.9$
- $\mu_{11} = 0.03$
- $\mu_{12} = -0.02$
- $\mu_{21} = 1.2$
- $\mu_{22} = 0.9$
- $d_1 = 0.3$
- $d_2 = 0.8$



Classification Rule

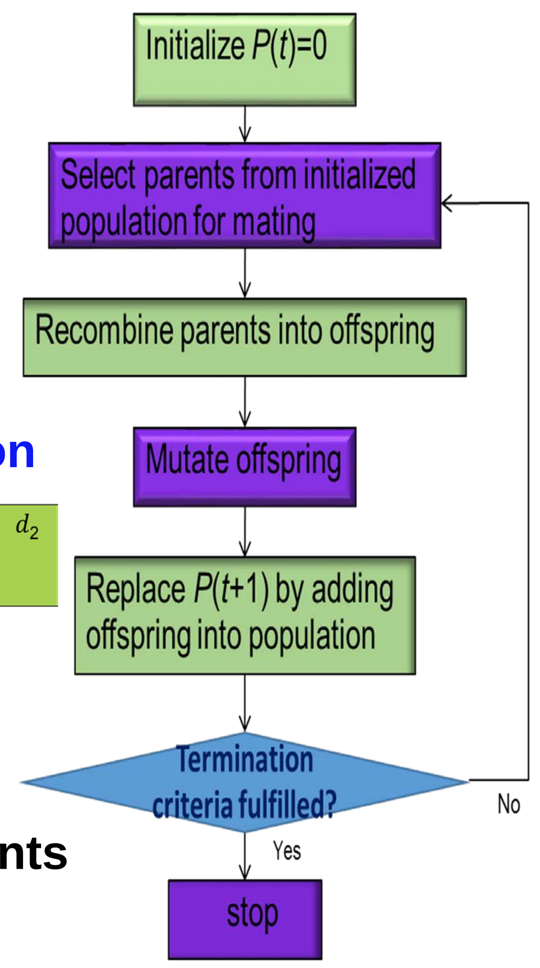
If C ≥ d_1 and C' ≤ d_2 ,the point is in class 1
Otherwise, the point is in class 2.

According to the classification rule, for the given data, we need to optimize parameters of Choquet integrals such that the misclassification rate is as small as possible. This optimization can be reached by a **genetic algorithm** numerically.

Genetic Algorithm

Involved Concepts:

- *Chromosome Representation*
- *Search Space*
- *Operators*
- *Objective function.*



Chromosome representation

a_1	a_2	b_1	b_2	μ_{11}	μ_{12}	μ_{21}	μ_{22}	d_1	d_2

Criteria for Finding Parents

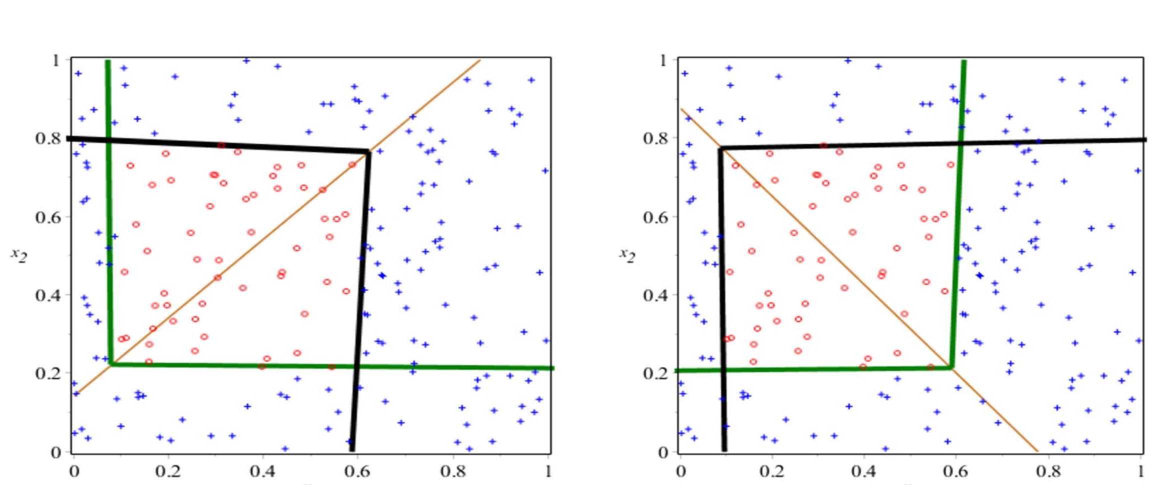
fitness function's value:

***e* = 1 / (1+(*m* / *l*))**

***l* : total number of data points**

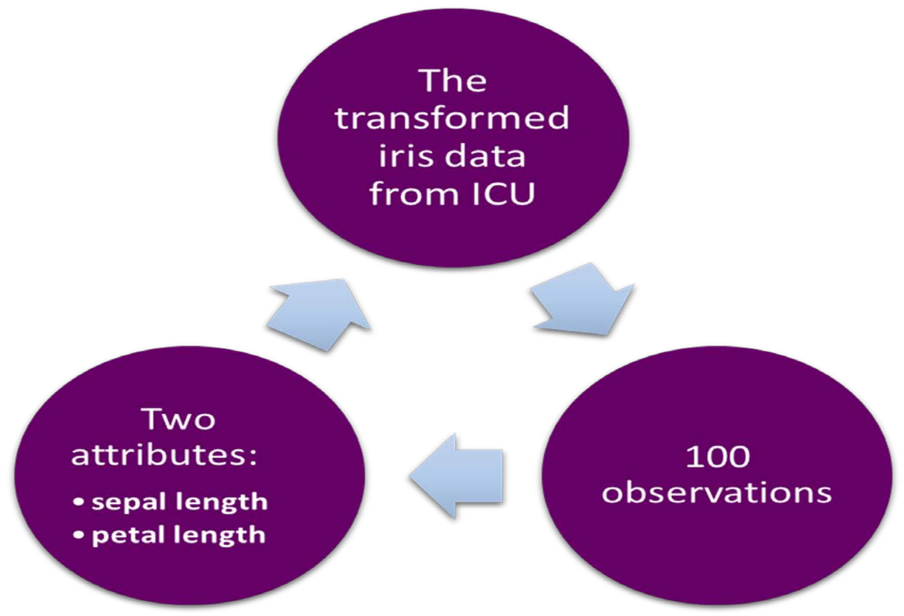
***m*: misclassified points**

Try on Classification (two outcomes)

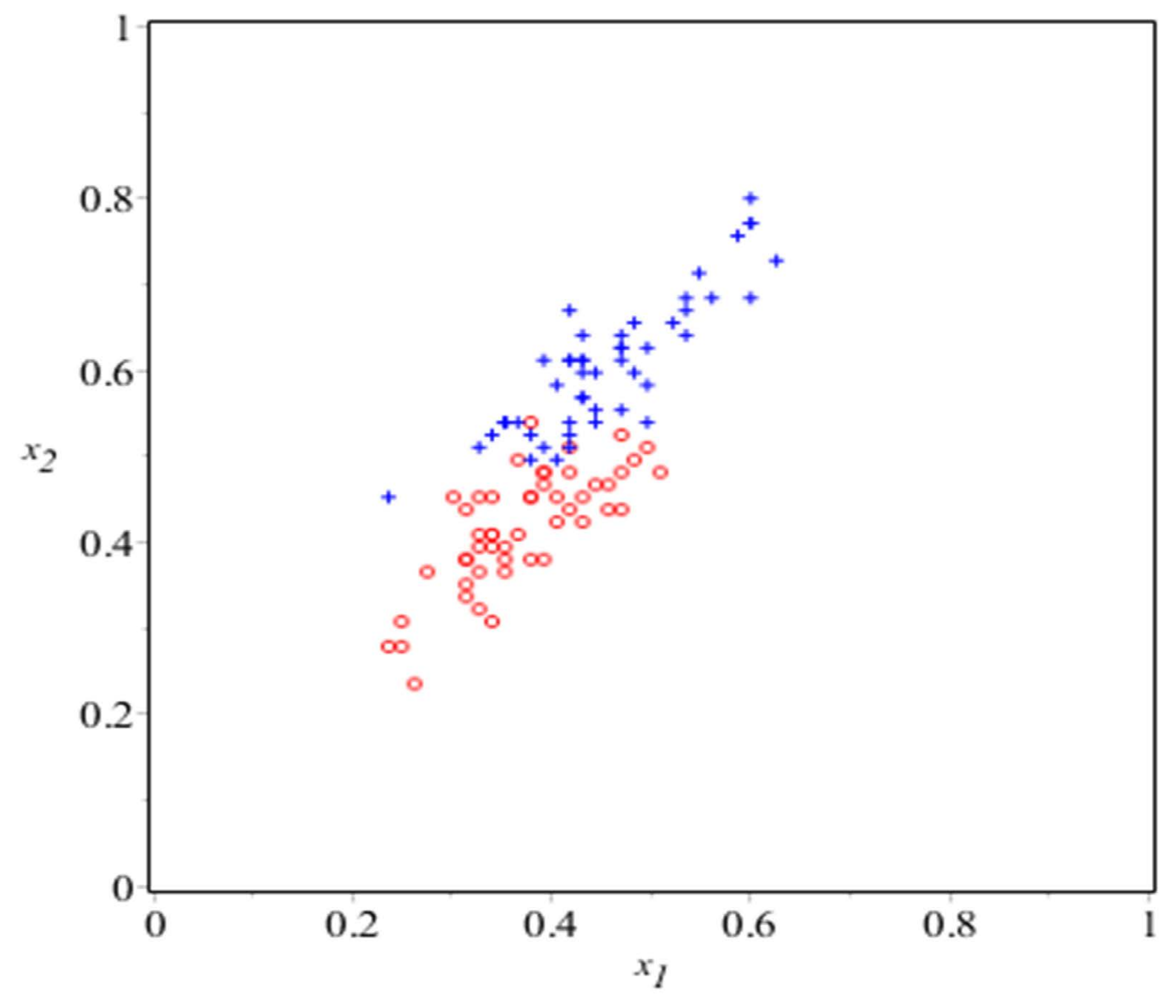


Real Iris Data

class 1 as *iris vesicolor* (red circle)
class 2 as *iris virginica* (blue corss)



Graph of Iris Data

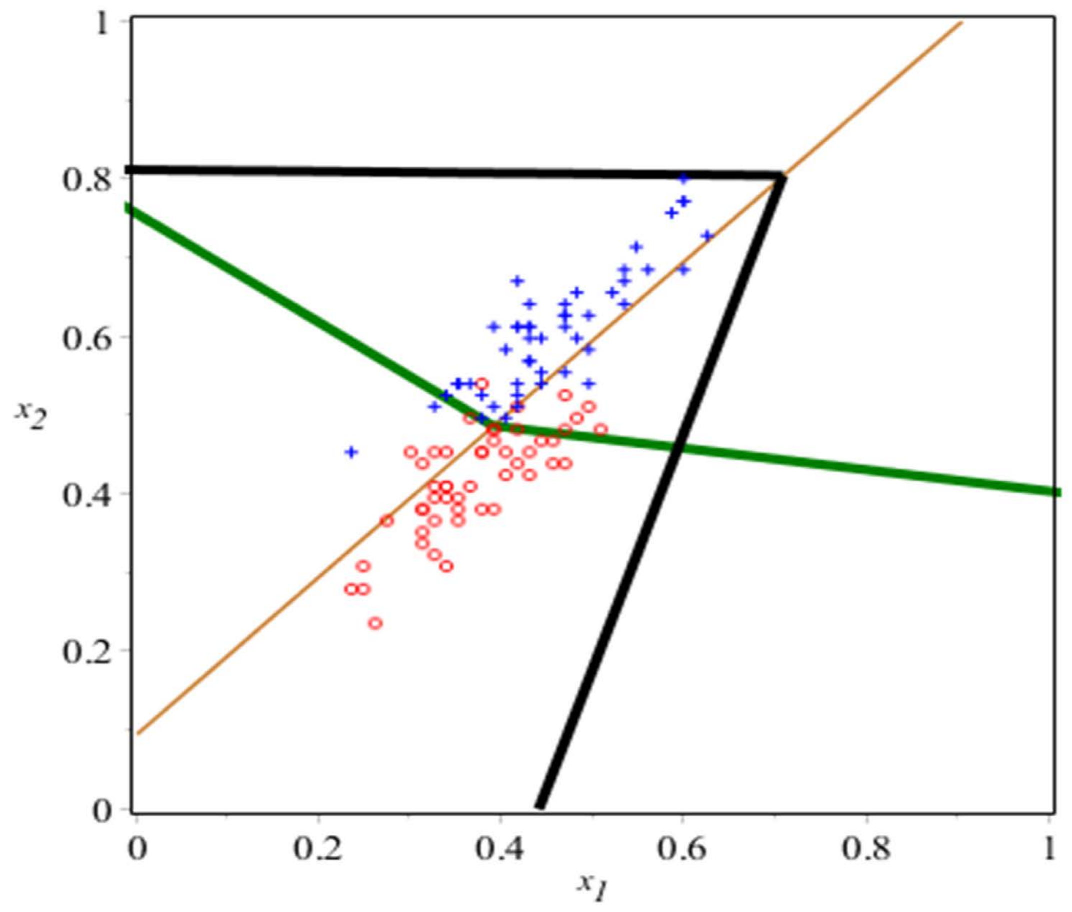


Source of Iris Data:
Retrieved from: http://archive.ics.uci.edu/ml/

Conclusions

- Better classification accuracy
- Good insight in intuitive understanding

Variable	Following Genetic Algorithm
a_1	0.1
a_2	0.005
b_1	0.99
b_2	1
μ_1	0.001
μ_2	0.59
v_1	1.5
v_2	0.99
d_1	0.49
d_2	0.81



References

[1] Z. Y. Wang, R. Yang, and K. S. Leung, *Nonlinear integrals and their applications in data mining,* pp. 238-263, July 2010.

[2] K. B. Xu, Z. Y. Wang, P. A. Heng, and K. S. Leung, "Classification by Nonlinear Integral Projections," in *IEEE Transactions on Fuzzy Systems*, vol. 11,no. 2, April 2003, pp 187-201.

[3] H. Fang, M. L. Rizzo, H. G. Wang, K. A. Espy, and Z. Y. Wang, "A New Nonlinear Classifier with a Penalized Signed efficiency measure Using Effective Genetic Algorithm," in H. Fang ET AL. IN Pattern Recognition 43 (2010).

[4] N. Yan, Z. Y. Wang, and Z. X. Chen, "Classification with Choquet Integral with Respect to Signed Non-Additive Measure," in Workshops of Seventh *IEEE International Conference* on Data Mining.

[5] M. Mitchell and X. Melanie, An Introduction to Genetic Algorithms, Cambridge, Mass., MIT Press, 1996.