Probabilistic Intervals of Confidence Interpretation of Adaptive Models

Norbert Jankowski

Department of Computer Methods, Nicholas Copernicus University

ul. Grudziądzka 5, 87–100 Toruń, Poland, phone: +48 56 6113307, fax: +48 56 621543

e-mail: Norbert.Jankowski@phys.uni.torun.pl, http://www.phys.uni.torun.pl/~norbert

What is the goal?

- High accuracy should not be the only goal of classification
- Important are also: alternatives diagnoses and their probability, evaluation of confidence
- Neural models — just the winner class — they work as *black boxes*.

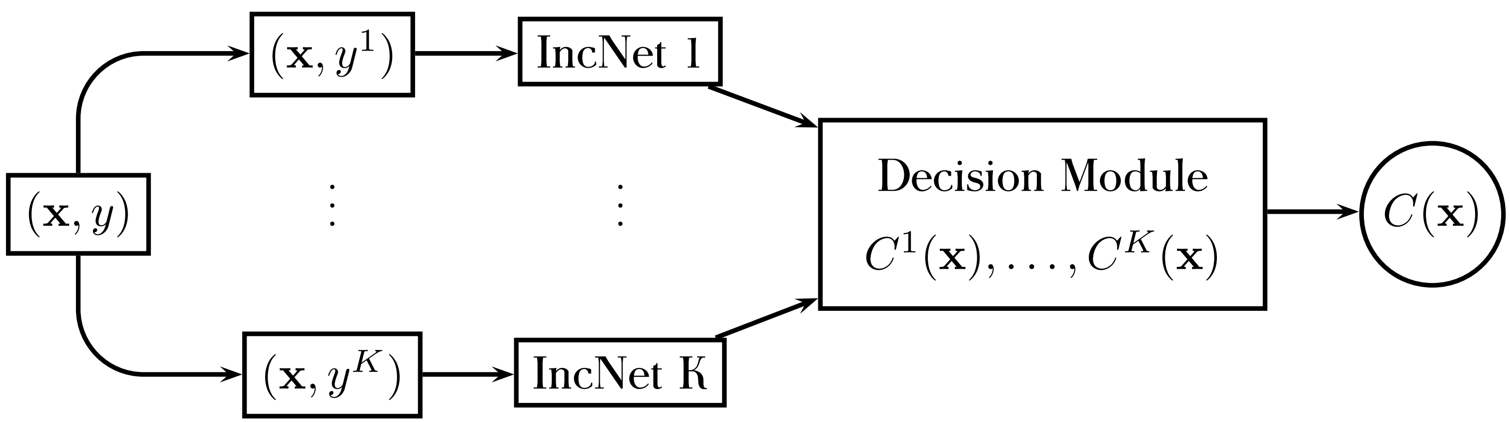
Probabilistic Confidence Intervals helps to:

- evaluate the certainty of the winning class and the importance of alternative classes
- compare the influence of each feature in classification of a given case, showing changes of the probability of all important classes
- visualize the class memberships of a given case and its neighborhood

Disadvantages of (crisp) logical rules

- Rules assign a given case to a class without any gradation which could give information on uncertainty of such classification
- Rules conditions use hyper-rectangular membership function and therefore shape of their decision borders are very limited
- Because of *rectangular shapes* rules may not cover the whole input space, leaving subspaces in which no classification is done
- Rules may also overlap producing ambiguous classification
- Logical rules are not reliable near decision borders

Incremental Network



Winning class:

$$C(\mathbf{x}) = \arg\max_i \; C^i(\mathbf{x})$$

Probability:

$$p(C^i|\mathbf{x}) = \frac{\sigma(C^i(\mathbf{x}) - \frac{1}{2})}{\sum_{j=1}^{K} \sigma(C^j(\mathbf{x}) - \frac{1}{2})}$$

The IncNet network was used because of its good performance — network structure is controlled by growing and pruning criterion to keep complexity of network similar to the complexity of data.

Confidence Intervals (CI)

⋆ Confidence intervals — calculated individually for a given input vector while

⋆ Logical rules are extracted for the whole *training set*.

⋆ In general such probability may be estimated by any trustworthy model.

Suppose that for a given vector $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ the highest probability $p(C^k|\mathbf{x}; M)$ is found for class k.

The confidence interval $[x^r_{min}, x^r_{max}]$ for the feature r is defined by

$$x^r_{min} = \min_{\bar{x}} \{C(\bar{\mathbf{x}}) = k \ \wedge \ \forall_{x_r > \hat{x} > \bar{x}} \ C(\hat{\mathbf{x}}) = k\} \tag{1}$$

$$x^r_{max} = \max_{\bar{x}} \{C(\bar{\mathbf{x}}) = k \ \wedge \ \forall_{x_r < \hat{x} < \bar{x}} \ C(\hat{\mathbf{x}}) = k\} \tag{2}$$

where

$$\bar{\mathbf{x}} = [x_1, \ldots, x_{r-1}, \bar{x}, x_{r+1}, \ldots, x_N], \quad \hat{\mathbf{x}} = [x_1, \ldots, x_{r-1}, \hat{x}, x_{r+1}, \ldots, x_N] \tag{3}$$

Confidence intervals for a given vector $\mathbf{x}$ measure maximal deviation from the value x_r, assuming all other feature values unchanged, that do not change classification of the vector.

Intervals with confidence level

should guarantee that *the winning* class k is considerably more probable than the most probable alternative class:

$$x_{min}^{r,\beta} = \min_{\bar{x}} \left\{ C(\bar{\mathbf{x}}) = k \ \wedge \ \forall_{x_r > \hat{x} > \bar{x}} \ C(\hat{\mathbf{x}}) = k \ \wedge \ \frac{p(C^k|\bar{\mathbf{x}})}{\max_{i \neq k} p(C^i|\bar{\mathbf{x}})} > \beta \right\} \quad (4)$$

$$x_{max}^{r,\beta} = \max_{\bar{x}} \left\{ C(\bar{\mathbf{x}}) = k \ \wedge \ \forall_{x_r < \hat{x} < \bar{x}} \ C(\hat{\mathbf{x}}) = k \ \wedge \ \frac{p(C^k|\bar{\mathbf{x}})}{\max_{i \neq k} p(C^i|\bar{\mathbf{x}})} > \beta \right\} \quad (5)$$

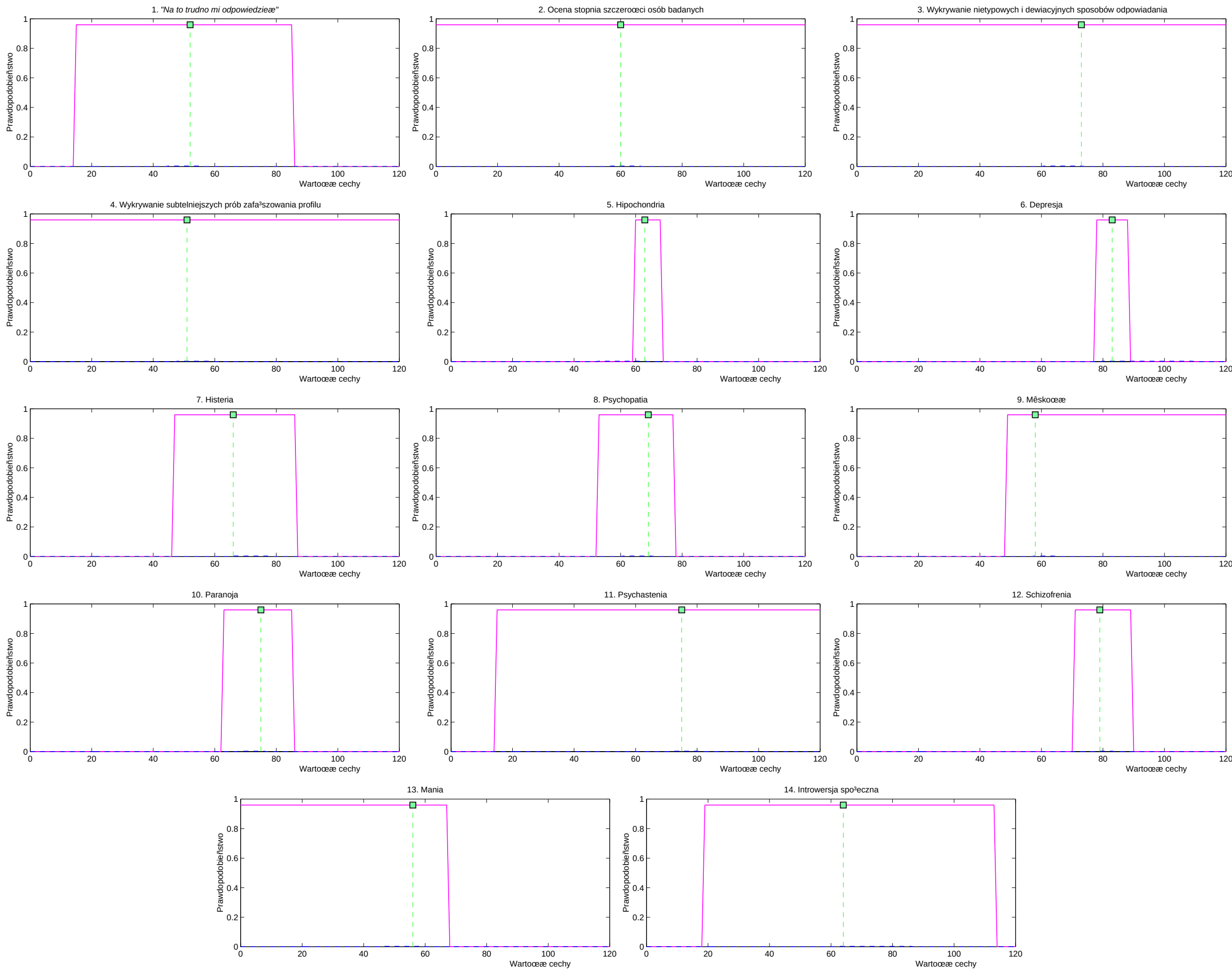


Figure 1: Reactive Psychosis.

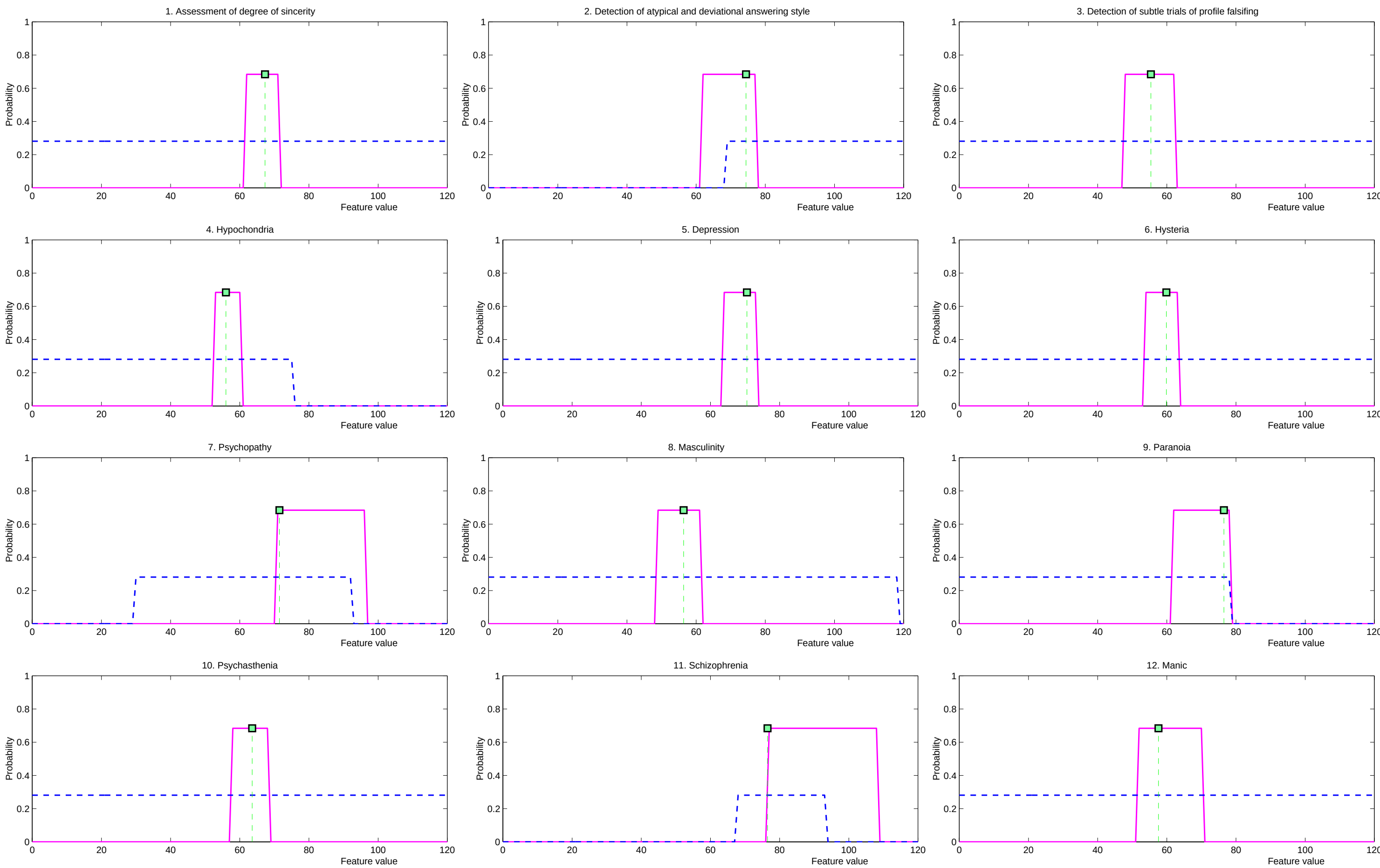


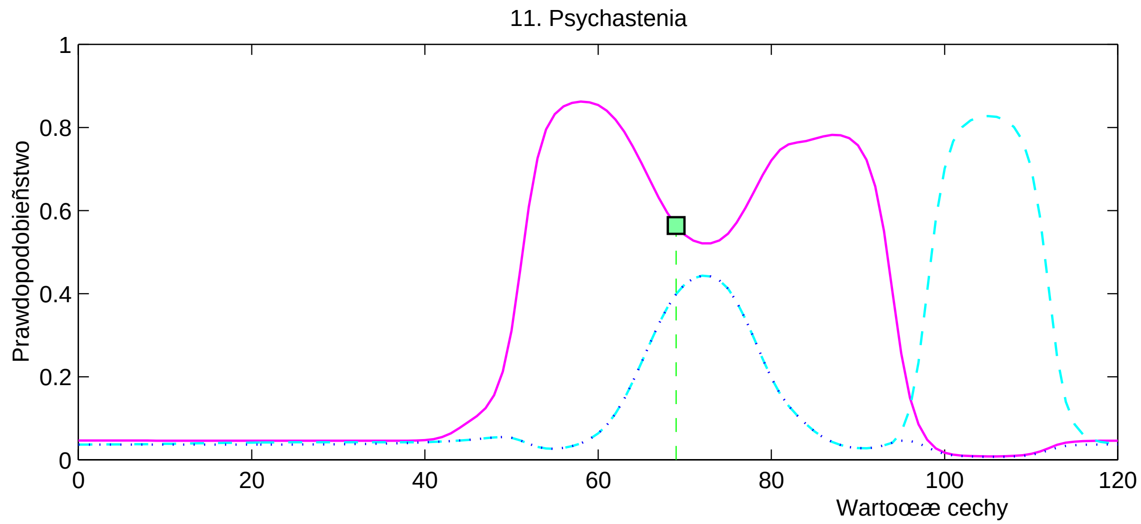
Figure 2: Class: Paranoia (prob. 0.68); alternative class: schizophrenia (prob. 0.28).

Probabilistic Intervals of Confidence (PIC)

For given vector $\mathbf{x}$ and feature r:

Class	Probability	# class
winner	$p(C(\mathbf{x})\|\bar{\mathbf{x}}(\mathbf{z}))$	$C(\mathbf{x})$
alternative I	$p(C^{k_2}\|\bar{\mathbf{x}}(\mathbf{z}))$	$k_2 = \arg\max_i \{p(C^i\|\mathbf{x}),\ C^i \neq C(\mathbf{x})\}$
alternative II	$p(C^{k_M}\|\bar{\mathbf{x}}(\mathbf{z}))$	$k_M = \arg\max_i \{p(C^i\|\bar{\mathbf{x}}(z)),\ C^i \neq C(\mathbf{x})\}$

$$\bar{\mathbf{x}}(z) = [x_1, \ldots, x_{r-1}, z, x_{r+1}, \ldots, x_N]$$



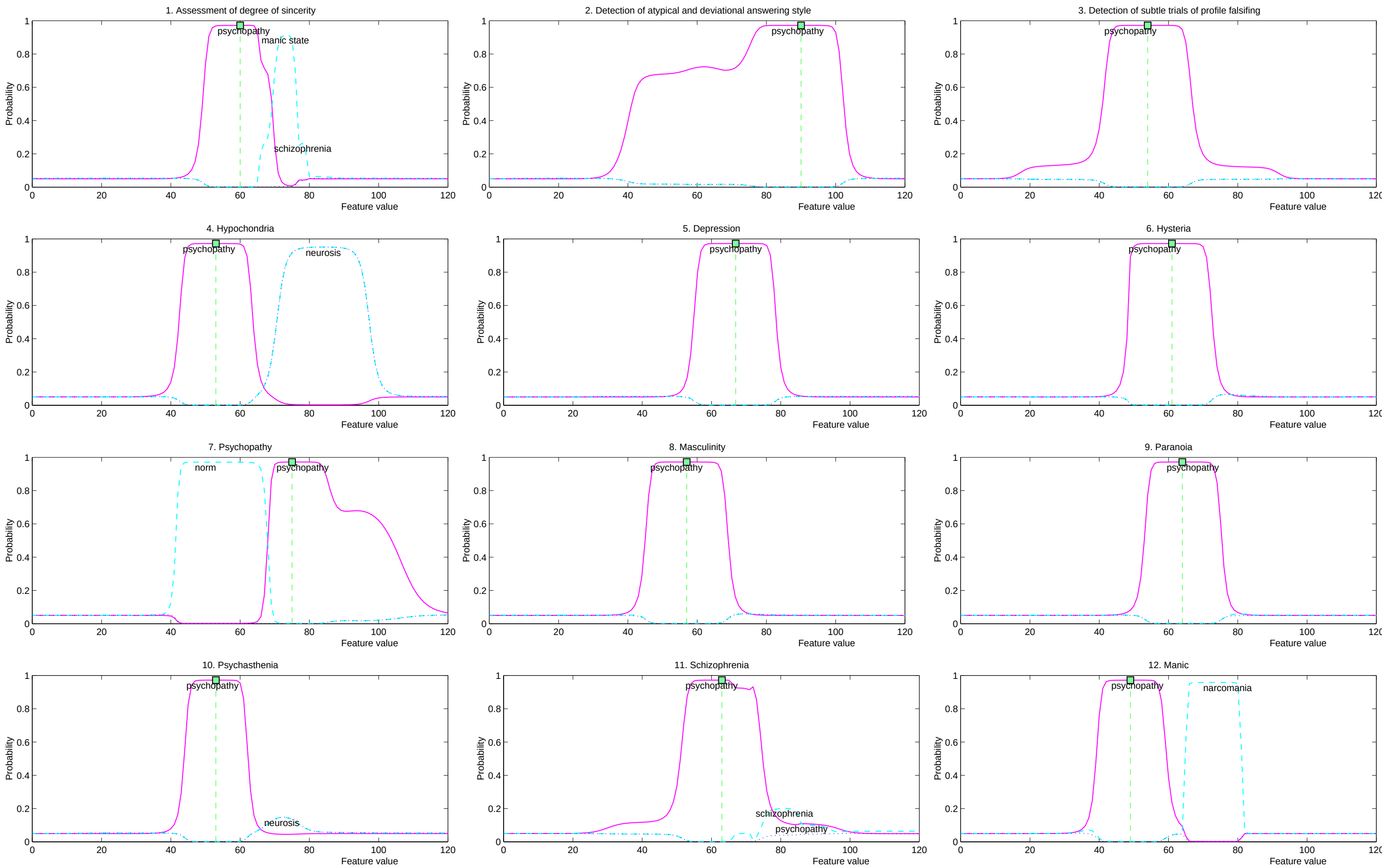


Figure 3: Class: Psychopathy (prob. 0.97); alternative class: neurosis (prob. 0.002).

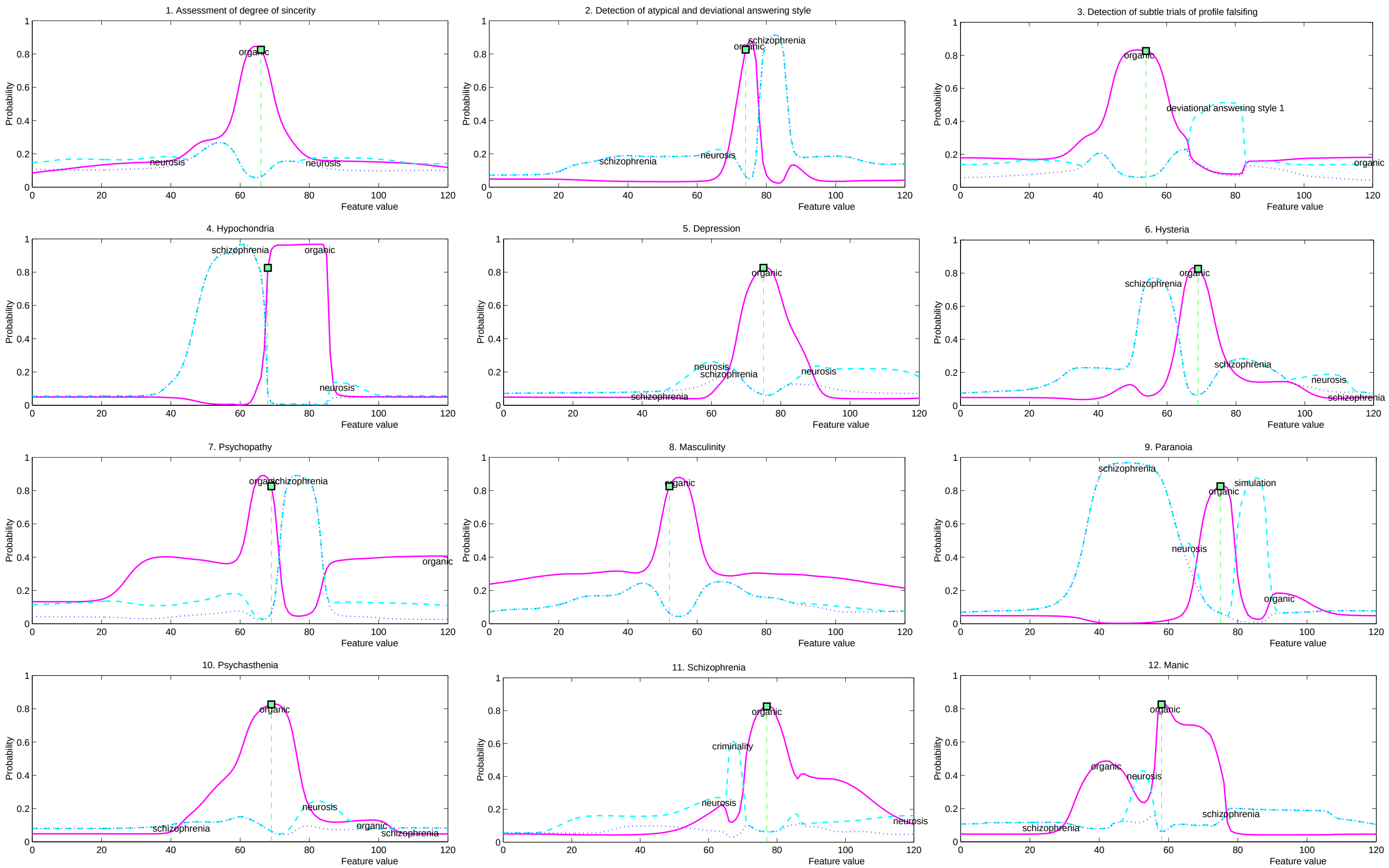


Figure 4: Organic (0.83), schizophrenia (0.062)

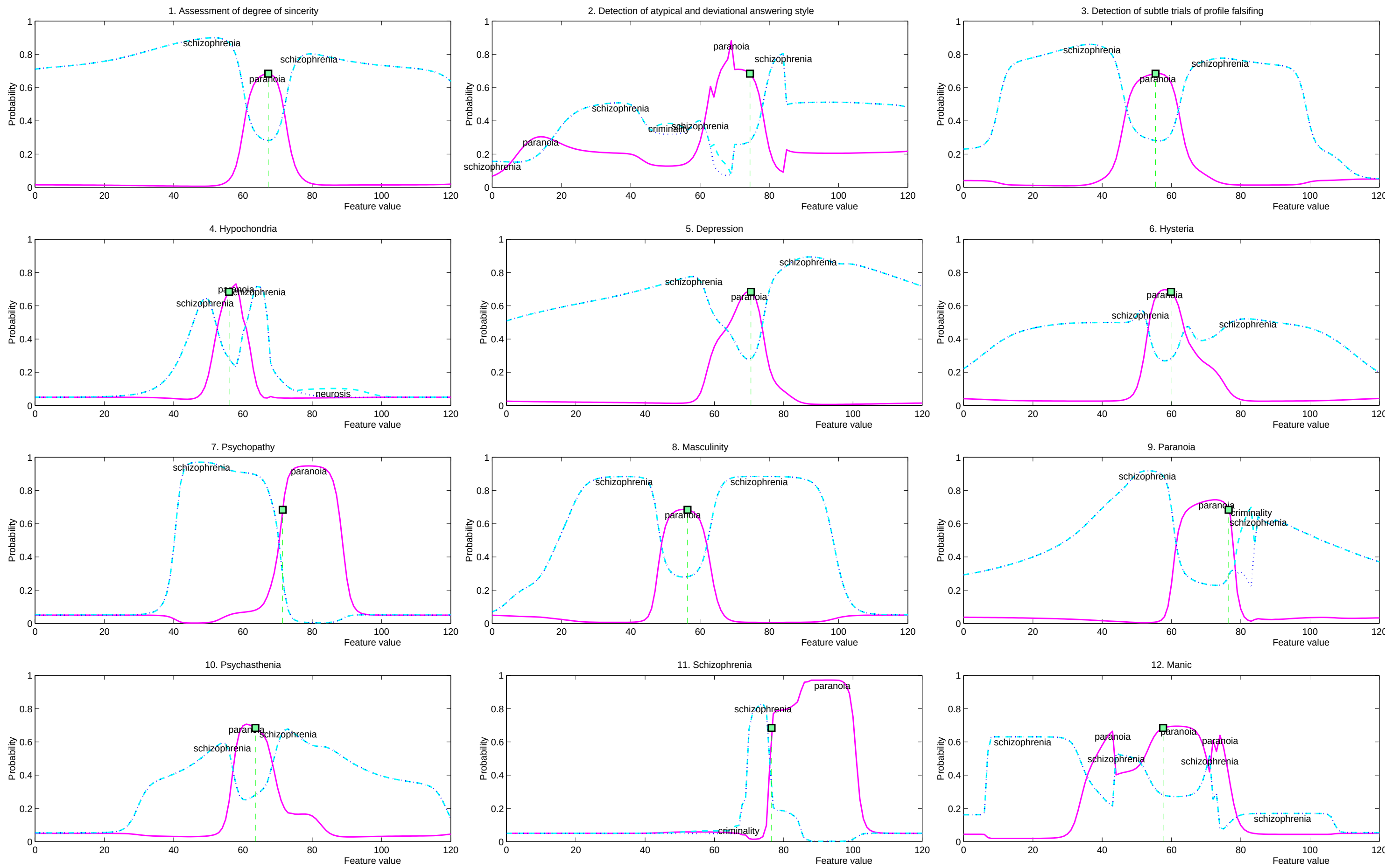


Figure 5: Class: Paranoia (prob. 0.68); alternative class: schizophrenia (prob. 0.28).

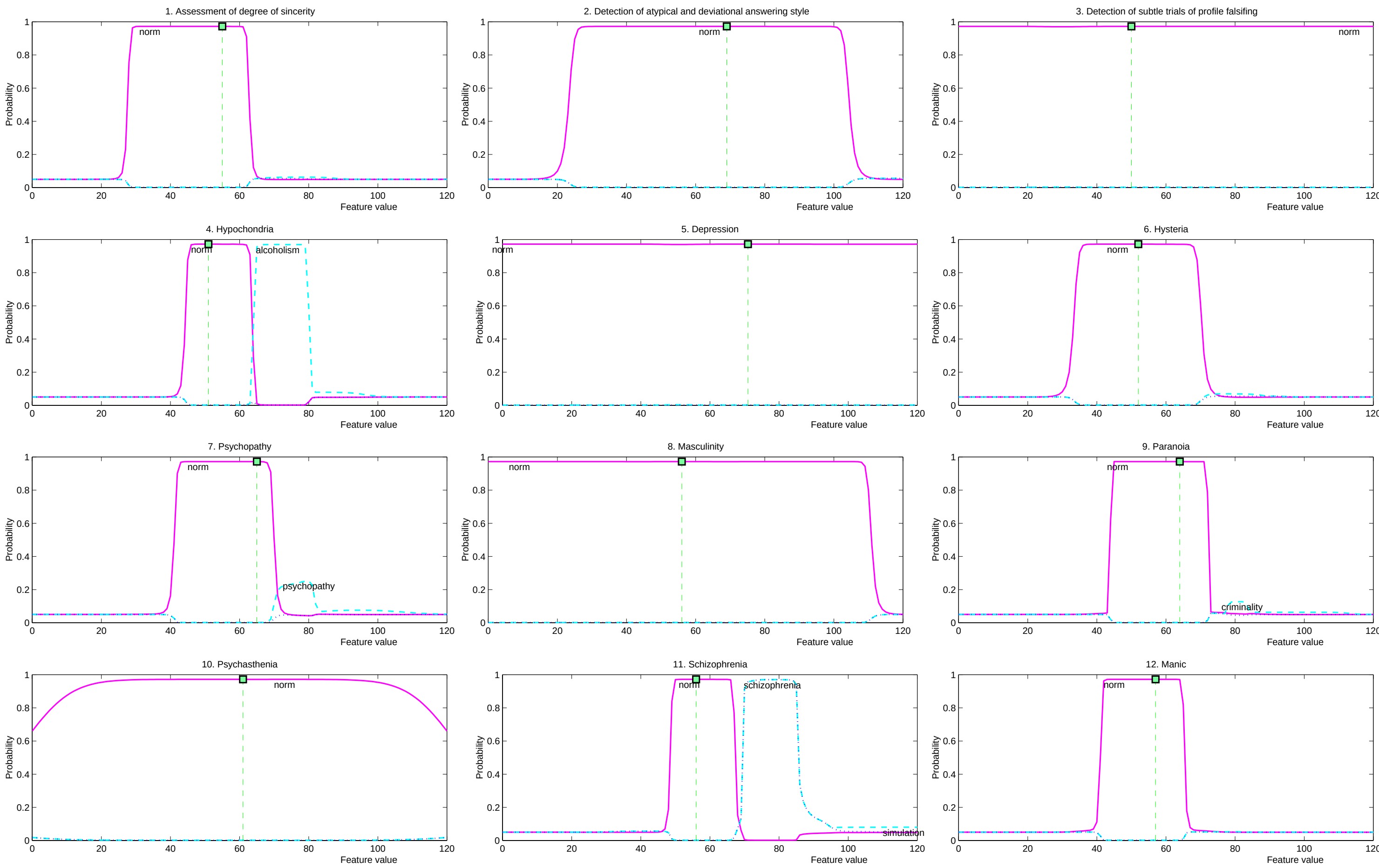


Figure 6: Class: Norm (prob. 0.97); non alternative class.

Description of previous pictures

Figures 3, 4, 5 and 6 show probabilistic intervals of confidence for two quite different patients (the first and the last scale has been omitted, therefore only 12 features are displayed). Little squares show the probability of the winning class corresponding to the measured input values of the psychometric scales. Figure 3 presents an easy case: the psychopathy has a large probability 0.97 and the case is quite far from any other alternative classes. The whole range of values, 0-120, is shown and an alternative class appears for features 1, 4, 7 and 12, but the confidence intervals are quite broad. Classification does not depend on the precise values of some features r (for example features 2, 3, 5, 6, etc) since there are no alternative classes in the whole range of values $\bar{\mathbf{x}}$ may take.

The second set of plots, Fig. 4, is more complex. The winner class, organic, has probability 0.83 while the alternative class, schizophrenia has probability 0.06. The analysis of plots shows that the values for scales 4 and 7 are close to the border and therefore both diagnoses are probable, and scales 4 & 7 are very important for diagnosis. Note that classification is not so simple although the probability is 0.83, because considered case lies so close the border of feature 4.

Case on Figure 5 is ambiguous too. The winner class, paranoia, has probability 0.68 while the alternative class, schizophrenia has probability 0.28. The analysis of plots shows that the values for scales 7 and 11 are close to the border and therefore both diagnoses are probable, and scales 7 & 11 are crucial for considered case.

Figure 6 describe typical case which belong to the ”norm” class.

Psychometric data classification

- Psychometric test: *Minnesota Multiphasic Personality Inventory*
- Test consist from over 550 questions
- 550 questions ➠ 14 features (control and clinic)

  hypochondria, depression, hysteria, psychopathy, masculinity, paranoia, psychasthenia, schizophrenia, manic, social introversion
- 20, 27 or 28 nosological types (classes)

  norm, neurosis, psychopathy, organic, schizophrenia, delusion, reactive psychosis, paranoia, manic state, criminality, alcoholism, etc.
- CV10 accuracy training with IncNet network is 93% (CV5 – 95.5%).

Conclusions

- PIC are new and very useful tools to support the process of diagnosis
- Information on winner and alternative classes is continuous and very precise
- Confidence interval shows neighboring alternative classes (if they exist)
- The distance from the case considered to decision borders may be analyzed in this way
- Analysis of complex cases, which often lie near the decision border, is much more reliable using probabilistic confidence intervals than logical rules
- It is very easy to find which features are important and which may be omitted

- Artificial neural networks may be interpreted using such tools, breaking the myth that neural networks are *black boxes*.