# PROBABILISTIC INTERVALS OF CONFIDENCE

**Norbert Jankowski**[1]
Department of Computer Methods
Nicolaus Copernicus University
ul. Grudziądzka 5, 87-100 Toruń, Poland

**Abstract:**
High accuracy should not be the only goal of classification: information concerning probable alternatives diagnoses, probability of these diagnoses, evaluation of confidence in classification, are also important. Neural models are used just to obtain the winner class but do not provide any justification for their recommendations – they work as *black boxes*. A method which determine confidence intervals and probabilistic confidence intervals is presented here. It helps to evaluate the certainty of the winning class and the importance of alternative classes. Probabilistic intervals are also useful to compare the influence of each feature in classification of a given case, showing changes of the probability of all important classes. Probabilistic confidence intervals help to visualize the class memberships of a given case and its neighborhood.

**Keywords:** Artificial neural networks, visualization, probabilistic intervals of confidence.

## 1 INTRODUCTION

The goal of diagnosis is not only to classify a given data. In real world application, such as in medicine and many other fields, classification process should be extended by analysis of alternative classes and comparison of their probabilities with the winner class. The analysis of influences of feature changes on these probabilities should allow to understand the importance of different features.

Most adaptive models, such as neural networks, fuzzy models or some machine learning methods, finish the diagnosis process just after classification, without any explanation or comparison between alternative classes. Some methods return information allowing to calculate probabilities of partitioning into different classes.

Rule extraction methods are an attempt at interpretation of knowledge from a *training set*. However, methods based on classical (crisp) rules have several disadvantages. First, such methods assign a given case to a class without any gradation which could give information on uncertainty of such classification. Second limitation of logical rules is that their conditions use hyper-rectangular membership function and therefore shape of their decision borders are very limited. In some cases, when more complex decision borders are necessary, the number of extracted rules is very big and rules become hard to use and interpret. 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 and assigning the same probability to alternative classes, while it may not be at all true. Thus rules are not certain on decision borders.

In the next section *confidence intervals (CI)* and *probabilistic intervals of confidence (PIC)* are introduced. Several advantages of PIC intervals are described, especially their usefulness as a visual interpretation method.

[1]E-mail address: Norbert.Jankowski@phys.uni.torun.pl, and www is: http://www.phys.uni.torun.pl/~norbert

# 2 PROBABILISTIC CONFIDENCE INTERVALS

An alternative way to go beyond logical rules introduced in [ 4] is based on *confidence intervals* and *probabilistic confidence intervals*. Confidence intervals are calculated individually for a given input vector while logical rules are extracted for the whole *training set*.

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 $k$. $p(C^i|\mathbf{x}; M)$ describe probability for model $M$ that given input vector $\mathbf{x}$ belong to class $i$. Let the function

$$C(\mathbf{x}) = \arg\max_i \; p(C^i|\mathbf{x}; M) \tag{1}$$

i.e. $C(\mathbf{x})$ is equal to the index $k$ of the most probable class for the input vector $\mathbf{x}$. The Incremental Network (IncNet) [4, 2, 3, 5] was used to compute probability $p(C^k|\mathbf{x}; M)$. In general such probability may be estimated by any trustworthy model. 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.

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

$$x^r_{min} \;=\; \min_{\bar{x}} \{C(\bar{\mathbf{x}}) = k \;\wedge\; \forall_{x_r > \hat{x} > \bar{x}}\; C(\hat{\mathbf{x}}) = k\} \tag{2}$$

$$x^r_{max} \;=\; \max_{\bar{x}} \{C(\bar{\mathbf{x}}) = k \;\wedge\; \forall_{x_r < \hat{x} < \bar{x}}\; C(\hat{\mathbf{x}}) = k\} \tag{3}$$

where

$$\bar{\mathbf{x}} = [x_1, \ldots, x_{r-1}, \bar{x}, x_{r+1}, \ldots, x_N], \quad \hat{\mathbf{x}} = [x_1, \ldots, x_{r-1}, \hat{x}, x_{r+1}, \ldots, x_N] \tag{4}$$

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. If the vector $\mathbf{x}$ lies near the class border the confidence intervals are narrow, while for vectors that are typical confidence intervals should be wide.

Intervals defined above may be extended by adding a confidence level which should guarantee that *the winning* class $k$ is considerably more probable than the most probable alternative class:

$$x^{r,\beta}_{min} \;=\; \min_{\bar{x}} \left\{ C(\bar{\mathbf{x}}) = k \;\wedge\; \forall_{x_r > \hat{x} > \bar{x}}\; C(\hat{\mathbf{x}}) = k \;\wedge\; \frac{p(C^k|\bar{\mathbf{x}})}{\max_{i \neq k}\, p(C^i|\bar{\mathbf{x}})} > \beta \right\} \tag{5}$$

$$x^{r,\beta}_{max} \;=\; \max_{\bar{x}} \left\{ C(\bar{\mathbf{x}}) = k \;\wedge\; \forall_{x_r < \hat{x} < \bar{x}}\; C(\hat{\mathbf{x}}) = k \;\wedge\; \frac{p(C^k|\bar{\mathbf{x}})}{\max_{i \neq k}\, p(C^i|\bar{\mathbf{x}})} > \beta \right\} \tag{6}$$

The $\beta$ factor determines the confidence level. Observation of changes in confidence intervals for different levels of $\beta$ may be quite informative. Comparison of probabilistic intervals for the winning class and alternative classes helps to estimate the likelihood of the winning class. Such method escapes the danger of relaying only on the decision borders of logical rules.

Next step beyond the above considerations is based on an observation *how* the probabilities of the winner and alternative classes change as a function of attribute values for different input dimensions. Displaying such probabilities the *probabilistic intervals of confidence* (PIC) are obtained.

Assuming that other features are held constant for a given case $\mathbf{x}$ three probabilities for each feature $r$ are important and will be visualized in analysis of a given case (cf. Fig. 1 and 2).

First probability (solid curve) is the probability of the **winning class** defined by

$$p(C(\mathbf{x})|\bar{\mathbf{x}}; M) \tag{7}$$

Note that such probability changes for different values of $\bar{\mathbf{x}}$ (Eq. 4).

The next probability displayed (a dotted curve) is the probability

$$p(C^{k_2}|\bar{\mathbf{x}}) \tag{8}$$

of the most probable **alternative class**, where class index $k_2$ is defined by

$$k_2 = \arg\max_i \{p(C^i|\mathbf{x}; M),\ C^i \neq C(\mathbf{x})\} \tag{9}$$

The $k_2$ class is determined for the point $\mathbf{x}$ only.

The third probability (dashed line) presents the probability

$$p(C^{k_M}|\bar{\mathbf{x}}) \tag{10}$$

of the most probable ***variable*** **alternative class** at the point $\bar{\mathbf{x}}$. The index $k_M$ is defined by

$$k_M = \arg\max_i \{p(C^i|\bar{\mathbf{x}}),\ C^i \neq C(\mathbf{x})\} \tag{11}$$

and may change, while index $k_2$ does not change.

These three probabilities carry all information about the case given for analysis, showing the stability of classification against perturbation of each feature and the importance of alternative classes in the neighborhood of the input $\mathbf{x}$. Probabilistic confidence intervals in action are showed in the next section.

# 3 PIC IN ACTION

Psychometric data will be used to show how the probabilistic intervals of confidence work.

In psychometric data classification problem each case (person) is assigned to a personality type using the data from Minnesota Multiphasic Personality Inventory (MMPI) test[ 1]. The MMPI test is one of the most popular psychometric tests designed to help in the psychological diagnoses. MMPI test consists of over 550 questions. Using the answers from each MMPI test 14 numerical factors are computed (by some arithmetic operations) forming the intermediate basis (**not** the final hypothesis) for the diagnosis.

Several data sets were collected and classified by psychologists. In this article two of those sets have been considered, the first with 27 classes and the second with 28 classes. Some classes concern men, and other women only. Each case can be classified as normal or belong to a disease such as neurosis, psychopathy, schizophrenia, delusions, psychosis, etc. Data sets consists of 1027 and 1167 examples respectively for 27 and 28 classes sets.

To illustrate some results conditional probabilities will be estimated using IncNet classifier [4, 2, 3, 5]. Figures 1 and 2 show probabilistic intervals of confidence for two quite

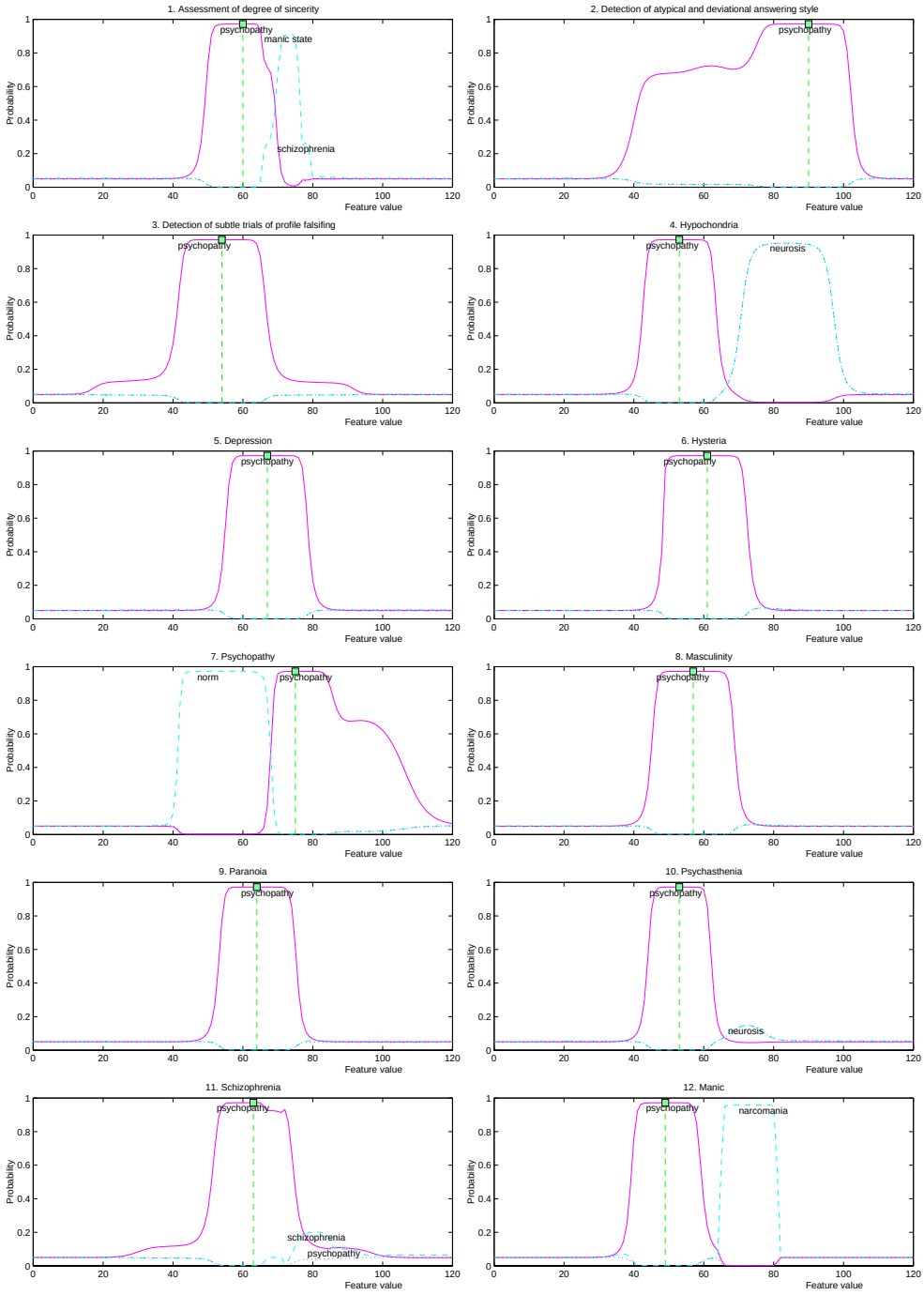


Figure 1: Class: Psychopathy (prob. 0.97); alternative class: neurosis (prob. 0.002).

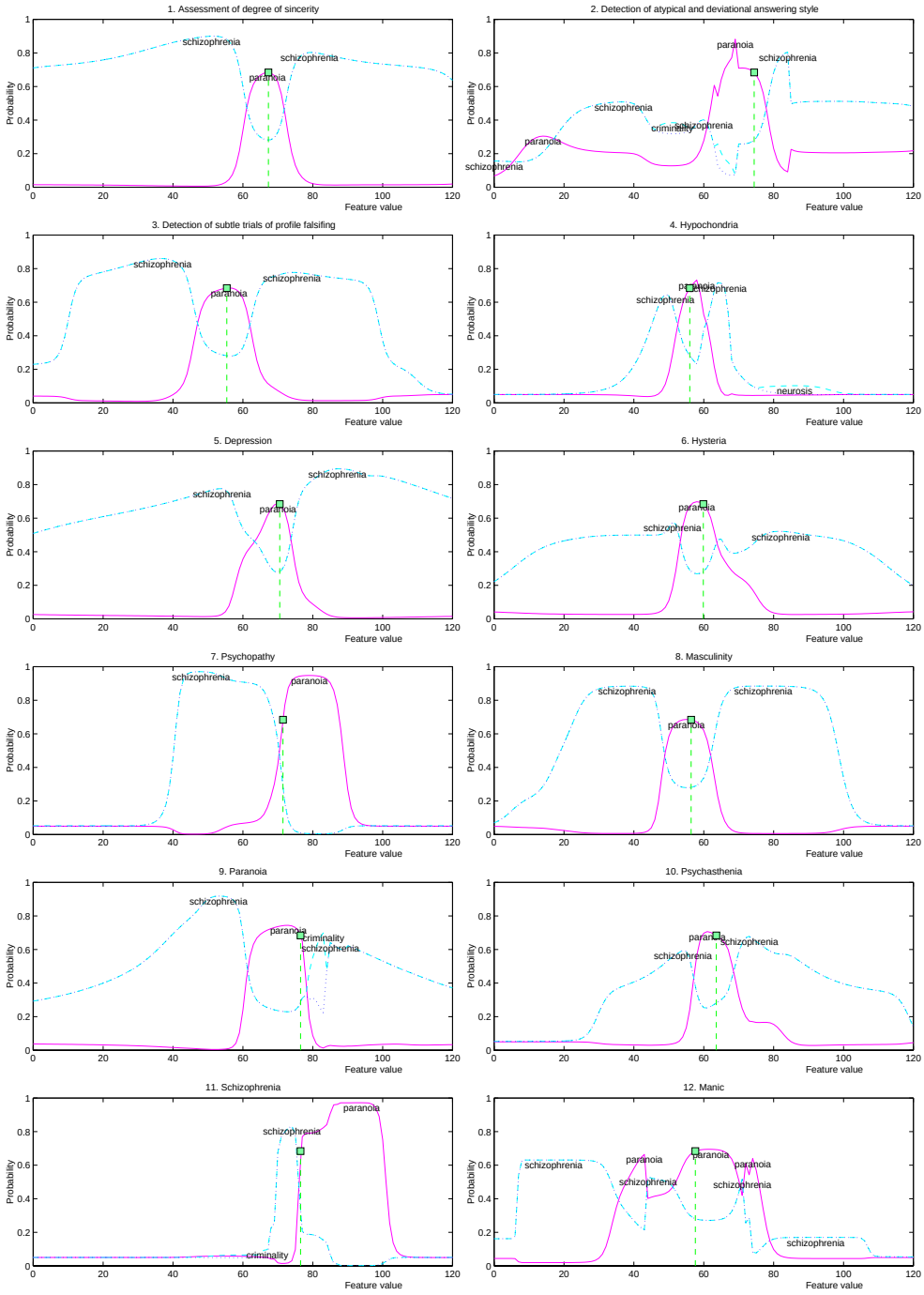


Figure 2: Class: Paranoia (prob. 0.68); alternative class: schizophrenia (prob. 0.28).

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 1 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. 2, is not so simple. 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.

# 4 CONCLUSIONS

Confidence intervals and probabilistic confidence intervals defined above are a new tool which may be very useful in the process of diagnosis. The most important is that confidence intervals and their probabilistic version are constructed (on-line) for a given case basing on previously estimated model. Information on winner and alternative classes is continuous and very precise in uncertainty estimation. 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 on 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.

Properties of probabilistic intervals of confidence make them a very useful diagnostic tools. Artificial neural networks may be interpreted using such tools, breaking the myth that neural networks are *black boxes*.

**Acknowledgments:** I would like to thank J. Gomuła and T. Kucharski for providing the psychometric data and comments.

# REFERENCES

[1] J. N. Buther, W. G. Dahlstrom, J. R. Graham, A. Tellegen, and B. Kaem-mer. *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. University of Minnesota Press, Minneapolis, 1989.

[2] N. Jankowski. Approximation and classification in medicine with IncNet neural networks. In *Machine Learning and Applications. Workshop on Machine Learning in Medical Applications*, pages 53–58, Chania, Greece, July 1999. (PDF).

[3] N. Jankowski. Approximation with RBF-type neural networks using flexible local and semi-local transfer functions. In *4th Conference on Neural Networks and Their Applications*, pages 77–82, Zakopane, Poland, May 1999. (PDF).

[4] N. Jankowski. *Ontogenic neural networks and their applications to classification of medical data*. PhD thesis, Department of Computer Methods, Nicholas Copernicus University, Toruń, Poland, 1999. (PDF).

[5] N. Jankowski and V. Kadirkamanathan. Statistical control of RBF-like networks for classification. In *7th International Conference on Artificial Neural Networks*, pages 385–390, Lausanne, Switzerland, October 1997. Springer-Verlag.