Pattern Recognition 77 (2018) 30–44

Contents lists available at ScienceDirect

Pattern Recognition

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

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

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

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

ARTICLE INFO

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

Keywords:
Anomaly detection
Classifier decision incongruence
Bayesian surprise

ABSTRACT

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

1. Introduction

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

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

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

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

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

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

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

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

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

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

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

In summary, the contributions of the paper include:

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

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

2. Related work

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

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

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