A decision cognizant Kullback-Leibler divergence

Moacir Ponti[a,*], Josef Kittler[b], Mateus Riva[a], Teófilo de Campos[b], Cemre Zor[b]

[a] *Institute of Mathematical and Computer Sciences — ICMC, University of São Paulo, São Carlos/SP, 13560-590, Brazil*
[b] *Centre for Vision, Speech and Signal Processing — CVSSP, University of Surrey, Guildford, GU2 7XH, UK*

Abstract

In decision making systems involving multiple classifiers there is the need to assess classifier (in)congruence, that is to gauge the degree of agreement between their outputs. A commonly used measure for this purpose is the Kullback-Leibler (KL) divergence. We propose a variant of the KL divergence, named decision cognizant Kullback-Leibler divergence (DC-KL), to reduce the contribution of the minority classes, which obscure the true degree of classifier incongruence. We investigate the properties of the novel divergence measure analytically and by simulation studies. The proposed measure is demonstrated to be more robust to minority class clutter. Its sensitivity to estimation noise is also shown to be considerably lower than that of the classical KL divergence. These properties render the DC-KL divergence a much better statistic for discriminating between classifier congruence and incongruence in pattern recognition systems.

Keywords Kullback-Leibler divergence, divergence clutter, classifier incongruence

1. Introduction

Decision making systems often benefit from the use of multiple classifiers [1]. As a part of a pattern recognition system, these classifiers can, for example, represent models trained with different sensors, trained with different sets of features, or also created in order to work in different levels of data abstraction [2]. In these scenarios the classifiers are designed to output similar probability estimates when predicting classes for an input. However, when the predictions diverge, we may have classifier incongruence.

Classifier incongruence and its applications have been the subject of studies in the last decade [3, 4, 5]. It may point to the presence of an unexpected event, or an unwanted particularity of one of the classifiers. As such, assessing classifier incongruence may be useful in controlling a classifier fusion process designed to enhance the decision-making system performance, or as an indicator of potential anomaly: incongruent support for a hypothesis provided by different sensor modalities, or by contextual and noncontextual classifiers, or generic and specific classifiers. Thus, there is interest in tools for measuring and detecting classifier incongruence.

Examples of applications include transfer learning from automatic interpretation of videos of tennis singles to tennis doubles, where the failure of the domain models to explain the observed data can be interpreted as a classifier incongruence [4]. In the detection of subcategories of objects in images it is possible to train a general classifier for some category, e.g. motorbike, and then specific classifiers for each known subcategory e.g. cross, road and sport bikes; if there is congruence among the classifiers then the object belongs to a known category; otherwise, a new subcategory is detected [6]. Another example is the out-of-vocabulary word detection scenario [7], in which a phoneme detector may have strong confidence for each observation (phoneme), but the classifier dealing with a whole sequence of phonemes rejects the hypothesis because the word corresponding to the phoneme sequence does not exist in the system vocabulary, indicating a probable out-of-vocabulary word rather than an error [4].

Incongruence may be detected by divergence, which measures the difference between two prob-

*Corresponding author
Email addresses: ponti@usp.br (Moacir Ponti), j.kittler@surrey.ac.uk (Josef Kittler), mateusriva@usp.br (Mateus Riva), t.decampos@st-annes.oxon.org (Teófilo de Campos), c.zor@surrey.ac.uk (Cemre Zor)

ability distributions — in the context of classifiers the aposteriori probability outcomes. A significant range of different divergence measures has been studied and organized [8]. These measures may have properties which make them uniquely suited to the solution of a particular problem, or for use in specific applications. However, with the exception of the work of Weinshall et al. [3] and Kittler et al. [4], interest in this field has not extended far into the study of divergences as a measure of classifier incongruence.

The Kullback-Leibler (KL) divergence [9] is a widely used information theoretic measure of the divergence between two probability distributions. It involves averaging the log ratio of the probabilities in the distribution, and due to its theoretical properties, it has been used in a wide range of pattern recognition fields such as dimensionality reduction [10], feature selection [11] and estimating prior class probabilities on training data [12]. It is shown to have connections to the statistical learning theory when used in the problem of regularized loss functions minimization [13]. Recent studies also use approaches based on the KL divergence in order to detect anomalies or rare events [14, 15]. In the context of classification, we highlight a classifier selection method using KL minimization to aggregate class posterior probabilities [16], a study on the reliability of classifiers outputs [17], and the use of probabilistic kernels for generative/discriminative learning [18].

KL divergence is also the classical tool to detect incongruence between two classifiers [3], each of which compute the posteriori class probabilities to make a decision. It is coined Bayesian surprise by Itti and Baldi [5]. However, the KL divergence treats all class probabilities in the same way. It does not give any special consideration to the dominant hypothesis which are of particular interest in classification scenarios. In multiclass problems, the averaging over the nondominant classes introduces a clutter which can seriously distort the measurement of the intrinsic classifier incongruence as defined by the dominant classes identified by the two classifiers.

We propose a modified version of KL divergence, referred to as decision cognizant Kullback-Leibler (DC-KL) divergence, which attempts to reduce the amount of clutter of the nondominant hypotheses by merging them into a single event. The aim of this paper is to demonstrate the beneficial properties of the new divergence in the context of measuring classifier incongruence. In order to achieve our aim we report a theoretical study of DC-KL, and a series of simulated experiments exploring the relationship between the regular KL and the proposed divergence as well as an experiment to study error sensitivity of both methods. We show both theoretical and empirical evidence that the DC-KL is more reliable than the regular KL, in particular scenarios involving many classes, while also providing a stronger framework for the definition of thresholds for congruence and incongruence, thus facilitating its use in a pattern recognition system. It also displays predictable behaviour when faced with noisy scenarios (such as sensor noise), which makes it better suited for real-world applications.

This paper is organized as follows: in Section 2, we describe the decision cognizant Kullback-Leibler divergence and its theoretical properties, in particular regarding the clutter, i.e. the influence of non-dominant hypothesis probabilities. In Section 3, we report a series of experiments in order to demonstrate the behaviour of the proposed method under different scenarios, including studies on clutter and error sensitivity. Finally, Section 4 is devoted to the conclusions and final remarks.

2. The Decision Cognizant Kullback-Leibler divergence

We shall consider a pattern recognition problem involving k classes in $\Omega = \{\omega_1, \cdots, \omega_k\}$. Based on pattern vectors $\mathbf{x}$ and $\mathbf{y}$, respectively, the classifiers compute the posterior class probabilities $P(\omega_i|\mathbf{x}), \forall i$ and $\tilde{P}(\omega_i|\mathbf{y}), \forall i$ and engage a Bayesian decision rule to effect the class assignment. Note that, $\mathbf{x}$ and $\mathbf{y}$ are vectors representing a given object, even though not necessarily by the same set of features or data source. P and $\tilde{P}$ relates, respectively, to the posterior probabilities of two different models when classifying an object.

We are concerned with the problem of measuring the incongruence of these two classifiers in supporting the respective hypotheses. 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 clarity, in the following discussion we shall drop the reference to specific instances $\mathbf{x}, \mathbf{y}$ and adopt a simplified notation for the class probabilities as P_i and $\tilde{P}_i$, i.e.

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

As discussed in Section 1, we shall be using the Kullback-Leibler divergence as our baseline. The K-L divergence $\tilde{P}_i$ from P_i is defined as:

$$D_K(P||\tilde{P}) = \sum_i \tilde{P}_i \log \frac{\tilde{P}_i}{P_i}. \tag{2}$$

Let $\frac{\tilde{P}_i}{P_i} = u_i$. Then it can alternatively be expressed using the following notation:

$$D_K(P||\tilde{P}) = \sum_i P_i \frac{\tilde{P}_i}{P_i} \log \frac{\tilde{P}_i}{P_i} = \sum_i P_i u_i \log u_i, \tag{3}$$

in which $u \log u$ is a convex function of variable u satisfying $u \geq 0$.

Inspecting Equations 2 and 3, the K-L divergence has the following properties:

1. It is assymetric, i.e: $D_K(P||\tilde{P}) \neq D_K(\tilde{P}||P)$.
2. It is unbounded.
3. It is decision agnostic, that is, the measure aggregates contributions from all the classes, regardless of the decision made by the classifiers.
4. It is nonnegative by virtue of the convexity property, as using Jensen's inequality $D_K(P||\tilde{P})$ can be bounded from below as:

$$\begin{aligned} D_K(P||\tilde{P}) \geq \left[\sum_i P_i u_i\right] \log \left[\sum_i P_i u_i\right] = \\ = \left[\sum_i \tilde{P}_i\right] \log \left[\sum_i \tilde{P}_i\right] = 0. \end{aligned} \tag{4}$$

Whether classifiers agree or disagree is in the first instance determined by their consensus regarding the dominant hypothesis. These are the classes identified by the classifiers as being most probable. Any differences regarding their support for non-dominant hypotheses would be deemed less important. Thus, ideally, we would like to use a measure which deemphasises the contribution of the non-dominant classes, which we refer to as *clutter*.

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 $\bar{\omega} = \{\Omega - \omega - \tilde{\omega}\}$. Let $\tilde{P}_{\bar{\omega}}$ and $P_{\bar{\omega}}$ be the sum of all posterior probabilities in $\bar{\omega}$ for each classifier, respectively. We thus define a new **decision cognizant Kullback-Leibler divergence**, D_D,

$$D_D(P||\tilde{P}) = \sum_{i\epsilon\{\omega,\tilde{\omega}\}} \tilde{P}_i \log \frac{\tilde{P}_i}{P_i} + \tilde{P}_{\bar{\omega}} \log \frac{\tilde{P}_{\bar{\omega}}}{P_{\bar{\omega}}}, \tag{5}$$

which retains the properties 1, 2 and 4 but it is no longer decision agnostic.

2.1. Clutter

The motivation for introducing the decision cognizant divergence is to reduce the contribution to the divergence measure made by the nondominant classes, referred to as clutter. Therefore it is pertinent to investigate the relationship between the clutter of the standard and decision cognizant KL divergences. For brevity, we will be denoting $D_K(P||\tilde{P})$ simply as D_K, and similarly for D_D. The clutter affecting the classical KL divergence is given by

$$D_{K_{\text{clutter}}} = \sum_{i\epsilon\bar{\omega}} \tilde{P}_i \log \frac{\tilde{P}_i}{P_i} \tag{6}$$

whereas the DC-KL clutter is given as

$$D_{D_{\text{clutter}}} = \tilde{P}_{\bar{\omega}} \log \frac{\tilde{P}_{\bar{\omega}}}{P_{\bar{\omega}}} \tag{7}$$

By virtue of the log sum inequality we have:

$$D_{K_{\text{clutter}}} \geq D_{D_{\text{clutter}}} \tag{8}$$

Thus the DC-KL clutter is always lower than the KL divergence clutter.

The difference between the clutters will be particularly accute in common scenarios where the posterior probabilities for non-dominant hypotheses are low, i.e. $P_i \approx 0$ for some $i \in \bar{\omega}$, in which case KL divergence can be dominated by a high term coming from such classes in the clutter, whereas in the decision cognizant form this effect is minimized.

It is also interesting to note that the decision cognizant clutter is a function of $\tilde{P}_{\bar{\omega}} \log \tilde{P}_{\bar{\omega}}$ plus a linear term of $\tilde{P}_{\bar{\omega}}$, which is parameterised by $\log P_{\bar{\omega}}$. Thus, in certain scenarios $D_{D_{\text{clutter}}}$ can assume values approaching infinity. This will occur when the residual probabilities $P_{\bar{\omega}}$ for one of the classifiers approaches zero. Even when two classifiers are congruent, but the relative strengths of their support for the dominant class differ, the clutter can induce misleading results even for the DC-KL divergence. However, for a given $P_{\tilde{\omega}}$ and $\tilde{P}_{\tilde{\omega}}$, the decision cognizant divergence clutter is deterministic. In contrast, classical divergence clutter is a function of the distribution of the constituting elements of $\tilde{P}_{\bar{\omega}}$ and $\tilde{P}_{\bar{\omega}}$, and this further fuzzifies the classifier incongruence measure landscape as chartered by the classical Kullback-Leibler divergence.

By analysing the behaviour of the two clutters in different scenarios we can easily demonstrate that the decision cognizant divergence clutter has superior properties. For instance, by differentiating Equation 7 with respect to $\tilde{P}_{\bar{\omega}}$ we find the condition for the lowest decision cognizant clutter to be $\tilde{P}_{\bar{\omega}} = \frac{P_{\bar{\omega}}}{e}$ (considering the natural logarithm) in which the decision cognizant divergence clutter will be $D_{D_{\text{clutter}}} = -\tilde{P}_{\bar{\omega}}$. Thus the lowest clutter value will vary from zero to minus the residual probability value of one of the classifiers. When the residual probabilities for both classifiers are comparable, $D_{D_{\text{clutter}}}$ will approach zero. Thus there is a spectrum of operating conditions when the clutter corrupting decision cognizant divergence will be low and will not hide the underlying value of classifier (in)congruence. However, even when the decision cognizant divergence clutter is low, the classical divergence clutter can assume values at infinity. This clearly demonstrates the advantageous properties of the decision cognizant divergence.

3. Simulation experiments

In order further to demonstrate the behaviour of the proposed decision cognizant Kullback-Leibler divergence and how it compares with the regular Kullback-Leibler divergence, two sets of simulation experiments are carried out.

First, we study strong/weak agreement/disagreement between two classifiers. In particular we are interested in how the confidence outcomes, i.e. the posterior class distribution of the classifiers, affect each divergence. In this set of simulations we also investigate the relative sensitivity of DC-KL and KL to estimation errors.

Second, we sample the space of posterior class probability distributions P and $\tilde{P}$ in order to produce a broader dataset. Then we compare both divergences in terms of their differences, the clutter and also their respective error sensitivity.

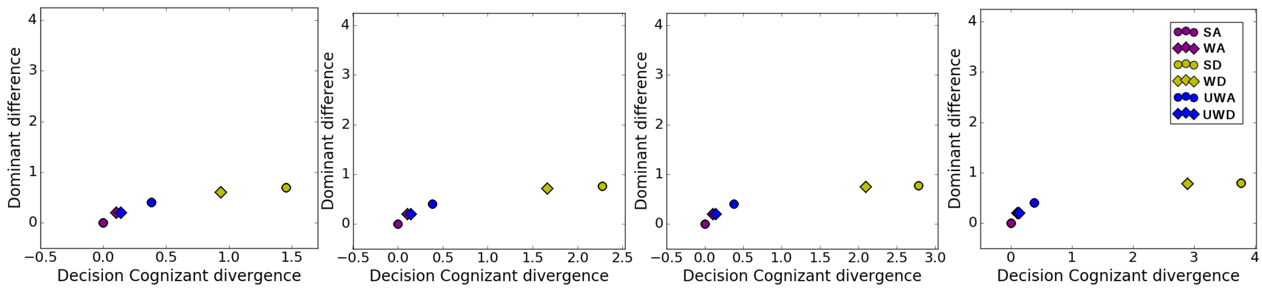
3.1. Case study experiments

We study controlled experiments for a different number of classes $k = \{3, 6, 10, 30\}$ and pairs of posterior probability vectors — one per classifier — with some fixed and arbitrary posterior probabilities for the dominant hypotheses ω and $\tilde{\omega}$. The following cases are investigated:

1. Agreement ($\omega = \tilde{\omega}$):
   – SA (strong agreement) $\tilde{P}_{\omega} = 0.8$, $P_{\omega} = 0.8$;
   – WA (weak agreement) $\tilde{P}_{\omega} = 0.8$, $P_{\omega} = 0.6$;
2. Disagreement ($\omega \neq \tilde{\omega}$) with $\tilde{P}_{\tilde{\omega}}$ fixed with a high probability and making $\tilde{P}_{\omega} = (1 - \tilde{P}_{\tilde{\omega}})/(k-1)$, so that it retains some amount of the remaining probability:
   – SD (strong disagreement) $\tilde{P}_{\tilde{\omega}} = 0.8$, $\tilde{P}_{\omega} = 0.2/(k-1)$ and $P_{\omega} = 0.8$, $P_{\tilde{\omega}} = 0.2/(k-1)$;
   – WD (weak disagreement) $\tilde{P}_{\tilde{\omega}} = 0.8$, $\tilde{P}_{\omega} = 0.2/(k-1)$ and $P_{\omega} = 0.6$, $P_{\tilde{\omega}} = 0.4/(k-1)$;
3. Uncertain scenarios (lower confidences for dominant hypothesis):
   – UWA (uncertain, weak agreement) $\tilde{P}_{\tilde{\omega}} = 0.8$, $P_{\omega} = 0.4$ with $\omega = \tilde{\omega}$;
   – UWD (uncertain, weak disagreement) $\tilde{P}_{\tilde{\omega}} = 0.4$, $\tilde{P}_{\omega} = 0.2$ and $P_{\tilde{\omega}} = 0.2$, $P_{\omega} = 0.4$ with $\omega \neq \tilde{\omega}$.

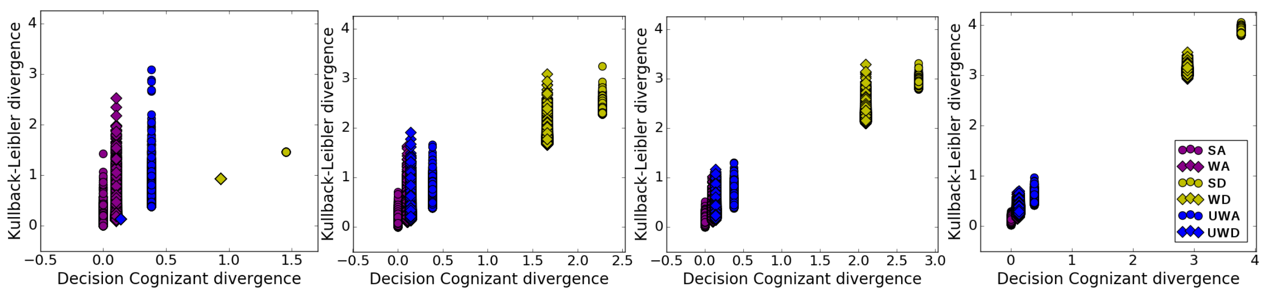
For each item above with fixed probabilities for ω and $\tilde{\omega}$, we produced 1000 probability vectors by randomly drawing values – using a uniform distribution – for the remaining non-dominant classes $\bar{\omega} = \{\Omega - \omega - \tilde{\omega}\}$, and normalizing them in order to assure unity sum. Three types of scatterplots are shown: (i) $D_D \times |P_{\omega} - \tilde{P}_{\omega}|$, which shows in Figure 1 how the decision cognizant divergence behaves regarding differences on a given dominant hypothesis; (ii) $D_D \times D_K$, showing a comparison of the range of divergence values for each scenario in Figure 2; and (iii) $D_D(\text{clutter}) \times D_K(\text{clutter})$, which shows in Figure 3 how the clutter influences each divergence. Note that there are some cases in which DC-KL and KL divergences are similar, but in general those produced by the former suffer from a large variance for a given scenario.

The first interesting result is the log-shaped curve obtained for values from lower to higher divergences, i.e. $D_D \times |P_{\omega} - P_{\tilde{\omega}}|$, in Figure 1, from congruent values (concentrated near zero) to incongruent values (spanning values above 0.3). As expected, the DC-KL was invariant to changes in clutter, while regular KL often showed high variance



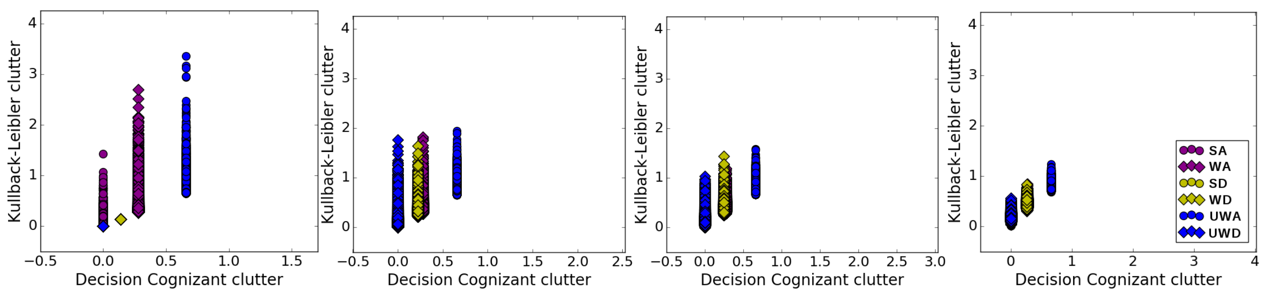
(a) $k = 3$ (b) $k = 6$ (c) $k = 10$ (d) $k = 30$

Figure 1: Scatter plot of the dominant hypothesis differences as a function of D_D for different number of classes: 3 (a), 6 (b), 10 (c) and 30 (d). The points refer to the cases of SA (strong agreement), PA (weak agreement), SD (strong disagreement), PD (weak disagreement), UPA (uncertain, weak agreement) and UPD (uncertain, weak disagreement).



(a) $k = 3$ (b) $k = 6$ (c) $k = 10$ (d) $k = 30$

Figure 2: Scatter plot for D_K as a function of D_D for different number of classes: 3 (a), 6 (b), 10 (c), and 30 (d). In (a), KL and DC-KL are similar for disagreement scenarios and therefore all fall in a single point in the scatter plot. The points refer to the cases of SA (strong agreement), PA (weak agreement), SD (strong disagreement), PD (weak disagreement), UPA (uncertain, weak agreement) and UPD (uncertain, weak disagreement).



(a) $k = 3$ (b) $k = 6$ (c) $k = 10$ (d) $k = 30$

Figure 3: Scatter plot for D_K(clutter) as a function of D_D(clutter) for different number of classes: 3 (a), 6 (b), 10 (c), and 30 (d). The points refer to the cases of SA (strong agreement), PA (weak agreement), SD (strong disagreement), PD (weak disagreement), UPA (uncertain, weak agreement) and UPD (uncertain, weak disagreement).

(see Figure 2). A closer look at how clutter influences the divergence shows that, in general, KL divergence hampers in particular the congruent cases due to its sensitivity to clutter variations.

Sensitivity to estimation error analysis for the case study. In order to study the sensitivity of each measure, Gaussian noise with zero mean and standard deviation $\sigma = 0.05 \cdot (1/\log(k))$ was added to each probability vector 100 times, generating 100 noisy versions and totaling 100,000 probability distributions for each scenario. Note that defining σ according to the number of classes was necessary in order to add a fair amount of noise while keeping the dominant hypothesis still valid. Considering the case-studies as a controlled scenario without noise in the labels, we want to make sure that after adding noise the following should still hold:

$$\arg\max_i P(\omega_i|\mathbf{x}) = \omega, \tag{9}$$

$$\arg\max_i \tilde{P}(\omega_i|\mathbf{y}) = \tilde{\omega}. \tag{10}$$

In order to illustrate how the probabilities are affected by the noise, in Figure 4 we plot lines connecting the class posterior probability distributions after adding noise multiple times as a way of visualizing the effect of noise. Each line represents a noisy instance of the posterior, showing the variance caused by the noise and how it increases uncertainty in the dominant classes.

For each k we compare the expected divergence (the one obtained in the noise-free data) with the estimates under noise by computing a histogram of the divergences on noisy data for: strong agreement (SA), weak agreement (WA), strong disagreement (SD) and weak disagreement (WD). The results of the error sensitivity experiments are shown in Figure 5 for 3 classes, Figure 6 for 6 classes, Figure 7 for 10 classes, and Figure 8 for 30 classes. For $k = 3$, because the divergences are different only by one term, the DC-KL divergence shows its advantages only in SA. The desired properties become clearer for $k > 3$.

An analysis of the results shows the robustness of DC-KL over the regular KL in particular under strong agreement (SA), but also for strong disagreement (SD) and weak disagreement (WD). In WA cases both DC-KL and KL behave similarly. In WD scenarios with $k > 3$, DC-KL is more robust to noise than regular KL, which in $k = 6$ produces lower values, towards congruence, while the actual state is incongruent (see Figure 6). In some disagreement scenarios the decision cognizant divergence can degrade to congruence in the presence of both noise and high uncertainty regarding the dominant hypotesis.

We believe the experimental evidence in the case study favors, overall, the decision cognizant over the regular Kullback-Leibler divergence. In the next section a more complete simulation is performed to analyze the behaviour of both methods.

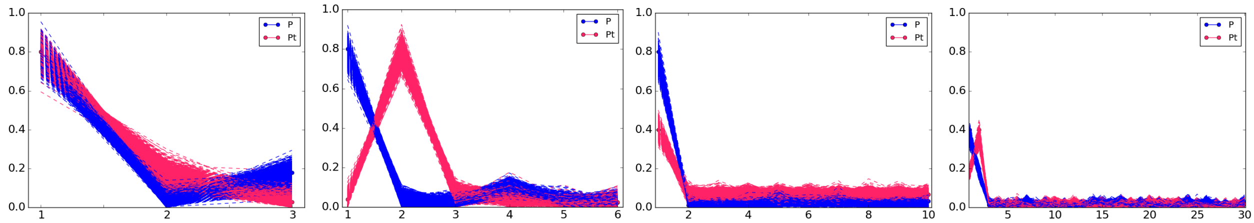
3.2. Experiments sampling over the space of posterior probability distributions

In order to analyse the performance of the DC-KL divergence more thoroughly, an investigation was conducted by sampling the posterior probability distribution space. This simulation can be considered a more complete analysis of the behaviour of the DC-KL divergence measure given different outcomes for the pair of classifiers.

The simulation involved two posterior probability vectors P and $\tilde{P}$ created by fixing the first two class probabilities using values in the range $[0.02, 0.98)$ with a step of 0.02, in order to cover all valid permutations that do not result in a zero probability value for any class. After the first probability (for class ω_1) is chosen, the available values for the second one are sampled in the range of $[0.02, 1.0 - P_{\omega_1})$ with step 0.02. The values for the non-dominant classes were not sampled, but randomly drawn from a uniform distribution, and normalized so that the vector sums up to 1. For each fixed combination, 10 different non-fixed class sets were drawn, so that the effects of randomly generating probabilities could be reflected in the results. Thus, a total of 1.382.976 probability vector pairs were created for the simulation.

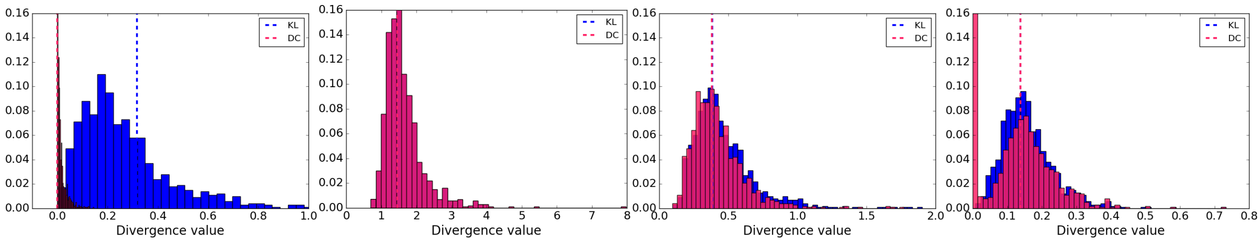
3.2.1. Exploration by sampling the probability space

Similarly to the controlled experiments, the following scatterplots are shown to characterize the divergences over the probability distribution space: (i) $D_D \times D_K$ in Figure 9 and (ii) $D_D(\text{clutter}) \times D_K(\text{clutter})$, which shows in Figure 10 how the clutter influences each divergence. In order to visualize the scatterplots, five scenarios were arbitrarily assigned to colors: strong agreement, when $\omega = \tilde{\omega}$ and $P_\omega, \tilde{P}_{\tilde{\omega}} \geq 60\%$; strong disagreement, when $\omega \neq \tilde{\omega}$ and $P_\omega, \tilde{P}_{\tilde{\omega}} \geq 60\%$; weak agreement,



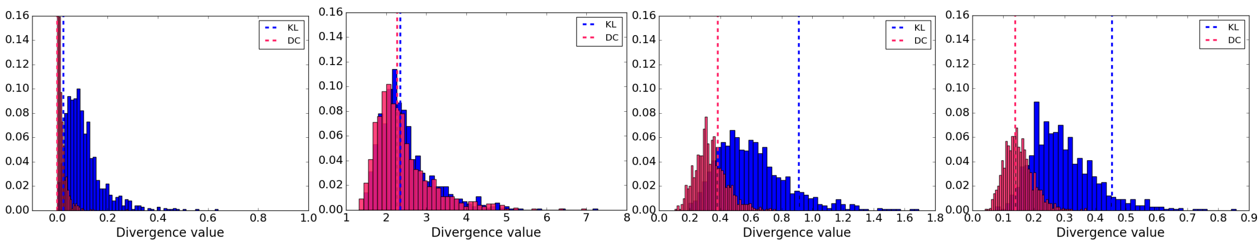
(a) $k = 3, \omega = \tilde{\omega}$ (b) $k = 6, \omega \neq \tilde{\omega}$ (c) $k = 10, \omega = \tilde{\omega}$ (d) $k = 30, \omega \neq \tilde{\omega}$

Figure 4: Examples of noise added to probability distributions – probabilities on the vertical axis and classes on the horizontal axis: (a) strong agreement with 3 classes, (b) strong disagreement with uncertainty involving 6 classes, (c) weak agreement with 10 classes (c) weak disagreement involving 30 classes.



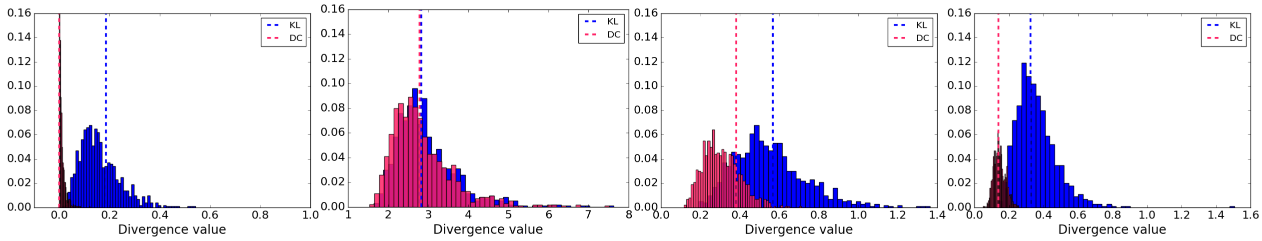
(a) $\omega = \tilde{\omega}$ SA (b) $\omega = \tilde{\omega}$ WA (c) $\omega \neq \tilde{\omega}$ SD (d) $\omega \neq \tilde{\omega}$ WD

Figure 5: Error sensitivity results for 3 classes, showing the histograms of divergences obtained after applying noise: (a) strong agreement – SA, (b) weak agreement – WA, (c) strong disagreement – SD; and (d) weak disagreement – WD. The vertical lines are divergence values computed over noise-free data.



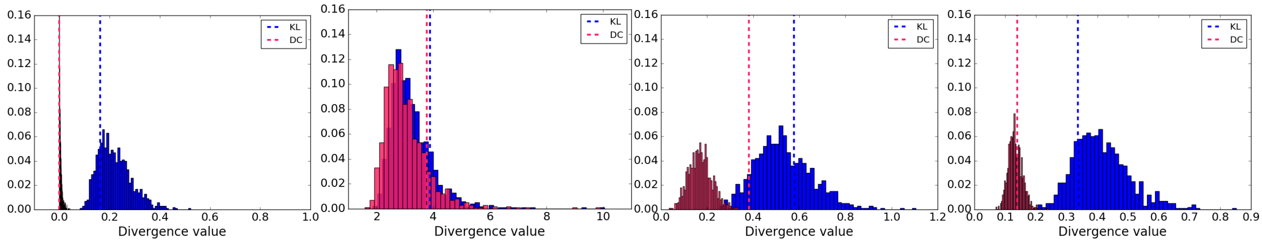
(a) $\omega = \tilde{\omega}$ SA (b) $\omega = \tilde{\omega}$ WA (c) $\omega \neq \tilde{\omega}$ SD (d) $\omega \neq \tilde{\omega}$ WD

Figure 6: Error sensitivity results for 6 classes, showing the histograms of divergences obtained after applying noise: (a) strong agreement – SA, (b) weak agreement – WA, (c) strong disagreement – SD; and (d) weak disagreement – WD. The vertical lines are divergences values computed over noise-free data.



(a) $\omega = \tilde{\omega}$ SA (b) $\omega = \tilde{\omega}$ WA (c) $\omega \neq \tilde{\omega}$ SD (d) $\omega \neq \tilde{\omega}$ WD

Figure 7: Error sensitivity results for 10 classes, showing the histograms of divergences obtained after applying noise: (a) strong agreement – SA, (b) weak agreement – WA, (c) strong disagreement – SD; and (d) weak disagreement – WD. The vertical lines are divergences values computed over noise-free data.



(a) $\omega = \tilde{\omega}$ SA (b) $\omega = \tilde{\omega}$ WA (c) $\omega \neq \tilde{\omega}$ SD (d) $\omega \neq \tilde{\omega}$ WD

Figure 8: Error sensitivity results for 30 classes, showing the histograms of divergences obtained after applying noise: (a) strong agreement – SA, (b) weak agreement – WA, (c) strong disagreement – SD; and (d) weak disagreement – WD. The vertical lines are divergences values computed over noise-free data.

when $\omega = \tilde{\omega}$ and $P_\omega, \tilde{P}_{\tilde{\omega}} \geq 40\%$ but the requirements for strong agreement are not met; weak disagreement, when $\omega \neq \tilde{\omega}$ and $P_\omega, \tilde{P}_{\tilde{\omega}} \geq 40\%$ but the requirements for strong disagreement are not met; and uncertainty, for all remaining combinations of values. These scenarios are meant to be used as a visual guide of easily recognizable scenarios in the plots of Figures 9 and 10.

These results reinforce the findings of the case study, showing that a clear, class-independent threshold for congruence can be established for the DC-KL divergence, for an arbitrarily decided notion of congruence, while the regular KL divergence may output similar values for agreement and disagreement. In the $k = 3$ scenario, it is easy to see that the measures only differ when the classifiers agree on the dominant class, which is a natural conclusion of grouping the clutter together. As the class count increases, the regions previously defined remain within the same range of values for D_D, something that D_K cannot reliably achieve.

Based on these and the case study results for the DC-KL measure, we have established that any $D_D \leq 0.3$ can be considered congruent. The threshold for incongruence, on the same basis, can be established at $D_D \geq 0.7$. Note that defining such thresholds becomes more challenging with the KL divergence, as can be seen in Figure 9, if one draws a horizontal line, cutting the space of possible outcomes for D_K, there is a stronger confusion among the possible scenarios for a given divergence value.

In Figure 10 the results show what was expected: the stronger the effect of the dominant classes, the less clutter present. In some strong agreement scenarios, the value of the clutter alone can go over 1.5 for the regular KL divergence, while the decision cognizant one presents much more reasonable clutter for the same scenarios, never crossing 1.0.

3.2.2. Sensitivity analysis of estimation error

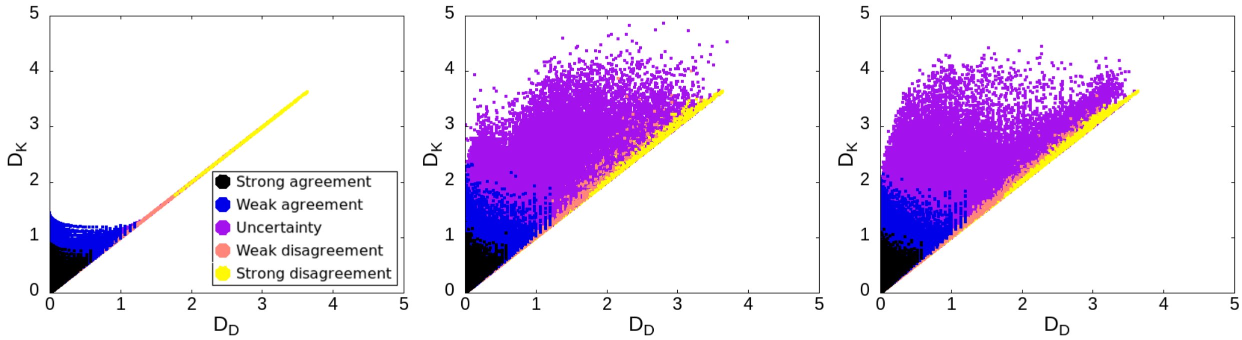
The sensitivity to estimation errors was investigated by choosing all probability vectors whose divergence measure was close to a desired point and adding Gaussian noise with zero mean and $\sigma = 0.05 \cdot (1/\log(k))$ to each of these probability vectors 300 times. Note again that defining the σ according to the number of classes was necessary in order to keep the dominant hypothesis still valid. However, because this dataset – differently from the case studies – spans the whole probability space, we cannot guarantee that the dominant classes of the noisy vectors will always be the same as of the true vector. This effect make it possible to produce incorrect labels when the original estimates are already uncertain.

The error sensitivity results for $k = 3, 10$ and 30 classes are shown in Figure 11 for *congruent* values, sampled around 0.15, which is the mean of the congruent interval $0 \leq D_D \leq 0.3$, in Figure 12 for *uncertain* values (for which the state of congruence or incongruence is unclear), sampled around 0.5, the mean of the interval $0.3 < D_D < 0.7$, and Figure 13 for *incongruent* values, sampled around 1.2, the densest point for $D_D \geq 0.7$.

As the number of classes increases, all histograms display the same effects: they become narrower and their means shift closer to zero. For the 30 class scenario, on Figure 13 (c), it is possible to see that the incongruent sample $D_D = 1.2$ can even cross the threshold into the uncertainty region after the addition of noise, with a tail on the congruent interval. This reflects both the properties of the Kullback-Leibler divergence itself (as it is dependent on the value of the dominant class and may change significantly as the noise affects them) and of our choice of noise generation, which tends to increase uncertainty by shifting up low probability values, while decreasing the probability of dominant hypothesis. In fact, the true cases which tended to produce congruent results had either $P_{\tilde{\omega}}$ or $\tilde{P}_\omega$ close to 5%. Adding noise to these low probability values would have a significant impact on the resulting divergence value.

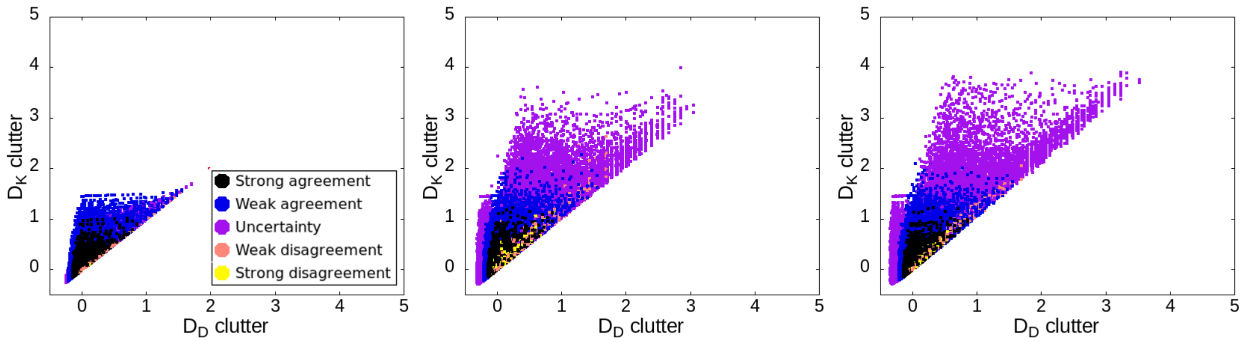
However, it is safe to say that the measure is robust with regards to noise added to a truly congruent probability vector pair. Figure 11 demonstrates that the vast majority of noised samples remain within the defined threshold.

Finally, note that the shift of the mean correlates with regard to the noise and the number of classes. For instance in the 30 class scenario, the mean shifts from 0.15 to 0.1, from 0.5 to 0.4 and from 1.2 to 0.9. In order to study the behaviour of this shift we sampled the distribution shift and fitted a polynomial function $f(x) = a \cdot x^2 + b \cdot x + c \log(x) + d$. We found $a \approx 0$, and with a low least squares fitting error, the following function describes well how a divergence x shifts under noise: $f(x) = 0.63x + 0.07 \log(x) + 0.13$. This indicates that, by having some knowledge about the noise, it is possible to estimate how it would change the divergence outputs, offering a mechanism for compensating for its effect.



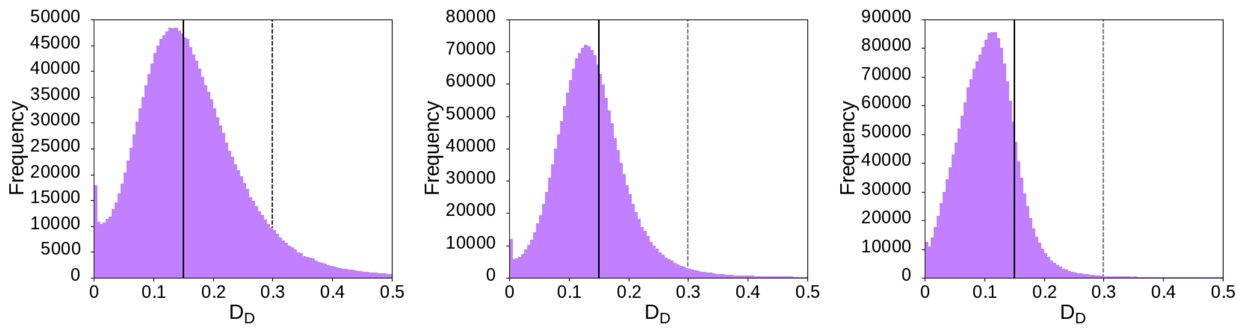
(a) $k = 3$ (b) $k = 10$ (c) $k = 30$

Figure 9: Scatter plot for D_K as a function of D_D for different number of classes: 3 (a), 10 (b) and 30 (c).



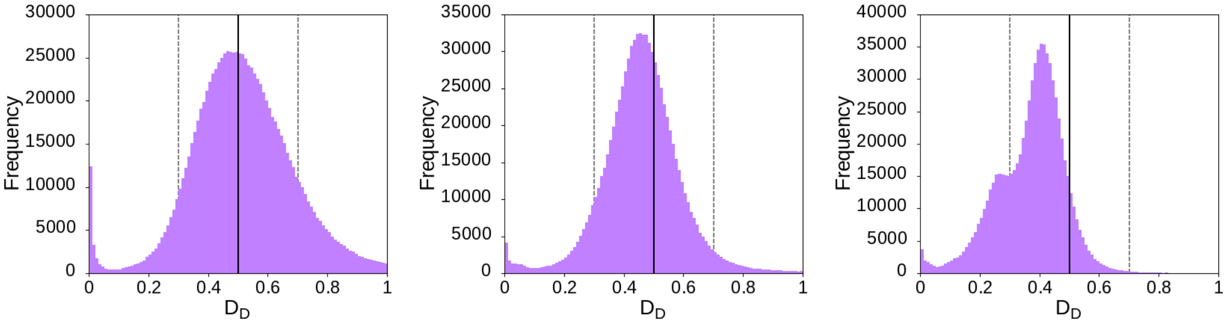
(a) $k = 3$ (b) $k = 10$ (c) $k = 30$

Figure 10: Scatter plot for D_K(clutter) versus D_D(clutter) for different number of classes: 3 (a), 10 (b) and 30 (c).



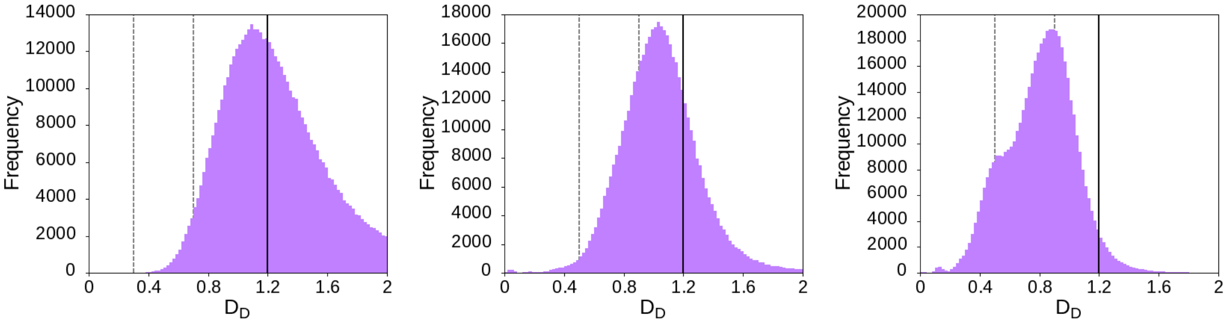
(a)$k = 3$ (b) $k = 10$ (c) $k = 30$

Figure 11: Error sensitivity results: (a) 3 classes, (b) 10 classes, and (c) 30 classes. The vertical dashed line shows the previously defined threshold for congruence (0.3). The vertical solid line is the true divergence value $D_D = 0.15$.



(a)$k = 3$ (b) $k = 10$ (c) $k = 30$

Figure 12: Error sensitivity results: (a) 3 classes, (b) 10 classes, and (c) 30 classes. The vertical dashed lines show the previously defined thresholds for congruence and incongruence (0.3 and 0.7, respectively). The vertical solid line is the true divergence value $D_D = 0.5$.



(a)$k = 3$ (b) $k = 10$ (c) $k = 30$

Figure 13: Error sensitivity results: (a) 3 classes, (b) 10 classes, and (c) 30 classes. The vertical dashed lines show the previously defined thresholds for congruence and incongruence (0.3 and 0.7, respectively). The vertical solid line is the true divergence value $D_D = 1.2$.

3.3. Impact on practical applications and future work

In this paper we focused on theoretical aspects and simulated using the entire posterior probability subspace the conditions under which the proposed measure provides a more principled way to define thresholds for congruence and incongruence. As mentioned in Section 1, there are several applications in which detecting (in)congruence is useful such as domain anomaly detection [4], subclass detection [6] and speech recognition, in particular the out-of-vocabulary word detection [7]. As both the theory and the empirical evidence shows, DC-KL would benefit in particular scenarios with multiple classes and noisy data. Examples of such cases are the use of divergence to assess fusion of multiple classifiers with uncertain estimates due to noisy data [19] and the use of classifier diversity to generate pattern recognition systems that are more robust to noise [20]. The DCKL divergence can also replace the KL divergence when evaluating probability estimates over time [21] with more stability regarding clutter variations.

4. Conclusions

We set out to investigate a measure of divergence which could be better suited for detecting classifier incongruence than the KL divergence, by diminishing the impact of non-dominant classes — or clutter — on the final measure. This is based on the fact that classifiers are designed to output dominant classes. Our decision cognizant measure was shown to behave in a much more predictable and desirable way when compared with the regular KL divergence in this context. In particular the results point to the possibility of establishing much clearer boundaries between congruence and incongruence. Additionally, the DC-KL divergence is capable of detecting partial agreement — when classifiers disagree, while supporting the opposing dominants with relatively high probability values. In contrast, the regular KL often lacks this capability.

One drawback of the decision cognizant KL divergence is its lack of robustness to noise when faced with incongruent cases. This is a characteristic inherited from the regular KL divergence, but in a different shape: the decision cognizant measure tends to estimate values closer to zero, misclassifying incongruent cases, while the regular measure tends to estimate values closer to a specific, non-zero point, misclassifying congruent cases. Care must be taken in the definition of thresholds for congruency and incongruency when faced with a context where noise is a significant issue. We believe that the simulations spanning the probability space provide evidence that DC-KL divergence will be more robust then KL divergence in general, but real applications are still to be investigated. Also, future work can explore the new divergence from the point of view of domain anomaly and classifier diversity.

Acknowledgements

This work was carried out as part of EPSRC project "Signal processing in a networked battlespace" under contract EP/K014307/1 and "FACER2VM" reference EP/N007743/1. The EPSRC financial support is gratefully acknowledged. We also would like to thank FAPESP for the financial support (grants 2015/24652-2 and 2015/13504-2).

References

[1] J. Kittler, M. Hatef, R. P. Duin, J. Matas, On combining classifiers, Pattern Analysis and Machine Intelligence, IEEE Transactions on 20 (3) (1998) 226–239.

[2] M. Ponti Jr, Combining classifiers: from the creation of ensembles to the decision fusion, in: 24th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), IEEE, 2011, pp. 1–10.

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

[4] 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 (2014) 1,14.

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

[6] D. Coppi, T. de Campos, F. Yan, J. Kittler, R. Cucchiara, On detection of novel categories and subcategories of images using incongruence, in: Proceedings of International Conference on Multimedia Retrieval, ICMR '14, ACM, New York, NY, USA, 2014, pp. 337:337–337:344.

[7] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, J. Cernocky, Combination of strongly and weakly constrained recognizers for reliable detection of OOVS, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008, pp. 4081–4084. doi:10.1109/ICASSP.2008.4518551.

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

[9] S. Kullback, R. A. Leibler, On information and sufficiency, The Annals of Mathematical Statistics (1951) 79 – 86.

[10] K. T. Abou-Moustafa, F. D. L. Torre, F. P. Ferrie, Pareto models for discriminative multiclass linear dimensionality reduction, Pattern Recognition 48 (5) (2015) 1863–1877. doi:10.1016/j.patcog.2014.11.008.

[11] J. M. Sotoca, F. Pla, Supervised feature selection by clustering using conditional mutual information-based distances, Pattern Recognition 43 (6) (2010) 2068–2081. doi:10.1016/j.patcog.2009.12.013.

[12] T. F. Li, An efficient algorithm to find the MLE of prior probabilities of a mixture in pattern recognition, Pattern Recognition 29 (2) (1996) 337–339. doi:10.1016/0031-3203(95)00079-8.

[13] J. Honorio, T. Jaakkola, A unified framework for consistency of regularized loss minimizers, in: Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014, pp. 136–144.

[14] W. Wang, B. Zhang, D. Wang, Y. Jiang, S. Qin, L. Xue, Anomaly detection based on probability density function with Kullback–Leibler divergence, Signal Processing 126 (2016) 12 – 17. doi:10.1016/j.sigpro.2016.01.008.

[15] J. Xu, S. Denman, C. Fookes, S. Sridharan, Detecting rare events using Kullback–Leibler divergence: A weakly supervised approach, Expert Systems with Applications 54 (2016) 13 – 28. doi:10.1016/j.eswa.2016.01.035.

[16] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, Dynamic classifier selection for one-vs-one strategy: Avoiding non-competent classifiers, Pattern Recognition 46 (12) (2013) 3412–3424. doi:10.1016/j.patcog.2013.04.018.

[17] J. Barranquero, J. Díez, J. J. del Coz, Quantification-oriented learning based on reliable classifiers, Pattern Recognition 48 (2) (2015) 591–604. doi:10.1016/j.patcog.2014.07.032.

[18] N. Bouguila, Bayesian hybrid generative discriminative learning based on finite liouville mixture models, Pattern Recognition 44 (6) (2011) 1183–1200. doi:10.1016/j.patcog.2010.12.010.

[19] F. Breve, M. Ponti-Junior, N. Mascarenhas, Multilayer perceptron classifier combination for identification of materials on noisy soil science multispectral images, in: XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007), IEEE, 2007, pp. 239–244.

[20] J. A. Sáez, M. Galar, J. Luengo, F. Herrera, Tackling the problem of classification with noisy data using multiple classifier systems: Analysis of the performance and robustness, Information Sciences 247 (2013) 1–20.

[21] S. H. Mallidi, T. Ogawa, H. Hermansky, Uncertainty estimation of DNN classifiers, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2015, pp. 283–288.