18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

Gradual vs. Binary Conflicts in Bayesian Networks Applied to Sensor Failure Detection

Max Krüger
Hochschule Furtwangen (HFU)
University of Applied Sciences
Robert-Gerwig-Platz 1, D-78120 Furtwangen, Germany
E-mail: max.krueger@hs-furtwangen.de

***Abstract*—Bayesian Networks have various applications including medical and technical diagnosis, financial scoring, and target behavior/pattern recognition. Bayesian Classification Networks fuse evidence from heterogeneous and homogeneous sources and calculate classification results. For many reasons, pieces of evidence from different sources can carry apparently contradicting information and in these cases are called conflicting evidence. Diagnostic sensor failure-tests for application in Bayesian classification processes may be based on a binary conflict definition or a gradual conflict-level measure. This paper investigates four different failure-tests: (1) binary Conflict Binomial, (2) binary Conflict Ratio, (3) gradual Average Conflict, and (4) gradual Gauss Conflict, with (3) and (4) being new failure-test proposals. In a comparative air surveillance simulation, the detection performance of these diagnostic sensor failure-tests is evaluated and compared.**

Keywords: Conflicting evidence in Bayesian Networks, gradual and binary conflict measures, classification, identification, diagnostic failure-test, sensor failure detection, detection performance, simulation, air surveillance systems

I. Introduction

Applications of classification are widely spread in different areas, including medical diagnosis, financial credit scoring, and classification of Chilean wines [1]. In technical fields, pattern recognition and technical monitoring are typical applications [2]. Civil and military air surveillance systems need classification functionality to assign an identity to each tracked object, see [3], [4], [5]. In this context, classification is also called *identification* and typically requires automated assistance, compare [6].

Formally, classification can be defined as the task of making "[...] a decision or forecast [...] on the basis of currently available information [...], in which each new [classification] case must be assigned to one out of a set of predefined *classes* on the basis of observed *attributes* or *features*" [7, p. 1].

Bayesian Networks in general, and Naïve Bayes Networks in particular provide an established approach to classification, see e.g., [2, ch. 2.11], [8, ch. 8], [9, pp. 205-210]. Performing classification with Bayesian Networks, for each considered object probabilities of possible resulting classes are calculated based on observations of its attributes or features. These observations are called *evidence*, *feature values*, *findings*, or *source declarations* and are typically provided by a number of heterogeneous sources, compare [8, pp. 265-267].

Findings from different sources are called *conflicting evidence*, if they carry reliable but substantially different information, that taken alone, indicate different classification results, compare [8, pp. 174-179], [10]. Conflicting evidence may result from deficits and inaccuracies in modeling, or may have technical reasons, such as flawed or inaccurate sensor measurements, or failure of raw data evaluation [8, pp. 174-179], [11].

In our previous work [12], [13], binary indication of a source in conflict (or not) was used to detect sensor failures in Bayesian classification. It was shown, that active sensors with failures are involved in a significantly increased number of conflicts, a fact that can be used for implementation of automatic diagnostic sensor failure-tests, see [12], [13]. Our present contribution examines, whether a gradual conflict measure providing a continuous conflict value is able to further improve diagnostic performance of the previously provided sensor failure-tests.

The outline of this paper is as follows: Section II provides a short introduction into classification based on Bayesian Networks and introduces binary and gradual conflicts. Sensor failure detection's context is given in Section III, where in particular different diagnostic tests based on binary and gradual conflicts are defined. Following, in Section IV measures of diagnostic failure-tests' performance, a simulation scenario, and a setup for a comparative simulation are described. In Section V the simulation results are presented and discussed. Finally, conclusions are provided in Section VI and some future work is outlined.

II. Conflicts in Bayesian Networks

Bayesian Networks are an established mean to perform classification, see [8, ch. 8], [14, p. 727]. Naïve Bayes Classifiers [8, pp. 266-267] are special Bayesian Networks, that stand out by simplicity and amazing performance, see [15]. For this reason, we only consider Naïve Bayes Networks in this contribution. But note, that all definitions, concepts and in particular all diagnostic failure-tests can be transferred to general Bayesian Networks in a straight-forward manner.

Naïve Bayes Networks consist of a set C of all possible classification results, and finite sets S_i of all possible evidence of source S_i. These particular networks realize a classifier function $\mathrm{cl} : S_1 \times \ldots \times S_N \to C$ [8, p. 265].

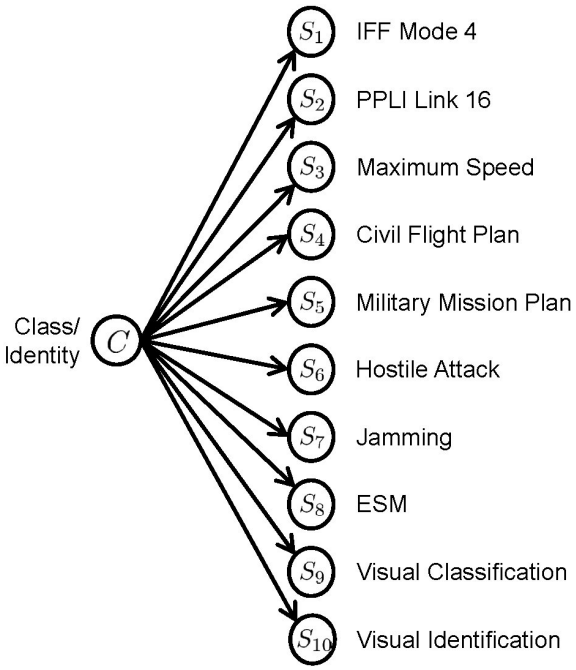

Figure 1. Naïve Bayes Network for Classification [12, slightly modified]

S_i and C also represent the states of discrete random variables, which are nodes of the Naïve Bayes Network. For simplicity reasons, we likewise denote these random variable by S_i and C, whereat each variable takes, with a certain probability, one out of a finite number of states $d_{i,1}, \ldots, d_{i,N_i} \in S_i$ and $c_1, \ldots, c_M \in C$, respectively. Note, that states of random variables are node-dependent and mutually exclusive. Figure 1 shows the graph of an exemplary Naïve Bayes Network for classification, taken from [12].

Inference with Bayesian Networks is based on application of the *Theorem of Bayes*

$$p(c_i|d_1, ..., d_N) = \frac{p(d_1, ..., d_N|c_i) \cdot p(c_i)}{\sum_{j=1}^{N} p(d_1, ..., d_N|c_j) \cdot p(c_j)} \,. \qquad (1)$$

By this theorem, posterior probabilities of classification results $c_1, \ldots, c_M \in C$ are calculated, based on given pieces of evidence $d_1 \in S_1, \ldots, d_N \in S_N$ provided by N sensors, e.g., those given in Figure 1.

Taking an application expert's viewpoint, 'conflicting evidence' is dissonant information from several sources [10], whereat 'dissonance' means "[...] the extent to which information is explicitly contradictory or conflicting" [16, criterion II.8.2.1.5]. A classical example is a very reliable '*valid IFF (Identification Friend or Foe) mode 4 response*' of an aircraft, which would be in conflict with an '*attack on own forces*' declaration by an other source.

From a technical perspective, the experts' definition is too vague for implementation. Technical definition approaches for conflicting evidence reach back to [17] in 1976. More deeply in context of Bayesian Networks, conflicts have been considered in the 1990s by [18], [11], [19], and [8, pp. 174-179]. Besides formal definition approaches, this research pointed out, that a discrepancy between model and declared pieces of evidence [8, p. 99] is the major characteristic of conflicts' occurrence. Typically, these discrepancies can be traced back to rare cases, a model not covering the actual situation, or flaws and inaccuracies of sensor measurement and raw data evaluation. Comparison of technical conflict definitions with the intuitive understanding of application domain experts were investigated in [10].

Primarily, all of these approaches try to define a criterion for a conflict being present or not. We call this a binary conflict. In contrast, a *gradual conflict* is supposed to measures the actual conflict level between sources on a continuous scale. In the following we will provide more formal definitions for each type.

Binary Conflicts

Apparently, the most established definition approach of conflicts is according to [8, pp. 175-176], [11]. Based on this approach, the slightly modified 'Coherence Conflict Measure' [10] leads to a conflicts' definition considering one source S_i compared to all other sources: Given declared pieces of evidence $d_1 \in S_1, \ldots, d_N \in S_N$, source S_i is according to [10] in conflict with all other sources, iff

$$\frac{\left(\sum_{j=1}^{M} p(d_i|c_j)\right) \cdot \left(\sum_{j=1}^{M} p(d_{-i}|c_j)\right)}{M \cdot \sum_{j=1}^{M} p(d_1, ..., d_N|c_j)} > (1 + \varepsilon_{coh}) \qquad (2)$$

holds, with $d_{-i} := (d_1, ..., d_{i-1}, d_{i+1}, ..., d_N)$ for convenient notation. Small fluctuations are meant to be suppressed by a threshold value $\varepsilon_{coh} > 0$.

This conflict's definition based on Equation (2) considers the discrepancy between declaration d_i of source S_i and all other sources' combined declarations. As the approach in [8, pp. 175-176], [11], it is based on the idea, that *coherent pieces of evidence support each other*, expressed by the inequality $\frac{p(d_i) \cdot p(d_{-i})}{p(d_1, ..., d_N)} < 1$, compare [11]. Thereby Equation (2) states in essence, that in a conflict case the joint occurrence $d_1, ..., d_N$ is less likely than the occurrence of d_i and d_{-i}, if considered independently of each other. More details can be found in [10]. Note, that using this binary conflict definition, it can always be determined, whether a conflict is present or not.

Gradual Conflicts

Likewise based on the Coherence Conflict Measure [10] the actual conflict level of source S_i can be measured for declared pieces of evidence $d_1 \in S_1, \ldots, d_N \in S_N$ by the conflict-level

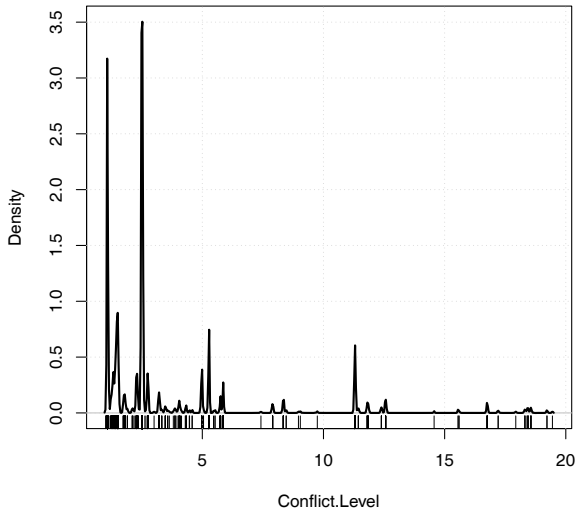

Figure 2. Density Estimation of Sources S_1: IFF Mode 4

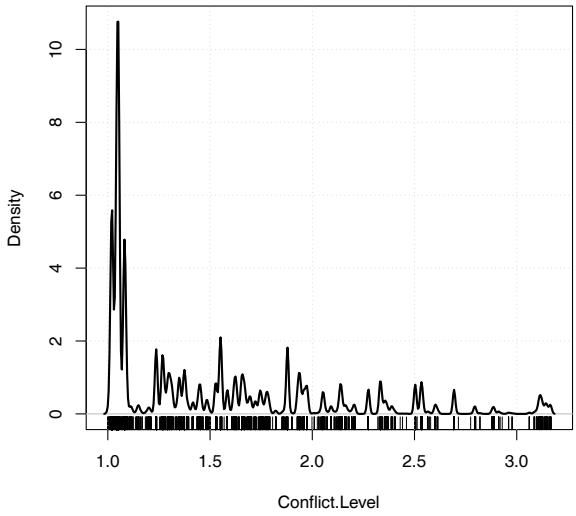

Figure 3. Density Estimation of Sources S_{10}: Visual Identification

function $\mathrm{conf}_{\mathrm{coh}} : S_1 \times \ldots \times S_N \rightarrow [1, +\infty)$ with

$$\mathrm{conf}_{\mathrm{coh}}(d_1, \ldots, d_N) := \max\left(1, \frac{\left(\sum_{j=1}^{M} p(d_i|c_j)\right) \cdot \left(\sum_{j=1}^{M} p(d_{-i}|c_j)\right)}{M \cdot \sum_{j=1}^{M} p(d_1, ..., d_N|c_j)}\right). \quad (3)$$

The interpretation of the fraction in Equation (3) is basically the same as for binary conflicts. As soon as the fraction value is bigger than 1.0, it is interpreted as the conflict-level of source S_i. Note that $\mathrm{conf}_{\mathrm{coh}}(d_1, \ldots, d_N) = 1.0$ indicates coherent findings and that no threshold for conflict-level detection is needed.

In order to get an impression of the distribution of conflict-level values, Figure 2 and Figure 3 show density estimations of source S_1 and S_{10}. The values are taken from the simulation described in Section IV. Obviously, conflict-level values are not normal-distributed.

III. Sensor Failure Detection

Bayesian application can be applied to civil and military air surveillance with hundreds of tracks to be classified every hour. To a large number of tracks the same classification task must be applied, as shown in Figure 4. According to [6], there is a strong need for automated assistance, in particular with respect to monitoring of sensor failures, because in each classification task different heterogeneous sensors are involved. Note, that formally a source is a sensor in combination with its data evaluation component, but in this paper we use the terms sensor and source synonymously.

A sensor failure may have many reasons, ranging from sensor flaws, inaccurate measurements and evaluation failures on the technical side (see [8, pp. 174-179], [11]), through an inadequate configuration to intentional deception or jamming in military applications [13]. A sensor failure can have different facets, e.g., by providing always the same, deviating, random or no declarations at all [12].

A. Sensor Failure Diagnostics

In case, no source-internal integrated failure detection functionality is available, comparison of declarations by means of conflicts' concept can be used in building a comprehensive approach of sensor failure detection. This idea addresses in particular deviating and random behavior of failing sources, whereas always the same or missing declarations can be easily detected by other approaches [13].

Subsequently, four approaches of sensor failure detection are described. The first and second are based on binary conflicts, and have been recently investigated in [12], [13]. Both other approaches are new, and rely on gradual conflict-level. It is expected, that by use of the gradual conflict measure, characteristics of conflicts caused by sensor failures can be better detected.

As sketched in Figure 4, the considered approaches of diagnostic sensor failure-tests evaluate the previous n classification cases in a sliding test-window. This is done by either monitoring actual number n_i and frequency level $l_i \in [0, 1]$ of binary conflicts for each source S_i, or the actual mean value $\overline{v}_i \geq 0$ of all conflict-level values greater than 1.0 for each source S_i in context of gradual-conflict based approaches. These values are compared with reference values l_i^* and $\overline{v}_i^*$, recorded in normal operation phases.

B. Conflict Binomial Failure-Test (CBF-Test)

Following [12], a one-sided Binomial test is applied in the Conflict Binomial Failure-Test, using conflict or no-conflict as underlying Bernoulli trial with the reference frequency level l_i^* as success probability. The probability $\alpha_i \in [0, 1]$ of source

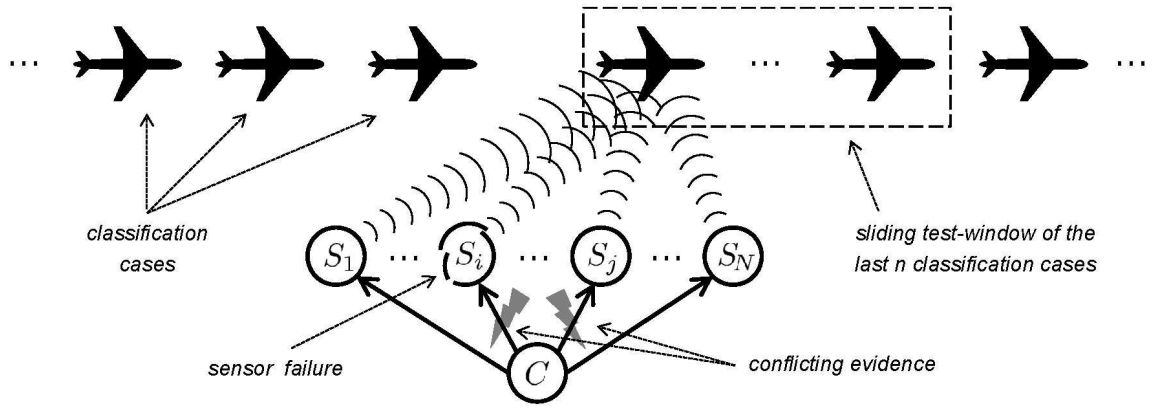

Figure 4. Sensor Configuration and Classification Cases [13, slightly modified]

S_i being involved in n_i or more conflicts is calculated under the assumption of no failing source by

$$\alpha_i = \sum_{j=n_i}^{n} \binom{n}{j} (l_i^*)^j \cdot (1 - l_i^*)^{n-j}. \tag{4}$$

Then, according to [12] a sensor failure is indicated, iff

$$\alpha_i \leq r_{\text{bin}} \tag{5}$$

holds for a Binomial-detection threshold $r_{\text{bin}} \in (0, 1)$, which can be adjusted.

C. Conflict Ratio Failure-Test (CRF-Test)

The Conflict Ratio Failure-Test [12] compares actual conflict ratio l_i and reference conflict ratio l_i^* of source S_i by indicating a sensor failure, iff

$$\frac{l_i}{l_i^*} > r_{\text{ratio}} \tag{6}$$

for a given adjustable ratio-detection threshold $r_{\text{ratio}} \in (0, \infty)$.

D. Average Conflict Failure-Test (ACF-Test)

Based on actual average $\overline{v}_i$ and reference average $\overline{v}_i^*$ of all gradual conflict-level values of source S_i that are greater than 1.0, the Average Conflict Failure-Test indicates a source failure, iff

$$\frac{\overline{v}_i}{\overline{v}_i^*} > r_{\text{aver}} \tag{7}$$

for given average-detection threshold $r_{\text{aver}} \in (0, \infty)$, which also can be adjusted.

E. Gauss Conflict Failure-Test (GCF-Test)

In Section II it was denoted, that conflict-level values of a source S_i are not normal-distributed. Nevertheless, the reference average $\overline{v}_i^*$ is approximately normal-distributed for sufficiently large samples. Mean μ_i and variance σ_i^2 of $\overline{v}_i^*$ can be determined in normal operation phases. Based on the well-known one-sided *Z-Test*, for the actual average $\overline{v}_i$ the probability $\beta_i \in [0, 1]$ with

$$\beta_i = p_{N(\mu_i, \sigma_i^2)}(X \geq \overline{v}_i) \tag{8}$$

can be calculated. A sensor failure is indicated, iff

$$\beta_i \leq r_{\text{Gauss}} \tag{9}$$

is true for a Gauss-detection threshold $r_{\text{Gauss}} \in (0, 1)$.

Note, that all four diagnostic sensor failure-tests use a adjustable sliding test-window, that covers the n previous classification cases, shown in Figure 4. In general, these diagnostic tests are applicable in sensor configurations with heterogeneous and homogeneous sensors, which provide evidence on tracked objects in short time intervals, compare [13].

IV. Comparative Simulation

In [12], [13] performance and applicability of the binary-conflict based Conflict Binomial and Conflict Ratio Failure-Test have been investigated. The main focus of this paper are the newly proposed gradual-conflict based Average and Gauss Conflict Failure-Tests as compared to the binary-based tests. The following subsection is based on the description in [13] and recapitulates adequate performance measures:

A. Diagnostic Tests' Performance Measures

Diagnostic sensor failure-tests can be evaluated and compared by certain performance measures taken from medical statistics, as found in [20, pp. 342-349]: Sensitivity and Specificity cover a technical perspective on diagnostic performance, since in concrete classification cases, true reference is typically not available.

- **Sensitivity** is the probability of detecting a source that is failing. Therefore, sensitivity measures success probability of a diagnostic failure test in detecting defect sources, compare [20, p. 340], [12].
- **Specificity** is the probability of proper marking a faultless sources as impeccable, compare [20, p. 340], [12]. The false alarm rate is equal to $1-$Specificity.

In typical applications, there is a strong need to detect as many failing sources as possible, i.e., a high sensitivity, while minimizing the false alarm rate at the same time. Unfortunately, these are opposing objectives [13].

From the operational viewpoint, the following measures are likewise interesting, because they support interpretation of concrete test results [13]:

- **Positive Predictive Value** (PPV) is the probability, that a source with failure indication is truly defect, compare [20, p. 342], [12]. Hence, reliability of failure indication is measured by PPV, which for example may lead to downgrading the classification process by switching off sources by operators [13].
- **Negative Predictive Value** (NPV) is the probability, that a source without failure indication is truly impeccable, compare [20, p. 342], [12]. This measure supports operators judgement in case of suspicious (e.g., deviating) source behavior while no failure indication is shown [13].

The probability of a correct diagnostic failure-test result is an overall performance measure and is called **Accuracy** [13]. The portion of defect sources in an application is named **Prevalence of Source Failure** and is needed for appropriate interpretation of the other measures [20, p. 342], [12]. Obviously, all these performance measures are related to each other, but each describes a particular aspect of diagnostic sensor failure-test's performance.

B. Application Scenario and Simulation Settings

A classical application scenario from air surveillance is used for our comparative simulation, see [5, pp. 54-57], [21]. With the Bayesian Classification Network in Figure 1 the following identities according to the *Extended Basic Identity Object Class* (EBIOC) [22] were determined: *Own Force Military*, *Own Force Civil*, *Non-aligned Military*, *Non-aligned Civil*, *Enemy Force Military*, and *Enemy Force Civil*. This scenario was initially taken from [23].

For performance evaluation and comparison of diagnostic sensor failure-tests, the simulation program from [12] and [13] is used with the following settings:

- Observability probability p_{obs} of source declarations with value $p_{obs} = 0.3$ and
- Probability p_{act} of defect sources' activities with value $p_{act} = 0.3$.

The threshold of binary conflict detection, compare Subsection III-A, was set to its standard value $\varepsilon_{coh} = 0.052$. Additional technical details of the simulation approach can be found in [12]. Each diagnostic sensor failure-test was applied with different thresholds:

- Binomial-detection threshold: $r_{\text{bin}} = 0.0001$ to 1.0
- Ratio-detection threshold: $r_{\text{ratio}} = 0.0$ to 15.0
- Average-detection threshold: $r_{\text{aver}} = 0.0$ to 15.0
- Gauss-detection threshold: $r_{\text{Gauss}} = 0.00001$ to 1.0

Comparative simulation runs were performed with variations of these thresholds to compare all four diagnostic sensor failure-tests to each other, with a typical sliding-test-window size $n = 100$. Additionally, application of each sensor failure-test was simulated with different values $n = 50, 75, 100, 150$ to investigate performance's dependency on the sliding window size. The simulation results are discussed in the following section.

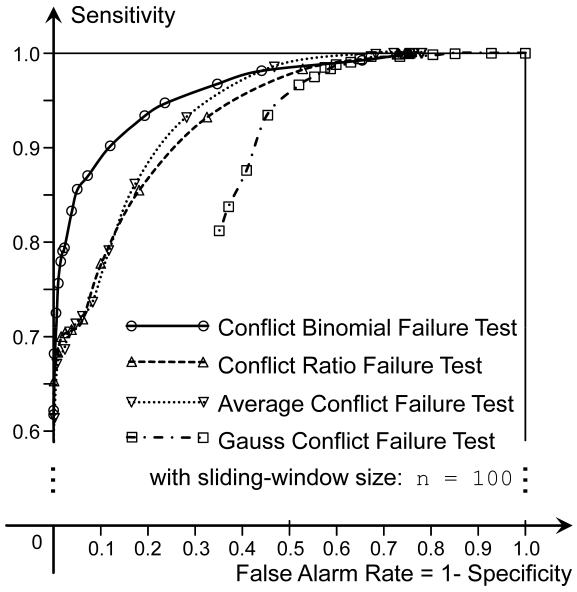

Figure 5. ROC Curves of All Failure-Tests with $n = 100$

V. Results

Diagnostic tests' operating range is defined as (typical) threshold settings with sensitivity and specificity values both above 70%. Within the operating range at window size $n = 100$, the Conflict Binomial Failure-Test (CBF-Test) shows accuracies from 77.3% to 98.2%, Conflict Ratio Failure-Test's (CRF-Test) values lie between 68.8% and 97.0%, and the Average Conflict Failure-Test (ACF-Test) achieves accuracies of $72.9\% - 95.3\%$. It was not possible to find a Gauss-detection threshold for the Gauss Conflict Failure-Test (GCF-Test) with sensitivity and specificity being in the operating range simultaneously. Best observed accuracy value of the GCF-Test is 65.7%. Prevalence values of sensor failures range within $4.9\% - 5.1\%$ in all performed simulation runs.

Receiver Operating Characteristics Results

Detection reliability of diagnostic tests is described by the opposing performance measures sensitivity and specificity, compare Section IV. *Receiver Operating Characteristic* (ROC) curves [2, pp. 49-51], [20, p. 348] visualize the trade-off between sensitivity and false alarm rate (=1-specificity) [13]. Sensitivity is denoted on the vertical axis, false alarm rate on the horizontal axis. In ROC-curve diagrams, different diagnostic tests can be compared simultaneously, whereat the upper left corner is the optimal combination of sensitivity and specificity values, each being maximal [13].

Figure 5 displays the ROC-curves of all diagnostic failure tests at a sliding-test-window size of $n = 100$. Within the operating range, the CBF-Test outperforms the other diagnostic failure tests. The gradual-conflict based ACF-Test and the binary-conflict based CRF-Test show similar results, with the former slightly surpassing. Performance of the GCF-Test is

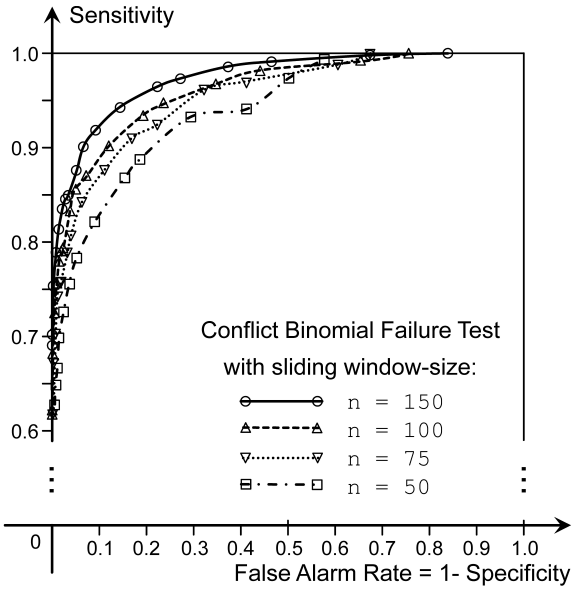

Figure 6. ROC Curves of Conflict Binomial Failure Tests with $n = 50 - 150$

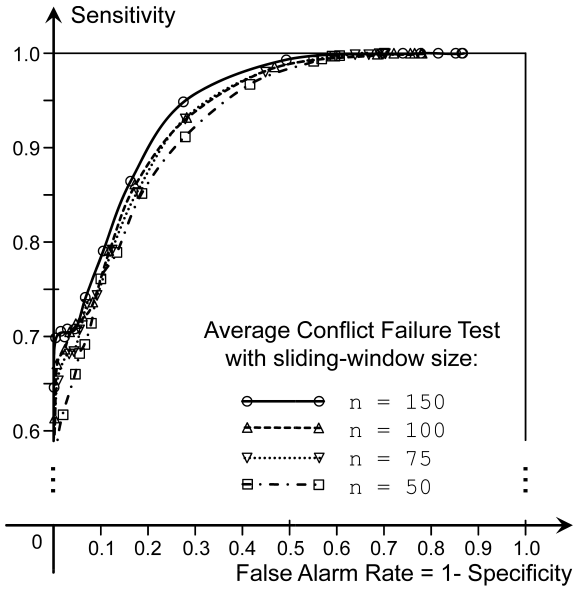

Figure 8. ROC Curves of Average Conflict Failure-Tests with $n = 50 - 150$

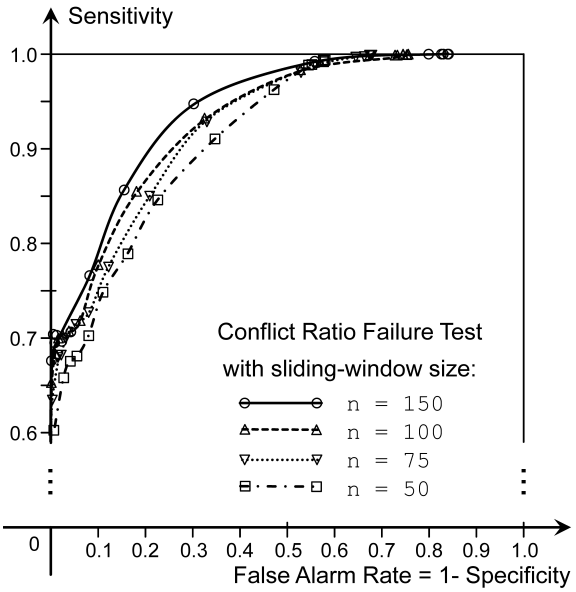

Figure 7. ROC Curves of Conflict Ratio Failure-Tests with $n = 50 - 150$

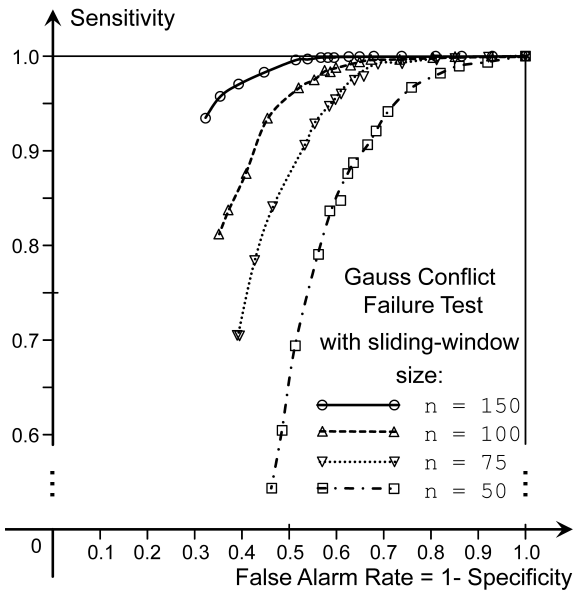

Figure 9. ROC Curves of Gauss Conflict Failure-Tests with $n = 50 - 150$

far below the other diagnostic failure-tests, false alarm rates below 30% are not achievable.

A good Binomial-detection threshold setting of the CBF-Test shows a (sensitivity|specificity)-performance value combination of $(90.2\%|88.0\%)$. For the CRF-Test, a good threshold yield the $(85.5\%|81.9\%)$ combination, and the ACF-Test provides a $(86.2\%|82.8\%)$ sensitivity-specificity combination for an appropriate setting. The best threshold setting of the GCF-Test yields only $(81.2\%|64.9\%)$.

In each of the Figures 6, 7, and 8 the ROC-curves of one diagnostic test is shown, dependent on different sliding-test-window sizes of $n = 50, 75, 100, 150$. Within the operating range, larger window sizes improve ROC-performance for all diagnostic test approaches. The CBF-Test benefits most of a larger window size, enabling a $(91.8\%|90.8\%)$ sensitivity-specificity combination. A significant dependency on larger window sizes is also present for the CRF-Test, with a $(85.7\%|84.5\%)$ best-performance values combination. The

ACF-Test has $(86.5\%|84.7\%)$ as optimal ROC-performance for $n = 150$. Note, that it shows only low dependency on the window-size, e.g., by a $(85.2\%|81.3\%)$ combination for size $n = 50$.

On a weak level, the GCF-Test likewise strongly benefits from large window sizes, as displayed in Figure 9. The weak performance of this test is probably due to the assumption of a normal-distributed average sum of conflict-level values, discussed in Subsection III-E and Section II. This prerequisites seems not to be met adequately, because of too small samples. Besides the apparent lower limits of the false alarm rates, this also explains, why the GCF-Test strongly benefits from larger window sizes as well as from higher observability values p_{obs} of sources and higher probability values p_{act} of defect sources' activities, compare Subsection IV-B.

A preliminary summary of the investigation regarding ROC-performance, i.e., sensitivity and specificity, the binary-conflict based Conflict Binomial Failure-Test shows the best diagnostic performance. But it requires higher computational costs and significantly depends on sliding-test-window size. Gradual-conflict based Average Conflict Failure-Test and binary-conflict based Conflict Ratio Failure-Test also show good results, which are similar to each other. But it is remarkable, that the ACF-Test has a low dependency on window size, compared to CRF- and CBF-Test, a fact that may be very relevant in practical implementations.

Positive vs. Negative Predictive Performance Results

Trade-off between positive and negative predictive performance of a sensor failure-test is displayed in a PPV vs. NPV-curve diagram [13] in Figure 10 with a window size $n = 100$. PPV-values are found on the vertical, NPV-values on the horizontal axis. In our simulation, we used (realistic) low prevalence values of failing sources, so only the rightmost part of the curve is relevant.

In our simulation, comparisons' results of the predictive performance values PPV and NPV are similar to ROC-performance results: The CBF-Test clearly outperforms ACF- and CRF-Test but at the price of higher computational costs. Gradual-conflict based ACF-Test and the CRF-Test, based on binary-conflicts produce almost identical positive and negative predictive values, wherewith the low dependency on window size of the ACF-Test gets more relevant. GCF-Test lies far behind.

Summing up, its very good performance make the binary-conflict based Conflict Binomial Failure-Test to be operational users' preference. But size of the sliding test-window affects the reaction time of diagnostic sensor failure-tests. Therefore, in certain scenarios smaller window sizes might be an important criterion, making the gradual-conflict based Average Conflict Failure Test the better choice.

VI. Conclusions

In this contribution four sensor failure-tests in Bayesian Classification Networks, based on gradual and binary conflicts are investigated. Except for the Gauss Conflict Failure-Test,

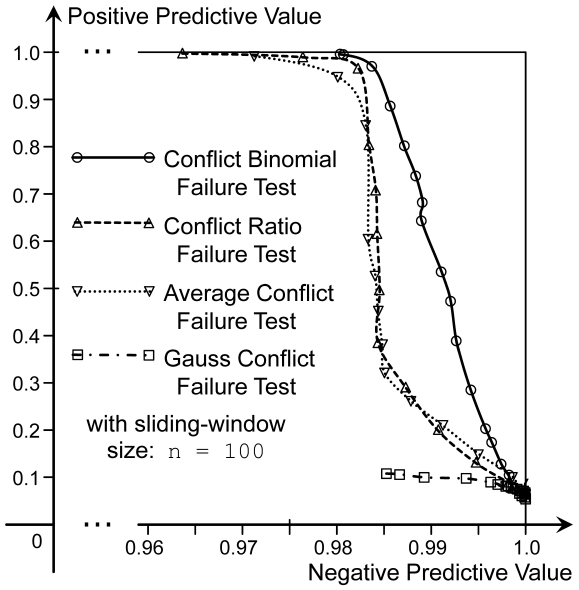

Figure 10. PPV vs. NPV Curves of All Failure-Tests with $n = 100$

the Conflict Binomial, Conflict Ratio, and Average Conflict Failure-Test showed their suitability to adequately detect failing sources in a simulated air surveillance scenario. Based on higher computational costs, the Conflict Binomial Failure-Test shows best performance results. Both other tests have performance results and a computational effort similar to each other. But the newly proposed Average Conflict Failure-Test has only low dependency on the test-window size, leading into shorter reaction time of diagnostic sensor failure-tests. All concepts and diagnostic tests are directly transferable to general Bayesian Networks.

Future work in this topic might address the prerequisite of the Gauss Conflict Failure-Test by finding a better description of the conflict-levels' probability distribution as basis of hypothesis testing. In addition, an approach to better discrimination between one and several simultaneous sensor failures would further improve practical applicability.

References

[1] O. Pourret, P. Naïm, and B. Marcot, Eds., *Bayesian Netwworks – A Practical Guide to Applications*. Chichester (UK): John Wiley & Sons, Ltd, 2008.

[2] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York, Chichester, Weinheim, Brisbane, Singapore, Toronto: John Wiley & Sons, Inc., 2001.

[3] D. L. Hall and S. A. McMullen, *Mathematical Techniques in Multisensor Data Fusion*, 2nd ed. Boston, London: Artech House Publishers, 2004.

[4] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Boston, London: Artech House Publishers, 1999.

[5] E. Waltz and J. Llinas, *Multisensor Data Fusion*. Boston, London: Artech House Publishers, 1990.

[6] M. Krüger and N. Kratzke, "Monitoring of Reliability in Bayesian Identification," in *Proceedings of the 12th International Conference on Information Fusion (Seattle (WA), USA 06 - 09 July 2009)*. ISIF, July 2009.

[7] D. Michie, D. Spiegelhalter, and C. Taylor, *Machine Learning, Neural and Statistical Classification*, D. Michie, D. Spiegelhalter, and C. Taylor, Eds. New York, London, Toronto, Sydney, Tokyo, Singapore: Ellis Horwood Limited, 1994.

[8] F. V. Jensen and T. D. Nielsen, *Bayesian Networks and Decision Graphs*. New York: Springer Science, 2007.

[9] H. Mitchell, *Multi-Sensor Data Fusion – An Introduction*. Berlin, Heidelberg, New York: Springer-Verlag, 2007.

[10] M. Krüger, "Measures of Conflicting Evidence in Bayesian Networks for Classification," in *Proceedings of the 16th International Conference on Information Fusion (Istanbul, 09 - 12 July 2013)*. ISIF, July 2013, pp. 1574–1581.

[11] K. B. Laskey, "Conflict and suprise: Heuristics for model revision," in *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, 1991, pp. 197–204.

[12] M. Krüger, "Detection of Failing Sensors by Conflicting Evidence in Bayesian Classification," in *Proceedings of the 17th International Conference on Information Fusion (Salamanca, 07 - 10 July 2014)*. ISIF, July 2014.

[13] M. Krüger, "Sensor Failure-Test in Repeatedly Performed Bayesian Identification Tasks in Air Surveillance," in *Proceedings of the 9th Workshop on Sensor Data Fusion (SDF): Trends, Solutions, Applications (Bonn, 08 - 10 October 2014)*. IEEE, October 2014.

[14] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge (MA), USA: The MIT Press, 2009.

[15] H. Zhang, "The Optimality of Naive Bayes," in *Proceedings of the 17th International FLAIRS Conference (Miami Beach, Florida, USA)*. Florida Artificial Intelligence Research Society, 2004.

[16] Evaluation of Techniques for Uncertainty Representation Working Group (ETURWG), "Evaluation Criteria: II. URREF Ontology (ver. 1)," http://eturwg.c4i.gmu.edu/?q=URREF_Ontology (accessed on 23.02.2013), International Society for Information Fusion (ISIF), 2011.

[17] J. Habbema, "Models diagnosis and detection of diseases," in de Dombal et al., editors, *Decision Making and Medical Care*. North-Holland, 1976, pp. 399–411.

[18] F. Jensen, B. Chamberlain, T. Nordahl, and F. Jensen, "Analysis in HUGIN of data conflict," in *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*. Boston (MA): Association for Uncertainty in Artificial Intelligence, 1990, pp. 546–554.

[19] Y.-G. Kim and M. Valtorta, "On the detection of conflicts in diagnostic Bayesian networks using abstraction," in P. Besnard and S. Hanks, editors, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, 1995, pp. 362–367.

[20] J. L. Peacock and P. J. Peacock, *Oxford Handbook of Medical Statistics*. New York: Oxford University Press, 2011.

[21] P. van Gosliga and H. Jansen, "A Bayesian Network for Combat Identification," in *RTO IST Symposium on 'Military Data and Information Fusion' (Prague, Czech Republic, 20-22 October 2003)*, ser. RTO Meeting Proceedings MP-IST-040. NATO Research & Technology Organization, March 2004.

[22] *STANAG 4162 (edition 2): Identification Data Combining Process (IDCP)*, NATO Standardization Agency (NSA), Brussels, 2009, (NATO unclassified).

[23] M. Krüger and D. Hirschhäuser, "Source Conflicts in Bayesian Identification," in *INFORMATIK 2009: Im Focus das Leben*, ser. Lecture Notes in Informatics (vol. 154). Gesellschaft für Informatik e.V. (GI), 2009, pp. 2485–2490.