Pattern Recognition 81 (2018) 601–614

Contents lists available at ScienceDirect

Pattern Recognition

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

Conditional classifiers and boosted conditional Gaussian mixture model for novelty detection

Reza Mohammadi-Ghazi[a], Youssef M. Marzouk[b], Oral Büyüköztürk[a,*]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
[b] Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

ARTICLE INFO

Article history:
Received 27 June 2017
Revised 4 March 2018
Accepted 23 March 2018
Available online 26 March 2018

Keywords:
Novelty detection
Mixture models
Graphical models
Conditional dependence
Conditional density
Additive modeling
Boosting
False positive

ABSTRACT

Novelty detection is an important task in a variety of applications such as object recognition, defect localization, medical diagnostics, and event detection. The objective of novelty detection is to distinguish one class, for which data are available, from all other possible classes when there is insufficient information to build an explicit model for the latter. The data from the observed class are usually represented in terms of certain features which can be modeled as random variables (RV). An important challenge for novelty detection in multivariate problems is characterizing the statistical dependencies among these RVs. Failure to consider these dependencies may lead to inaccurate predictions, usually in the form of high false positive rates. In this study, we propose conditional classifiers as a new approach for novelty detection that is capable of accounting for statistical dependencies of the relevant RVs without simplifying assumptions. To implement the proposed idea, we use Gaussian mixture models (GMM) along with forward stage-wise additive modeling and boosting methods to learn the conditional densities of RVs that represent our observed data. The resulting model, which is called a boosted conditional GMM, is then used as a basis for classification. To test the performance of the proposed method, we apply it to a realistic application problem for analyzing sensor networks and compare the results with those of alternative schemes.

1. Introduction

Novelty detection is an unsupervised learning task which aims to identify the unknown or inconsistent samples from a set of training data [1,2]. This has variety of applications, especially in healthcare and medical diagnostics [3,4], network security [5,6], image and video processing [7–9], monitoring of industrial and mechanical systems [10–13], and analyzing sensor networks [14].

From the pattern recognition point of view, novelty detection can be viewed as a one-class classification. Important methods in this regard are distance-based approaches such as k-nearest neighbors [15], one-class support vector machines (SVM) [16], graph embedded one-class classifiers [17], neural networks [18], density estimation and clustering [19], random processes [20], and decision tree based techniques such as one-class random forests [21]. For comparative evaluation of various novelty detection approaches, the readers are referred to [22]. These methods are usually most effective when the dependencies of random variables (RV), which represent the observed data, are weak. However, in some applications such as saliency detection in image processing [7], analyzing sensor networks in structural health monitoring (SHM) [11,23,24], and considering a patient's family medical history and genetic tests for medical diagnostics [25], the statistical dependencies of the relevant RVs can be strong. Neglecting dependencies in such applications may considerably affect the inference results in various ways, such as increasing false positive rates.

Previous studies [7,8,23,24] typically used probabilistic graphical models to encode the dependencies of RVs that are used to model the observed data. One main challenge in using these models is learning their structures, which is impossible through standard methods with the available information in novelty detection [26]. Therefore, the application of such models is limited to the problems where the graph structure can be intuitively learned [24] such as saliency detection in image processing [8]. This limitation motivates the objective of this research, which is to develop a novelty detection technique that can account for dependencies of RVs with general, nonintuitive dependence structure.

To address this objective, we introduce conditional classifiers for novelty detection. In doing so, our paper makes the following contributions: (1) We formulate a classification method that can encode arbitrary dependence structure among the RVs in novelty detection problems without any simplifying assumptions;

* Corresponding author.
E-mail addresses: rezamg@mit.edu (R. Mohammadi-Ghazi), ymarz@mit.edu (Y.M. Marzouk), obuyuk@mit.edu (O. Büyüköztürk).

https://doi.org/10.1016/j.patcog.2018.03.022
0031-3203/

Table 1
Notations that are used in this study.

Notation	Definition
n	Number of components in the system
m	Number of training sample points
i	Dummy index $\in \{1, \ldots, n\}$
j	Dummy index $\in \{1, \ldots, m\}$
x_i	Bernoulli RV that can take on values in $\{-1, +1\}$
x_i	A realization of x_i that indicates the state of the i^{th} component of the system
$\mathbf{y}_i$	Real-valued RV to model the feature vectors corresponding to the observations from the i^{th} component
d	Dimensionality of feature vector $\mathbf{y}_i$, i.e., $\mathbf{y}_i \in \mathbb{R}^d$
y_{ij}	The realization of $\mathbf{y}_i$ corresponding to the j^{th} observation
$\mathbf{Y}$	The set of all feature vectors, i.e., $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$
$\mathbf{Y}$	A realization of $\mathbf{Y}$
$p^{(b)}_{\mathbf{Y}}$	Baseline distribution of $\mathbf{Y}$
$\mathbf{Y}^{(b)}_j$	j^{th} sample that is drawn from $p^{(b)}_{\mathbf{Y}}$
$\mathbf{S}_{i,\mathbf{Y}}$	A set that contains the feature vectors corresponding to the system's intact components, except the i^{th} one, according to the observation set $\mathbf{Y}$
$p^{(b)}_{\mathbf{y}_i\|\mathbf{S}_{i,\mathbf{Y}}}$	Baseline conditional distribution of features from the observations of the i^{th} component given the observations of the intact components
p^*_α	The corresponding likelihood threshold of the high density region of $p^{(b)}_{\mathbf{y}_i\|\mathbf{S}_{i,\mathbf{Y}}}$ for significance level α

(2) We develop an implementation of the proposed technique, named boosted conditional Gaussian mixture model (BC-GMM); (3) We experimentally evaluate the proposed method in an SHM application problem with various sensor network configurations and damage scenarios, and compare the results with other techniques.

This paper is structured as follows: First, Section 3 describes the problem and our research approach. Then, the proposed method of conditional classifiers is explained in Section 4 followed by presenting the implementation of the method, BC-GMM, in Section 5 and a discussion on the algorithm's convergence and computational complexity in Section 6. The result of the experimental evaluation of the proposed method and its comparison with other novelty detection algorithms are presented in Section 7. Finally, we conclude by a summary of our findings and a discussion on future research directions.

2. Notations and terminologies

Throughout this paper, RVs are denoted by sans-serif fonts (e.g., x) and deterministic quantities such as realizations of RVs are denoted using serif fonts (e.g., x). For vectors and matrices we use bold lowercase letters and bold uppercase letters, respectively. For example, $\mathbf{y}$ is a random vector and $\mathbf{A}$ is a deterministic matrix.

To exclude particular entries from a set we use "\" followed by another set that contains the entries to be excluded. As an example, consider $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\}$ and let $S = \{\mathbf{z}_2\}$; then, $\mathbf{Z} \setminus S = \{\mathbf{z}_1, \mathbf{z}_3\}$.

Probability density functions are denoted by p with a subscript denoting the RV; e.g., p_{y} is the probability density of the RV y. The probability of an event e is denoted by $\mathbb{P}(e)$.

Finally, note that in this paper, we use *normal* as the opposite of *novel*, e.g., a novel sample point with respect to a baseline distribution is the one that is NOT normal with respect to that distribution. For Gaussian distributions we always use the term *Gaussian*. Table 1 summarizes the essential notation we use in this paper. Some of the notation in this table is associated with a system that has multiple components and certain properties. For more information about such systems, readers are referred to Section 4.

3. Problem description and research approach

The main question that novelty detection aims to answer is as follows. Given m independent realizations of a RV $\mathbf{y}$ from an unknown distribution with density $p_{\mathbf{y}}$, is a new input $\mathbf{y}$ drawn from the same distribution [27]?

Typically, this problem is solved by approximating $p_{\mathbf{y}}$, or the range of $\mathbf{y}$ (i.e., the support $\text{supp}(p_{\mathbf{y}})$ of the probability density), via methods such as density estimation and SVM [28]. These approaches may be well suited to univariate problems; however, they may result in high false positive rates in multivariate cases due to the statistical dependencies among RVs [7]. To clarify this challenge, consider a SHM problem where a mechanical system is instrumented by n sensors in order to detect the occurrence of possible damages. Let $\mathbf{y}_i$ be the feature vector extracted from the structural response measured by the i^{th} sensor, $i \in \{1, \ldots, n\}$. Also assume that for each $\mathbf{y}_i$, there exists a baseline distribution that represents the expected behavior of the intact structure at that sensor location. Then, the monitoring objective is to continuously analyze new measurements $\mathbf{y}_i$ so that, if damage occurs, we can detect and localize it, based on sensor locations where the responses deviate from their baseline state. Due to the connectivity among various sensor locations, via structural elements of the mechanical system, responses at these locations are dependent. Therefore, the occurrence of damage may affect responses at sensor locations within the close neighborhood of the damage *and* at other sensor locations which are further away. Damage localization may thus be blurred [24]. Various other applications, such as saliency detection in image processing, suffer from the same issue and solutions require the dependencies among RVs to be taken into account when making inferences [7,8].

Our proposed approach for considering the statistical dependencies of RVs is to use their conditional densities instead of their marginals, i.e., to use the high density region of $\text{supp}(p_{\mathbf{y}_i|S_i})$ instead of $\text{supp}(p_{\mathbf{y}_i})$, where S_i is a subset of $\{\mathbf{y}_1, \ldots, \mathbf{y}_{i-1}, \mathbf{y}_{i+1}, \mathbf{y}_n\}$. A rationale for this proposition relies on the possibility of reducing the uncertainty of a RV by conditioning it on other RVs. This can be understood by using the notion of entropy, $H(\cdot)$, as a measure of uncertainty. It is known that the conditional entropy of an RV is always less than or equal to its marginal entropy, i.e., $H(\mathbf{y}_i|S_i) \leq H(\mathbf{y}_i)$ [29–31]. Therefore, if S_i is appropriately chosen, we expect the performance of the proposed approach to be at least as good as the univariate approach that considers the marginal densities of RVs as the basis for making inference. The details of the proposed method and how to choose S_i are explained in the next two sections.

4. Conditional Classifiers

Consider a system with n components where each can take two possibles states, namely *normal* and *novel*. Let x_i, $i \in \{1, \ldots, n\}$ be a Bernoulli RV that can take on values in $\{-1, +1\}$ such that $x_i = +1$ and -1 indicate that the i^{th} component of the system is normal and novel, respectively. Let $\mathbf{y}_i \in \mathbb{R}^d$ be a feature vector from some observation of the i^{th} component of the system. Consider the two