A Measure of Surprise for Incongruence Detection

J Kittler, C Zor

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, GU2 7XH, U.K.
Email: {J.Kittler, C.Zor}@surrey.ac.uk

Keywords: Anomaly detection, Bayesian surprise, incongruence detection

Abstract

The ability to detect unexpected events improves dramatically when more than one expert is involved in decision making. Incongruence between two or more experts is indicative of something unusual and its measuring has applications in domains such as anomaly detection and multimodal decision systems. In this paper, we propose a new classifier incongruence measure, which overcomes the critical shortcomings of those existing in the literature. An experimental study has been carried out showing the advantageous properties of the proposed measure including its relatively low sensitivity to estimation noise, under the assumption of constrained Gaussian distribution. For different noise-free measure values corrupted with different levels of noise, we show that it is possible to determine classifier incongruence thresholds at given levels of statistical significance.

1 Introduction

Many data interpretation systems involve multiple classifiers. Examples include classifier ensembles which are designed to enhance the classification system performance, or multimodal systems, where the gain in performance is achieved by combining complementary sources of sensor information about the objects being classified, as well as hierarchical interpretation systems which engage low non-contextual classifiers the output of which is then combined at a higher level using contextual decision making. The information flow and its fusion in such multiple classifier systems can be hard wired. However, it is becoming evident that further performance improvements can be realised by intelligent processing of the outputs of the respective classifiers and associated information, such as data quality, decision confidence, and classifier incongruence, which can be used for information fusion control purposes and other applications.

Among these control measures, classifier incongruence is a relatively new tool for analysing the decision making process, which gauges the consistency of classifier outputs. When incongruence is detected, it warrants further investigation, as normally all component classifiers should provide a coherent support for a particular decision. In anomaly detection, for instance, incongruence of noncontextual and contextual classifiers may be indicative of a particular nuance of anomaly. For example, in speech recognition, incongruence between phoneme classifiers and word classifiers could indicate an "out of vocabulary" event [3]. In multimodal systems, incongruence between different modalities could signify a sensor failure, a spoofing attempt, or signal corruption. Incongruence monitoring and detection then provides a very useful mechanism for triggering an appropriate control action.

Measuring incongruence involves two discrete probability distributions. If the two distributions are similar, then the classifier outputs would be considered congruent. Hence, incongruence could be detected by defining a suitable similarity metric. This suggests that histogram similarity measuring techniques could be adapted for measuring classifier incongruence, although there are not yet any 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 [13]. These tests are the extensions of some well-known techniques that are mainly used for calculating the goodness-of-fit between an empirical and a reference distribution. Examples are Chi-square, Kolmogorov-Smirnov [4], Cramér-von-Mises [5, 6], and Anderson-Darling [7] tests. We will 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.

The Bayesian surprise measure (*BS*) [1], which calculates the Kullback-Leibler distance between two probability distributions, is specifically suggested for measuring incongruence, and is the key existing technique used in practice. However, this similarity measure is decision agnostic as it does not ignore the terms associated with vestigial a posteriori probabilities of the classes that are unlikely to be selected by the adopted decision rules. This gives rise to a lot of irrelevant jitter in the final value of the measure. Moreover, the measure diverges to infinity and is dependent on the reference expert, giving two different values according to which classifier is selected as the reference.

To overcome the shortcomings of *BS*, an alternative measure, delta (Δ^*), is proposed in [2]. This measure focuses only on

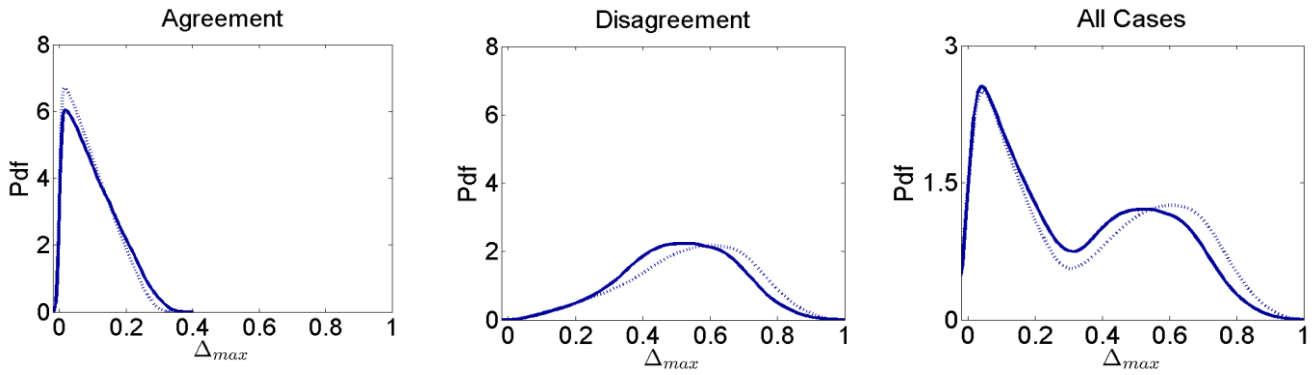


Figure 1: Probability density functions (pdf.s) of Δ_{max} for expert agreement on the most probable hypothesis (a), for disagreement (b), and for all cases (c). 3 class problems are indicated by dashed lines, and 6 class problems with solid lines.

the most probable classes identified by the respective classifiers and is confined to the interval [0,1]. Furthermore, it is symmetric with respect to the expert selected as reference. However, its disadvantage is that it does not explicitly take into account label switching, that is when the identity of the most probable classes selected by the two classifiers differs. This may result in the delta surprise measure having higher values for when the most probable classes are the same, than when they differ.

In this paper, we propose a new way of measuring incongruence by updating Δ^* in such a way that the surprise value is magnified if the two experts support distinct dominant hypotheses than if they support the same. This measure is named as Δ_{max}. The formulation and technical details of Δ_{max} are given in Section 2, and the experimental analysis including error sensitivity are provided in Section 3. Finally, in Section 4, a discussion of the findings are provided with conclusions drawn.

2 An alternative incongruence measure: Δ_{max}

Let $\tilde{\mathrm{P}}(\omega_\mathrm{j}|\mathrm{x})$ and $\mathrm{P}(\omega_\mathrm{j}|\mathrm{x})$; $\mathrm{j} = 1, \ldots, \mathrm{m}$ denote the a posteriori probabilities estimated by two experts (classifiers) about the association of input data x to class ω_j. Their incongruence can be detected by measuring the Kulback-Leibler divergence between their respective class a posteriori probability distributions, by considering the output by one of the experts as a reference. This measure is referred as the Bayesian Surprise (BS) [1] and is given as

$$BS = \sum_{j=1}^{m} \tilde{\mathrm{P}}(\omega_\mathrm{j}|\mathrm{x}) \log \frac{\tilde{\mathrm{P}}(\omega_\mathrm{j}|\mathrm{x})}{\mathrm{P}(\omega_\mathrm{j}|\mathrm{x})} \qquad (1)$$

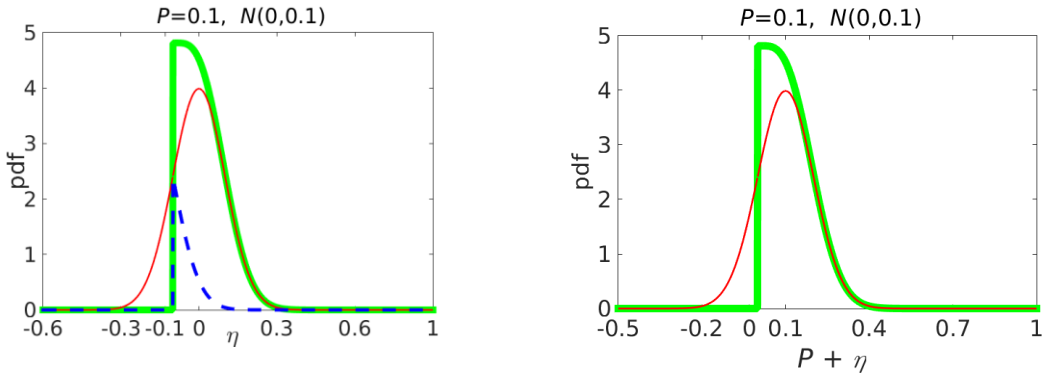
BS can be interpreted as a measure of dissimilarity: High values of the measure suggest a big difference in the a posteriori probability distributions, flagging incongruence between the classifier outputs. The measure will tend to zero if the distributions are identical or similar.

Although *BS* is the most commonly used technique for measuring classifier incongruence, there exist some issues associated with it. First of all, the measure goes to infinity for $\mathrm{P}(\omega_\mathrm{j}|\mathrm{x})$ going to zero. This can bring about false alarms of incongruence. Moreover, *BS* is not symmetric in terms of selecting different experts as reference, as different references yield distinct values of surprise/incongruence. It should also be noted that *BS* accumulates contributions from all classes, including those that can be interpreted as noise, and therefore is strongly affected by estimation errors.

So as to deal with the problems associated with *BS*, delta measure (Δ^*) was proposed in [2]. Δ^* is symmetric, confined to a fixed interval ([0,1]) and focuses on the dominant hypotheses flagged by the two experts, namely $\mu = \arg\max_\omega \mathrm{P}(\omega|\mathrm{x})$ and $\tilde{\mu} = \arg\max_\omega \tilde{\mathrm{P}}(\omega|\mathrm{x})$. This eliminates the noise contributions from the non-dominant classes. The formulation of Δ^* is given as

$$\Delta^* = \frac{1}{2}\{|\mathrm{P}(\mu|\mathrm{x}) - \tilde{\mathrm{P}}(\mu|\mathrm{x})| + |\tilde{\mathrm{P}}(\tilde{\mu}|\mathrm{x}) - \mathrm{P}(\tilde{\mu}|\mathrm{x})|\} \qquad (2)$$

However, Δ^* has one undesirable property. At times when $\tilde{\mu} = \mu$ (when the favoured hypotheses of the two experts are the same), the second term of the measure becomes identical to the first and causes a doubling of the difference between the two a posteriori probability values. This may result in masking the case where the two favoured hypotheses differ, which is more of a surprise.



(a) (b)

Figure 2: Distributions of noise (a) and a posteriori estimates (b) for $P = 0.1$, $q(\eta) = N(0,0.1)$

In order to overcome the shortcomings of the existing incongruence measures, we formulate a new measure, Δ_{max}, as follows:

$$\Delta_{max} = \frac{1}{2}\max \left\{ \begin{matrix} \left[\left|\mathrm{P}(\mu|\mathrm{x}) - \widetilde{\mathrm{P}}(\mu|\mathrm{x})\right| + \delta(\mu,\tilde{\mu})\left|\widetilde{\mathrm{P}}(\tilde{\mu}|\mathrm{x}) - \widetilde{\mathrm{P}}(\mu|\mathrm{x})\right|\right], \\ \left[\left|\widetilde{\mathrm{P}}(\tilde{\mu}|\mathrm{x}) - \mathrm{P}(\tilde{\mu}|\mathrm{x})\right| + \delta(\mu,\tilde{\mu})\left|\mathrm{P}(\mu|\mathrm{x}) - \mathrm{P}(\tilde{\mu}|\mathrm{x})\right|\right] \end{matrix} \right\} \quad (3)$$

where the delta function (δ) is defined as equal to 0 if $\tilde{\mu} = \mu$ and 1 otherwise.

Δ_{max} is defined as the maximum of the basic surprise over the two reference classifier identities, as the notion of surprise is dependent on the reference classifier. It can be observed that the terms with the multiplier δ vanish when the two classifiers favour the same dominant hypothesis. This helps boosting the surprise value when two experts disagree on the favoured hypotheses. With this characteristic of Δ_{max}, it becomes preferable to Δ^* while the advantageous properties of Δ^* such as being confined to interval [0,1] and focusing on the dominant hypotheses by avoiding jitter from non-dominant classes.

In Figure 1, probability density functions (pdf.s) of Δ_{max} are provided for different cases. Figure 1-a depicts the pdf.s for the scenarios where the two experts flag the same dominant hypotheses (cases of label agreement): Distributions for 3 class problems are indicated by dashed lines, and for 6 class problems by the solid curve. Figure 1-b shows distributions for the cases of label disagreement, and Figure 1-c the aggregate distributions for all cases (combination of label agreement and disagreement). Note that the distributions are obtained using uniformly sampled 10^6 different combinations of classifier a posteriori probability outputs, P and $\widetilde{\mathrm{P}}$.

It can be deduced from Equation (3) that the upper limit for surprise is $[\,1 - (\,1/m\,)\,]/2$ for the label agreement case, where m denotes the number of classes. This value increases with m and converges to 0.5 at infinity. The fact that the surprise upper limit and m are directly proportional can be confirmed by comparing the solid and dashed lines in Figure 1-a. As for the case of label disagreement, high values of surprise can be observed to become less likely for larger m.

Using the pdf. information obtained for Δ_{max}, we will experimentally analyse its statistical properties in Section 3, after characterising the associated estimation noise. This will be followed by a discussion and a conclusion based on the findings in Section 4.

3 Experimental Analysis

In this section, we initially characterise the estimation noise that will be affecting the true a posteriori probabilities, P and $\widetilde{\mathrm{P}}$, and hence trigger a cumulative effect on value of the output surprise measure. This is followed by an experimental analysis that aims to investigate the statistical properties of Δ_{max} in the presence of the characterised noise.

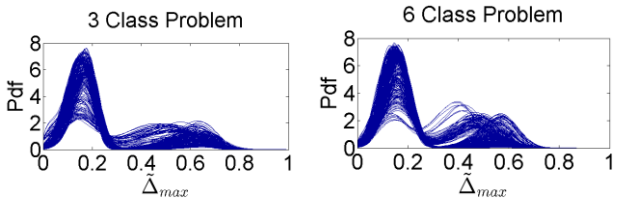
3.1 Characterising the estimation noise

Let us define the estimation errors associated by the true a posteriori probability values, $\mathrm{P}(\omega|\mathrm{x})$ and $\widetilde{\mathrm{P}}(\omega|\mathrm{x})$, as $\eta_\omega(\mathrm{x})$ and $\tilde{\eta}_\omega(\mathrm{x})$ and refer to their probability density functions as $q(\eta)$ and $\tilde{q}(\eta)$, respectively. For simplicity, we will assume that $q(\eta)$ and $\tilde{q}(\eta)$ are normal distributions with zero mean and standard deviation σ while satisfying the conditions

$$\sum_{i=1}^{m} \eta_i(\mathrm{x}) = 0 \quad (4)$$

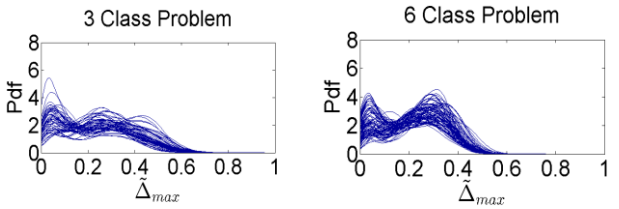
and

$$0 \le \eta_\omega(\mathrm{x}) + \mathrm{P}(\omega|\mathrm{x}) \le 1 \quad (5)$$



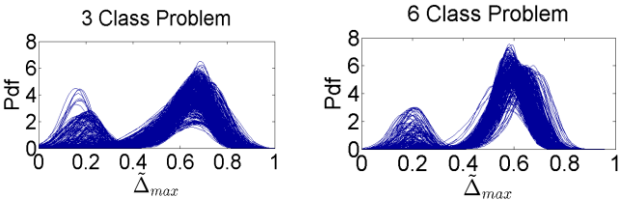
(a) (b)

Figure 3: Pdf. curves of $\tilde{\Delta}_{max}$ for the label agreement case, obtained for $\Delta_{max} = 0.2$, corrupted by noise $p(\eta)$, for 3 class problems (a), and 6 class problems (b)



(a) (b)

Figure 4: Pdf. curves of $\tilde{\Delta}_{max}$ for the label disagreement case, obtained for $\Delta_{max} = 0.2$, corrupted by noise $p(\eta)$, for 3 class problems (a), and 6 class problems (b)



(a) (b)

Figure 5: Pdf. curves of $\tilde{\Delta}_{max}$ for the label disagreement case, obtained for $\Delta_{max} = 0.7$, corrupted by noise $p(\eta)$, for 3 class problems (a), and 6 class problems (b)

Note that the condition in Equation (5) requires the normality assumption for $q(\eta)$ to break down for a posteriori probabilities close to zero or one. Hence, we will assume that the tail of the Gaussian, which is cut off by any of these constraints, flips over by mirror imaging with respect to the line of symmetry placed at the cut-off point. The flipped tail is then added to the existing distribution.

The resulting error distribution, $p(\eta, \mathrm{P})$, which is dependent on the noise-free posterior P, then becomes a folded Gaussian close to the boundaries (cut-off points), and approximates into a standard Gaussian distribution in the middle of the range, such that

$$if\ \mathrm{P} \leq 0.5$$
$$p(\eta, \mathrm{P}) = \begin{cases} 0 & \eta < \mathrm{P} \\ q(\eta) + q(-2\mathrm{P} - \eta) & \eta \geq -\mathrm{P} \end{cases}$$

$$if\ \mathrm{P} > 0.5$$
$$p(\eta, \mathrm{P}) = \begin{cases} 0 & \eta > 1 - \mathrm{P} \\ q(\eta) + q(2 - 2\mathrm{P} - \eta) & \eta \leq 1 - \mathrm{P} \end{cases} \qquad (6)$$

Let us analyse an example boundary scenario to visualize $p(\eta, \mathrm{P})$ for $\mathrm{P} = \mathrm{P}(\omega|\mathrm{x}) = 0.1$ and $q(\eta) = N(0, 0.10)$. In Figure 2-a, $q(\eta)$ is represented by a thin solid line. The thick solid line illustrates $p(\eta, \mathrm{P})$; obtained by folding the tail of q at the cut-off point of $-P = -0.1$, as indicated by the dashed line, and adding it to the unfolded distribution. On the other hand, in Figure 2-b, the thick solid line illustrates the probability density function $r(s(\mathrm{x}))$ of the estimate $\mathrm{s}(\mathrm{x}) = \mathrm{P}(\omega|\mathrm{x}) + \eta_\omega(\mathrm{x})$. It should be remembered that r is obtained as a convolution of the distributions of P and η, that is

$$r(s(\mathrm{x})) = \int_{-\infty}^{\infty} \delta(s(\mathrm{x}) - \mathrm{P} - \lambda)\, p(\lambda, \mathrm{P})\mathrm{d}\lambda \qquad (7)$$

In Figure 2-b, $r(s(\mathrm{x}))$ can be observed to satisfy the condition in Equation (5). Finally, the thin line in Figure 2-b is provided for convenience and depicts what $r(s(\mathrm{x}))$ would look like if the constraint in Equation (5) did not exist.

The probability estimation errors will accumulate estimation errors on any surprise measure. Due to the fact that the proposed surprise measure, Δ_{max}, involves summation over at most two classes, it should be considered as more robust to noise compared to *BS*, which is formulated as a sum involving all classes.

3.2 Statistical analysis of Δ_{max}

In this section, we conduct empirical studies of the effect of the a posteriori class probability estimation errors on the distributions of Δ_{max}, using noise as characterised in Section 3.1

We parameterise scenarios by varying noise-free surprise measure values, and for each choice, study the impact of noise: For a given noise-free surprise measure value, all possible pairs of the probability distributions P and $\tilde{\mathrm{P}}$, which output this value in Equation (3), are recorded. In our experiments, these pairs are selected from a pool of 10^6 combinations for 3 and 6 class problems. The process of selecting probability distribution pairs involves the cases of label agreement and disagreement in the most probable hypothesis. This is so as to account for $\delta = 0$ and $\delta = 1$ as given in Equation (3).

On the selected P and $\tilde{\mathrm{P}}$ pairs, addition of noise is then carried out. Noise terms are drawn from the distribution p as given in Equation (6). Note that in these experiments, σ is set to 0.10. The resulting distributions of noisy Δ_{max} (denoted as $\tilde{\Delta}_{max}$) are obtained from the corrupted P and $\tilde{\mathrm{P}}$.

Using the Δ_{max} distributions provided in Figure 1, a few representative (noise-free) Δ_{max}values have been selected to perform the analysis. In the initial set of experiments, probability distribution functions of $\tilde{\Delta}_{max}$ are obtained for the label agreement case where $\Delta_{max} = 0.2$ and the results are depicted in Figure 3-a and Figure 3-b for 3 and 6 class problems respectively. Similarly, Figure 4 presents the results for the label disagreement case for $\Delta_{max} = 0.2$, and Figure 5 for $\Delta_{max} = 0.7$. Note that the distributions illustrated are approximations acquired from discrete observations.

An interesting observation can be made from these results about the effect of the number of classes, m, on the surprise measure distributions obtained. For both surprise measures and all scenarios of label agreement and disagreement, it is shown that the distributions are less likely to have realisations towards both ends of the [0,1] range for $m = 6$ compared to $m = 3$. However, the amount of shift towards the ends of the range for different m is dependent on specific values of noise-free Δ_{max} values; e.g. for the disagreement case, the difference is high when $\Delta_{max} = 0.7$ whereas it is low for $\Delta_{max} = 0.2$.

It is important to note here that in practice, we will not know the characteristics of the underlying scenarios, i.e. the exact values of $(P, \tilde{P})$ pairs that generate a particular Δ_{max}. A practically more useful analysis, we should integrate over the various scenarios while taking their prior probability of occurrence into account. This analysis can be accomplished by obtaining a plot of the area under the tail of the $\tilde{\Delta}_{max}$ distribution as a function of threshold.

The rationale for this integration can be explained using a simple example. Looking at Figure 3-a, it can be observed that using a threshold of 0.6 for surprise detection can leave an important portion of some distribution curves out and cause false alarms. However, it turns out that the cases with large under-the-tail areas for this threshold are not likely to occur with high probability, e.g. they only happen when the estimation noise causes a label change. In others words, the contribution of these cases to the probability of false alarm is expected to be low, and this can only be evaluated by a cumulative analysis involving the likelihoods of scenario occurrence.

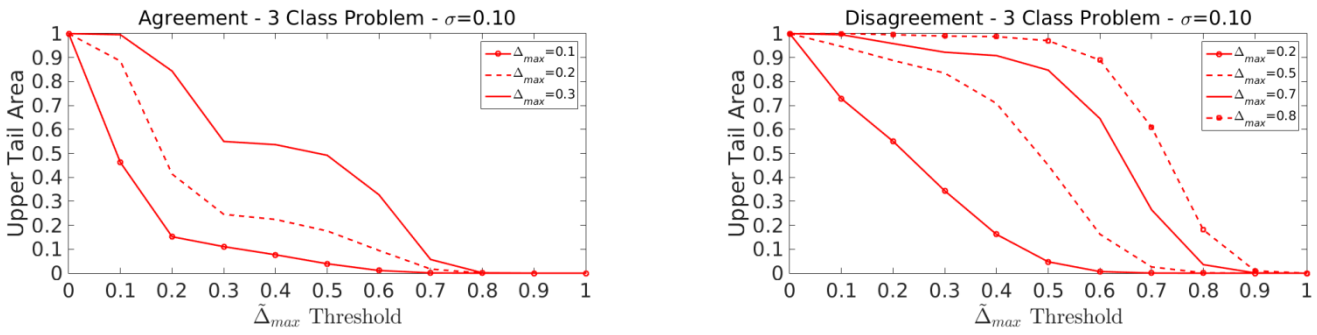
In this set of experiments, the average size of the upper tail area of $\tilde{\Delta}_{max}$ pdf.s obtained from the 10^6 many $(P, \tilde{P})$ pairs is calculated for given threshold points. This is achieved by taking the likelihood of the distributions into consideration. Note that the estimates of the area under the tail are parameterised by noise level.

In Figure 6 and Figure 7, the resulting graphs depicting upper tail area (% over the total area)) versus threshold are given for 3 and 6 class problems respectively. In each figure, the graphs in the first column correspond to the case of label agreement, whereas the second column applies to disagreement. Note that $\sigma = 0.1$ and the results for different fixed noise-free surprise values are shown using different line types.

A comparison of Figure 6-a with Figure 7-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 noise-free Δ_{max} values.

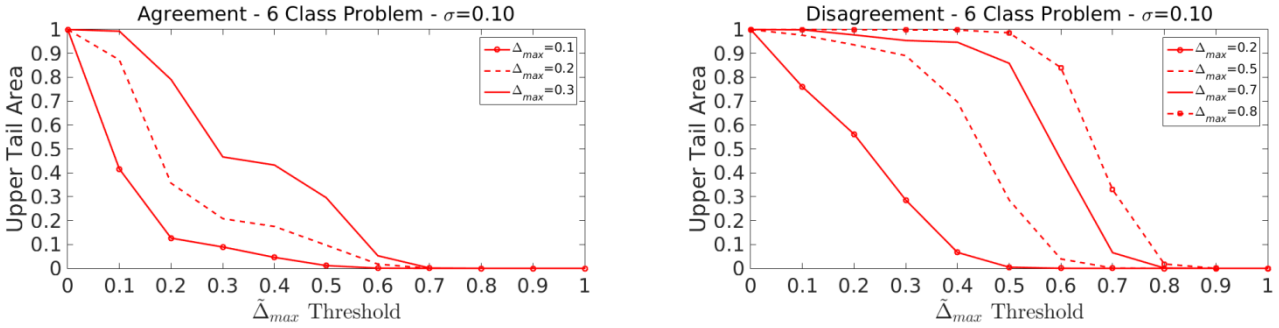
For the case of label disagreement, let us analyse, for instance, the scenario in which noise-free $\Delta_{max} = 0.5$ by comparing Figure 6-b and Figure 7-b. It has been previously shown in Section 2 that the spread of the surprise distribution towards both ends of the [0,1] range is greater for $m = 3$ than for $m = 6$. This characteristic is also reflected in the respective area under the tail curves. For example, for the threshold 0.6, the upper tail area is just under 0.2 for $m = 3$, whereas it is close to zero for $m = 6$.

In Figure 6-a and Figure 7-a, a threshold greater than or equal to 0.7 can be observed to cover more than 90% 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. Looking at Figure 7-b and Figure 7-b to analyse the case of label disagreement, we can see that in order to be able to label the scenarios with noise-free



(a) (b)

Figure 6: Upper tail area size versus $\tilde{\Delta}_{max}$ threshold for different noise-free Δ_{max}. Given for 3 class problems under the scenarios of classifier label agreement (a) and disagreement (b)



(a) (b)

Figure 7: Upper tail area size versus $\tilde{\Delta}_{max}$ threshold for different noise-free Δ_{max}. Given for 6 class problems under the scenarios of classifier label agreement (a) and disagreement (b)

$\Delta_{max} = 0.7, 0.8$ as incongruence with $\sim 90\%$ confidence, a smaller threshold (e.g. a threshold around 0.5) is required. As m decreases, the minimum value of the threshold needed to cover incongruent scenarios resulting from high Δ_{max} also increases. Note that this finding is in line with the observations on the spread of surprise distribution depending on m as given above.

These experiments suggest that it is possible to effectively use the proposed surprise measure for incongruence detection with a choice of a suitable threshold according to the nature of the problem, and shed light on how to select a threshold. The use of a threshold greater than 0.7 is suggested for problems where the true negative rate is of importance, and a smaller threshold closer to 0.5 should be preferred for systems requiring low false positive rates.

4 Conclusions

In this study, the problem of classifier incongruence detection for multiple classifier systems has been addressed. We have pointed out the disadvantages and deficiencies of the existing methods used for measuring classifier incongruence, and proposed the use of a new measure to overcome these problems. An experimental study of the proposed Δ_{max} measure has been conducted to investigate its statistical properties under the presence of estimation noise. The study was carried out for various scenarios defined in terms of the noise-free measure values. The area under-the-tail of the distribution of the Δ_{max} measure parameterised by various thresholds has been calculated to guide the selection of a suitable incongruence detection threshold.

Acknowledgements

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

References

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

[2] 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. On Pattern Analysis and Machine Intelligence*, **35**, pp. 1–4, (2013).

[3] 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*, pp. 1757–1760 (2007).

[4] F. J. Massey. "The Kolmogorov-Smirnov test for goodness of fit". *Journal of the American Statistical Association*, **46(253)**, pp. 68–78, (1951).

[5] H. Cramèr. "On the composition of elementary errors: First paper: Mathematical deductions". *Scandinavian Actuarial Journal*, **1**, pp. 13–74, (1928).

[6] R. von Mises. "Wahrscheinlichkeit, statistik und wahrheit". (1928).

[7] T. W. Anderson, D. A. Darling. "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes". *The Annals of Mathematical Statistics*, **23(2)**, pp. 193–212 (1952).