

Delta divergence: A novel decision cognizant measure of classifier incongruence

Josef Kittler *Life Member, IEEE* and Cemre Zor

Abstract—Disagreement between two classifiers regarding the class membership of an observation in pattern recognition can be indicative of an anomaly and its nuance. As in general classifiers base their decision on class a posteriori probabilities, the most natural approach to detecting classifier incongruence is to use divergence. However, existing divergences are not particularly suitable to gauge classifier incongruence. In this paper, we postulate the properties that a divergence measure should satisfy and propose a novel divergence measure, referred to as *Delta divergence*. In contrast to existing measures, it is decision cognizant. The focus in Delta divergence on the dominant hypotheses has a clutter reducing property, the significance of which grows with increasing number of classes. The proposed measure satisfies other important properties such as symmetry, and independence of classifier confidence. The relationship of the proposed divergence to some baseline measures is demonstrated experimentally, showing its superiority.

Keywords f-divergences, total variation distance, divergence clutter, classifier incongruence

I. INTRODUCTION

Divergence in information theory has been intensively studied and researched over the last six decades. On one hand the massive interest in the subject has been driven by the diversity of applications where divergence plays the key role as an objective function. On the other hand the investigation of the underlying theoretical properties of divergence has motivated the discovery of new measures with tailor made characteristics that are fine tuned for specific applications. This dual drive has produced extensive families of divergences which are encapsulated in the generic expressions presented e.g. in [10], with many specific examples listed in the review paper, e.g. [19]. We shall provide a very brief overview of these developments in Section II and give representative examples in Section III-A.

The key designation of divergences is to measure differences between two probability distributions. These distributions can relate to discrete random variables such as symbols in communication systems, or continuous random variables when comparing, for example, two density functions. The differences can also stem from comparing an empirical distribution of some data, and its parametric model. In decision making applications the two distributions could be a posteriori class probability functions of observations to be classified. The nuances of

these different applications call for divergences of different properties and the existing spectrum of divergence measures bears witness to the endeavors in the field reported over the decades.

In this paper we focus on the use of divergence to measure incongruence of two classifiers. The problem arises in complex decision making systems which often perform sensor data classification tasks using multiple classifiers. Examples of such systems include classifiers processing different modalities of data, ensemble of classifiers aiming to improve classification performance, or hierarchical classification systems where the base classifiers at one level feed their outputs to a contextual classification level. At this decision level the context provided by neighbouring objects is used to improve performance, or derive structural interpretation of the input data. These multiple classifiers voice their opinions about a given set of hypotheses, expressed in terms of a posteriori class probability for each possible outcome.

In decision making systems engaging multiple classifiers, one would normally expect all the classifiers to support the same hypothesis. A classifier disagreement usually signifies something abnormal; a subsystem malfunction, a sensor data modality being absent, or some anomalous event or situation in the observed scene. It is therefore desirable to monitor classifier outputs with the aim of detecting ‘surprising’ classifier incongruence as a trigger for a deeper investigation of its possible causes.

In information theory the magnitude of surprise is intimately linked to the probabilities of the outcome of an experiment. In the decision making context considered in this paper the experiment outcome is the true class membership of a given observation (i.e. finding out which class hypothesis is correct). For outcomes of low probability the surprise is huge, whereas for events that are certain (with probability approaching unity) the surprise is null. The conventional way of measuring the amount of information learnt from an outcome with probability P is using the logarithm of the inverse of P . The information gain from an experiment is then measured by averaging over all the possible outcomes.

In the case of classifier incongruence we are interested in measuring the information gain from an outcome involving two or more classifiers. For the sake of simplicity, in this paper we shall consider two classifiers only. More specifically, we have two random variables representing class identities, with their distributions, and the question is whether the classifiers agree in supporting the various class hypotheses, or disagree. The nature of information gain from an experiment changes to a comparison of the respective probabilities of possible

The authors are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. E-mail: {J.Kittler, C.Zor}@surrey.ac.uk

This work was carried out as part of EPSRC project “Signal processing in a networked battlespace” under contract EP/K014307/1. The EPSRC financial support is gratefully acknowledged.

outcomes. Congruent classifiers would have similar probability distributions over classes, whereas for incongruent the distributions would be different.

Measuring the information gain from an experiment involving two classifiers is different from quantifying the gain from learning the outcome involving a single classifier. What matters in the case of two classifiers is their comparison. Even if the information gain associated with an experiment involving a single classifier is huge, if two classifiers have the same a posteriori class probability distribution, they will be congruent.

A common criterion used for comparison of the distributions of two random variables is divergence. The most popular divergence measure is the Kullback-Leibler (K-L) divergence, referred to by Itti [13] as *Bayesian surprise* measure. It has been used as a measure of classifier incongruence by Weinshall [32], but it is not ideal for a number of reasons:

- 1) If the distributions are different, the value of incongruence will depend on the actual class probability distributions, rather than on probability differences only.
- 2) Measure is asymmetric, i.e. its value depends on which of the two classifier distributions is used as a reference.
- 3) Its values are unbounded, which makes it difficult to set a threshold on congruence.
- 4) In multi class problems the nondominant classes contribute to clutter, which makes the divergence very noisy.

Some of the above drawbacks have been addressed by alternative divergence measures discussed in Section II. The symmetrized K-L divergence recovers the symmetry property. The Jensen-Shanon divergence [20] is both, symmetric, and bounds the range of its values to the interval $[0, 1]$. However, neither of these measures address properties 1 and 4. In search for more suitable candidates one can consider the general family of f -divergences [19]. It includes, the divergences based on the Renyi α -entropies [27], of which the commonly used Shannon entropy - the basis of K-L divergence - is a special case for $\alpha = 1$. Another interesting member is, for instance, the α -entropy, for $\alpha = \infty$, defined entirely by the probability of the most likely hypothesis, which is used for decision making by each classifier. This choice would avoid the problem of clutter in 4, but this particular property migrates to the associated α -divergence family in an undesirable way by focusing on the maximum ratio of a posteriori probabilities, which can emanate from nondominant hypotheses. This can potentially provide a highly misleading information about classifier incongruence.

In this paper we address the problem of measuring classifier incongruence by first introducing the mathematical framework and our baseline - the Kullback-Leibler divergence. This classical information theory divergence is critically assessed in the context of classifier incongruence detection. The critical analysis allows us to identify the properties that a divergence should possess to be able to serve as a measure of classifier incongruence effectively. A brief overview of the options offered by existing tools, and their ability to satisfy the incongruence measure properties identified provides the motivation for a new measure, called *Delta divergence*. Its basis is total variation distance, but we eliminate the clutter

by noting that classifier congruence assessment involves only at most three outcomes of material interest: the two classes predicted by the two classifiers, plus the possibility that the true class is neither of the two. The proposed divergence is a function of the absolute value of the difference of the a posteriori class probabilities estimated by the respective classifiers for the dominant hypotheses. It is shown to exhibit all the required properties, i.e. being bounded, symmetric, decision cognizant, and decision confidence independent. The relationship of the proposed divergence with state-of-the-art classifier incongruence measures highlight its advantages which are also confirmed experimentally by showing the effect of clutter on the Kullback-Leibler divergence, as well as on other baseline measures.

The paper is organised as follows. The related literature is briefly reviewed in Section II. Section III-A introduces the mathematical framework and analyses the properties of KL divergence from the point of view of detecting classifier incongruence. As an outcome of this analysis the properties required by any measure of classifier incongruence are postulated in Section III-B. After a brief discussion of the properties of other existing tools for measuring classifier incongruence a new divergence is proposed in Section III-C and its properties established in Section III-D. The novel, *decision cognizant* formulation of the classifier incongruence detection problem mitigates the clutter generated by nondominant class hypotheses. This is first shown analytically and later demonstrated experimentally in Section III-D. In Section IV we discuss the relationship of the proposed divergence with some baseline criteria as well as with the recently advocated heuristic measures of classifier incongruence. Section V draws the paper to conclusions.

II. RELATED WORK

The introduction of the concept of divergence is attributed to Jeffreys [14] who proposed it as a measure for comparing the likelihood of two competing hypotheses in statistical hypothesis testing. Jeffrey's divergence is defined as the difference between the means of the log likelihood ratio computed respectively under the two hypotheses. However, earlier references to the notion of divergence can be traced back to Mahalanobis [21] in his work on measures for comparing two statistical populations, and Bhattacharyya [2] who proposed to measure the distance between two distributions using the cosine of the angle between the vectors whose components are constituted by the square root of the values of the associated two probability distributions. The Bhattacharyya coefficient is closely related to the Hellinger distance (see e.g. in [23]) which dates as far back as 1909.

In spite of the above credits, the key impetus of the intensive study of the topic over the last six decades was the information theoretic notion of divergence proposed by Kullback and Leibler [18]. Inspired by the seminal work of Shannon [29] on information theory, Kullback and Leibler conceived divergence as the relative gain in information received from an experiment involving two probability distributions relating to the same random variable.

In their original paper the authors define divergence as the mean information for discrimination between two competing hypotheses. They point out a link between divergence and Fisher's information [8], and therefore the relevance of the information theoretic notion of divergence to statistical estimation theory. The paper also establishes basic properties of K-L divergence, including its nonnegativity and the conditions that would need to be satisfied for divergence to exhibit the property of transformation invariance.

One of the factors constraining the use of the K-L divergence involving probability densities is the requirement that the probability distributions are absolutely continuous. To overcome this problem, Lin [20] proposed an alternative, the Jensen-Shannon divergence, which mitigates this problem and renders his measure more generally applicable.

The information theoretic framework inspired immense interest in theoretical properties of K-L divergence, and led to its generalisation using other entropy functions such as α -entropy of Renyi [27], which includes the K-L divergence as a special case. An even broader generalisation was proposed by Csiszár [5] under the name of f-divergences. The family of f-divergences is defined by various choices of convex functions of the likelihood ratio of the respective probability distribution values associated with the alternative hypotheses [23], [19]. The family is included in the class of yet more general divergences known as Bregman divergences, see e.g. [30].

The properties of the numerous divergences have been intensively studied by many authors [5], [26], [19]. The studies investigate divergence measure characteristics such as boundedness, finiteness, additivity for independent observations, behaviour under transformation [25], symmetry, sensitivity to outliers, treatment of inliers, uniqueness, range, behaviour in the case of the two distributions being orthogonal [9], convergence of quantised divergences [12], relationships between divergences and their mutual bounds. For instance some divergence measures are less amenable to analytical simplification, and mutual bounds are useful to compare them with those measures that can be analytically developed for certain types of distributions (e.g. K-L divergence for normal distributions). There is interest in establishing the existence of metric properties, as well as topological and geometric properties. The study of topological and geometric properties of f-divergences by Csiszár in [6], [7] led to the advocacy of perimeter divergences [22] and their generalisation proposed by Österreicher and Vajda [24].

An interesting overview of the properties of f-divergences is presented in [19]. The authors provide elegant derivations of the well known properties based on the Taylor expansion of f-divergences, rather than by resorting to the commonly adopted approach based on Jensen's inequality. The subject of properties of divergence measures continues to generate interest even now, especially in the context of specific applications [28].

In communication systems, divergence is used to measure, for example, communication channel distortion rates and to optimise channel and source coding (see e.g. [28]). Similarly, divergences play a role in optimising the quality of audio and video material compression for storage and archival purposes.

Information-theoretic divergences have application not only in communication systems, but many diverse areas. In statistics, divergence measures have been used for the analysis of contingency tables [11] and for estimating the parameters of model distributions [14], gauging the consistency of observations with a hypothesised probability distribution model [9], and comparing true distributions with their approximations [4], as well as for comparing stochastic processes over time using divergence rates [10].

In the context of statistical decision making, Kailath [15] investigated the relative merits of divergence and Bhattacharyya distance as surrogate criteria for error probability in signal selection for signal detection. In a similar vein, Boeke [3] studied divergence as a criterion for feature selection in pattern recognition and Toussaint [31] advocated its use instead of error probability for pattern classification. The use of divergence instead of classification error probability may have computational advantages. Most of all, the results in the literature are normally applicable to two class pattern recognition problems only, but some of the divergences, such as the Jensen-Shannon divergence [20] support extension to multiclass cases, including error bounds. Bregman divergences have also been used for non supervised pattern classification and for data analysis based on clustering [1].

In this paper, our focus is on the application of divergences for detecting classifier incongruence. Closest to this particular interest is the use of K-L divergence for gauging classifier incongruence by Weinshall et al [32]. They adopted K-L divergence following Itti and Baldi [13] who used it as an objective measure of surprise experienced by subjects reacting to a stimulus induced by the content of a test video. In their experiments divergence was used to compare prior belief captured in terms of a prior distribution, with a new stimulus represented by a posterior distribution. They referred to the K-L divergence in this context as 'Bayesian surprise' measure. Some of the deficiencies of the K-L divergence as a measure of classifier incongruence were addressed by heuristic measures proposed in [16], [17]. In the next section we provide a more principled basis for classifier incongruence detection and develop a novel measure, referred to as Delta divergence, which satisfies the set of desirable properties identified for this specific application.

III. DELTA DIVERGENCE

We start the discussion by introducing the necessary mathematical notation. We then revisit the classical K-L divergence to establish a baseline and to point out some of its deficiencies from the point of view of measuring classifier incongruence. This will allow us to define the notion of classifier incongruence and postulate the properties a divergence measure should possess to support this particular application. We then consider the spectrum of available divergences to identify a suitable candidate and develop it to a novel classifier incongruence measure that is classifier decision cognizant and reflects the specified properties.

A. Baseline

Let us consider a pattern recognition problem where the object or phenomenon to be recognised is represented by a pattern vector $\mathbf{x}$ belonging to one of mutually exclusive classes $\omega_i, i = 1, \dots, m$. Given observation $\mathbf{x}$, we shall denote the aposteriori probability of its membership in class ω_i as $P(\omega_i|\mathbf{x})$. The automatic assignment of pattern vector $\mathbf{x}$ to one of the classes is carried out by a classifier employing an appropriate decision function. Regardless of the type of machine learning solution, we shall assume that the classifier effectively computes the aposteriori class probabilities $P(\omega_i|\mathbf{x}), \forall i$ and engages a Bayesian decision rule to effect the class assignment.

Let us assume that for the same object or phenomenon there is another classifier which is basing its opinion about the object's class membership on its set of aposteriori class probabilities $\tilde{P}(\omega_i|\mathbf{y}), \forall i$, this time based on observation $\mathbf{y}$. The observation could be the same as $\mathbf{x}$ but, in general, $\mathbf{y}$ can be distinct. We are concerned with the problem of measuring the congruence of these two classifiers in supporting the respective hypotheses given the observations $\mathbf{x}$ and $\mathbf{y}$. In essence we have two probability distributions, and the classifiers would be deemed congruent if the two probability distributions agree, and incongruent, if the two probability distributions are different. For the sake of simplicity and notational clarity, in the following we shall focus on a specific instance $\mathbf{x}, \mathbf{y}$ and drop referring to these observations explicitly, using a shorthand notation for the class probabilities as P_i and $\tilde{P}_i$, i.e.

$$P_i = P(\omega_i|\mathbf{x}) \quad \tilde{P}_i = \tilde{P}(\omega_i|\mathbf{y}) \quad \forall i \quad (1)$$

The basic concept in information theory is the notion of self-information. It conveys the amount of information we gain by observing an event ω which occurs with probability $P(\omega)$. If the probability of occurrence is high, i.e. close to one, we learn very little when the event occurs. However, when the probability $P(\omega)$ is low, the amount of information we gain is huge. Accordingly, self-information $I(\omega)$ is defined as

$$I(\omega) = -\log P(\omega) \quad (2)$$

which takes values from the interval $[0, \infty]$. $I(\omega)$ is referred to as 'surprisal', as it quantifies the surprise of seeing a particular outcome.

In general, when an experiment has a number of possible outcomes $\omega_i, i = 1, \dots, m$, the uncertainty associated with the experiment is expressed in terms of the average information gain from observing the outcome. Let P_i be the probability distribution over the events ω_i . The information gain $h(P)$ is defined as

$$h(P) = -\sum_{i=1}^m P_i \log P_i \quad (3)$$

$h(P)$ is known as entropy. It is interesting to note that, as a result of the averaging process, the contribution to entropy made by events with small probability values is low, as

$$\lim_{x \rightarrow 0} x \log x = 0 \quad (4)$$

Rather than measuring the information gained from an experiment, here we are interested in assessing the degree of agreement between two probability distributions P and $\tilde{P}$ estimated over a set of hypotheses $\Omega = \{\omega_i, i = 1, \dots, m\}$ by two different classifiers to gauge whether the classifiers agree in supporting a particular hypothesis or not. This can be achieved by comparing relative uncertainties associated with the two probability distributions P and $\tilde{P}$. A disagreement in their opinion about the identity of an object being classified would be considered surprising. We therefore need a measure of surprise which compares these two distributions. The classical measure suggested for this purpose is the K-L divergence

$$D_K = \sum_i \tilde{P}_i \log \frac{\tilde{P}_i}{P_i} \quad (5)$$

coined Bayesian surprise by Itti [13], and used for measuring classifier incongruence by Weinshall [32].

B. Notion of classifier incongruence

We know that classifiers compute class aposteriori probabilities to make a decision, and that these probabilities must be involved in the definition of classifier incongruence. However, the notion of classifier incongruence is far from self evident. It is not crisply defined as, for instance, classifier error, or a particular shade of colour. If the class probabilities output by two classifiers are similar, then we would agree that the classifiers are congruent. However, by how much can they differ before they cease to be congruent? If incongruence is like 'distance', then the concept is clearly a continuum, rather than a discrete property, and the dichotomy between congruence and incongruence can only be defined by an appropriate threshold. However, what gauge should be used as an incongruence measure?

To answer these questions and to develop a suitable metric, we shall consider the classical K-L divergence (5) in more detail by elaborating a few special cases that should give us insight regarding the essence of congruence/incongruence. First of all, let us start with the simplest case when all the aposteriori class probabilities generated by the two classifiers are identical. In such a scenario the K-L divergence D_K will be zero, flagging the status of congruence of the two decision making experts. Next, let us consider the case when the classifiers agree on the dominant hypothesis, i_{dom} , and support it with identical strength, i.e. $P_{i_{dom}} = \tilde{P}_{i_{dom}}$. Clearly, the contribution to the K-L divergence due to the dominant class would be zero. We would probably all agree that in such situation the classifiers would be congruent. Yet the support for the nondominant hypotheses, which we shall refer to as clutter, given by the two classifiers

$$D_K = \sum_{i, i \neq i_{dom}} \tilde{P}_i \log \frac{\tilde{P}_i}{P_i} \quad (6)$$

could be substantially different from zero, giving potentially a high value to the K-L divergence. It is apparent, that for a given threshold, the K-L divergence may give rise to false rejections of congruent classifier outputs.

As the next scenario, we shall investigate the case when the two classifier disagree on the dominant hypothesis, but support the nondominant hypotheses in an identical way. Denoting the respective dominant hypotheses by i_{dom} and $\tilde{i}_{dom}$, the K-L divergence in this case will be

$$D_K = P_{i_{dom}} \log \frac{P_{i_{dom}}}{\tilde{P}_{i_{dom}}} + P_{\tilde{i}_{dom}} \log \frac{P_{\tilde{i}_{dom}}}{\tilde{P}_{\tilde{i}_{dom}}} \quad (7)$$

Note in (7) that the value of K-L divergence in this ‘zero clutter’ case will depend on the actual dominant class probabilities, reflecting the surprisal value in the relative information gained.

Some of these properties are evident from the scatter plots in Figure 1 and 2 which show values of D_K as a function of the difference between the aposteriori probabilities computed by the two classifiers for the dominant hypothesis μ selected for one of the classifiers. These values are generated by sampling the space of aposteriori class probability distributions P_i and $\tilde{P}_i$, constrained by the choice of P_μ and $\tilde{P}_\mu$. In total one million distributions of P and $\tilde{P}$ have been drawn for a six class and a three class case. We can see that for every choice of the difference, the K-L divergence takes values from zero to infinity, even for $|P_\mu - \tilde{P}_\mu| = 0$. The scatter plot makes it clear that K-L divergence cannot naturally distinguish the state of classifier incongruence from classifier congruence. This is primarily due to the contribution to K-L divergence made by the nondominant hypotheses. We note that the contribution of the nondominant classes becomes more significant for a bigger number of classes as the size of the space of probability distributions increases with m . Thus, with increasing number of classes, the false positive rate, i.e. the probability of congruent classifiers being deemed incongruent by D_K will be magnified. At the other end of the spectrum, even for $|P_\mu - \tilde{P}_\mu| = 1$, which definitely signifies incongruence, D_K can assume value close to zero. Thus D_K has the capacity to flag a lot of false negative incongruences.

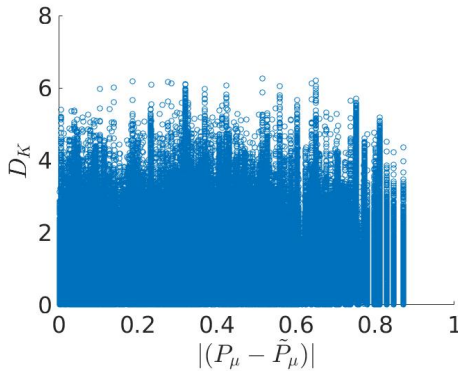


Fig. 1. Scatter plot of Kullback-Leibler divergence comparing the outputs of two classifiers sampled from a population of probability distributions defined over six classes, constrained by a given difference of posterior class probabilities obtained for the dominant class selected by one of the classifiers.

There are a number of conclusions that can be drawn from this analysis. First of all we can see that while ‘perfect’ congruence is independent of the actual values of aposteriori class probabilities of the two distributions, as they are identical,

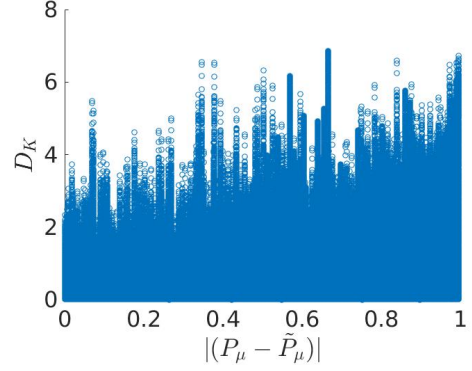


Fig. 2. Scatter plot of Kullback-Leibler divergence comparing the outputs of two classifiers sampled from a population of probability distributions defined over three classes, constrained by a given difference of posterior class probabilities obtained for the dominant class selected by one of the classifiers.

in the case of general congruence and incongruence scenarios, the magnitude of the D_K measure will exhibit strong dependence on the input probability distribution values. The clutter induced by nondominant classes will create ambiguity, that will degrade the separability of notionally congruent and incongruent classifier cases. It should also be noted that the value of K-L divergence will depend on the class probability distributions used as a reference. If we choose $\tilde{P}_i$ instead of P_i , the observed incongruence value will be different. This is not a useful property for applications where the notion is conceptually symmetric. Also the values of the incongruence measure should be confined to a bounded interval to facilitate the setting of a suitable threshold to dichotomise congruence and incongruent cases.

From these observation the following desirable properties of the ideal measure of classifier incongruence are beginning to emerge:

- 1) Overriding focus on dominant hypotheses
- 2) Independence of surprisal content
- 3) Minimum clutter effect
- 4) Symmetry (independence of the choice of distribution as a reference)
- 5) Bounded range of incongruence measure values

Properties 1 and 3 are linked, and suggest that the required measure should concentrate on the dominant hypotheses, and suppress the effect of nondominant classes. Thus the measure we seek should be decision cognizant. Property 2 suggests that classifier incongruence should be a function of differences in aposteriori class probabilities rather than some function of their respective values. The choice of a divergence measure should exhibit symmetry Property 4 and yield values which are bounded, as specified by Property 5. In the following subsection we shall identify a suitable starting point and develop a novel divergence measure which satisfies the above postulated properties.

C. Delta divergence measure

Our aim is to develop a divergence that will have all the above stated properties when used as a classifier incongruence

measure: namely boundedness, symmetry, being clutter free, and ideally also of low sensitivity to probability estimation errors. Heuristic attempts at finding incongruence gauging measures satisfying these properties were presented in [16] and [17]. The key idea in these two papers is to focus on dominant classes as identified by the two classifiers and ignore all the other hypotheses. More specifically, let $\omega = \arg \max_i P_i$ and $\tilde{\omega} = \arg \max_i \tilde{P}_i$. These decision dependent measures are defined in [16] and [17] respectively as

$$\Delta^* = \frac{1}{2} [|P_\omega - \tilde{P}_\omega| + |\tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}}|] \quad (8)$$

and

$$\Delta_{max} = \frac{1}{2} \max\{|P_\omega - \tilde{P}_\omega| + \delta\{\omega, \tilde{\omega}\}|\tilde{P}_{\tilde{\omega}} - \tilde{P}_\omega|, |\tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}}| + \delta\{\omega, \tilde{\omega}\}|P_\omega - P_{\tilde{\omega}}|\} \quad (9)$$

where $\delta\{\omega, \tilde{\omega}\}$ is defined as

$$\delta\{\omega, \tilde{\omega}\} = \begin{cases} 0 & \text{if } \omega = \tilde{\omega} \\ 1 & \text{if } \omega \neq \tilde{\omega} \end{cases} \quad (10)$$

In contrast to these heuristic techniques, our objective is to develop a classifier incongruence measure with a solid theoretical underpinning by demanding that it is a proper divergence. The appropriate toolbox for measuring incongruence between two discrete probability distributions is the family (h, ϕ) of functions

$$h[\sum_i \phi(P_i, \tilde{P}_i)] \quad (11)$$

with $h()$ and $\phi(P_i, \tilde{P}_i)$ being polynomial, logarithmic, polylogarithmic, quasi-polynomial, or quasi-polylogarithmic functions [10], or convex functions [5]. This family includes Bregman divergences [30]

$$D_B = \sum_i [f(P_i) - f(\tilde{P}_i) - (P_i - \tilde{P}_i)f'(\tilde{P}_i)] \quad (12)$$

the Csiszar f-divergences [5] reviewed in [23] and [19]

$$D_C = \sum_i P_i f\left(\frac{\tilde{P}_i}{P_i}\right) \quad (13)$$

and the Renyi divergences [27] parameterised by α

$$D_R = \frac{1}{\alpha - 1} \log\left[\sum_i P_i \left(\frac{\tilde{P}_i}{P_i}\right)^\alpha\right] \quad (14)$$

For an overview the reader is referred to [10].

Armed with the toolbox, the key question of interest to us is which member of the family would exhibit the properties that reflect the notion of classifier incongruence discussed in Section III-B. We already established in Section III-B that the Kullback-Leibler divergence does not. The Jensen-Shannon divergence [20]

$$D_J = \frac{1}{2} \sum_i [P_i \log \frac{2P_i}{P_i + \tilde{P}_i} + \tilde{P}_i \log \frac{2\tilde{P}_i}{P_i + \tilde{P}_i}] \quad (15)$$

confines its values to a bounded interval, and is symmetric. However, the contributions to divergence generated by a difference in probabilities for a particular hypothesis are a function of the probabilities themselves, which does not satisfy Property 2 in Section III-B. Most importantly, all the measures,

including Jensen-Shannon divergence, are affected by the divergence clutter injected by weakly supported hypotheses. This clutter is also likely to aggravate the sensitivity of these divergence measures to noise.

Herein we set to develop an incongruence measure which is a member of the family of divergences in (11). This is the most general family of divergences which has the potential to source the starting point of our development. We start by choosing

$$h(z) = z \quad (16)$$

in (11) and opting for the family of f-divergences in (13). The requirement that our starting point satisfies the property that the contribution to divergence is dependent purely on differences in probabilities, rather than their actual values retains only the *total variation distance* from this family of f-divergences, defined as

$$D_T = \frac{1}{2} \sum_i P_i \left| \frac{\tilde{P}_i}{P_i} - 1 \right| = \frac{1}{2} \sum_i |\tilde{P}_i - P_i| \quad (17)$$

This measure is symmetric and bounded, taking values from the interval $[0, 1]$.

The measure in (17) is still affected by clutter of non-dominant classes. The effect of clutter can significantly be reduced by the following argument: When we compare the outputs of two classifiers, there are only three outcomes of interest: the dominant class ω identified by the classifier with probability distribution P , the dominant class $\tilde{\omega}$ identified by the other classifier, and neither of the two, in other words $\hat{\omega} = \Omega - \omega - \tilde{\omega}$. We thus define a new decision cognizant divergence D_Δ , which we name Delta divergence, as

$$D_\Delta = \frac{1}{2} \left[\sum_{i \in \{\omega, \tilde{\omega}\}} |\tilde{P}_i - P_i| + |\tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}}| \right] \quad (18)$$

Noting that the outcome $\hat{\omega}$ arises with the complement probabilities we can develop Delta divergence, D_Δ , in (18) further by considering the cases when the labels of the dominant classes identified by the two classifiers agree and when they disagree.

1) *Label agreement*: When the labels agree, i.e. $\omega = \tilde{\omega}$, the complement probabilities for the event that the true class is not ω are $1 - \tilde{P}_\omega$ and $1 - P_\omega$. Then the Delta divergence in (18) can be expressed

$$D_\Delta = \frac{1}{2} [| \tilde{P}_\omega - P_\omega | + | 1 - \tilde{P}_\omega - 1 + P_\omega |] = |P_\omega - P_\omega| \quad (19)$$

In other words, the classifier incongruence can be measured simply by comparing the probabilities of the dominant hypothesis output by the two classifiers.

2) *Label disagreement*: When the dominant labels identified by the two classifiers disagree, the probabilities of the event $\hat{\omega}$ that neither of the two dominant classes is the true class are given as

$$\begin{aligned} P_{\hat{\omega}} &= 1 - P_\omega - P_{\tilde{\omega}} \\ \tilde{P}_{\hat{\omega}} &= 1 - \tilde{P}_\omega - \tilde{P}_{\tilde{\omega}} \end{aligned} \quad (20)$$

In this scenario the Delta divergence becomes

$$\begin{aligned} D_{\Delta} &= \frac{1}{2} [|\tilde{P}_{\omega} - P_{\omega}| + |P_{\omega} - \tilde{P}_{\omega}| + \\ &\quad + |P_{\tilde{\omega}} - P_{\tilde{\omega}} + \tilde{P}_{\omega} - P_{\omega}|] \\ &= \frac{1}{2} [|A| + |B| + |A - B|] \end{aligned} \quad (21)$$

Note that the terms A and B can either be both positive, or one of them positive and the other negative. It can be easily shown that it is impossible for both terms to be positive. Consider, for instance, the case $A < 0$, i.e. $\tilde{P}_{\omega} - P_{\omega} < 0$. Then, since $P_{\tilde{\omega}} < P_{\omega}$ (ω being the dominant class for classifier with distribution P), and $\tilde{P}_{\tilde{\omega}} > \tilde{P}_{\omega}$ ($\tilde{\omega}$ being the dominant class for classifier $\tilde{P}$) we have

$$0 < P_{\tilde{\omega}} - \tilde{P}_{\tilde{\omega}} < P_{\omega} - \tilde{P}_{\omega} \quad (22)$$

The positivity of A when B is negative can be shown in the same way.

Now suppose A is negative. Then in $A - B$ in (21) is also negative, and its absolute value is equal $|A + B| = |A| + |B|$. If, on the other hand, B is negative, then $-B$ is positive, and the absolute value of $A - B$ will again equal $|A| + |B|$. Thus when one of the terms, A and B is negative, Delta divergence will be

$$D_{\Delta} = \frac{1}{2} [|A| + |B|] = \frac{1}{2} [|\tilde{P}_{\omega} - P_{\omega}| + |P_{\omega} - \tilde{P}_{\omega}|] \quad (23)$$

When both A and B are positive, the term $A - B$ is either positive, or negative, depending on the relationship of A and B . If $A > B$, then the difference will be positive and we can ignore the absolute value operation, i.e. $|A - B| = A - B$. If $A < B$, then the difference will be negative and $|A - B| = B - A$. Thus we can write for D_{Δ} in (21)

$$D_{\Delta} = \begin{cases} A & \text{if } A \geq B \\ B & \text{if } A < B \end{cases} \quad (24)$$

3) *Delta divergence overview*: Combining the results for these scenarios yields a surprisingly simple divergence measure for gauging classifier incongruence, i.e.

$$D_{\Delta} = \begin{cases} |\tilde{P}_{\omega} - P_{\omega}| & \omega = \tilde{\omega} \\ \max\{|\tilde{P}_{\omega} - P_{\omega}|, |\tilde{P}_{\tilde{\omega}} - P_{\omega}|\} & \omega \neq \tilde{\omega} \quad \begin{cases} A \geq 0, B \geq 0 \\ A < 0, B \geq 0 \\ A \geq 0, B < 0 \end{cases} \\ \frac{1}{2} [|\tilde{P}_{\omega} - P_{\omega}| + |\tilde{P}_{\tilde{\omega}} - P_{\omega}|] & \omega \neq \tilde{\omega} \quad \begin{cases} A < 0, B < 0 \end{cases} \end{cases} \quad (25)$$

In other words, the incongruence measure is defined either by the maximum absolute value difference between the probabilities output by the two classifiers for the respective dominant hypotheses or by the average of these differences. This, in fact, applies even when the dominant hypotheses are the same.

The measure has attractive properties. It is zero, whenever the aposteriori probabilities for the shared dominant class are identical, regardless of the differences in the distribution of the residual probability mass over all the other classes. As it always involves the difference of two probability values it is symmetric. Also, its sensitivity to estimation errors should be very low. It has a monotonic transition between the function values for the label agreement and label disagreement cases. In fact, as we move from the label agreement to the label disagreement case, when another class for the second classifier

begins to assume the dominant role, Delta divergence will continue increasing (potentially by a step change) by virtue of the growing difference between the dominant class probability of the first classifier and the support for this hypothesis voiced by the second classifier.

4) *Two class case*: In the two class case, when the classifiers agree on the dominant hypothesis ω , Delta divergence is given as

$$D_{\Delta} = \frac{1}{2} [|P_{\omega} - \tilde{P}_{\omega}| + |1 - P_{\omega} - 1 + \tilde{P}_{\omega}|] = |P_{\omega} - \tilde{P}_{\omega}| \quad (26)$$

In the label disagreement case, the set of nondominant hypotheses is empty. Hence the Delta divergence has just two terms that are identical to those in (26). Thus the general formula for D_{Δ} in the case of agreement and disagreement is as given in (26).

D. Properties of Delta divergence

In this section we briefly review the properties of Delta divergence and verify that it satisfies the characteristics specified in Section III-B. In addition we shall determine the conditions under which the proposed divergence measure is a metric. This particular property is interesting in the context of assessing incongruence of more than two classifiers.

- **Decision cognizance property**: The Delta divergence proposed in (25) is defined in terms of the aposteriori class probabilities associated with the dominant hypotheses identified by the two classifiers. The measure therefore focuses only on the dominant class hypotheses as required by property 1 in Section III-B.
- **Surprisal independence**: The proposed divergence is defined in terms of differences in aposteriori class probabilities of the dominant hypotheses, rather than their respective values. Thus the value of delta divergence is independent of the base level of these probabilities, and consequently of the surprisal values.
- **Robustness to clutter**: The advantage of Delta divergence over total variation distance can be demonstrated by comparing the contributions of the nondominant hypotheses to these two measures. In the case of Delta divergence the implicit contribution to ‘clutter’ is given by $\frac{1}{2} |P_{\tilde{\omega}} - \tilde{P}_{\tilde{\omega}}|$ where $\tilde{\omega}$ represents the set of nondominant classes. In the case of total variation distance the ‘clutter’ contribution becomes

$$\frac{1}{2} \sum_{i \in \tilde{\omega}} |P_i - \tilde{P}_i| \quad (27)$$

Rearranging the clutter contribution to Delta divergence we have

$$\frac{1}{2} |P_{\tilde{\omega}} - \tilde{P}_{\tilde{\omega}}| = \frac{1}{2} \left| \sum_{i \in \tilde{\omega}} [P_i - \tilde{P}_i] \right| \leq \frac{1}{2} \sum_{i \in \tilde{\omega}} |P_i - \tilde{P}_i| \quad (28)$$

Thus the sensitivity of Delta divergence to clutter is significantly lower than that of total variation distance. It is interesting to note, that if the first two terms in (21) are considered as ‘pure incongruence measure’ (PIM) and

the last term as a group clutter, $D_{\Delta_{clutter}}$, then from (25) we conclude

$$D_{\Delta_{clutter}} = \begin{cases} \frac{1}{2} |\tilde{P}_{\omega} + \tilde{P}_{\omega} - P_{\omega} - P_{\omega}| & \begin{cases} \tilde{P}_{\omega} - P_{\omega} \geq 0 \\ P_{\omega} - \tilde{P}_{\omega} \geq 0 \end{cases} \\ \frac{1}{2} \times PIM & elsewhere \end{cases} \quad (29)$$

This shows that the contributed group clutter is equal to the magnitude of pure incongruence measure in most cases. When the labels of the dominant hypotheses selected by the classifiers disagree, and the difference between the probability for the top ranking hypothesis rendered by the supporting classifier relative to the other classifier is nonnegative, the group clutter equals one half of the difference of the two differences. Alternatively, the clutter is equal to the difference between the support for the union of the two hypotheses. Thus in this particular case the clutter is proportional to the difference between the residual probability masses associated with the non-dominant classes.

The superiority of Delta divergence D_{Δ} over D_T from the clutter point of view is also evident from the experimental results shown in Figure 3 and 4. The figures present the scatter plot of clutter injected in the total variation divergence D_T against the clutter of D_{Δ} , as defined on the right hand side and left hand side of (28) respectively. The values of the clutters are computed by sampling the populations of aposteriori class probability distributions of the two classifiers for three-class and six-class problems as described in Section III-B. It should be noted that the clutter affecting D_T in the three class case is less severe than in the six class case, because the scope for cluttering in the former case is considerably limited. In the two class case it disappears altogether. By the same token, in pattern recognition problems involving a large number of classes, the induced clutter can dominate the value of total variation divergence and make it impossible to detect classifier incongruence reliably.

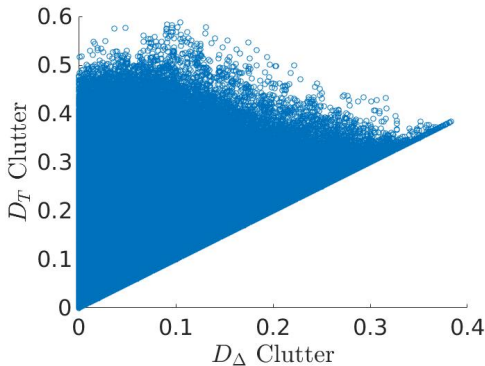


Fig. 3. Scatter plot of D_T clutter against D_{Δ} clutter, affecting the outputs of two classifiers using these two divergences. The values are computed for samples from a population of probability distributions defined over six classes.

- **Bounded range:** Inspecting (25), it is evident that its values satisfy $0 \leq D_{\Delta} \leq 1$. Hence Delta divergence is bounded to interval $[0, 1]$ in compliance with Property 4 of Section III-B.

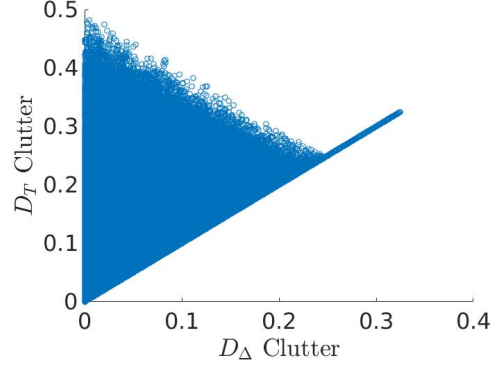


Fig. 4. Scatter plot of D_T clutter against D_{Δ} clutter, affecting the outputs of two classifiers using these two divergences. The values are computed for samples from a population of probability distributions defined over three classes.

- **Symmetry:** As (25) involves only differences of aposteriori class probabilities, D_{Δ} is symmetric in compliance with property 5 of Section III-B.
- **Metric property:** The total variation distance, D_T , from which the proposed divergence has been developed is a metric. This can easily be checked by considering three classifiers A, B, C with probability distributions $P, \tilde{P}$ and $\hat{P}$ respectively. The sum of variation distances D_{AB} and D_{BC} can be written as

$$\begin{aligned} D_{AB} + D_{BC} &= \sum_i [|P_i - \tilde{P}_i| + |\tilde{P}_i - \hat{P}_i|] \geq \\ &\geq \sum_i [|P_i - \tilde{P}_i + \tilde{P}_i - \hat{P}_i|] = \\ &= \sum_i [|P_i - \hat{P}_i|] = D_{AC} \end{aligned} \quad (30)$$

The metric property does not extend to D_{Δ} because of the clutter reducing operation of merging all nondominant hypotheses into a single event, as the resulting sets for the three classifiers can be different. However, in the two class case when the set of nondominant hypotheses is empty, the Delta divergence (18) will degenerate to the total variation distance (17), and the incongruence measure will become a metric.

IV. RELATIONSHIP OF D_{Δ} TO OTHER MEASURES

As we have developed Delta divergence from the total variation distance it is pertinent to elaborate the key differences between these two divergences. The main distinguishing feature of Delta divergence is the way it deals with clutter. Let us denote by Ω^+ the set of dominant hypotheses identified by the two classifiers, which will have a single element for label agreement and two elements for label disagreement. The complement set Ω^- is constituted by all the nondominant hypotheses, i.e. $\Omega^- = \Omega - \Omega^+$ and the probability of one of its members being the true class is $P_{\Omega^-} = \sum_{i \in \Omega^-} P_i$ and $\tilde{P}_{\Omega^-} = \sum_{i \in \Omega^-} \tilde{P}_i$ respectively for the two classifiers. Referring to (17), we can express D_{Δ} as

$$\begin{aligned} D_{\Delta} &= \frac{1}{2} [\sum_{i \in \Omega^+} |P_i - \tilde{P}_i| + |P_{\Omega^-} - \tilde{P}_{\Omega^-}|] \leq \\ &\leq \frac{1}{2} [\sum_{i \in \Omega^+} |P_i - \tilde{P}_i| + \sum_{i \in \Omega^-} |P_i - \tilde{P}_i|] = D_T \end{aligned} \quad (31)$$

Thus $D_{\Delta} \leq D_T$, with equality only for the two class case $m = 2$. Even for $m = 3$ the total variation distance will be

greater than Delta divergence because the set of nondominant hypotheses will contain more than one element in the case of label agreement.

The relationship between these two divergences is shown in Figures 5 and 6 which plot values of D_T against D_Δ . It should be noted that for every value of D_Δ there are many possible values of D_T , as already shown in Section III-D. These have been identified by sampling the probability distributions P and $\tilde{P}$ for a fixed value of D_Δ , as described in Section III-B, to obtain the scatter plots in Figure 5 and 6. It is apparent from the plots that for $m = 6$ the distribution scenarios are much less heavily constrained than for $m = 3$, resulting in much greater differences in the values of D_T and D_Δ . Let us consider any case with $m \geq 3$, and a given threshold, D_{Δ_T} , dichotomising congruence and incongruence. It is apparent that for any $D_\Delta \leq D_{\Delta_T}$ the total variation divergence D_T can take almost any value from the $[0, 1]$ range, which means that due to clutter it will be impossible to discriminate between congruent and incongruent cases.

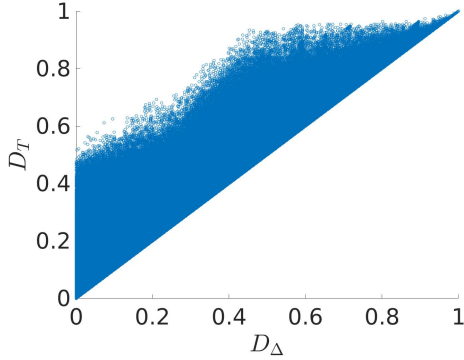


Fig. 5. Scatter plot of total variation divergence D_T against Delta divergence D_Δ , comparing the outputs of two classifiers sampled from a population of probability distributions defined over six classes.

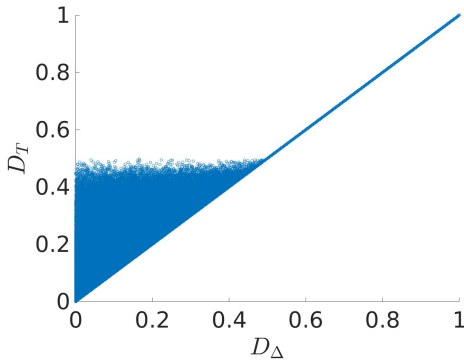


Fig. 6. Scatter plot of total variation divergence D_T against Delta divergence D_Δ , comparing the outputs of two classifiers sampled from a population of probability distributions defined over three classes.

Next we compare Delta divergence with K-L divergence. The same experiment, involving the sampling of the space of probability distributions P and $\tilde{P}$ for fixed values of D_Δ , is conducted and the results plotted in Figures 7 and 8 for a six class and three class problems. We note a number of

observations: First of all the range of values assumed by the K-L divergence is much greater, which would make it difficult to set a suitable threshold between classifier congruence and incongruence. The unbounded range reflects the dependence of K-L divergence on the surprisal values of the additive terms in the expression for the K-L divergence. In the plots the observed values greater than $D_K = 8$ are not shown to avoid undue data compression. We also note that the greater the number of classes, the greater the variation of K-L divergence values caused by the contribution of the clutter of nondominant classes. The clutter is responsible for a significant overlap of K-L divergence values for the classifier congruence and classifier incongruence cases. This can be seen by drawing horizontal lines cutting the scatter plots at different K-L divergence thresholds and noting the resulting distributions (data scatters). For instance setting the threshold to $D_K = 3$ will retain many cases with a high value of D_Δ in the congruent category, leading to underdetection of incongruence. Lowering the threshold to, say, 0.75 will miss many cases with low value of Delta divergence, resulting in a high proportion of false positives.

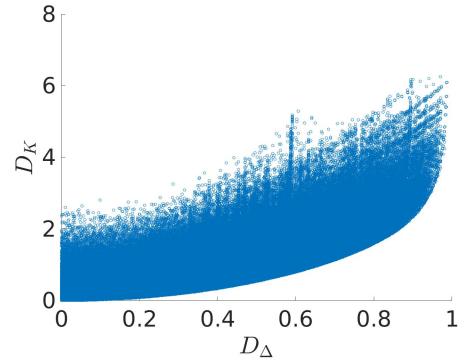


Fig. 7. Scatter plot of Kullback Leibler divergence D_K against Delta divergence D_Δ , comparing the outputs of two classifiers sampled from a population of probability distributions defined over six classes.

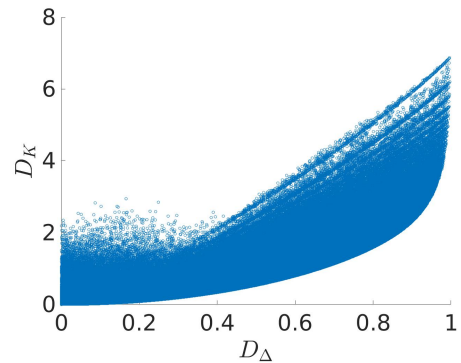


Fig. 8. Scatter plot of Kullback Leibler divergence D_K against Delta divergence D_Δ , comparing the outputs of two classifiers sampled from a population of probability distributions defined over three classes.

It is also interesting to compare Delta divergence with the heuristic measures of classifier incongruence in (8) and (9) which are based on intuition, rather than information theoretic

foundations. It can be easily verified that Δ^* in (8) is exactly the same as Delta divergence in (25) in all cases with the exception of positive values of A and B in label disagreement. In other words we can express D_Δ as

$$D_\Delta = \begin{cases} \max\{|\tilde{P}_\omega - P_\omega|, |\tilde{P}_\omega - P_\omega|\} & \omega \neq \tilde{\omega} \\ D^* & \text{else-where} \end{cases} \quad \begin{cases} \tilde{P}_\omega - P_\omega \geq 0 \\ P_\omega - \tilde{P}_\omega \geq 0 \end{cases} \quad (32)$$

In the case of label agreement also (9) is identical to Delta divergence. In the case of label disagreement, a number of situations may arise. When all the terms in Δ_{max} are positive, we can drop the absolute value operations and after rearrangement Δ_{max} becomes

$$\Delta_{max} = \frac{1}{2} \max\{P_\omega + \tilde{P}_\omega - 2\tilde{P}_\omega, P_\omega + \tilde{P}_\omega - 2P_\omega\} = \frac{1}{2}[P_\omega + \tilde{P}_\omega] - \min\{\tilde{P}_\omega, P_\omega\} \quad (33)$$

When either $P_\omega - \tilde{P}_\omega < 0$ or $\tilde{P}_\omega - P_\omega < 0$ it can easily be shown that

$$\Delta_{max} = \begin{cases} \frac{1}{2}[P_\omega + \tilde{P}_\omega] - P_\omega & P_\omega - \tilde{P}_\omega < 0 \\ \frac{1}{2}[P_\omega + \tilde{P}_\omega] - \tilde{P}_\omega & \tilde{P}_\omega - P_\omega < 0 \end{cases} \quad (34)$$

Thus in contrast to Delta divergence, in the case of label disagreement Δ_{max} picks as incongruence value the average of the probabilities associated with the top hypotheses of the two classifiers minus a term corresponding to one of the probabilities assigned by each classifier to the dominant class identified by the other classifier. The key point to note is that neither of the heuristic measures, Δ^* and Δ_{max} , is a member of the family of divergence functions.

In Figures 9, 10, 11, 12 we present the scatter plots showing the relationship of Δ^* and Δ_{max} to D_Δ for the cases of the three class and six class problems. The first two figures show that Δ^* is never greater than D_Δ . This means that its clutter properties are even better than those of D_Δ . Thus the advantage of D_Δ being a proper divergence has been achieved at the expense of slight deterioration of its sensitivity to clutter.

In the case of the scatter plots for Δ_{max} , it is notable that the variability of this measure is much greater than that of Δ^* . In particular, for a given D_Δ , the values of Δ_{max} can be both greater or lower than D_Δ . Thus Δ_{max} is less crisp, i.e. in its measurement space there is a greater ambiguity (overlap) between the states of classifier congruence and incongruence. In other words there will be a greater propensity for incongruence detection error. Comparing the six class plot with the three class one suggests that this ambiguity grows with the number of classes.

V. CONCLUSIONS

The problem of detecting classifier incongruence was addressed in the paper. It involves comparing the output of two classifiers to gauge the level of agreement in their support for a particular decision. As, in general, the output of a classifier is a probability distribution over the admissible hypotheses, classifier incongruence detection basically involves a comparison of these distributions. The existing classifier incongruence

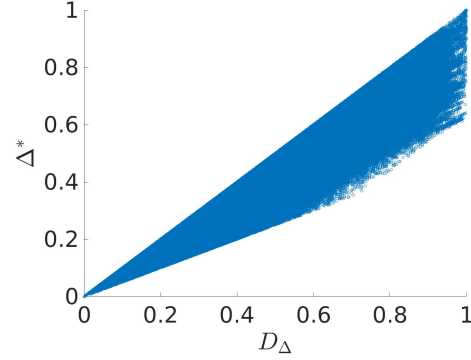


Fig. 9. Scatter plot of Δ^* against Delta divergence D_Δ , comparing the outputs of two classifiers sampled from a population of probability distributions defined over six classes.

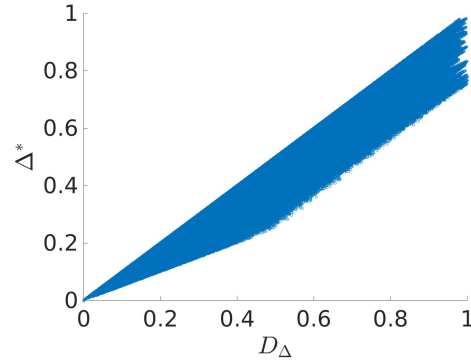


Fig. 10. Scatter plot of Δ^* against Delta divergence D_Δ , comparing the outputs of two classifiers sampled from a population of probability distributions defined over three classes.

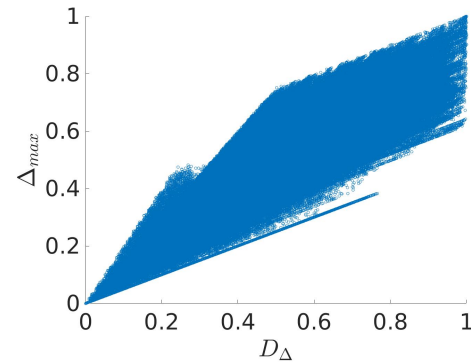


Fig. 11. Scatter plot of Δ_{max} against Delta divergence D_Δ , comparing the outputs of two classifiers sampled from a population of probability distributions defined over six classes.

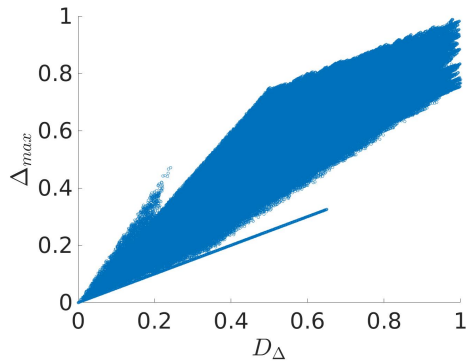


Fig. 12. Scatter plot of Δ_{max} against Delta divergence D_{Δ} , comparing the outputs of two classifiers sampled from a population of probability distributions defined over three classes.

measures advocated in the literature include the Bayesian surprise (K-L divergence) [13], or the Delta measures (Δ^* and Δ_{max}) introduced in [16], [17]. Unfortunately, the former has a number of undesirable properties and the latter two are heuristic.

Measuring differences between two probability distributions is a standard problem in information theory and statistics. The key tool for this purpose is divergence. Many different divergence functions have been proposed in the literature, each exhibiting different properties. In order to adopt or develop a suitable measure for detecting classifier incongruence it is of paramount importance to understand the properties required for this particular application. We argued that a classifier incongruence measure should focus on differences in the classifier support for the dominant hypotheses, be bounded, symmetric, insensitive to surprisal, and insensitive to clutter induced by nondominant hypotheses.

The list of required properties postulated in the paper can be considered as an important contribution in its own right. However, in the context of the paper, this was just a prerequisite for the main task of developing a principled method of measuring classifier incongruence. A review of existing divergences established that none of them fully satisfied the list of requirements. We adopted the total variation divergence as a starting point, because of its insensitivity to surprisal values. We then reformulated the problem of comparing two probability distributions by grouping all the nondominant classes into a single event. This allowed us to develop the total variation measure into a novel divergence, called *Delta divergence*, which is classifier decision cognizant. As a result of this reformulation, the proposed measure is less sensitive to clutter induced by nondominant hypotheses. By studying the characteristics of the proposed measure we demonstrated that it satisfied all the required properties.

Finally, we conducted a number of experiments showing the relationship of the proposed Delta divergence to baseline classifier incongruence measures, and its robustness to clutter. The experiments confirmed its superiority as a measure of classifier incongruence.

REFERENCES

- [1] A Banerjee, S Merugu, I S Dhillon, and J Gosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705 – 1749, 2005.
- [2] A Bhattacharyya. On a measure of divergence between two statistical populations defined by the probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.
- [3] D E Boekee and J C A van der Lubbe. Some aspects of error bounds in feature selection. *Pattern Recognition*, 11(5-6):353 – 360, 1979.
- [4] C K Chow and C N Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Information Theory*, IT-14(3):462–467, 1968.
- [5] I Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229 – 318, 1967.
- [6] I Csiszár. On topological properties of f-divergences. *Studia Scientiarum Mathematicarum Hungarica*, 2:329 – 339, 1967.
- [7] I Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146 –158, 1975.
- [8] R A Fisher. Theory of statistical estimation. *Proc. Cambridge Philosophical Society*, 22:700–725, 1925.
- [9] I Frýdlová, I Vajda, and V Kús. Modified power divergence estimators in normal models - simulation and comparative study. *Kybernetika*, 48(4):795 – 808, 2012.
- [10] V Girardin and L Lhote. Rescaling entropy and divergence rates. *IEEE Trans. Information Theory*, 61(11):5868 – 5882, 2015.
- [11] D V Gokhale and S Kullback. *Information in contingency tables*. Marcel Dekker, New York, 1978.
- [12] P Harremoës and I Vajda. On the Bahadur-efficient testing of uniformity by means of entropy. *IEEE Trans. Information Theory*, 54(1):321 – 331, 2008.
- [13] L Itti and P F Baldi. A principled approach to detecting surprising events in video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 631–637, 2005.
- [14] H Jeffreys. An invariant form for the prior probability in estimating problems. *Proc. Royal Soc., Ser. A*, 186:453–461, 1946.
- [15] T Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. on Communication Technology*, COM-15(1):52 – 60, 1967.
- [16] J Kittler, W Christmas, T de Campos, D Windridge, F Yan, J Illingworth, and M Osman. Domain anomaly detection in machine perception: A system architecture and taxonomy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35:1,14, 2014.
- [17] J Kittler and C Zor. A measure of surprise for incongruence detection. In *Intelligent Signal Processing*. IET, 2015.
- [18] S Kullback and R A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79 – 86, 1951.
- [19] F Liese and I Vajda. On divergences and informations in statistics and information theory. *IEEE Trans. Information Theory*, 52(10):4394 – 4411, 2006.
- [20] J Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Information Theory*, 37:145 151, 1991.
- [21] P C Mahalanobis. On a measure of divergence between two statistical populations defined by the probability distributions. *Journal and Proceedings of Asiatic Society of Bengal*, 26:541 – 588, 1930.
- [22] F "Osterreicher. On a class of perimeter-type distances of probability distributions. *Kybernetika*, 32(4):389–393, 1996.
- [23] F Osterreicher. Csiszár's f-divergence - basic properties, 2002.
- [24] F "Osterreicher and I Vajda. A new class of metric divergences and on probability spaces and its applicability in statistics. *Ann. Inst. Statist. Math.*, 55(3):639–653, 2003.
- [25] M C Pardo and I Vajda. About distances of discrete distributions satisfying the data processing theorem of information theory. *IEEE Trans. Information Theory*, 43(4):1288 – 1293, 1997.
- [26] M C Pardo and I Vajda. On asymptotic properties of information-theoretic divergences. *IEEE Trans. Information Theory*, 49(7):1860 – 1868, 2003.
- [27] A Rényi. On measures of entropy and information. In *4th Berkeley Symp. on Mathematical Statistics and Probability*, pages 547 – 561. University of California Press, 1961.
- [28] I Sason. Tight bounds for symmetric divergence measures and a refined bound for lossless source coding. *IEEE Trans. Information Theory*, 61(2):701 – 707, 2015.
- [29] C E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379 – 423; 623 –656, 1946.

- [30] W Stummer and I Vajda. On Bregman distances and divergences. *IEEE Trans. Information Theory*, 58(3):1277 – 1288, 2012.
- [31] G T Toussaint. Probability of error, expected divergence, and the affinity of several distributions. *IEEE Trans. Systems, Man and Cybernetics*, SMC-8:482 – 485, 1978.
- [32] D Weinshall, A Zweig, H Hermansky, S Kombrink, F W Ohl, J Ane-meuller, J-H Bach, L Van Gool, F Nater, T Pajdla, M Havlena, and M Pavel. Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34:1886–1901, 2012.