Comments and questions can be sent to: michael@cs.uu.nl

ELSEVIER

International Journal of Medical Informatics 46 (1997) 103–112

International Journal of Medical Informatics

Original article

Robustness metrics for measuring the influence of additive noise on the performance of statistical classifiers

M. Egmont-Petersen [a,*], J.L. Talmon [b], A. Hasman [b]

[a] *Department of Biophysics, Maastricht University, P.O.B. 616, NL-6200 MD Maastricht, The Netherlands*
[b] *Department of Medical Informatics, Maastricht University, Maastricht, The Netherlands*

Received 3 March 1997; received in revised form 25 May 1997; accepted 14 July 1997

Abstract

This paper presents a novel quality measure called robustness. The robustness measure quantifies the influence of measurement noise in the attribute values on the credibility of the classification of a case. It is assumed that the type of distribution of the noise-generating process is known. It is not simple to measure the robustness in the general situation where the noise-free distribution of the attributes is unknown. Therefore, two approximations are suggested and compared with the robustness measure based on the noise-free distribution of the attributes. The usefulness of the suggested robustness measure is explored in a simulation experiment. © 1997 Elsevier Science Ireland Ltd.

Keywords: Robustness measure; Quality assessment; Quality metrics; Bayesian classifier; Measurement noise; Remeasuring interval; Repeated measurements; Quality insurance

1. Introduction

Statistical classifiers have been developed for various medical classification tasks including diagnosis and therapy. The classification of a case can be based on many different types of attributes such as biochemical assays, bioelectric signals (EMG, ECG, EEG, etc.), patient history and clinical signs and symptoms. Some attributes are measured under uncertainty, as noise is inherent in the measurement process. Biochemical assays, for example, are usually contaminated by measurement noise.

Often, the amount of measurement noise can be influenced through quality control programs or simply by repeated measure-

* Corresponding author. Tel.: +31 43 3881665/58; fax: +31 43 3672287; e-mail: michael.egmont@bf.unimaas.nl

PII S1386-5056(97)00062-2

ments. Thereby, it is possible to increase the credibility of a tentative diagnosis based on attributes that are contaminated by measurement noise. Improving the 'signal-to-noise-ratio' (SNR), however, has its price. Measurement noise may only have influence on the decision when the measured value is observed close to a decision boundary relative to the variance of the measurement noise. When the measurement is far from any decision boundary, remeasuring the same sample will not lead to a change in classification.

In general, an object is represented by a set of attribute values that are observed from some underlying distribution. Due to measurement noise, the observed values differ from the true (but unknown) values. When a statistical classifier uses attributes that are measured with noise and the type of distribution of the noise generating process as well as its parameters are known, it is possible to quantify the uncertainty of a classification in relation to the measurement noise of the attributes. In this context we define:

> The robustness of a classification of a case is the probability that the case would obtain the same class label if the (unknown) true attribute values were known.

The concept of robustness of classifications was first introduced by Brender et al. suggesting it can be measured for a set of cases and hence is a property of a classifier [1]. The probability that a different class label can be obtained when one or more attributes is remeasured varies from case to case as this probability depends on how close the case is to the decision boundary. Our definition of robustness is a property of a classification of a particular case. Robustness resembles confidence as defined by Willard and Critchfield [2]. They measure the confidence of a classification of a case in relation to the variance of the parameters of the classifier. The difference is that the robustness measure assesses the uncertainty of a classification given the uncertainty inherent in the noisy measurements while the confidence measure relates the classification of a case to the uncertainty with which the parameters of the classifier were estimated.

In the following, we restrict ourselves to situations where the attributes are real valued. We assume that the distribution of the measurement noise in each attribute is Gaussian with a known variance. Firstly, we give a mathematical definition of the robustness of a classification. Secondly, we analyze two two-class problems in the special situations: (1) where the attributes are normally distributed (unimodal); and (2) where the class-conditional distributions are Gaussians (bimodal). Unbiased estimation of the robustness requires knowledge of the type and parameters of the distribution of the true, noise-free, attribute values. In the general case, this information is seldomly available. We therefore suggest two metrics that approximate the robustness measure. We will show to what extent these approximations result in biased estimates of the robustness for the unimodal and bimodal situations. Next, we discuss how the robustness measure can be used to guide the improvement of the SNR by means of repeated measurements. It is analyzed how often an attribute value has to be remeasured to ensure a classification with sufficient robustness.

2. Measuring the robustness

A classification task can be defined as a mapping from an n-dimensional attribute space to a c-dimensional class space. Let the c classes be characterized by the class-condi-

tional probability density functions (PDFs) of the true, noise-free, attribute values $p_t(\boldsymbol{t}|\omega_j)$, $j=1,\ldots,c$[1]. Let the corresponding PDF be defined as $p_t(t)=\Sigma_{j=1,\ldots,c} P(\omega_j)p_t(\boldsymbol{t}|\omega_j)$. Denote with $p(\boldsymbol{o}|\boldsymbol{t})$ the PDF of the measurement noise: the distribution of observable attribute values $\boldsymbol{o}$ given the true attribute values $\boldsymbol{t}$.

In the following, we will assume that the measurement process induces Gaussian noise with zero mean and that the noise in one attribute is independent of the noise in the other attributes as well as of the true attribute values $\boldsymbol{t}$:

$$p(\boldsymbol{o}|\boldsymbol{t}) = (2\pi)^{-n/2}\left|\sum_m\right|^{-0.5} \times \exp\left(-0.5(\boldsymbol{o}-\boldsymbol{t})^T\sum_m^{-1}(\boldsymbol{o}-\boldsymbol{t})\right) \quad (1)$$

with Σ_m a diagonal matrix, with entry (i,i), $i=1,\ldots,n$, the variance of the measurement noise of attribute i.

The PDF $p_o(\boldsymbol{o})$ of the measured (noisy) attribute values is given by the convolution of $p(\boldsymbol{o}|\boldsymbol{t})$ with $p_t(\boldsymbol{t})$ [3]

$$p_o(\boldsymbol{o}) = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} p(\boldsymbol{o}|\boldsymbol{t})p_t(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} \quad (2)$$

Knowledge about which true attribute values $\boldsymbol{t}$ could lead to the observed values $\boldsymbol{o}$ would allow us to compute the robustness. Using Bayes' rule we obtain

$$p(\boldsymbol{t}|\boldsymbol{o}) = \frac{p(\boldsymbol{o}|\boldsymbol{t})p_t(\boldsymbol{t})}{p_o(\boldsymbol{o})} \quad (3)$$

According to our definition, the robustness of a case classified as class j is given by

$$\varrho = \int_{S_j} p(\boldsymbol{t}|\boldsymbol{o})\,\mathrm{d}\boldsymbol{t} \quad (4)$$

[1] The subscript 't' indicates that $p_t(\boldsymbol{t}|\omega_j)$ is the PDF of the attribute values without measurment noise.

with S_j the region in the input space for which the classifier assigns class label j.

Simplifying to a situation with one attribute t, we may rewrite the robustness measure as

$$\varrho = \int_{S_j} p(t|o)\,\mathrm{d}t = \int_{S_j} \frac{p(o|t)p_t(t)}{p_o(o)}\,\mathrm{d}t \quad (5)$$

Using the fact that $p_o(o)$ is constant, the robustness can also be written as

$$\varrho = \frac{\int_{S_j} p(o|t)p_t(t)\,\mathrm{d}t}{\int_{-\infty}^{\infty} p(o|t)p_t(t)\,\mathrm{d}t} \quad (6)$$

It is clear that the robustness of a classification based on a contaminated observation o, is the probability that o is located in the range S_j.

3. Robustness in two simple situations

Many (intermediate) medical decisions rely on the interpretation of one attribute value. In different clinical disciplines, the specialist has to make decisions based on noisy measurements. Whenever he thinks that no credible decision can be made based on the available information, he can collect additional information or remeasure the quantity. The latter decision is sensible when the noisy measurement is close to a decision boundary as in that case remeasuring the quantity can lead to a different decision.

The attribute used can often be modelled with either a unimodal or a bimodal distribution. In the unimodal case, one often wants to identify whether the measurement exceeds a certain threshold. This can be the case, e.g. for diseases that gradually develop. In the bimodal situation the decision is whether a case belongs to one or another category, assuming that the case is from either of the two

distributions. Such situations occur when a clinical condition is either present or not. Bimodally distributed attributes could for example be clinical chemical tests where each group may often be characterized by a normal distribution.

3.1. The unimodal situation

A simple classification task is to discriminate two groups with an attribute t that is unimodally distributed. When the attribute is normally distributed, it can be shown that $p(t|o)$ is a Gaussian PDF. Let $p_t(t)$ be

$$p_t(t) = (2\pi\sigma_t^2)^{-0.5} \exp\left(-\frac{(t-\mu_t)^2}{2\sigma_t^2}\right) \tag{7}$$

with σ_t^2 the variance and μ_t the mean of the noise-free attribute values. Using the fact that the convolution of two Gaussian densities is also a Gaussian density [3], $p_o(o) = p(o|t)*p_t(t)$, with a variance equal to the sum of their variances, $p(t|o)$ becomes

$$p(t|o) = \frac{(2\pi\sigma_m^2)^{-0.5} \exp\left(-\frac{(o-t)^2}{2\sigma_m^2}\right)(2\pi\sigma_t^2)^{-0.5} \exp\left(-\frac{(t-\mu_t)^2}{2\sigma_t^2}\right)}{(2\pi(\sigma_t^2+\sigma_m^2))^{-0.5} \exp\left(-\frac{(o-\mu_t)^2}{2(\sigma_t^2+\sigma_m^2)}\right)} \tag{8}$$

which simplifies to

$$p(t|o) = (2\pi\sigma_{mt}^2)^{-0.5} \times \exp\left(-\frac{\left(t-\frac{o+\mu_t\sigma_m^2/\sigma_t^2}{1+\sigma_m^2/\sigma_t^2}\right)^2}{2\sigma_{mt}^2}\right) \tag{9}$$

with

$$\sigma_{mt}^2 = \frac{\sigma_m^2\sigma_t^2}{\sigma_m^2+\sigma_t^2} \tag{10}$$

The standard deviation of the distribution $p(t|o)$, σ_{mt}, is smaller than the standard deviation of $p(o|t)$, σ_m. The mean of $p(t|o)$

$$\frac{o+\mu_t\sigma_m^2/\sigma_t^2}{1+\sigma_m^2/\sigma_t^2} \tag{11}$$

is different from o, the measured attribute value, when $o \neq \mu_t$.

In the situation where the two classes are discriminated by a threshold λ, $S_1 = [-\infty, \lambda]$ and $S_2 = [\lambda, \infty]$, the robustness of a classification based on the measurement that indicated class label 1, $o \in S_1$, is given by

$$\varrho = \int_{-\infty}^{\lambda} p(t|o)\,dt \tag{12}$$

with p$(t|o)$ as defined in Eq. (9).

Fig. 1 shows the density $p(t|o)$, $o = 1$, $\mu_t = 0$, $\sigma_t^2 = 1.0$, for three different noise levels, $\sigma_m^2 = 0.0009$, 0.01, 0.09. The graph clearly shows that the deviation between the mean of $p(t|o)$ and o increases as a function of the noise level.

Fig. 2 shows the robustness for different values of o, $\mu_t = 0$ and the same three different noise levels as used above, $\sigma_t^2 = 1.0$. The threshold λ is set to 1.

An interesting observation is that the robustness of classifications on both sides of the boundary do not approach the same value when the measured attribute value approaches λ from the left and right side, respectively. This is explained by the fact that the mean of $p(t|o)$ is different from o and hence

$$\lim_{\varepsilon\to 0}\int_{-\infty}^{\lambda} p(t|\lambda-\varepsilon)\,dt \neq \lim_{\varepsilon\to 0}\int_{\lambda}^{\infty} p(t|\lambda+\varepsilon)\,dt$$

Another observation in Fig. 2 is that for observations smaller than λ the robustness

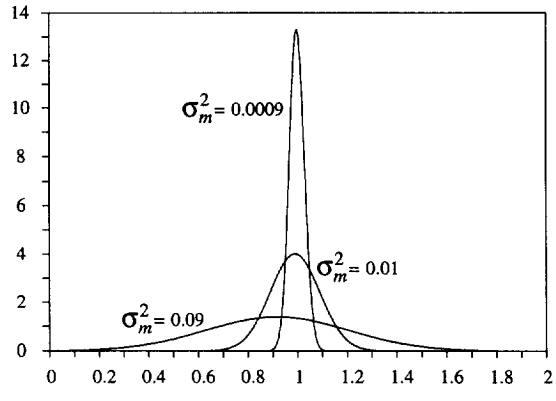

Fig. 1. The distribution $p(t|o)$ with $o = 1$, $\mu = 0$, for three different noise levels.

for a high level of measurement noise is not always smaller than the robustness for a low noise level, i.e. the robustness curves cross. For high noise levels, the area of the upper tail of $p(t|o)$ beyond λ becomes significant, even when o is at a considerable distance from λ. For measurements close to λ, the mean of $p(t|o)$ for high noise levels will be further away from λ than for lower noise levels. When o is close enough to λ, this difference in means outweighs the difference in variance and the area of the upper tail of $p(t|o)$ for a small noise level will

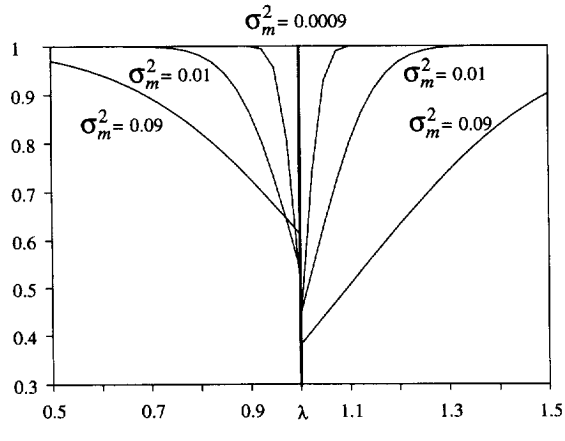

Fig. 2. The robustness computed for three different noise levels. The threshold separating the two classes $\lambda = 1$.

become larger than the area of the upper tail of $p(t|o)$ for higher noise levels.

3.2. Approximations

In practice, the densities $p_t(t)$ and $p_o(o)$ are often unknown. The density $p_o(o)$ can be approximated by a parametric distribution or estimated from a set of cases using, for example, histograms or kernel functions. The density $p_t(t)$ is, however, more problematic to estimate as it involves a deconvolution of the density $p_o(o)$ by $p(o|t)$. It is well-known that such computations are in many cases instable [4].

Rather than estimating $p_t(t)$, we assume that this density can be approximated by $p_o(t)$, the PDF of the noisy measurement. Using this in Eq. (3), we obtain

$$p^*(t|o) = \frac{1}{c(o)} \frac{p(o|t)p_o(t)}{p_o(o)} \tag{14}$$

with

$$c(o) = \int_{-\infty}^{\infty} \frac{p(o|t)p_o(t)}{p_o(o)} \tag{15}$$

such that $p^*(t|o)$ is a probability density function. It is easily shown that in the unimodal Gaussian case this approximation boils down to using Bayes' formula (Eq. (3)) in which $p_t(t)$ is replaced by $p_o(t)$ and $p_o(o)$ is substituted by a Gaussian that has the same mean as $p_o(o)$ but with a variance of $\sigma_t^2 + 2\sigma_m^2$. The approximation results in robustness values with twice the measurement noise.

The robustness metric ϱ^* is computed by integrating over $p^*(t|o)$

$$\varrho^* = \int_{S_j} p^*(t|o)\, dt \tag{16}$$

Fig. 3 shows the ratio ϱ^*/ϱ in the situation where $p_t(t)$ is a normal distribution, $\lambda = 1$. The graph illustrates that the bias increases as o approaches the decision boundary λ. For a variance $\sigma_m^2 = 0.09$, the induced bias is at most 2.5%. Note also the asymmetry around λ.

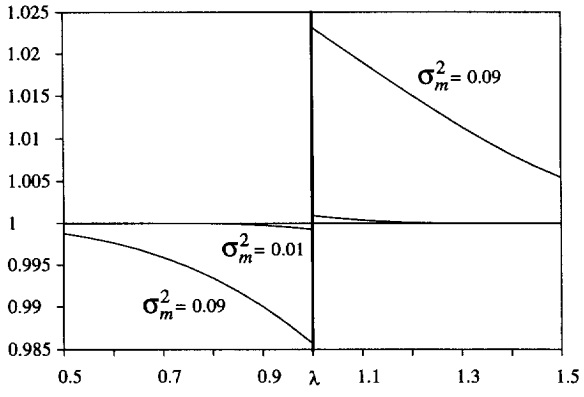

Fig. 3. The ratio of the robustness and its approximation, ϱ^*/ϱ, computed for two noise levels.

Another rough approximation would be to compute of the robustness using $p(o|t)$:

$$\varrho^{**} = \int_{S_j} p(o|t)\,\mathrm{d}t \tag{17}$$

Fig. 4 shows the ratio ϱ^{**}/ϱ in a setting similar to Fig. 3. The graph illustrates that ϱ^{**} is a much more biased estimate of the robustness than ϱ^*. Therefore, in the unimodal situation, we recommend to use the approximation ϱ^* which results in maximally 2.5% bias for a SNR (defined as σ_t^2/σ_m^2) equal to 10.

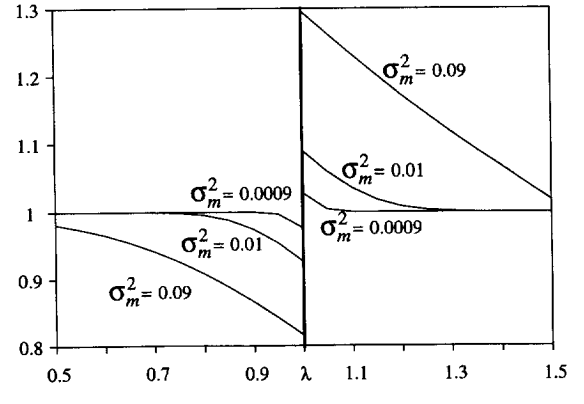

Fig. 4. The ratio ϱ^{**}/ϱ between the robustness and the second approximation computed for three different noise levels.

3.3. The bimodal situation

Another typical classification problem in medicine is the discrimination of two groups where each group is unimodally distributed. For a two-class problem with the class-conditional densities $p_t(t|\omega_1)$ and $p_t(t|\omega_2)$, $p_t(t) = P(\omega_1)p_t(t|\omega_1) + P(\omega_2)p_t(t|\omega_2)$ is the density of the attribute without measurement noise and $P(\omega_1)$ and $P(\omega_2)$ the prior probabilities of the two classes. The robustness is computed by integrating t over the density

$$p(t|o) = \frac{p(o|t)\sum_{j=1}^{2} P(\omega_j)p_t(t|\omega_j)}{\sum_{j=1}^{2} P(\omega_j)p_o(o|\omega_j)} \tag{18}$$

This can be written as

$$\frac{P(\omega_1)p(o|t)p_t(t|\omega_1)}{p_o(o)} + \frac{P(\omega_2)p(o|t)p_t(t|\omega_2)}{p_o(o)} \tag{19}$$

and

$$\frac{P(\omega_1)p_o(o|\omega_1)\dfrac{p(o|t)p_t(t|\omega_1)}{p_o(o|\omega_1)}}{p_o(o)} + \frac{P(\omega_2)p_o(o|\omega_2)\dfrac{p(o|t)p_t(t|\omega_2)}{p_o(o|\omega_2)}}{p_o(o)} \tag{20}$$

Using Eq. (3), this expression can be written as

$$\frac{P(\omega_1)p_o(o|\omega_1)p(t|o,\omega_1)}{p_o(o)} + \frac{P(\omega_2)p_o(o|\omega_2)p(t|o,\omega_2)}{p_o(o)} \tag{21}$$

which simplifies to

$$p(t|o) = \frac{p(t|o,\omega_1)}{1+\dfrac{P(\omega_2)p_o(o|\omega_2)}{P(\omega_1)p_o(o|\omega_1)}} + \frac{p(t|o,\omega_2)}{1+\dfrac{P(\omega_1)p_o(o|\omega_1)}{P(\omega_2)p_o(o|\omega_2)}} \tag{22}$$

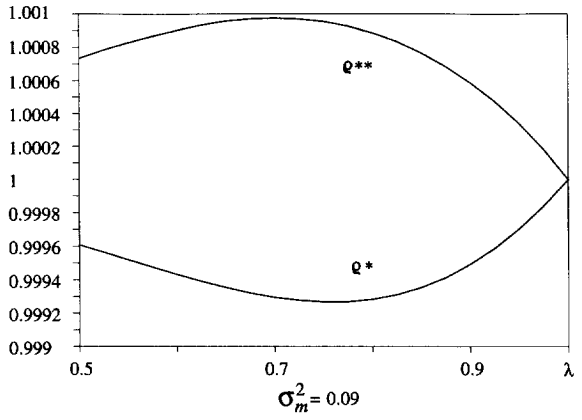

Fig. 5. The relation between the robustness and the approximations suggested ϱ^*/ϱ and ϱ^{**}/ϱ computed for the highest noise level in the symmetric bimodal situation.

It can be seen that the distribution of t given an observation o is a weighted sum of the class-conditional distributions $p(t|o, \omega_1)$ and $p(t|o, \omega_2)$ each of which corresponds to the density derived in Eq. (9). The two distributions $p(o|\omega_1)$ and $p(o|\omega_2)$ and the prior probabilities determine the two weights. When o is observed far from the decision boundary one weight will be close to 1 and the other close to 0. Nearby the minimal error boundary, however, both weights approach 1/2.

In this situation, we have used the same approximations as in the previous section. Fig. 5 shows the ratios ϱ^*/ϱ and ϱ^{**}/ϱ in the bimodal situation where $p_t(t|\omega_1)$ and $p_t(t|\omega_2)$ are normal distributions, $\lambda = 1$, $\sigma^2_{t|\omega_1} = \sigma^2_{t|\omega_2} = 1$, $\sigma^2_m = 0.09$. Only observations on one side of the decision boundary λ are shown. The two curves are symmetric round the decision boundary as the prior probabilities are equal, $P(\omega_1) = P(\omega_2)$, as are the variances of the two Gaussian distributions $p(t|o, \omega_1)$ and $p(t|o, \omega_2)$. In this situation, the robustness computed by ϱ, ϱ^* and ϱ^{**} are symmetrical around the decision boundary. The graph in Fig. 5 indicates that even for small signal-to-noise ratios the bias induced by the two metrics can be neglected, it is maximally 1‰.

Other analyses indicated that when the two class-conditional distributions have different variances, or when the prior probabilities were unequal, the approximation with the metric ϱ^{**} deviates quicker than the approximation with the metric ϱ^*. When the prior probability of one of the classes becomes vary small, the unimodal situation is approached.

4. Application

The robustness measure can be useful in two situations. Firstly, it can be used to identify which attributes are the most critical for the classification of a case. Secondly, the robustness measure can be used as a means for quality improvement. It is possible to increase the robustness of a classification by repeated measurements.

It is well-known that the average of r independent measurements approaches a normal distribution with a variance σ^2_m/r. The PDF $p(o|t)$ becomes

$$p(\bar{o}|t) = (2\pi\sigma^2)^{-0.5} \exp\left(-\frac{(\bar{o}-t)^2}{2\sigma^2}\right) \qquad (23)$$

with $\bar{o}$ the average of the r measurements. Substituting the density $p_o(o)$ with the density $p_{o|r}(\bar{o})$ using Bayes' rule, the robustness of the classification based on $\bar{o}$ can be computed from

$$\varrho = \int_{S_j} \frac{p(\bar{o}|t)p_t(t)}{p_{o|r}(\bar{o})}\,\mathrm{d}t \qquad (24)$$

or in a similar manner using one of the two approximations.

We performed three simulation experiments with a unimodally distributed attribute. The purpose was to investigate how

often the attribute has to be remeasured to ensure a classification with a robustness higher than β. In the first experiment, the following parameter settings were used: $\mu_t = 0$, $\sigma_t^2 = 1$, $\lambda = 1$, $\sigma_m^2 = 0.09$ and $\beta = 0.975$. The measured attribute value o was systematically varied from O to $3\sigma_t^2$ in steps of 0.01. For each value of o, we computed the robustness. When the robustness of a classification was less than β, we generated 1000 realizations of possible values from $p(t|o)$. For each of these realizations, a value was drawn from $p(o|t)$ and the average $\bar{o}$ was computed. New values were drawn from $p(t|o)$ and $\bar{o}$ recomputed until a classification with a robustness higher than β could be obtained. So each value of t resulted in a distribution of the number of necessary remeasurements. This procedure is schematically shown in Fig. 6. We plotted the 5, 50 and 95-percentiles of these 1000 estimates of n for each value of o.

In the second experiment, the parameter settings were: $\mu_t = 0$, $\sigma_t^2 = 1$, $\lambda = 2$, $\sigma_m^2 = 0.09$ and $\beta = 0.975$. In the third experiment, the parameter settings were: $\mu_t = 0$, $\sigma_t^2 = 1$, $\lambda = 1$, $\sigma_m^2 = 0.01$ and $\beta = 0.975$.

Quality control experiment procedure

```
for o=0 to 3*var(t) step 0.01 do
  for i=1 to 1000 do
    t=draw(p(t|o))
    o'=draw(p(o|t))
    n=1
    while robustness(o')<β and n<1000 do
      o''=draw(p(o''|t))
      o'=(n×o'+o'')*(n+1)
      n=n+1
    end while
    print(o,n)
  end for i
end for o
```

Fig. 6. Procedure followed in quality control experiments.

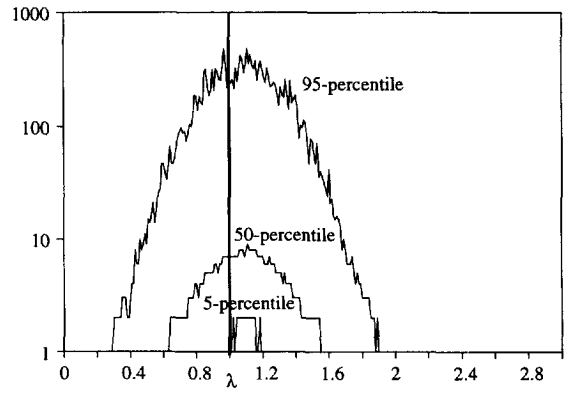

Fig. 7. The 5, 50 and 95-percentiles of the number of repeated measurements necessary to obtain a robust classification as a function of o. Experiment 1.

Figs. 7–9 indicate how often one has to remeasure the attribute in the three situations to obtain a robust classification. Note the logarithmic scale of the ordinate. The first simulation indicates that for $0.8 \leq o \leq 1.4$, the 50-percentile of the measurements is already larger than three. The second simulation gave a similar result; when $1.9 \leq o \leq 2.5$, the 50-percentile of the measurements is already larger than three. In the third simulation, the corresponding interval is $0.9 \leq o \leq 1.1$.

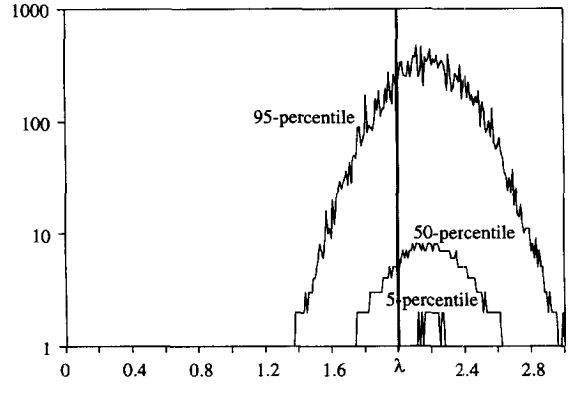

Fig. 8. Number of repeated measurements in second simulation experiment.

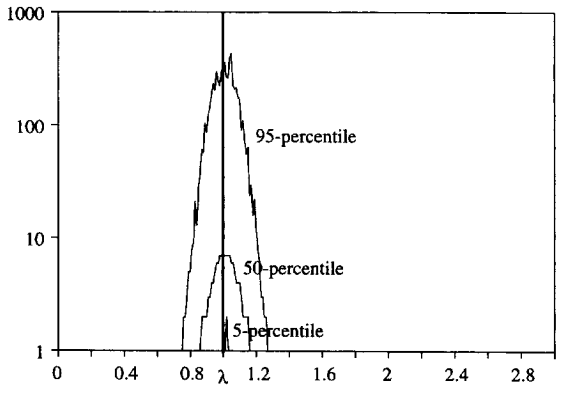

Fig. 9. Number of repeated measurements in simulation 3.

In general, this type of simulations give insight in for which ranges of o remeasurement is sensible for specific measurement costs. On each side of the class boundary there is an interval bounded by, at one side observations that always result in robust classifications. To the other side of this interval are values of o for which a robust classification is unlikely to be obtained even when the attribute is remeasured. In the latter situation, a decision should rely on other sources of information. We define these two intervals $I_{\text{left}} = [\gamma_{\text{left}}, \eta_{\text{left}}]$ and $I_{\text{right}} = [\gamma_{\text{right}}, \eta_{\text{right}}]$ as the remeasuring intervals of the noisy attribute o, with $\gamma_{\text{left}}, \eta_{\text{left}}, \gamma_{\text{right}}, \eta_{\text{right}} \in (\mu_t, \infty)$ when $\lambda > \mu_t$. The remeasuring intervals I_{left} and I_{right} are determined by the variances σ_t^2 and σ_m^2, the decision boundary λ and the minimally required robustness, β. When an attribute is observed in either I_{left} or I_{right}, it makes sense to remeasure it. The bounds of I_{left} and I_{right} can for a specific measurement situation be determined through simulations.

The three figures show that the maxima of the percentile curves appear at values larger than the decision boundary λ. This effect is caused by the 'regression to the mean' effect. When one extreme observation is made, the probability of obtaining again an observation that is just as extreme, is smaller than 0.5. Determining the remeasuring interval of an attribute using the same type of simulations as performed here, takes this effect into account.

5. Discussion

We have introduced a measure to quantify the influence of measurement noise on the classification result. The measure is in general difficult to compute as it involves a deconvolution of the probability density function of the noisy attribute values. We proposed two metrics that approximate the robustness measure. Analyses indicated that in the situations where an attribute is unimodally or bimodally distributed, one of the metrics ϱ^* had a bias that is smaller than 2.5% for a SNR = 10. The second robustness metric ϱ^{**} is only a good approximation in the bimodal situation when the two class conditional distributions have about equal variances and identical prior probabilities. We have only investigated the bias of the two metrics when the distributions are normal. In the general case, where the distributions are unknown, the unimodal or bimodal situations may serve as an idealized model.

The experiments indicate that there is one (in general possibly more) interval(s) on each side of the decision boundary where remeasuring the attribute value makes sense. The limits of the remeasuring intervals depend on the measurement imprecision as well as the costs associated with remeasuring an attribute.

We assumed in this paper that the variance of the measurement noise σ_m^2 is known. In practice, σ_m^2 has to be estimated using repeated measurements. In most clinical chemistry laboratories, for example, it is common to estimate the measurement noise of the

biochemical assays as part of the quality control procedures. In such a situation, reliable estimates of σ_m^2 are available and the robustness of a diagnosis based on one or more noisy measurements can be computed. If the attribute is for example the pixel intensity in a digital radiograph, the measurement noise can be estimated as the intensity variation in a homogeneous part of the image (e.g. the background).

6. Conclusion

We have presented a novel measure for the quality aspect of the classification of a case called robustness. It expresses the probability that the true, noise-free, attribute value would result in the same class label as the observed ones. We examined the properties of the robustness for two classification problems that can be seen as prototypical for medical application in, e.g. biochemistry. As the theoretically correct approach involves a deconvolution, which is computationally problematic, we propose two approximations in the form of metrics. We examined the effect of these approximations for the classification problems. We concluded that for classification tasks that are often based on identifying whether a case has occurred in the upper or lower tail of a distribution, the approximation based on the observed attribute distribution gives credible results. For a two class problem, the approximation solely based on the distribution of the measurement noise was also a good approximation. In a practical measurement situation, one may benefit from the robustness measure in two ways. Firstly it makes it possible to relate the credibility in a classification to the measurement imprecision. Secondly, the robustness measure can be used to identify remeasuring intervals on each side of a decision boundary for which remeasuring the attribute makes sense. If a measurement is too close to the boundary, one might be better off performing another test.

Although the robustness measure was defined in the general situation with n attributes that are observed under noisy measurement conditions, we have analyzed the approximations only in the situation where the classification depends on one attribute. Measuring the robustness of a classifier based on more than one attribute is in practice not trivial for nonlinear classifies such as neural networks. They discern the classes by n-dimensional hypersurfaces (the hidden nodes) and one may have to take recourse to Monte Carlo integration to estimate the robustness for more attributes. In such a situation, the rough approximation ϱ^{**} may be a practical solution to obtain a first estimate of the robustness.

References

[1] J. Brender. P. McNair, H. Raun, J. Nolan, S. Vingtoft, Metaknowledge as a means for quality management in knowledge-based systems, in: R. O'Moore, S. Bengtsson, J.R. Bryant, J.S. Bryden (Eds.), Proceedings of Medical Informatics in Europe 1990, Lecture Notes in Medical Informatics, Springer Verlag, Berlin, 1990, pp. 360–368.

[2] K.E. Willard, G.C. Critchfield, Probabilistic analysis of decision trees using symbolic algebra, Decision Making 6 (1986) 93–100.

[3] E. Parzen, Modern Probability Theory and its Applications, Wiley, New York, 1960.

[4] R.C. Gonzalez, R.E. Woods, Digital image processing Addison Wesley, Reading, 1992.