Pattern Recognition 77 (2018) 30–44

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Error sensitivity analysis of Delta divergence - a novel measure for classifier incongruence detection

Josef Kittler[a,*], Cemre Zor[a,*], Ioannis Kaloskampis[b], Yulia Hicks[b], Wenwu Wang[a]

[a] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK
[b] School of Engineering, Cardiff University, Queens Buildings, The Parade, Cardiff, UK

ARTICLE INFO

Article history:
Received 15 February 2017
Revised 24 October 2017
Accepted 30 November 2017
Available online 5 December 2017

Keywords:
Anomaly detection
Classifier decision incongruence
Bayesian surprise

ABSTRACT

The state of classifier incongruence in decision making systems incorporating multiple classifiers is often an indicator of anomaly caused by an unexpected observation or an unusual situation. Its assessment is important as one of the key mechanisms for domain anomaly detection. In this paper, we investigate the sensitivity of Delta divergence, a novel measure of classifier incongruence, to estimation errors. Statistical properties of Delta divergence are analysed both theoretically and experimentally. The results of the analysis provide guidelines on the selection of threshold for classifier incongruence detection based on this measure.

1. Introduction

Many sensor data analysis systems involve multiple classifiers to interpret input data, which leads to improved performance by virtue of exploiting complementary information derived from multiple modalities of sensing, multiple representations, contextual information, and hierarchical structuring of the interpretation process. In addition to increased performance, an important corollary of involving multiple experts in decision making is the ability to flag anomalies by looking for discrepancy between their outputs, referred to as incongruence.

Anomaly detection, i.e. finding patterns in data that do not conform to expected normal behaviour [1], has been studied in many areas including statistical signal processing and pattern recognition [2–7], as well as a wide variety of applications, such as intrusion detection for cyber-security [8–11], surveillance [12,13], video-based crowd-behaviour analysis [14–16] and fault detection in sensor systems [17,18]. A large number of techniques have been developed for this problem, including the methods based on e.g. classification, clustering, statistical modelling, among many others, as surveyed by Chandola et al. [1], Markou and Singh [6,7], and Patcha and Park [19]. The basic approach to anomaly detection adopted in all these techniques is to compare incoming data against a reference model that embodies normality. This approach is also known as outlier detection.

Despite this effort, the development of good models of normality for diverse applications is not without challenges. Moreover, detecting anomalies in multiple classifier systems raises additional issues. It has been argued in [20] that in order to identify and distinguish the multifaceted nature of anomaly and take appropriate control actions, a more complex system consisting of several other mechanisms are needed in addition to outlier detection. They include data quality assessment, classifier decision confidence estimation and classifier incongruence detection [20]. Among these mechanisms, classifier incongruence detection, in other words measuring the disagreement between the classifiers embodied in the system, is of paramount importance. It helps to differentiate between certain types of anomalous events such an out-of-context event, where an event is unexpected, a rare event, where a given configuration of components occurs very infrequently, or an unknown structure [20]. This mechanism is the subject and focus of this paper.

A simple example of anomaly detection using incongruence is out-of-vocabulary word detection in speech recognition [21]. A speech recognition system would typically involve a hierarchical decision making strategy based on the outputs of noncontextual and contextual classifiers. Noncontextual classifiers operating at a low level of representation attempt to identify phonemes based on the speech content, whereas contextual classifiers combine this

* Corresponding authors.
E-mail addresses: j.kittler@surrey.ac.uk (J. Kittler), c.zor@surrey.ac.uk (C. Zor), KaloskampisI@cardiff.ac.uk (I. Kaloskampis), HicksYA@cardiff.ac.uk (Y. Hicks), w.wang@surrey.ac.uk (W. Wang).

https://doi.org/10.1016/j.patcog.2017.11.031

low level symbolic representation with prior knowledge to segment and recognise larger semantic units such as words. Implicitly, in this complex decision making process, we get two opinions about the identity of each phoneme: one derived from the contextual classifier and one from its noncontextual counterpart. For successful speech understanding, we do not necessarily need to be concerned with the low level interpretation process. However, by monitoring the outputs of both contextual and noncontextual classifiers we may glean very useful information which could enable us to qualify the failure of the speech recognition system to interpret input data. For instance, if the low level classifier makes confident decisions about the identity of the phonemes, but a sequence of the detected phonemes does not produce a meaningful output, the system may be encountering an out-of-vocabulary word. Discerning such nuances in sensor data interpretation would allow us to act accordingly. This, however, requires a reliable method of classifier incongruence detection which can spot and discriminate disagreements in classifier opinions about one or more hypotheses.

Detecting incongruence can be formulated as a statistical hypothesis testing problem [6]. This typically involves some proposition, referred to as a null hypothesis and a test statistics. If the outcome of the test statistics is consistent with its known distribution model, then the null hypothesis is accepted. An outlier of that distribution would lead to the hypothesis rejection. An observation is considered an outlier at a given level of significance, i.e. if the test statistics value exceeds a threshold corresponding to some vestigial probability, such as 5% or 1%. Accordingly, the proposition in incongruence detection is that two classifier outputs are congruent. If the test statistics exceeds a threshold corresponding to the required level of significance then the hypothesis is rejected, that is the classifier outputs are deemed incongruent. Let us emphasise here that measuring classifier incongruence is meaningful only when a dominant class probability output by a classifier exceeds a certain confidence level and there is sufficient margin between the probabilities of the dominant class and the next strongest class.

Clearly the test statistics is a crucial component of a hypothesis testing process. The choice not only influences its statistical properties, but also how faithfully it reflects the concept tested. For instance, the throw of a coin and counting the number of heads in testing whether the coin is biased introduces a statistical element in the test process. A much more transparent test would consist in looking at both sides of the coin, which would immediately, in unambiguous terms, establish whether the coin is biased or not. It is the choice of the experiment of repeated trials, and the head count, which makes the hypothesis testing more difficult than it needs to be, and injects randomness in the experimental outcome. Moreover, this particular choice only reflects the phenomenon to be tested indirectly, rather than in the most transparent way possible.

A classical classifier incongruence test statistic is the Kullback–Leibler (KL) divergence known as Bayesian surprise [22]. However, it has recently been pointed out that this measure has some deficiencies. In particular in multiclass problems, it has been shown to be unpredictably affected by the probabilities of nondominant classes (referred to as clutter) and a variant of the KL divergence, referred to as Decision–Cognizant KL (DC-KL) divergence has been proposed instead [23]. Some other undesirable properties of KL type divergence, induced by its log function, have been rectified by the recently proposed Delta divergence [24]. However, the key question not addressed so far, is whether the superior theoretical properties of Delta divergence are robust to estimation errors. For example, in multiple classifier fusion, sensitivity to errors changed the ranking of the product and sum fusion rules, although the former is founded on sound theoretical principles.

The aim of this paper is to investigate error sensitivity of Delta divergence as a measure of classifier incongruence. The study includes a theoretical analysis of a few special cases to gain intuitive feeling for the behaviour of Delta divergence in noisy conditions. A more comprehensive investigation is carried out by simulation studies where the space of class a posteriori probabilities is sampled to estimate the probability distribution of noise-free Delta divergence values for various scenarios. The samples of the a posteriori probability distributions are then corrupted by estimation errors and their impact on Delta divergence is measured experimentally. The aggregation of the statistical distributions of Delta divergence over different scenarios and the distribution of noise-free Delta divergence values produces the final test statistics distribution which can be used to determine appropriate classifier incongruence detection thresholds. Although the simulation studies are limited by the assumptions made regarding the estimation noise, their main merit is to give the reader a better understanding of the behaviour of Delta divergence. For practical purposes we propose guidelines for incongruence detector design, given a training set of class probability estimates. The design procedure is illustrated on a problem of detecting incongruence of noncontextual and contextual classifiers developed to recognise action and activity in breakfast dataset videos.

In summary, the contributions of the paper include:

- An error sensitivity analysis of Delta divergence utilising marginalisation of the test statistics over different scenarios
- Estimation of the statistical distribution of Delta divergence as a basis for classifier incongruence threshold selection
- Guidelines for classifier incongruence threshold selection in practical anomaly detection systems

The paper is structured as follows. The background and related work are the subjects of Section 2. In Section 3, Delta divergence is introduced as a novel classifier incongruence measure and its properties are related to the Bayesian surprise measure which is used as a baseline both theoretically and experimentally. The statistical properties of the proposed measure are investigated in Section 3.1. In Section 4, a discussion on how to determine the classifier incongruence threshold is carried out via experimental analysis on synthetic and real data. Finally, in Section 5, the main results of this study are summarised and the paper is drawn to conclusion.

2. Related work

The idea of using classifier incongruence for anomaly detection has been advocated by Weinshall et al. in [25]. As in [25], we consider just two decision making experts, classifying the data into one of m possible categories. Let $\tilde{P}(\omega_j|\mathbf{x})$ and $P(\omega_j|\mathbf{x}),\ \ j = 1, \ldots, m$ denote the a posteriori probabilities associated with the hypothesis that model ω_j explains the input data, x, which have been estimated by the two experts. If the two distributions are identical or similar, then the classifier outputs would be considered congruent. For measuring incongruence, Weinshall et al. [25] advocated the adoption of Itti's Bayesian surprise measure [22] originally proposed for detecting content changes in video. In particular, by considering the a posteriori class probability distribution output by one of the experts as a reference, one can detect incongruence by calculating

$$D_K = \sum_{j=1}^{m} \tilde{P}(\omega_j|\mathbf{x}) \log \frac{\tilde{P}(\omega_j|\mathbf{x})}{P(\omega_j|\mathbf{x})} \tag{1}$$

which is basically the Kullback–Leibler divergence between the two distributions.

The Kullback–Leibler divergence primarily measures the similarity between the two probability distributions through an inverse relationship. If the distributions are identical, or similar, the measure will tend to zero. A high value of the measure would indicate differences in the a posteriori probabilities, and therefore high incongruence between the classifier outputs. There are other information theory divergences that could be used for the same purpose [26,27].

Alternatively, one could adapt any statistical measure of similarity between two distributions and use it as a test statistic for detecting classifier incongruence. More specifically, mapping the classes onto consecutive numbers (bins) will create two discrete probability distribution functions, resembling normalised histograms, which sum up to unity. This analogy suggests that well-known criteria, namely histogram similarity measures, mainly used for calculating the goodness-of-fit between an empirical and a reference distribution, could be adapted for the purpose of measuring classifier incongruence, although there are no reported attempts in the literature to adopt them for this purpose. A comprehensive analysis of the tests that can be used for measuring the similarity between two histograms can be found in [28]. Examples are Chi-square, Kolmogorov-Smirnov [29], Cramér-von-Mises [30,31], and Anderson-Darling [32] tests; Geometric test using Bhattacharyya distance, and likelihood-ratio and likelihood-value tests. We plan to investigate the applicability of these histogram matching methods to the problem of incongruence detection in the future, but here we are focusing on the established state of the art methodology of incongruence detection constituted by the Bayesian surprise measure.

It should be noted that the term *measures of surprise* in Bayesian analysis also refers to test statistics developed for outlier detection. This confusing terminology relates to the classical notion of anomaly detection where instead of measuring the similarity between two probability distributions, the aim is to compare a single observation with the hypothesised distribution model [33–39]. Recently in [40], some state-of-the-art measures of surprise in Bayesian analysis have been thoroughly analysed and modifications have been proposed. However, these techniques are not relevant to the topic addressed in this paper.

Accordingly, Itti's Bayesian surprise [22] and its decision cognizant variant DC-KL [23] are the key existing technique for assessing classifier incongruence in the literature. Thus, we shall adopt them as a reference for our deliberation. The issues with the Bayesian surprise measure can be listed as follows:

1. It goes to infinity for any hypothesis ω for which $P(\omega|\mathbf{x}) \rightarrow 0$ while $\tilde{P}(\omega|\mathbf{x}) \neq 0$. This can occur even for insignificant hypotheses and result in producing false alarms of incongruence.
2. The measure is not symmetric, in a sense that if we use the distribution of $P(\omega|\mathbf{x})$ as a reference instead of $\tilde{P}(\omega|\mathbf{x})$, we will get a different value of the divergence.
3. The divergence function may produce the same value for completely different scenarios and may diverge to infinity. Hence, it is difficult to assess which values imply congruence / incongruence, and define a suitable threshold.
4. The measure is classifier decision agnostic. In other words, all hypothesis (classes) are involved in the calculation of the surprise.
5. By virtue of Property 4, it is also strongly affected by estimation errors on probabilities $P(\omega|\mathbf{x})$ and $\tilde{P}(\omega|\mathbf{x})$.

In contrast, DC-KL is decision cognizant, that is the measure ignores all the terms associated with the classes that are not selected by the decision rule. The main argument for ignoring the contribution of the classes with non maximum posterior is that first of all they contribute with a lot of irrelevant jitter to the value of the similarity measure. This contamination is proportional to the number of hypotheses. In other words, in multi hypotheses problems, this background jitter potentially can bury the useful information, i.e. the probability differences for the classes selected by the decision rule. The elimination of this clutter impacts favourably also on Property 5. However, both KL and DC-KL share Properties 1–3 which limit their ability to distinguish between classifier congruence and incongruence robustly. Let us illustrate the limitation on the real data application discussed in Section 4, which is concerned with action and activity recognition videos.

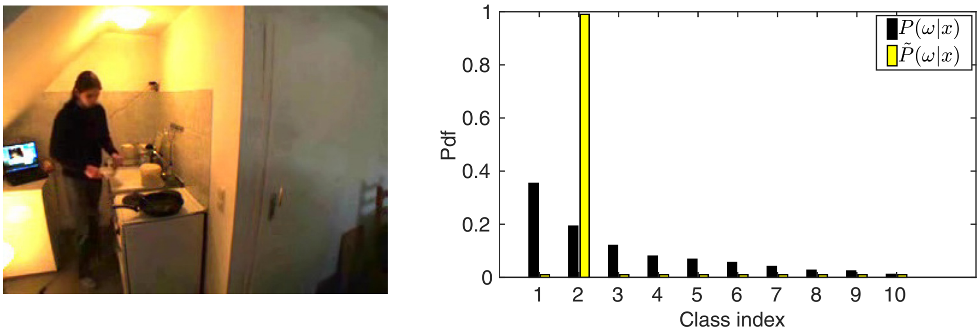
Breakfast dataset [41] is used for performing action and activity recognition from breakfast scenario videos, and is comprised of 10 activities and 52 action classes. In our approach, the action in each segment of a video is interpreted by a noncontextual and a contextual classifier, the latter taking into account the complete sequence of actions to identify the breakfast scenario activity captured by the video. As an example, for the video segment represented by the key frame shown in Fig. 1(a), the top ten hypotheses output by the two classifiers are shown in Fig. 1(b). The classifiers are clearly incongruent. Yet the corresponding KL and DC-KL incongruence values, $\tilde{D}_D = \tilde{D}_K = 1.63$, are very low in the context of the normal range of values of these test statistics shown in the histograms in Fig. 2(a) and (b), respectively. The histograms have been computed on a training set outputs of the two classifiers described in detail in Section 4.

To avoid the problems associated with KL and DC-KL, we have previously proposed alternatives, which not only focus on the dominant hypotheses flagged by the two experts [20,42], but have the additional advantage over [23] that their values are confined to a finite range of [0, 1]. Although the methods in [20,42] have attractive properties, their main disadvantage is that they are heuristic. Overcoming this shortcoming, in a recent paper [24] we have proposed a novel divergence, called Delta divergence (D_Δ), which exhibits all the desirable properties of a test statistic ideally suited for detecting classifier incongruence. Moreover, it is a proper information theoretic divergence, with all the advantages of a measure underpinned by information theory. Note that in [23], a detailed theoretical and experimental analysis demonstrates the superiority of Delta divergence over KL divergence.

The rest of this paper focuses on Delta divergence. The aim is to verify that the attractive properties of Delta divergence are robust to estimation errors on the class probabilities output by the two classifiers. We investigate the sensitivity of D_Δ both analytically and experimentally. Moreover, we show how the empirical distribution of this novel incongruence measure could provide a basis for selecting an appropriate classifier incongruence detection threshold at a given level of statistical significance. Note that in practice, the only observable information are classifier outputs which are already subject to estimation errors. For such scenarios, we propose practical incongruence detection guidelines and illustrate their use on a real data application concerned with action and activity recognition in breakfast scenario videos.

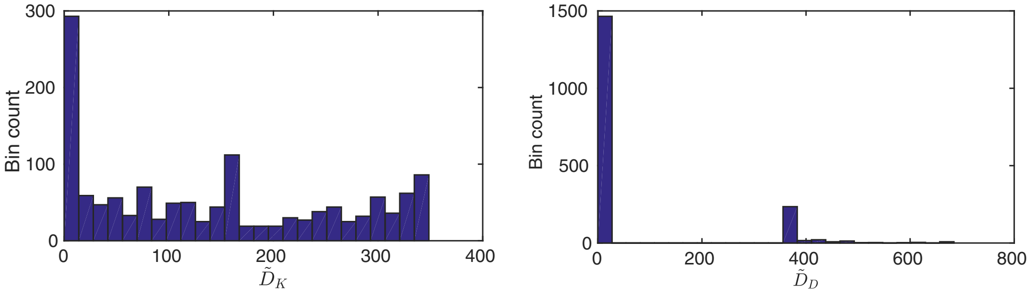
3. Statistical properties of D_Δ

Delta divergence, proposed in [24], has been developed from f-divergence [27], known as variation distance, by merging all the non-dominant class hypotheses into a single set. This preserves the nature of the measure as a proper divergence of differences between two probability distributions, but has the beneficial effect of reducing the "clutter" injected by the terms associated with the non-dominant hypotheses. The positive impact of this clutter reducing modification grows with the number of classes. Let us denote the dominant hypotheses identified by two classifiers by $\tilde{\mu} = \arg\max_\omega \tilde{P}(\omega|\mathbf{x})$ and $\mu = \arg\max_\omega P(\omega|\mathbf{x})$. Also, for the sake of notational simplicity, in the following, we shall drop making explicit references to specific observation $\mathbf{x}$ and denote the a poste-



(a) (b)

Fig. 1. (a) Key frame taken from an example Breakfast dataset segment (b) Probability distribution values belonging to the contextual and non-contextual classifiers given for a sample taken from the Breakfast dataset, for which $\tilde{D}_K = \tilde{D}_D = 1.63$.



(a) (b)

Fig. 2. Histograms of Bayesian surprise (KL) (a) and Decision-cognizant Bayesian surprise (DC-KL) (b) for the Breakfast dataset.

riori class probabilities $P(\omega|\mathbf{x})$ simply as P_ω, and $\tilde{P}(\omega|\mathbf{x})$ simply as $\tilde{P}_\omega$. Delta divergence is defined as

$$D_\Delta = \begin{cases} |P_\mu - \tilde{P}_\mu| & \mu = \tilde{\mu} \\ max\left\{|\tilde{P}_{\tilde{\mu}} - P_{\tilde{\mu}}|, |P_\mu - \tilde{P}_\mu|\right\} & \begin{cases} \mu \neq \tilde{\mu} \\ P_\mu - \tilde{P}_\mu \geq 0 \\ \tilde{P}_{\tilde{\mu}} - P_{\tilde{\mu}} \geq 0 \end{cases} \\ |\tilde{P}_{\tilde{\mu}} - P_{\tilde{\mu}}| + |P_\mu - \tilde{P}_\mu| & \begin{cases} \mu \neq \tilde{\mu} \\ \mathrm{sgn}(P_\mu - \tilde{P}_\mu) \neq \\ \mathrm{sgn}(\tilde{P}_{\tilde{\mu}} - P_{\tilde{\mu}}) \end{cases} \end{cases} \qquad (2)$$

The focus of Delta divergence (D_Δ) given in (2) is solely on differences between a posteriori probabilities of dominant classes (most probable classes identified by the two classifiers). When the two classifiers agree on the identity of the dominant hypothesis, Delta divergence measures only the difference between the corresponding a posteriori class probabilities. When they disagree, and the signs of the differences differ, Delta divergence equals the sum of the absolute values of the respective differences. When the labels disagree, and both of the differences of the a posteriori class probabilities are positive, it picks the maximum of the absolute values of these differences.

Apart from clutter reduction, D_Δ has a number of other attractive properties. It is independent of the actual values of a posteriori class probabilities, and therefore of their surprisal content. In other words, classifier incongruence measurement is not modulated by the likelihood of the dominant hypotheses. The measure is bounded and symmetric. In Section 4 we show that the robustness to clutter also reduces the sensitivity of Delta divergence to a posteriori class probabilities estimation error. All these characteristics jointly make Delta divergence ideal for gauging classifier incongruence.

D_Δ takes values from the interval [0, 1]. In order to provide insight into the frequency of occurrence of its values, we sample the space of different combinations of class probability distributions outputs (P and $\tilde{P}$) uniformly, and make a note of the resulting incongruence measure values after they enter the calculation defined in (2). We then identify the scenarios in which classifiers agree on the most probable hypothesis, or disagree (cases of label agreement and disagreement) separately, and create histograms, on which averaging over bins and normalization is applied to end up with probability distributions. The graphs given in Fig. 3 are estimated using a total number of 10^6 of such probability distribution pairs for problems involving a number of classes equal to 3 and 6. Fig. 3(a) shows the probability density functions of D_Δ for the cases of label agreement, Figure 3-b shows distributions for the cases of label disagreement; and Figure 3-c depicts the aggregate distributions for all cases (combination of label agreement and disagreement). In each subfigure, 3 class problems are indicated by dashed lines, whereas 6 class problems are indicated by the solid curves. Note that the indicated values of m are selected for illustration purposes and the trend for other values follow in accordance with the following analysis.

The effect of the number of classes, m, on the incongruence measure distribution can be observed by comparing the solid and dashed lines in Fig. 3. As m increases from 3 to 6, high values of incongruence become more likely for the label agreement case. This can also be deduced from (2); the upper limit for incongruence can be shown to equal $[1 - (1/m)]$. Note that as m goes to infinity, this value becomes equal to 1. A related observation for this case is the decrease in the likelihood of observing incongruence values close to zero when m increases. In the case of label disagreement, contrary to the findings for agreement, the realizations of lower D_Δ values are more probable for $m = 6$ than $m = 3$. Accordingly, for high D_Δ values, the probability densities are lower for $m = 6$ compared to $m = 3$. Note that the upper limit of the disagreement case is equal to 1 for all m, as a result of the second condition in (2).

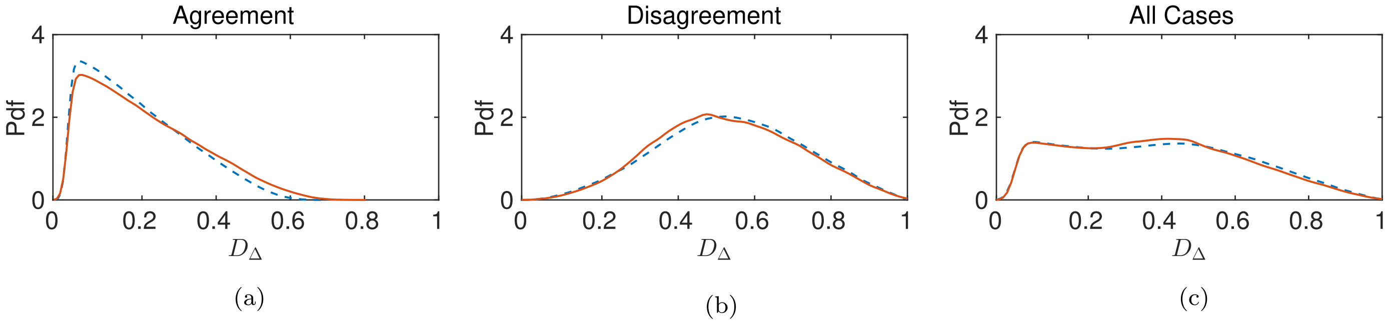


Fig. 3. Probability density functions (pdf) of D_Δ for classifier label agreement on the most probable hypothesis (a), for classifier label disagreement (b), and for all cases (c). Dashed lines indicate the distributions obtained for 3 class problems, and solid lines for 6.

Combining the two sets of observations for the label agreement and disagreement cases, it can be concluded that their corresponding distributions get shifted towards each other as m increases. This means that the bigger m is, the more difficult it becomes to tell if an obtained/measured incongruence value emerges from a scenario of agreement in the most probable hypothesis, or disagreement. On the other hand, for smaller m, the overall incongruence distribution has a higher variation within the range [0, 1]. The effect of the incongruence distribution on hypothesis thresholding is going to be further discussed in Section 4.3.

3.1. Error sensitivity

In reality, the a posteriori probabilities for the various hypotheses will be estimated by the two classifiers subject to estimation errors. The aim of the error sensitivity study is for the reader to get a feel for the effect of these estimation errors on the properties of Delta divergence. The intention is not to provide a comprehensive theoretical analysis, but instead consider a few simple cases where analysis is possible to gain intuitive idea of the impact of estimation errors. The subsequent simulation studies explore the scenario landscape more thoroughly, but it should be noted that even here the aim of the study is more educational than to present definitive findings. The justification for this is that in practice we will not have access to ground truth class probabilities, neither to estimation errors, and a more practical methodology will be required to design a class incongruence detector. Such a design methodology will be presented in Section 4.6 and its application illustrated in Section 4.6.1.

Let us denote the estimates of $P(\omega|\mathbf{x})$ and $\tilde{P}(\omega|\mathbf{x})$ by $P(\omega|\mathbf{x}) + \eta_\omega(\mathbf{x})$ and $\tilde{P}(\omega|\mathbf{x}) + \tilde{\eta}_\omega(\mathbf{x})$ respectively, where $\eta_\omega(\mathbf{x})$ and $\tilde{\eta}_\omega(\mathbf{x})$ are the estimation errors. We refer to the probability density functions of these errors as $q(\eta)$ and $\tilde{q}(\eta)$ accordingly.

For the sake of simplicity, we shall assume that $q(\eta)$ and $\tilde{q}(\eta)$ are normal distributions with zero mean and standard deviation σ. However, it should be emphasized that estimation errors have to satisfy the conditions

$$\sum_{\omega=1}^{m} \eta_\omega(\mathbf{x}) = 0 \tag{3}$$

and

$$0 \le \eta_\omega(\mathbf{x}) + P(\omega|\mathbf{x}) \le 1 \tag{4}$$

Thus, as probabilities have to be nonnegative as well as not exceeding unity, the normality assumption for $q(\eta)$ has to break down for a posteriori probabilities close to zero or one. In order to satisfy these constraints, we shall simply assume that the tail of the Gaussian, constrained by any of the conditions, is clipped; and the remaining part of the distribution is normalized to have under the curve area equal to 1. Dropping again explicit references to observation $\mathbf{x}$, for a noise-free posterior P, the resulting error distribution, $p(\eta, P)$, becomes

$$p(\eta, P) = \begin{cases} 0 & if \quad \begin{cases} \eta < -P \\ \eta > 1 - P \end{cases} \\ \left(\frac{1}{\int_{-P}^{1-P} q(\eta)}\right) q(\eta) & if \quad -P \le \eta \le 1 - P \end{cases} \tag{5}$$

An example is shown in Fig. 4 for $P = 0.1$ and $q(\eta) = N(0, 0.15)$. In Fig. 4(a), the thin solid line depicts $q(\eta)$. The thick solid line illustrates $p(\eta, P)$, obtained by clipping the tail of q at the cut off point, $-P = -0.1$, as indicated by the dashed line, followed by normalization. On the other hand, in Fig. 4(b), the thick solid line illustrates the probability density function $r(s)$ of the estimate $s = P + \eta$. It should be remembered that r is obtained as a convolution of the distributions of P and η, such that

$$r(s) = \int_{\lambda=-\infty}^{\infty} \delta(s - P - \lambda) p(\lambda, P) d\lambda \tag{6}$$

Finally, the thin line in Fig. 4(b) is provided for convenience and depicts what $r(s)$ would look like if the condition (4) did not exist.

The estimation errors corrupting class a posteriori probabilities will cause estimation errors on the computed incongruence values. It is evident that for incongruence measures involving summation over all the classes these probability estimation errors will create high background noise level which will make it difficult to measure incongruence (surprise) reliably. Hence, the proposed incongruence measure in (2), which involves summation over at most two classes (when $\mu \neq \tilde{\mu}$) should be considerably more robust to noise. Let us now investigate the statistical properties of D_Δ.

With the contamination by estimation errors, the incongruence measure can be expressed as

$$D_\Delta = \begin{cases} |P_\mu - \tilde{P}_\mu + \eta_\mu - \tilde{\eta}_\mu| & \mu = \tilde{\mu} \\ max\Big\{|\tilde{P}_{\tilde{\mu}} - P_{\tilde{\mu}} + \tilde{\eta}_{\tilde{\mu}} - \eta_{\tilde{\mu}}|, \\ \quad |P_\mu - \tilde{P}_\mu + \eta_\mu - \tilde{\eta}_\mu|\Big\} & \begin{cases} \mu \neq \tilde{\mu} \\ P_\mu - \tilde{P}_\mu \ge 0 \\ \tilde{P}_{\tilde{\mu}} - P_{\tilde{\mu}} \ge 0 \end{cases} \\ |\tilde{P}_{\tilde{\mu}} - P_{\tilde{\mu}} + \tilde{\eta}_{\tilde{\mu}} - \eta_{\tilde{\mu}}| + \\ \quad |P_\mu - \tilde{P}_\mu + \eta_\mu - \tilde{\eta}_\mu| & \begin{cases} \mu \neq \tilde{\mu} \\ \mathrm{sgn}(P_\mu - \tilde{P}_\mu) \neq \\ \mathrm{sgn}(\tilde{P}_{\tilde{\mu}} - P_{\tilde{\mu}}) \end{cases} \end{cases} \tag{7}$$

In the two class case, referring to (3), the estimation errors are not independent. However, as we consider problems involving several classes, we make the simplifying assumption that the probability estimation errors are statistically independent. The useful signal in each term defined by absolute value operators in (7), which is constituted by the difference of a posteriori class probabilities, is corrupted by the difference of the two probability estimation errors. As we assume that the errors are independent, the probability distribution $\tau(\nu)$ of their difference $\nu = \eta_\mu - \tilde{\eta}_\mu$, can be given by a convolution of the two component distributions, i.e.

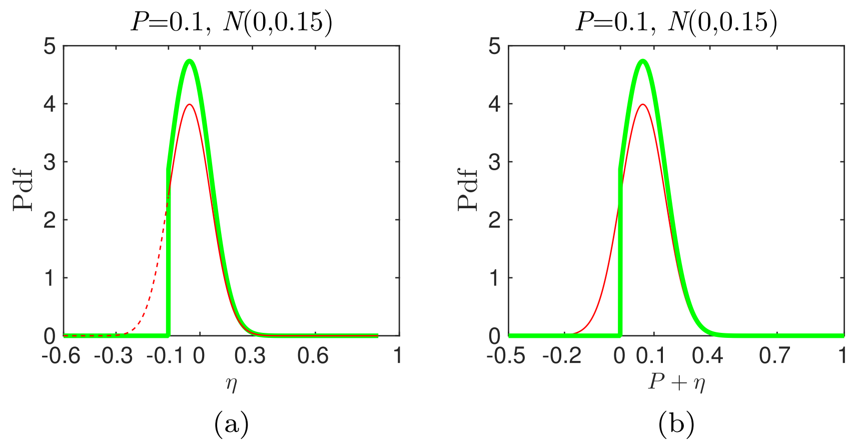


Fig. 4. Distributions of noise (a) and a posteriori estimates (b) for N(0, 0.15).

$$\tau(\nu) = \int_{-\infty}^{\infty} p(\nu-\lambda,0)\tilde{p}(-\lambda,0)d\lambda \tag{8}$$

without loss of generality (w.l.o.g) for all pairs of error terms. It would be difficult to perform an exhaustive bias and variance analysis of (7). However, to get a feel for the effect of the estimation errors, we shall consider a few special cases.

If the a posteriori probability of the most probable class for any expert is close to the cut-off points, then the corresponding estimation error distribution will result in tail clipping as given in (5) to satisfy (4). Any clipping affecting individual components would then show its effect on the joint error distribution defined by (8). Therefore, while computing the expected value of the incongruence measure given in (7), the absolute value operation in expectations would create additional bias of the estimated value as a result of clipping.

In order to keep the analysis simple, in the following few cases we will assume that no tail clipping of the error distributions occurs. In order to obtain closed forms, a further assumption that the identities of the most probable hypotheses do not change has been made. Note that these constraints are not invoked in the comprehensive experimental study given in Section 4.

Case 1: Both classifiers produce identical probability outputs for the most probable hypothesis

In this case, we assume that the expert probability outputs are identical for the most probable hypothesis before the addition of estimation noise. Hence,

$$D_\Delta = |\eta_\mu - \tilde{\eta}_\mu| = |\nu| \tag{9}$$

As no tail clipping occurs, the difference of errors will also be distributed normally with zero mean, but with variance $2\sigma^2$. The absolute value operation will result in D_Δ to have a half normal distribution with mean

$$E\{D_\Delta\} = 2\int_0^\infty \nu\tau(\nu)d\nu = \frac{2}{\sqrt{2\pi}\sqrt{2\sigma}}\int_{\nu=0}^\infty \nu\exp\{-\frac{\nu^2}{4\sigma^2}\}d\nu = \frac{2\sigma}{\sqrt{\pi}} \tag{10}$$

The implication of the result is that even when there is a 100% congruence between the classifiers, the incongruence measure will on average be nonzero, with the bias defined by the variance of the a posteriori probability estimation errors. The variance of the incongruence measure in this ideal case will be given by the variance of the half normal distribution, i.e.

$$var(D_\Delta) = 2\sigma^2(1-\frac{2}{\pi}) \tag{11}$$

Thus, the standard deviation σ_Δ of errors on D_Δ in this scenario is

$$\sigma_\Delta = \sigma\sqrt{\frac{2(\pi-2)}{\pi}} \tag{12}$$

The results in (10) and (12) have bearing on the selection of a threshold on the incongruence measure to detect unusual events.

Case 2: Both classifiers agree on the most probable hypothesis

In this scenario the incongruence measure is

$$D_\Delta = |P_\mu - \tilde{P}_\mu + \eta_\mu - \tilde{\eta}_\mu| \tag{13}$$

Assuming none of the component estimation noise values violates the axiomatic properties of probabilities, the true value of Delta divergence $a = |P_\mu - \tilde{P}_\mu|$ will be corrupted by a noise term with the distribution of a clipped Gaussian, rescaled by factor $1-\gamma = 1 - \frac{1}{2\sqrt{\pi}\sigma}\int_{-\infty}^{0}\exp\{-\frac{(\nu-a)^2}{4\sigma^2}\}d\nu$.

Now let us denote $\Delta P = P_\mu - \tilde{P}_\mu$. To determine the expected value of Delta divergence let us note that under the above assumptions the compound noise distribution $\tau(\nu)$ in (8) is symmetric. The argument $\Delta P + \nu$ can be either positive or negative. However, due to the symmetry induced by the absolute value operation, we need to consider only the case when the argument is positive, as the result for the negative argument will be exactly the same. In this scenario ΔP can be either positive or negative. In the first case, which will occur with probability $1-\gamma$, the contribution to the expected value will be $c_1 = \frac{1}{1-\gamma}\int_0^\infty \nu\tau(\nu-a)d\nu$. When ΔP is negative, the contribution to the mean will $c_2 = \frac{1}{\gamma}\int_0^\infty \nu\tau(\nu+a)d\nu$. The expected value will be given by the weighted sum of these two contributions, namely

$$E\{D_\Delta\} = (1-\gamma)c_1 + \gamma c_2 = \int_0^\infty \nu\tau(\nu-a)d\nu + \int_0^\infty \nu\tau(\nu+a)d\nu \tag{14}$$

This can be alternatively expressed as

$$E\{D_\Delta\} = \int_{-a}^\infty (\nu-a)\tau(\nu)d\nu + \int_a^\infty (\nu+a)\tau(\nu)d\nu \tag{15}$$

which after rearrangement becomes

$$E\{D_\Delta\} = 2\int_a^\infty \nu\tau(\nu)d\nu + a(1-2\gamma) \tag{16}$$

Noting that $\int_a^\infty \nu\tau(\nu)d\nu \geq a\gamma$, we find that the expected value in (16) will be positively biased, i.e.

$$E\{D_\Delta\} \geq a \tag{17}$$

For a given σ, this bias will diminish with increasing $a \leq 0.5$ and $\gamma \to 0$ as well. When $a = 0$ the bias will be equivalent to (10) of Case 1.

The positive bias of Delta divergence will suggest that the classifiers are less congruent than in reality. As a increases, the clipping will monotonically decrease, reducing the positive bias. For large enough differences in the support for the dominant hypothesis (larger a) provided by the two classifiers, the expected value of the incongruence measure will become unbiased, as the contribution of the first term of the expression in (16) will go to zero. This is because there will be no clipping caused by the absolute value operation at the boundary of 0, and the distribution of error differences $\tau(\nu)$ will remain Gaussian. At the same time the factor γ will also approach zero. In general, however, estimation error will be introducing a positive bias and the measured incongruence will appear to be stronger than its true underlying value (noise-free case).

When the distributions of estimation noise on the probabilities of the dominant hypothesis cease to be Gaussian due to the boundary constraint effects, the compound estimation noise distribution becomes complicated, rendering Case 2 intractable. In any case, the argument of the absolute value operation will be distributed according to $\tau(\nu)$ in (8). The inversion of the negative values of ν by the absolute value operation is likely to render the estimated magnitude of Delta divergence once again positively biased.

Case 3: Classifiers disagree on the most probable hypothesis

In this case, as the classifiers disagree on the most probable hypothesis, there is likely to be a gap between the a posteriori probabilities determined by the classifiers for class μ and $\tilde{\mu}$. Let us focus on the scenario where the signs of the probability distributions are positive. Under the assumption that the differences in the estimated a posteriori probabilities of the dominant hypotheses avoid clipping, the form of $\tau(\nu)$ will remain Gaussian for all error terms and the expected value of the incongruence measure will be

$$E\{D_\Delta\} = max\left\{|\tilde{P}_{\tilde{\mu}} - P_{\tilde{\mu}}|, |P_\mu - \tilde{P}_\mu|\right\} + b \qquad (18)$$

with the bias b dependent on the relationship between the arguments of the max operator and the estimation noise distributions, as discussed in **Case 2**. The limiting case of **Case 3** is when for one classifier the maximum a posteriori probability is equal to one while for the other it is zero, and vice versa. Then the estimation error distributions are subject to severe clipping. Note that in this case the estimation noise will tend to reduce the underlying difference between the a posteriori class probabilities and consequently, the expected value of ν will be negatively biased by an offset equal to the mean in (10)

$$E\{D_\Delta\} = [1 - E\{\nu\}] = 1 - \frac{2\sigma}{\sqrt{\pi}} \qquad (19)$$

Note that the effect of estimation noise will be studied experimentally in Section 4.

3.2. Incongruence measure thresholding

To flag incongruence between two classifiers, a suitable threshold must be selected for the incongruence measure. When there is complete agreement between the classifiers (i.e. Case 1), the threshold for the half normal error distribution, $|\eta_\mu - \tilde{\eta}_\mu|$, should be set, say, 3 standard deviations from the mean of the (unclipped) normal distribution. Recalling that the variance of the normal distribution of the compound noise is $2\sigma^2$, it follows that threshold T_Δ should satisfy

$$T_\Delta \geq 3\sqrt{2}\sigma = 4.24\sigma \qquad (20)$$

In practice the estimated a posteriori probabilities will be different. For instance, a contextual classifier is likely to have a sharper distribution of probabilities over the various hypotheses than a non contextual classifier. For a difference in a posteriori probabilities which would result in no error distribution folding and for absolute value operator that would cause no bias, i.e. $|P_\mu - \tilde{P}_\mu| = 3\sqrt{2}\sigma$, the threshold should be set at

$$T_\Delta \geq 3\sqrt{2}\sigma + 3\sqrt{2}\sigma = 8.48\sigma \qquad (21)$$

4. Experimental sensitivity analysis

The theoretical analyses presented in Sections 3.1–3.2 provide some insight into the incongruence measure distribution and hypothesis testing in the presence of noise. However, the basis it provides for selecting the test statistics threshold is incomplete for several reasons:

- In general, it is not possible to obtain closed forms.
- Each solution is for a specific scenario defined by the class probability distribution, the corresponding noise-free incongruence measure value, the level of noise, and its distribution, which changes dynamically as a function of the class probabilities for the dominant hypotheses.

The aim of the simulation studies designed and reported in this section is to obtain a more comprehensive picture of the properties of the test statistics and to develop a practical basis for setting an appropriate incongruence measure threshold. This will be achieved by

- conducting empirical studies of the effect of class probability estimation noise on the distributions of the proposed test statistic, which is parameterised by fixed noise-free incongruence measure values and the number of classes involved in decision making,
- exploring the variations of the test statistic distribution as a function of different scenarios giving rise to the same noise-free incongruence measure value,
- integrating the test statistic distribution over different scenarios, and
- integrating the test statistic distributions over a range of noise-free incongruence values deemed to reflect the state of the two classifiers being congruent.

The successive integrations will yield a resulting test statistic distribution which can be presented in terms of the area under its tail, facilitating the selection of a threshold that would meet a specified level of confidence in the acceptance of the hypothesis of classifier congruence based on the proposed measure.

In Section 4.1, we firstly consider an example scenario where the two classifiers estimate identical posterior distributions for the most probable hypothesis and there is no noise tail clipping. The results of this section are expected to confirm the theoretical findings analysed in Case 1 given in Section 3.1. In Section 4.2 we consider more general scenarios, parameterised by noise-free incongruence measure values and estimation noise statistics. Further experimental studies regarding hypothesis thresholding are carried out in Sections 4.3–4.5. Finally in Section 4.6, the practical implications and guidelines for incongruence detection are provided. This section also includes an example real data application for utilising the provided guidelines.

It should be mentioned that in all experiments, each of the probability distributions employed (P and $\tilde{P}$) has been created by uniform sampling. Specifically, for a given m class problem and an instance x, the a posteriori probability output belonging to class ω_n, $P(\omega_n|x)$, is obtained by drawing a random sample from within the range $\left[0, \left(1 - \sum_{y<n} P(\omega_y|x)\right)\right]$. Note that the upper limit is updated so that $\sum_\omega P(\omega|x) = 1$, and the probability belonging to the last class, ω_m, is assigned to $P(\omega_m|x) = 1 - \sum_{y<m} P(\omega_y|x)$ without

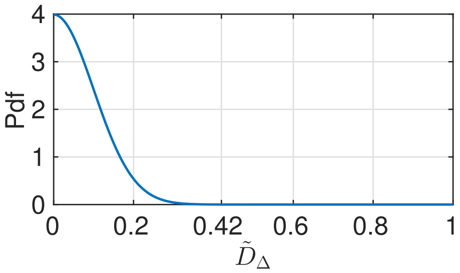


Fig. 5. Pdf curve for $\tilde{D}_{\Delta}$ obtained for identical classifier outputs for the most probable hypothesis, affected by $N(0, 0.10)$ noise with no tail clipping.

sampling. After creating 10^3 many P and $\tilde{P}$ distributions separately, the set of all possible combinations of $(P, \tilde{P})$ are used in the experiments. Hence, the total number of instances, x, is made to be equal to 10^6.

4.1. Identical class probability outputs for the most probable hypothesis

For this simple and somewhat unrealistic case, we assume that the underlying posterior probabilities output by the two classifiers are identical for the most probable hypothesis (i.e. $D_{\Delta} = 0$), and that the identity of the most probable hypothesis (label) does not change after the addition of the estimation noise (Case 1 of Section 3.1).

There is of course an infinite number of posterior class probability distributions which fit this specification. In this case study, we sample them subject to the constraint that the probability of the dominant hypothesis for any expert is sufficiently far away from the boundaries of their interval of support so as not to cause the estimation error distribution to have its tail clipped. The qualifying distributions, P and $\tilde{P}$, are then corrupted by zero mean Gaussian noise, and finally, incongruence measure distributions are acquired from the corrupted distributions. The noisy incongruence measures obtained are denoted as $\tilde{D}_{\Delta}$.

The resulting distribution of $\tilde{D}_{\Delta}$ given in Fig. 5, which is obtained for the standard deviation of the estimation noise $\sigma = 0.1$, supports the theoretical findings in Section 3.1. The curve is shown to be in the form of a half normal distribution as discussed in Section 3.1, and the use of any value greater than $4.24\sigma = 0.424$ as a surprise threshold is depicted to retain at least $\sim$99.7% of the distribution as given in Section 3.2. Note that the number of classes, m, does not have an effect in this particular case, as the terms to do with P and $\tilde{P}$ disappear from the calculation of surprise as shown in (9).

4.2. Distributions of $\tilde{D}_{\Delta}$ for arbitrary class posterior probability and estimation error distributions

In this set of experiments, we parameterise the scenarios by varying noise-free D_{Δ}, and study the impact of noise without applying restrictions on its characteristics such as tail clipping or label change.

Initially, for a given noise-free D_{Δ}, all possible pairs of the probability distributions P and $\tilde{P}$ which output this value from (2), are recorded. The process of selecting the probability distribution pairs takes the cases involving agreement and disagreement in the most probable hypothesis into account separately. As a second step, noise drawn from the distribution $p(\eta)$, which is obtained by regularising $N(0, \sigma)$ as given in (5), is added to the selected P and $\tilde{P}$ pairs. In these experiments, σ is set to 0.10. The resulting distributions of noisy $\tilde{D}_{\Delta}$ are acquired from the corrupted P and $\tilde{P}$.

Using the histograms given in Fig. 3, a few representative (noise-free) D_{Δ} values have been selected to perform the analysis. These values are 0.3 for the case of label agreement, and 0.3 and 0.7 for disagreement. The probability distribution functions of $\tilde{D}_{\Delta}$ obtained for the label agreement case are given in Fig. 6(a) and (b) for 3 and 6 class problems respectively. As for label disagreement, Fig. 7 presents the results for the fixed value of $D_{\Delta} = 0.3$, and Fig. 8 for $D_{\Delta} = 0.7$.

It can be observed for all scenarios of label agreement and disagreement that the peak of the noisy incongruence measure distributions appear at the value where the input noise-free measures are originally defined. However, the noise shows its effect throughout the [0,1] range and the intensity of this effect not only depends on the values of D_{Δ} and σ, but also on the number of classes, m. For greater m, the impact can be observed to be marginally smaller, and hence a narrower spread of the surprise within the range [0,1] is acquired.

4.3. Integration over scenarios

In this section, we concentrate on further experimental analysis regarding hypothesis testing, where the task is to find a threshold on our test statistic which would allow us to reject the hypothesis at a given level of significance.

The experimental analysis reported in Section 4.2 was based on a variety of incongruence measure probability distributions obtained for fixed input noise-free surprise values, sampled by our experimental procedure. However, as we will not know the characteristics of the underlying scenarios in practice, it is more appropriate to integrate over the various scenarios by taking their prior probability of occurrence into account. This integration can then be represented by a plot of the area-under-the-tail belonging to the $\tilde{D}_{\Delta}$ distribution as a function of threshold.

The rationale for this integration can be explained using a simple example. Looking at Fig. 8, it can be observed that a threshold of 0.5 can leave an important portion of some distribution curves out and cause false alarms during surprise detection. However, it may turn out that the cases with large lower tail areas for the given threshold may not be likely to occur with high probability, e.g. they might only happen when the estimation noise causes a label change. In other words, the contribution of these cases to the probability of false alarm might be expected to be low.

Hence, in this set of experiments, by taking the likelihood of the distributions into consideration, the average sizes of the upper tail areas (% over the total area) are gauged for given threshold points. Note that the area estimates are parameterised by noise level. In Figs. 9 and 10, the resulting graphs illustrating the upper tail area (%) versus threshold are given for 3 and 6 class problems respectively. In each figure, the results are obtained for different fixed noise-free surprise values and they are depicted using different line types. The graphs at the top row are acquired using noise distribution with standard deviation $\sigma = 0.05$, whereas at the bottom row with $\sigma = 0.1$. The first column corresponds to the results obtained from the case of label agreement, and the second column applies to disagreement.

Confirming the experimental results presented in Section 4.2, a comparison of Fig 9(a) with Fig. 10(a) shows that for any fixed surprise threshold, the upper tail area size is greater for 3 class problems ($m = 3$) compared to 6 classes ($m = 6$) in the label agreement case. This observation is valid for all values of σ and noise-free D_{Δ} values. For the case of label disagreement, let us analyse, for instance, the scenario in which noise-free $D_{\Delta} = 0.5$ and noise $\sigma = 0.05$ by comparing Fig. 9(b) and (b). The observation that the spread of the surprise distribution within the [0,1] range is greater for $m = 3$ than for $m = 6$ (as previously shown in Section 4.2) is again reflected in the respective area-under-the-tail curves. For ex-

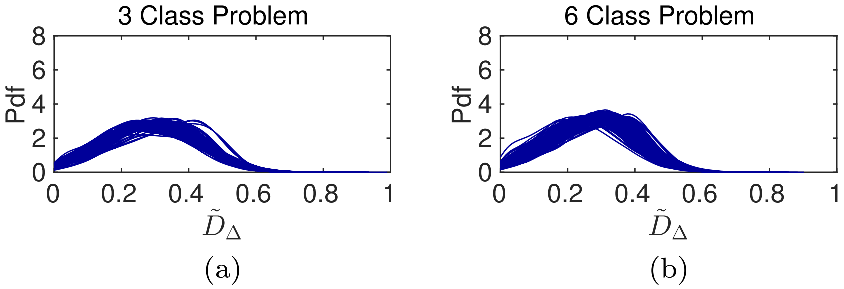


Fig. 6. Pdf curves of $\tilde{D}_\Delta$ for the case of label agreement, obtained for $D_\Delta = 0.3$ corrupted by noise $p(\eta)$, for 3 class problems (a) and 6 class problems (b).

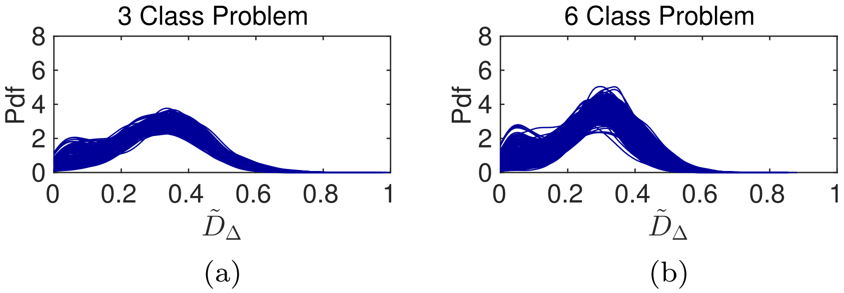


Fig. 7. Pdf curves of $\tilde{D}_\Delta$ for the case of label disagreement, obtained for $D_\Delta = 0.3$ corrupted by noise $p(\eta)$, for 3 class problems (a) and 6 class problems (b).

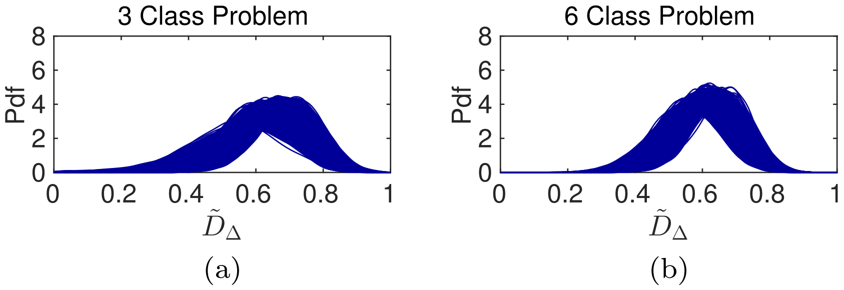


Fig. 8. Pdf curves of $\tilde{D}_\Delta$ for the case of label disagreement, obtained for $D_\Delta = 0.7$ corrupted by noise $p(\eta)$, for 3 class problems (a) and 6 class problems (b).

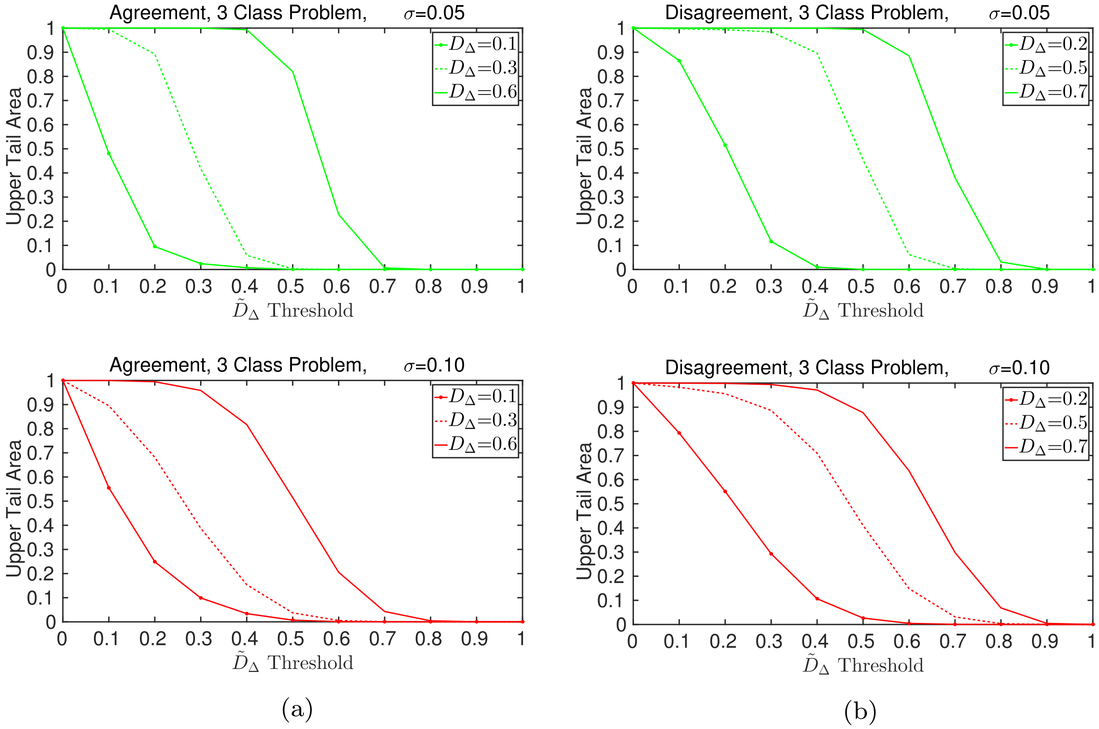


Fig. 9. Upper tail area size versus $\tilde{D}_\Delta$ threshold for different noise levels and different noise-free D_Δ. Given for 3 class problems under the scenarios of classifier label agreement (a), and disagreement (b).

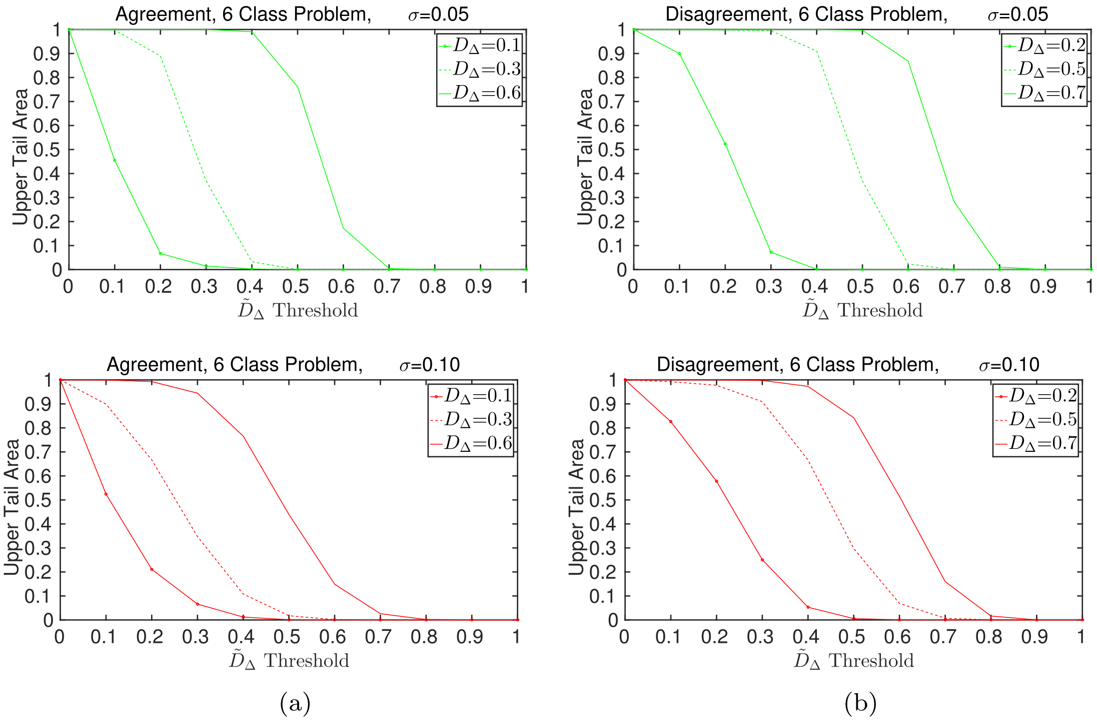


Fig. 10. Upper tail area size versus $\tilde{D}_\Delta$ threshold for different noise levels and different noise-free D_Δ. Given for 6 class problems under the scenarios of classifier label agreement (a), and disagreement (b).

ample, for $\tilde{D}_\Delta = 0.6$, the upper tail area is just under 0.1 for $m = 3$, whereas it is almost zero for $m = 6$.

In Figs. 9(a) and 10(a), a threshold around 0.7 can be observed to cover more than 95% of the lower tail areas for the label agreement cases in all scenarios. This means that almost all scenarios, which incorporate classifier agreement in the most probable hypothesis, will be perceived as congruence. However, it should be borne in mind that a scenario where there is high discrepancy between the probability outputs of two classifiers, giving rise to a high noise-free incongruence value, e.g. one greater than 0.5, should not necessarily be labeled as congruence even though there is label agreement regarding the most probable hypotheses identified by these classifiers. Hence, depending on the choice of a noise-free D_Δ cut-off for labelling congruence/incongruence, a more suitable threshold for $\tilde{D}_\Delta$ should be selected.

Let us say we are using the cut-off value of $D_\Delta = 0.5$ such that all noise-free surprise values below 0.5 are to be detected as congruence, and above this value for incongruence. Utilizing $\sigma = 0.05$, it can be observed from Figs. 9(a) and 10(a) that the threshold of $\tilde{D}_\Delta = 0.4$ labels all cases with $D_\Delta = 0.6$ as incongruence, and $D_\Delta = 0.1, 0.3$ as congruence with confidence around ∼95%.

Proceeding with $D_\Delta = 0.5$ cut-off value and $\sigma = 0.05$, and looking at Figs. 9(b) and 10(b) to analyse the case of label disagreement, it can be seen that employing a threshold of 0.4 (as in the case of label agreement) results in identifying the scenarios with noise-free $D_\Delta = 0.5, 0.7$ as incongruence, and scenarios with $D_\Delta = 0.2$ as congruence with ∼90% confidence.

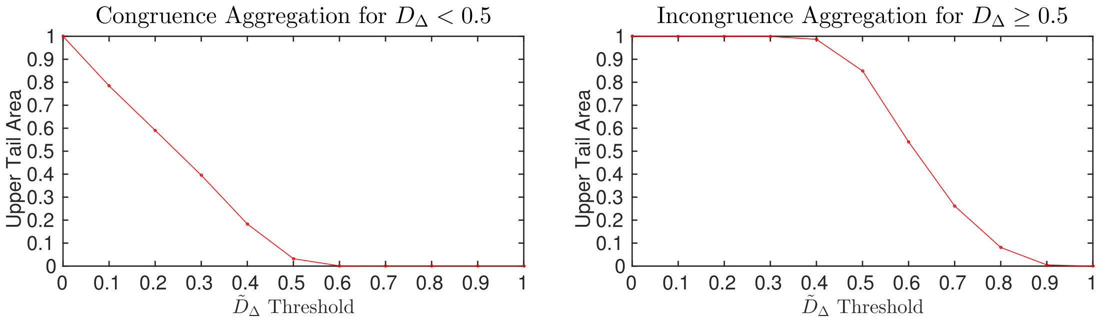
Although the findings in this section are of importance to give an insight into the effects of $\tilde{D}_\Delta$ for fixed D_Δ about the cases of agreement and disagreement separately, it should be noted that in practice it is not possible to know in advance the values of the noise-free incongruence measures or the nature of the problem (giving rise to label agreement or disagreement). Hence, in Section 4.4, we will be marginalizing over these concepts after selecting a cut-off value for the noise-free measure to define the congruence-incongruence boundary.

4.4. Integration over noise-free congruence values

In Section 4.3 we have integrated over various scenarios, each defined by a fixed noise-free surprise value, and presented the findings for the cases of label agreement and disagreement separately. Here, we further integrate these area-under-the-tail distributions by aggregating over all noise-free surprise values below 0.5 for congruence, and above for incongruence. This process takes the prior distributions of noise-free values into account and marginalises over the scenarios of label agreement/disagreement to reflect the use of the proposed measure in practice. Hence, the thresholds suggested as a result of the experiments in this section will be different to those from Section 4.3.

The results of the experiments are provided in Fig. 11 for a 6 class problem and for $\sigma = 0.05$. Fig. 11(a) indicates the confidence in the decision to accept the hypothesis that the two classifiers are congruent as a function of $\tilde{D}_\Delta$. It can be observed that, for instance, a threshold of 0.5 on the proposed measure would capture the classifier congruence cases at ∼95% confidence. Setting the threshold to 0.6 would raise the confidence level to ∼100%. However, the plot in Fig. 11(b) clearly indicates that we should not be too ambitious, as setting the threshold to yield high confidence levels for detecting classifier congruence will inevitably lead to unacceptable level of false negatives, i.e. declaring incongruent classifier outputs as congruent. For example, at 0.4 threshold we will correctly detect ∼100% of classifier incongruence instances, but this figure goes down to ∼80% for the threshold set at 0.5.

Thus choosing a suitable classifier incongruence detection threshold is a question of trade-off between low false positives and low false negatives. In this context, it is important to bear in mind, that in practical applications we will not normally be able to generate the area-under-the-tail curves for incongruence cases. The threshold selection will have to be based on such curves for classifier congruence cases only.



(a)

(b)

Fig. 11. Aggregate upper tail area versus $\tilde{D}_\Delta$ for $\sigma = 0.05$ and 6 classes. Aggregated over $D_\Delta < 0.5$ for congruence (a) and $D_\Delta \geq 0.5$ for incongruence (b).

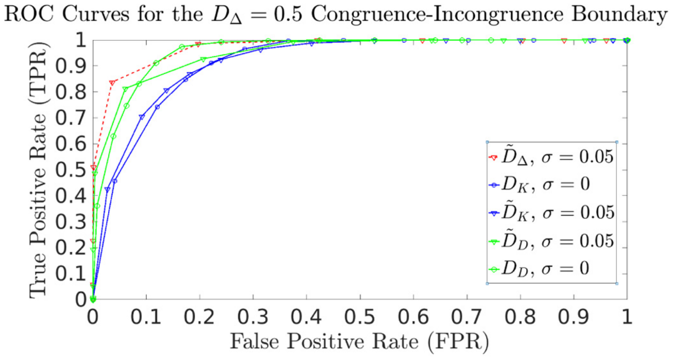


Fig. 12. ROC curves showing the capacity of $\tilde{D}_\Delta$, D_K, $\tilde{D}_K$, D_D and $\tilde{D}_D$ to separate the state of congruence from incongruence computed for $\sigma = 0.05$ and 6 classes.

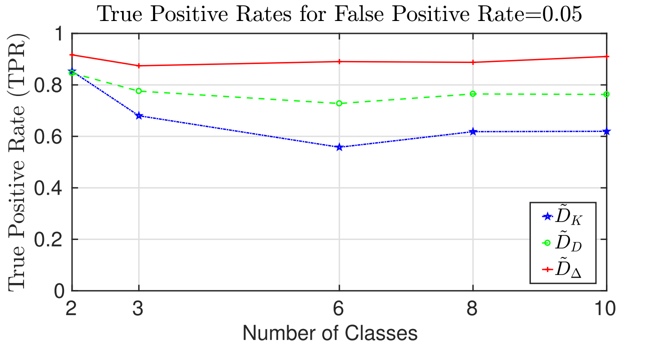


Fig. 13. True positive rates (TPR) calculated for a number of classes, after setting the false positive rate (FPR) to 0.05, for $\sigma = 0.05$.

4.5. Relationship of Delta divergence with KL and DC-KL under noise

The relationship of noise-free Delta divergence of Bayesian surprise (KL) was already shown and discussed in [24] as the main motivation for the development of the novel decision cognizant divergence and its validation. It is pertinent to investigate whether the favourable properties of Delta divergence vis-a-vis KL and its decision cognizant variant DC-KL are preserved even when the a posteriori class probability estimates are subject to errors.

In Fig. 12, we plot the Receiver Operating Characteristic (ROC) curves of the noisy Delta divergence ($\tilde{D}_\Delta$), noisy Bayesian surprise ($\tilde{D}_K$) and noisy DC-KL ($\tilde{D}_D$) for the 6 class problem and $\sigma = 0.05$. The ROC curves are computed by setting the boundary between congruence and incongruence at $D_\Delta = 0.5$. The figure also shows the ROC curve for noise-free Bayesian surprise and DC-KL measures, given as D_K and D_D respectively.

The results demonstrate that the ROC curves for the noise-free Bayesian surprise and DC-KL measures are quite remote from the top left corner (perfect separation) due to clutter, although DC-KL shows better performance than KL. In the presence of estimation noise, the areas under the ROC curves for $\tilde{D}_K$ and $\tilde{D}_D$ are much lower than that for $\tilde{D}_\Delta$. Moreover, as anticipated, the areas for $\tilde{D}_K$ and $\tilde{D}_D$ are also smaller than that for D_K and D_D. Note that with increasing levels of noise, the under-the-curve area sizes decrease, but the ranking of the measures is maintained. It is also interesting to mention that the area under the ROC curve for $\tilde{D}_\Delta$ is larger than the area related to the noise-free D_K and D_D as given in Fig. 12 for $\sigma = 0.05$, and this observation still holds for $\sigma = 0.1$.

In order to show the supremacy of Delta divergence over KL and DC-KL for a varying number of classes, in Fig. 13, we show the results of an experiment where we set the confidence level for false positives to 0.05 and calculate the corresponding true positive rates of the given measures for 2,3,6,8,10 and 15 classes. The measurements are performed for $\sigma = 0.05$. It can be observed that TPR for Delta divergence is better than that of DC-KL for all number of classes, and DC-KL outperforms KL except for 2 classes, where DC-KL becomes identical to KL as there is no 'clutter" class in this scenario. Note that the plots remain approximately constant for higher number of classes than 10 (not shown in the figure).

4.6. Practical implications

The theoretical analysis and the simulation studies presented in the paper are intended to provide an intuitive insight for the properties of the proposed incongruence measures. However, in practice, setting the decision thresholds is more likely to be based on empirical distributions of $\tilde{D}_\Delta$ estimated on some anomaly free content. This can simply be achieved by histogramming the incongruence measure values computed from the estimated class a posteriori probabilities on a stream of training sensor data. From such a histogram the graph relating the upper tail area to threshold on $\tilde{D}_\Delta$, similar to that plotted in Fig. 11(a), can be determined and a suitable threshold selected, corresponding to a given level of confidence. This is a very pragmatic approach, as it makes use of the posterior probabilities for all the hypotheses estimated by the data interpretation system. We do not need the ground truth values of these probabilities, neither noise estimates.

The selected threshold can be tested on an independent set of anomaly free data of the same quality to check for false positives. This again is realistic, and can be done without any ground truth annotation of the validation data.

Commonly, the decision threshold would be based on a desired level of confidence that the classifier outputs are congruent. This is compatible with the standard methodology of statistical hypothesis testing for outliers in statistical anomaly detection [6]. It is also consistent with the underlying philosophy that any anomaly de-

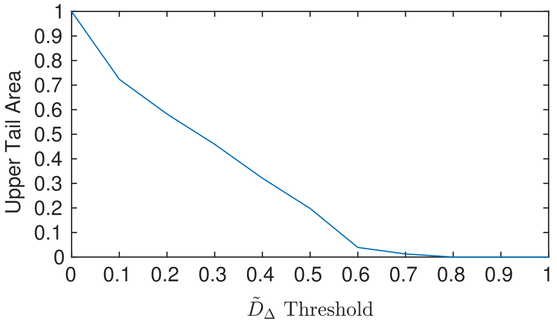


Fig. 14. Upper tail area for anomaly-free Breakfast dataset samples for training set.

tection system should be designed on anomaly free training data, as anomalies, by definition, are very rarely observed, and therefore cannot be used in training. However, in some cases a few anomaly observations may be available or even synthetically generated; anomalous objects or events could be inserted in the data, or alternatively, some object models could be removed from the model database. For example, a few items could be removed from the speech recognition system vocabulary, which would result in incongruence between the outputs of phoneme and word classifiers, indicating out of vocabulary word anomaly. The incongruence threshold level could then be checked for anomaly under-detection (false negatives) on realistic examples of anomalous, or at least incongruous, situations.

A set of guidelines to be utilized for measuring incongruence in practice can be given as follows:

1. Using an anomaly-free training set of sensor data, the a posteriori probabilities, which are computed by the classifiers for various hypotheses as part of the data interpretation process, are recorded.
2. The adopted incongruence measure values are computed from the probabilities obtained in Step 1, and their distribution is estimated.
3. The area under the tail of the distribution determined in Step 2 as a function of threshold on the test statistic is computed. This will produce a graph equivalent to the one shown in Fig. 11(a).
4. Using the plot derived in Step 3, a classifier incongruence hypothesis testing threshold is selected for a specified confidence level, as described in Section 4.3.

Note, if it is possible to create a validation set with synthetically injected anomalies, Steps 1-3 can be repeated so as to obtain an under the tail distribution equivalent to Fig. 11(b). In this scenario, it will be possible to compute a ROC curve similar to those provided in Fig. 12, and threshold selection can be made to reflect a suitable balance between false positives and false negatives.

Let us now demonstrate the use of these guidelines on the action recognition problem defined on the Breakfast dataset.

4.6.1. *Incongruence detection on Breakfast Dataset*

Breakfast dataset [41] is a current benchmark for action and activity recognition from videos, which comprises 10 activities related to breakfast preparation, performed by 52 different individuals in 18 different kitchens. Each activity consists of a number of action units, and 48 different action units are observed in total. For this dataset, the goal is twofold: (1) to recognise simple, primitive actions (such as *cut fruit, take bowl*), (2) to recognise high level, complex activities (such as *prepare salad*) by utilising the detected actions. In this section, we focus on the outputs of a contextual and a non-contextual classifier, to illustrate the design of a classifier incongruence detector based on the Delta divergence in a practical scenario.

We first extract low-level local features with improved dense trajectories (iDTFs) [43] and reduce their size to half (from 426 to 213 elements) with PCA. Using the training set defined by the experimental protocol for the Breakfast Dataset [41] we estimate a 16 mode Gaussian mixture model of the empirical distribution of the extracted features. The features are encoded to Fisher vectors [44] with the VLFeat toolbox [45]. Finally, L2 and power normalisations are applied to the Fisher vectors. The resulting Fisher vectors are of size $2 \times K \times D$, where K is the number of clusters of the GMM and D is the dimensionality of the PCA compressed iDTF descriptor. In our case, for $K = 16$ and $D = 213$ the size of each Fisher vector is 6816 dimensions. We reduce this size to 64 dimensions with a second PCA. Having obtained the reduced dimensionality Fisher vectors, we recognise actions in the dataset with the HTK toolkit [46].

The HTK toolkit performs a non-contextual action recognition. For each detected action, HTK provides its temporal extends (*i.e.* its start and end point within the video), its class (*e.g. pour water, stir milk*) and a detection score in the form of log-likelihood. The HTK toolkit contextual classifier performs activity recognition by utilising information regarding each action's neighbouring actions.

The contextual classifier partitions each video in the Breakfast dataset and assigns action labels to each of the resulting segments. The noncontextual classifier uses the segmentation information derived from the contextual classifier and also labels each segment individually. Delta divergence values are then computed for all segments, using the class probability values output by the classifiers. Afterwards, the set of all segments are divided into two random partitions for training and test, and anomalies are eliminated from

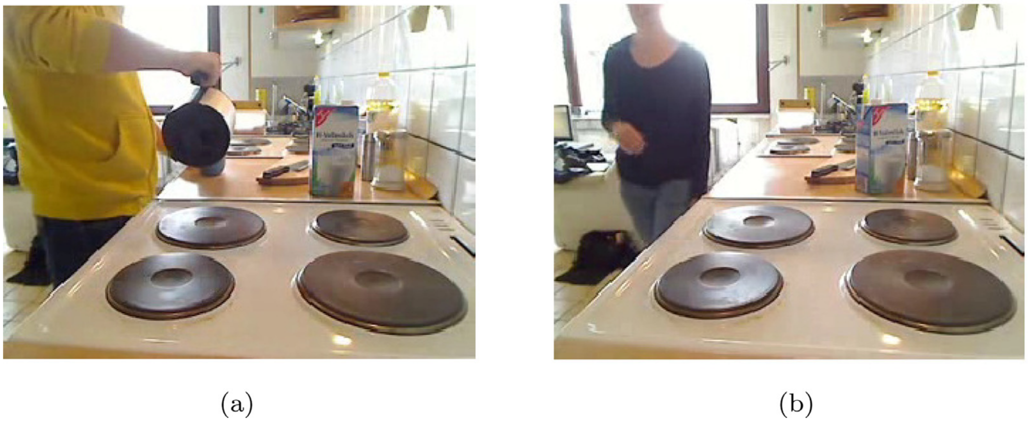


Fig. 15. Key frames belonging to the main action (a) and the background action (b), extracted from an example test sequence in Breakfast dataset.

the training set. By selecting monotonically increasing values of threshold for $\tilde{D}_{\Delta}$, the area under the tail of the Delta divergence distribution can be computed for the anomaly-free training set, as given in Fig. 14.

The operational threshold for incongruence detection can then be selected to produce an appropriate level of confidence in the acceptance of congruence hypothesis. Specifically, as an example, for the distribution in Fig. 14 we identify the threshold of 0.63 at 2.5% confidence level. The amount of the false negatives detected by this threshold in a separate test set is 2.6%, which is close to the set confidence level, as expected. Interestingly, the test set contains a few instances of classifier outputs producing incongruence value close to unity. An example of such a case is shown in Fig. 15.

This true incongruence flags a situation where the video happens to contain a main action sequence of coffee making, as demonstrated by the key frame in Fig. 15(a), and a secondary sequence which takes place at the background after the completion of the main action, as given by the key frame in Fig. 15(b). The contextual classifier recognises the final segment of this video, upon the completion of coffee making, as “no action”. However, the noncontextual classifier labels it as “take bowl”, as this is an action carried out by the background object at this time instance. Hence, each of the classifiers produces a sensible response, however they focus on different interpretations, and this disagreement is detected by the incongruence detector correctly.

5. Conclusion

We addressed the problem of classifier incongruence detection for decision making systems engaging multiple classifiers (contextual/noncontextual, multimodal). The problem has been cast as one of statistical hypothesis testing, with the focus of the paper directed on the choice of a suitable test statistics. It has been argued that the challenging nature of the classifier incongruence detection lies in the inherent fuzziness of the concept of incongruence, and the effect of estimation errors on the classifier outputs. After reviewing the deficiencies of the state-of-the-art methods for classifier incongruence detection, we carried out a theoretical and experimental investigation of a recently proposed measure, Delta divergence, with the aim of providing an intuitive feel for its behaviour. The simulation studies were designed to estimate the probability distribution of the test statistics for various scenarios defined in terms of noise-free classifier incongruence measure values and estimation error statistics. The area under the tail of the distribution for various thresholds on the test statistics can then be determined to illustrate the effect of estimation noise on incongruence threshold selection. Based on the theoretical findings, a set of guidelines have been developed for selecting classifier incongruence threshold in practice. The use of these guidelines has been illustrated on the problem of action and activity recognition in breakfast scenario videos recording the preparation of different types of dishes for breakfast.

As for future work, the analysis can further be expanded to account for scenarios where more than two decision making experts are taken into consideration. Moreover, it would be interesting to conduct an extensive comparative study of Delta divergence with other families of divergence measures such as Bregman [47] and Renyi [26] divergences. However these divergences would have to be extended to decision cognizant equivalents first.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) grant numbers EP/K014307/1 and EP/N007743/1 and the MOD University Defence Research Collaboration (UDRC) in Signal Processing.

References

[1] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, ACM Comput. Surv. 41 (3) (2009) 15:1–15:58.

[2] J.A. Quinn, M. Sugiyama, A least-squares approach to anomaly detection in static and sequential data, Pattern Recognit. Lett. 40 (2014) 36–40.

[3] L. Dong, S. Liu, H. Zhang, A method of anomaly detection and fault diagnosis with online adaptive learning under small training samples, Pattern Recognit. 64 (2017) 374–385.

[4] H. Wang, M. Tang, Y. Park, C. Priebe, Locality statistics for anomaly detection in time series of graphs, Signal Process. IEEE Trans. 62 (3) (2014) 703–717.

[5] V. Saligrama, E. Arias-Castro, R. Chellappa, A.O. Hero, R. Nowak, V.V. Veeravalli, Introduction to the issue on anomalous pattern discovery for spatial, temporal, networked, and high-dimensional signals, IEEE J. Sel. Top. Signal Process. 7 (1) (2013) 1–3.

[6] M. Markou, S. Singh, Novelty detection: a review-part 1: statistical approaches, Signal Process. 83:12 (2003) 2481–2497.

[7] M. Markou, S. Singh, Novelty detection: a review-part 2: neural network based approaches, Signal Process. 83:12 (2003) 2499–2521.

[8] W.L. Al-Yaseen, Z.A. Othman, M.Z.A. Nazri, Real-time multi-agent system for an adaptive intrusion detection system, Pattern Recognit. Lett. 85 (2017) 56–64.

[9] W. Hu, W. Hu, S. Maybank, Adaboost-based algorithm for network intrusion detection, IEEE Trans. Syst. Man Cybern., Part B (Cybern.) 38 (2) (2008) 577–583.

[10] J. Zhang, M. Zulkernine, A. Haque, Random-forests-based network intrusion detection systems, IEEE Trans. Syst., Man, Cybern., Part C (Appl. Rev.) 38 (5) (2008) 649–659.

[11] C. Zhou, S. Huang, N. Xiong, S.H. Yang, H. Li, Y. Qin, X. Li, Design and analysis of multimodel-based anomaly intrusion detection systems in industrial process automation, IEEE Trans. Syst., Man, Cybern. 45 (10) (2015) 1345–1360.

[12] M.J. Leach, E. Sparks, N.M. Robertson, Contextual anomaly detection in crowded surveillance scenes, Pattern Recognit. Lett. 44 (2014) 71–79.

[13] K. Ouivirach, S. Gharti, M.N. Dailey, Incremental behavior modeling and suspicious activity detection, Pattern Recognit. 46 (3) (2013) 671–680.

[14] R. Chaker, Z.A. Aghbari, I.N. Junejo, Social network model for crowd anomaly detection and localization, Pattern Recognit. 61 (2017) 266–281.

[15] O.P. Popoola, K. Wang, Video-based abnormal human behavior recognition - a review, IEEE Trans. Syst., Man, Cybern., Part C (Appl. Rev.) 42 (6) (2012) 865–878.

[16] Y. Yuan, J. Fang, Q. Wang, Online anomaly detection in crowd scenes via structure analysis, IEEE Trans. Cybern. 45 (3) (2015) 548–561.

[17] C. O'Reilly, A. Gluhak, M.A. Imran, S. Rajasegarar, Anomaly detection in wireless sensor networks in a non-stationary environment, IEEE Commun. Surv. Tutor. 16 (3) (2014) 1413–1432.

[18] H.H.W.J. Bosman, A. Liotta, G. Iacca, H.J. Wrtche, Anomaly detection in sensor systems using lightweight machine learning, in: 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 7–13.

[19] A. Patcha, J.-M. Park, An overview of anomaly detection techniques: Existing solutions and latest technological trends, Comput. Netw. 51 (12) (2007) 3448–3470.

[20] J. Kittler, W. Christmas, T. de Campos, D. Windridge, F. Yan, J. Illingworth, M. Osman, Domain anomaly detection in machine perception: A system architecture and taxonomy, IEEE Trans. Pattern Anal. Mach.Intell. 35 (2014) 1,14.

[21] H. Ketabdar, M. Hannemann, H. Hermansky, Detection of out-of-vocabulary words in posterior based ASR, in: Proceedings European Conference on Speech Communication and Technology: Interspeech, 2007, pp. 1757–1760.

[22] L. Itti, P.F. Baldi, A principled approach to detecting surprising events in video, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 631–637.

[23] M. Ponti, J. Kittler, M. Riva, T. de Campos, C. Zor, A decision cognizant Kullback-Leibler divergence, Pattern Recognit. 61 (2017) 470–478.

[24] J. Kittler, C. Zor, Delta divergence: anovel decision cognizant measure of classifier incongruence, 2016, arXiv:1604.04451.

[25] D. Weinshall, A. Zweig, H. Hermansky, S. Kombrink, F.W. Ohl, J. Anemeuller, J.-H. Bach, L.V. Gool, F. Nater, T. Pajdla, M. Havlena, M. Pavel, Beyond novelty detection: Incongruent events, when general and specific classifiers disagree, IEEE Trans. Pattern Anal. Machine Intell. 34 (2012) 1886–1901.

[26] A. Rényi, On measures of entropy and information, in: 4th Berkeley Symposium on Mathematical Statistics and Probability, 1961, pp. 547–561.

[27] F. Liese, I. Vajda, On divergences and informations in statistics and information theory, IEEE Trans. Inf. Theory 52 (10) (2006) 4394–4411.

[28] F.C. Porter, Testing consistency of two histograms, arXiv:0804.0380 (2008).

[29] F.J. Massey, The Kolmogorov-Smirnov test for goodness of fit, J. Am. Stat. Assoc. 46 (253) (1951) 68–78.

[30] H. Cramér, On the composition of elementary errors: first paper: mathematical deductions, Scand. Actuar. J. 1928 (1) (1928) 13–74.

[31] R. von Mises, Wahrscheinlichkeit, Statistik und Wahrheit, Berlin, 1928.

[32] T.W. Anderson, D.A. Darling, Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes, Ann. Math. Stat. 23 (2) (1952) 193–212.

[33] W. Weaver, Probability, rarity, interest, and surprise, Sci. Mon. 67 (6) (1948) 390.

[34] I. Guttman, The use of the concept of a future observation in goodness-of-fit problems, J. R. Stat. Soc. Ser. B (Methodol.) (1967) 83–100.

[35] D.B. Rubin, Bayesianly justifiable and relevant frequency calculations for the applies statistician, Ann. Stat. 12 (4) (1984) 1151–1172.

[36] I.J. Good, Surprise index, Encyclopedia of Statistical Sciences, 1988.

[37] I.J. Good, The surprise index for the multivariate normal distribution, Ann. Math. Stat. (1956) 1130–1135.

[38] M. Evans, Bayesian inference procedures derived via the concept of relative surprise, Commun. Stat. 26 (1997) 1125–1143.

[39] W. Weaver, Lady Luck: The Theory of Probability, Courier Dover Publications, 2012.

[40] M. Bayarri, J.O. Berger, Measures of surprise in Bayesian analysis, Duke University, 1997.

[41] H. Kuehne, A. Arslan, T. Serre, The language of actions: recovering the syntax and semantics of goal-directed human activities, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 780–787, doi:10.1109/CVPR.2014.105.

[42] J. Kittler, C. Zor, A measure of surprise for incongruence detection, Intelligent Signal Processing, IET, 2015.

[43] H. Wang, C. Schmid, Action recognition with improved trajectories, in: 2013 IEEE International Conference on Computer Vision, 2013, pp. 3551–3558, doi:10.1109/ICCV.2013.441.

[44] F. Perronnin, J. Sánchez, T. Mensink, Improving the Fisher Kernel for Large-Scale Image Classification, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 143–156.

[45] A. Vedaldi, B. Fulkerson, VLFeat: an open and portable library of computer vision algorithms, 2008, (http://www.vlfeat.org/).

[46] M.I.L. Machine Intelligence Laboratory of the Cambridge University Engineering Department, The hidden Markov model toolkit (HTK) (http://htk.eng.cam.ac.uk/), 2016.

[47] L.M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Comput.Math.Math.Phys. 7 (3) (1967) 200–217.

Josef Kittler (Ph.D., Sc.D.) is professor of machine intelligence at the Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering, University of Surrey. He conducts research in biometrics, video and image database retrieval, and cognitive vision. He published a Prentice Hall textbook on Pattern Recognition: A Statistical Approach, and more than 200 journal papers. He serves on the editorial board of several journals in pattern recognition and computer vision.

Cemre Zor is currently a research associate at the the Centre for Vision, Speech, and Signal Processing (CVSSP), University of Surrey/UK. She has previously obtained her M.Sc. and Ph.D. degrees from the same institution in 2008 and 2014, respectively. Her current research interests include anomaly detection, multiple classifier systems, bias and variance theory of classification.

Ioannis Kaloskampis is a research associate at Cardiff University working on anomaly detection in video. He obtained his M.Eng. degree from National Technical University of Athens in 2007 and his M.Sc. degree from University of Bristol in 2009. Between 2009 and 2013, he worked in the area of analysis of complex human behaviours in video for a Ph.D. award at Cardiff University. His research interests are in visual surveillance, biometric profiling, behaviour analysis and human-computer interaction.

Yulia Hicks joined the Cardiff School of Engineering as a lecturer in May 2004. Her main research interests are in the areas of computer vision and image processing. Dr. Hicks' other research interests lie in the areas of statistical modelling, modelling and tracking of human motion, joint modelling of facial articulation and speech and multi-modal signal processing.

Wenwu Wang is a reader in Signal Processing at the Centre for Vision, Speech and Signal Processing, University of Surrey. He is a member of the MOD University Defence Research Centre in Signal Processing (since 2009), a member of the BBC Audio Research Partnership (since 2011), an associate member of Surrey Centre for Cyber Security (since 2014), and a member of the MRC/EPSRC Microphone Network - Novel Applications of Microphone Technologies to Hearing Aids (since 2015).