Rare Fault Detection by Possibilistic Reasoning

Marc Thomas, Andreas Kanstein, and Karl Goser

Microelectronics Department,
University of Dortmund, 44221 Dortmund, Germany.
E-mail: thomas@luzi.e-technik.uni-dortmund.de

Abstract. Kernel based neural networks with probabilistic reasoning are suitable for many practical applications. But influence of data set sizes let the probabilistic approach fail in case of small data amounts. Possibilistic reasoning avoids this drawback because it is independent of class size.

The fundamentals of possibilistic reasoning are derived from a probability/possibility consistency principle that gives regard to relations. It is demonstrated that the concept of possibilistic reasoning is advantageous for the problem of rare fault detection, which is a property desired for semiconductor manufacturing quality control.

1 Introduction

Possibilistic reasoning allows the implementation of a possibility based classifier analogous to a classifier based on Bayes' theorem. The new concepts main advantage is found in problems of rare fault classification. The possibilistic reasoning approach is derived analogue to the probabilistic one. Possibilistic reasoning is based on a probability/possibility consistency principle different from the definition of Dubois and Prade [1].

Applications under investigation are dealing with fault detection in semiconductor fabrication such as defect density analysis, production line analysis, and tests of power semiconductor devices. In these cases faults are often limited to about a few percent [2]. Furthermore, problems are growing with separating different types of faults.

The possibility based classifier is implemented in a kernel based neural network similar to radial basis function networks that implement a suboptimal Bayesian classifier. Kernel based neural networks are a favoured technique for classifying tasks. They provide quick learning, an interpretable structure and robustness. These features result from their equivalence to certain fuzzy systems and from kernel adaptation by competitive learning of data clusters. That accounts for the growing interest in this neural network type.

2 Motivation

The usual way of implementing a classifier in a kernel based neural network is a radial basis function network (RBFN). The kernel neurons are trained by competitive learning to represent clusters of data. The composition of the activations of kernel neurons approximates a posteriori probabilities of classes. In this way the network implements a suboptimal Bayes classifier. This concept has been proven very powerful in many applications.

The main disadvantage of Bayesian classifiers is their inability to classify rare faults. If the size of the classes differ strongly, the interesting class representing faults is likely to be covered by probabilities of classes with a large number of data that represent the normal case. The Bayesian classifier might also fail due to badly estimated probability distributions if the number of samples is very low. Figure 1 displays both cases in drastic but illustrative examples. Misclassification occurs by choosing a disadvantageous criterion, because rare fault detection is made impossible by a dominating probability of the non-fault case. Low a priori probability can cause this effect, shown in Fig. 1(a). Misclassification also occurs by badly estimated probability distributions. Consider two adjacent uniform distributions, one small and one large. Estimating the probability distributions by very few examples can result in a dominating situation as shown in Fig. 1(b).

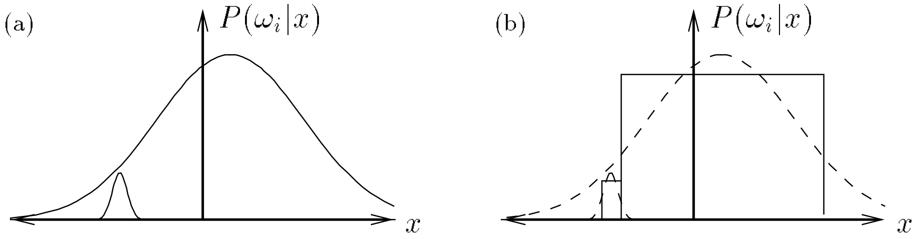

Fig. 1. Misclassification of rare faults in Bayesian classifiers due to dominating probability distributions. (a) Overlapping of distributions due to low a priori probability of rare faults. (b) Bad estimation of adjacent uniform distributions of different size due to very few samples of data.

These problems are very likely to appear in semiconductors manufacturing. Because of processing complexity the number of different fault classes is very large while the amount of data of single classes is low. The interest in the analysis of data that is collected in these processes is growing. The new concept of possibilistic reasoning implemented in a kernel based neural network is a promising approach because of the features listed above. In the following, the derivation of a decision rule from the definition of possibility distributions and a consistency principle is presented.

3 Possibilistic Approach

The set M of possible elements x includes all events that can ever appear and can be defined by probability values.

$$M := \{x \in X \mid P(x) > 0\} \quad .$$

The membership function of element x is given by:

$$\mu_M(x) = \begin{cases} 1 \text{ for } P(x) > 0 \\ 0 \text{ for } P(x) = 0 \end{cases} .$$

Because the set M is not known it has to be estimated using given data. Considering the uncertainty in this process, continuous possibility values in [0;1] are introduced. The similarity to fuzzy sets is obvious.

Definition 1. The mapping $\mathit{\Pi}\colon X \to [0;1]$, $X \subseteq \mathbb{R}^d$ is a possibility distribution.

A distinction between continuous and discrete universes, as necessary for probabilities, is not needed for these distributions. In the following continuous possibility distributions are denoted as $\pi(x)$.

To use possibilities as probabilities are in Bayes' theorem a connection between both has to be defined. The known consistency principles of Zadeh [3] or Dubois and Prade [1] are based on absolute values and do not regard relations as preferable for classification. The presented consistency keep this in mind.

Definition 2. A possibility and a probability distribution are *relational consistent*, iff

$$\forall\, x_1, x_2 \in \mathbb{R}^d : \quad \pi(x_1) > \pi(x_2) \Rightarrow p(x_1) \geq p(x_2) \ .$$

Equivalent to this definition is $\forall\, x_1, x_2 \in \mathbb{R}^d : p(x_1) \geq p(x_2) \Rightarrow \pi(x_1) > \pi(x_2)$. The properties of probabilities, the common concept of both and the relational consistency lead to the well known t-norm for intersection and s-norm for union of sets.

By using probabilities a distinction between the normal and the conditional case is necessary. To facilitate a decision as in Bayes classification, conditional possibility is required. Because t-norm and s-norm operators for intersection and set union of possibilities are not based on the property of dependent or independent events. Then the following relation can be used: $\mathit{\Pi}(A \cap B) = \tau(\mathit{\Pi}(A|B), \mathit{\Pi}(B))$; ($\tau$ is t-norm). Then $\mathit{\Pi}(A|B)$ can be calculated by $\mathit{\Pi}(B|A)$, $\mathit{\Pi}(A)$ and $\mathit{\Pi}(B)$ for a bijective operator or special cases.

Possibilistic Classification

The decision criterion is derived by minimizing the error possibility given by $\mathop{\mathrm{S}}_{i \neq r} \mathit{\Pi}(\omega_i)$ with $\{\omega_i\}$ all decisions and ω_r the resulting decision. Using the possibilistic approach analogous to Bayes decision theory, this criterion results in the following decision rule:

$$\text{Choose } \omega_r \text{ with } \max_{\omega_i \in \mathit{\Omega}} \tau \left(\mathop{\mathrm{S}}_{j} \tau(K(\mathbf{x}, \mathbf{m}_{ij}, \mathbf{s}_{ij}), \mathit{\Pi}(\omega_{ij}|\omega_i)), \mathit{\Pi}(\omega_i) \right) \ .$$

$\mathit{\Pi}(\omega_{ij}|\omega_i)$ and $\mathit{\Pi}(\omega_i)$ are the a priori possibilities and are free parameters of the classifier. τ is a t-norm and S denotes an s-norm. The trained clusters ω_{ij} given by centers $\mathbf{m}_{ij}$ and widths $\mathbf{s}_{ij}$ are specified by the kernel function $K()$

which computes the possibility $\pi(\mathbf{x}|\omega_{ij})$ of a vector $\mathbf{x}$ belonging to the cluster ω_{ij} of class ω_i. In contrast to probabilistic density estimation the kernel function need only be monotonous in distance. The used clustering algorithm is an online version of k-means with kernel function as distance measure. The process of classification, especially the clustering part is described in detail in [4].

4 Example

To demonstrate the target of possibilistic classification, an artificial dataset is used. The data are given by two overlapping gaussian distributions with one of them having a five times larger standard deviation and ten times greater amount of samples.

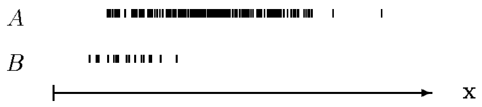

Fig. 2. Two overlapping gaussian distributions A and B

An RBF network, recognizing two clusters, gives a global error of 9.1% in hold-one-out. Best result by possibilistic reasoning is a value of 15.9%, keeping the a priori possibilities $\Pi(\omega_i)$ fixed to one. But examination of RBF-classification shows that always the bigger class A is determined. This is in final no classification for all inputs at all. Using possibilistic decision making results in an error of 16.9% in class A and 10% in B. This allows usable classification despite of a larger global error.

5 Conclusion

The presented approach for classification based on possibilistic reasoning allows to detect rare classes while keeping the facility of an implementation in a kernel based neural network. As the classical radial basis function approach stays as a special case of this network, the operation range of that neural network method is extended. The approach demonstrated here facilitates to handle the named task in semiconductor environment.

Acknowledgment: This work is supported by the "Deutsche Forschungsgemeinschaft (DFG)", grant number Go-379/11 and a grant in the Collaborative Research Center on Computational Intelligence (SFB 531).

References

1. D. Dubois and H. Prade, *Fuzzy sets and systems: Theory and Applications*, vol. 144 of *Mathematics in science and engineering*. Boston: Academic Press, 1980.
2. Ö. Hallberg, "Facts and fiction about the reliability of electronics," in *Proceedings of ESREF'95*, (Bordeaux, France), pp. 39–46, 1995.
3. L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," in *Fuzzy Sets and Systems 1*, pp. 3–28, North-Holland, 1978.
4. A. Kanstein and K. Goser, "Dynamic learning of radial basis functions for fuzzy clustering," in *Proceedings of IWANN'95*, (Málaga, Spain), pp. 513–518, 1995.
5. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: J. Wiley, 1973.
6. J. Moody and C. Darken, "Learning with localized receptive fields," in *Proc. 1988 Connectionist Summer School*, (San Mateo), pp. 133–143, Morgan Kaufmann, 1988.
7. E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, pp. 1065–1076, 1962.