

Integrated Clustering and Anomaly Detection (INCAD) for Streaming Data

Sreelekha Guggilam¹, Syed M. A. Zaidi², Varun Chandola^{1,2}, and Abani K. Patra¹

¹ Computational Data Science & Eng.,

² Computer Science & Eng., University at Buffalo, State University of New York (SUNY)
 {sreelekh,szaidi2,chandola,abani}@buffalo.edu

Abstract. Most current clustering based anomaly detection methods use scoring schema and thresholds to classify anomalies. These methods are often tailored to target specific data sets with “known” number of clusters. The paper provides a streaming clustering and anomaly detection algorithm that does not require strict arbitrary thresholds on the anomaly scores or knowledge of the number of clusters while performing probabilistic anomaly detection and clustering simultaneously. This ensures that the cluster formation is not impacted by the presence of anomalous data, thereby leading to more reliable definition of “*normal vs abnormal*” behavior. The motivations behind developing the INCAD model [17] and the path that leads to the streaming model is discussed.

Keywords: Anomaly detection, Bayesian non-parametric models, Extreme value theory, clustering based anomaly detection

1 Introduction

Anomaly detection heavily depends on the definitions of expected and anomalous behaviors [14, 20, 22]. In most real systems, observed system behavior typically forms natural clusters whereas anomalous behavior either forms a small cluster or is weakly associated with the natural clusters. Under such assumptions, clustering based anomaly detection methods form a natural choice [7, 10, 18] but have several limitations.

Firstly, clustering based methods usually require baseline assumptions that are often conjectures and generalizing them is not always trivial. This leads to inaccurate choices for model parameters such as the number of clusters or the thresholds that are required to classify anomalies. Score based models have thresholds that are often based on data/user preference. Such assumptions result in models that are susceptible to modeler’s bias and possible over-fitting.

Secondly, setting the number of clusters has additional challenges when dealing with streaming data, where new behavior could emerge and form new clusters. Non-stationarity is inherent as data evolves over time. Moreover, the data distribution of a stream changes over time due to changes in environment, trends or other unforeseen factors [13, 21]. This leads to a phenomenon called *concept drift*, due to which an anomaly detection algorithm cannot assume any fixed distribution for data streams. Thus, there arises a need for a definition of an anomaly that is dynamically adapted.

Thirdly, when anomaly detection is performed post clustering [2, 12], the presence of anomalies gives a skewed (usually slight) definition of traditional/normal behavior. However, since the existence of anomalies impacts the clustering as well as the definition of the ‘*normal*’³ behavior, it

³ Non-anomalous behavior is described as “normal” behavior. Should not be confused with Gaussian/Normal distribution.

seems counter-intuitive to classify anomalies based on such definitions⁴. To avoid this, simultaneous clustering and anomaly detection needs to be performed.

Table 1. Comparison with other anomaly detection methods

	Neural Networks	LOF	KNN	Kmeans – Kernel Function Based	Gaussian Model Based	INCAD
Clustering Based	✗	✗	✓	✓	✗	✓
Multi-dimension	✓	✗	✓	✓	✓	✓
Unsupervised	✗	✓	✓	✓	✓	✓
Non-parametric	✗	✗	✗	✗	✓	✓
Adaptable to streaming settings	✗	✗	✗	✗	✗	✓
Adaptive thresholds	✗	✗	✗	✗	✗	✓
Probabilistic scoring	✗	✗	✗	✗	✓	✓

In addition to the above challenges, extending these assumptions to the streaming context leads to a whole new set of challenges. Many supervised [8, 16] and unsupervised anomaly detection techniques [4, 8, 15, 18] are offline learning methods that require the full data set in advance for data mining which makes them unsuitable for real-time streaming data. Although supervised anomaly detection techniques may be effective in yielding good results, they are typically unsuitable for anomaly detection in streaming data [16]. We propose a method called Integrated Clustering and Anomaly Detection (INCAD), that couples Bayesian non-parametric modeling and extreme value theory to simultaneously perform clustering and anomaly detection. Table 1 summarizes the properties of INCAD vs other strategies for anomaly detection. The primary contributions of the paper are as follows:

1. **Generalized anomaly definition with adaptive interpretation** The model definition of an anomaly has dynamic interpretation allowing anomalous behaviors to evolve into normal behaviors and vice versa. This definition not only evolves the number of clusters with an incoming stream of data (using non-parametric mixture models) but also helps evolve the classification of anomalies.
2. **Combination of Bayesian non-parametric models and extreme value theory (EVT)** The novelty of the INCAD approach lies in blending extreme value theory and Bayesian non-parametric models. Non-parametric mixture models [19], such as *Dirichlet Process Mixture Models* (DPMM) [3, 25, 28], allow the number of components to vary and evolve during inference. While there has been limited work that has explored DPMM for the task of anomaly detection [26, 29], they have not been shown to operate in a streaming mode or ignore online updates to the DPMM model. On the other hand, EVT gives the probability of a point being anomalous which has a more universal interpretation, in contrast to the scoring schema with user-defined thresholds. Although EVT’s definition of anomalies is more adaptable for streaming data sets [1, 11, 27], fitting an extreme value distribution (EVD) on a mixture of distributions or even multivariate distributions is challenging. This novel combination brings out the much-needed aspects in both the models.
3. **Extension to streaming settings** The model is non-exchangeable which is well suited to capture the effect of the order of data input and utilize this dependency to develop streaming adaptation.

⁴ Clustering and defining “normal/traditional” behavior in presence of anomalies develop in skewed and inconsistent results.

4. **Ability to handle complex data generative models** The model can be generalized to multivariate distributions and complex mixture models.

2 Motivation

2.1 Assumptions on *Anomalous* Behavior

One of the key drivers in developing any model are the model assumptions. For INCAD model, we assume that the data has multiple “*normal*” as well as “*anomalous*” behaviors. These behaviors are dynamic with a tendency to evolve from “*anomalous*” to “*normal*” and vice versa. Each such behavior (normal/anomalous) forms a sub-population that can be represented using a cluster. These clusters are assumed to be generated from a family of distributions whose cluster proportions and cluster parameters are generated from a non-parametric distribution.

There are two distinct differences between normal and anomalous data that must be identified: (a) “*Anomalous*” instances are different from tail instances of “*normal*” behavior and need to be distinguished from them. They are assumed to be generated from distributions that are different from “*normal*” data. (b) The distributions for anomalous data result in relatively fewer instances in the observed set.

As mentioned earlier, clustering based anomaly detection methods could be good candidates for monitoring such systems, but require the ability to allow the clustering to *evolve* with the streaming data, i.e., new clusters form, old clusters grow or split. Furthermore, we need the model to distinguish between anomalies and extremal values of the “*normal*”.

Thus, non-parametric models that can accommodate infinite clusters are integrated with extreme value distributions that distinguish between anomalous and non-anomalous behaviors.

We now describe the two ingredients that go into our model namely, mixture models and extensions of EVT.

2.2 EVT and Generalized Pareto Distribution in Higher Dimensions

Estimation of parameters for extreme value distributions in higher dimensions is complex. To overcome this challenge, Clifton et al. [9] proposed an extended version of the generalized Pareto distribution that is applicable for different multi-model and multivariate distributions. For a given data $X \in \mathbb{R}^n$ distributed as $f_X : X \rightarrow Y$ where $Y \in \mathbb{R}$ is the image of the pdf values of X , let $Y \in [0, y_{max}]$ be the range of Y where $y_{max} = \sup(f_X)$. Let,

$$G_Y(y) = \int_{f_Y^{-1}([0,y])} f_X(x)dx \tag{1}$$

where $f_Y^{-1} : Y \rightarrow X$ is the pre-image of f_X given by, $f_Y^{-1}([0, y]) = \{x | f_X(x) \in [0, y]\}$. Then G_Y is in the domain of attraction of generalized Pareto distribution (GPD) G_Y^ξ for $y \in [0, u]$ as $u \rightarrow 0$ given by,

$$G_Y^\xi(y) = \begin{cases} 1 - (1 - \xi(\frac{y-\nu}{\beta})^{-1/\xi}), & \xi \neq 0 \\ 1 - \exp(-\frac{y-\nu}{\beta}), & \xi = 0 \end{cases} \tag{2}$$

where, ν, β and ξ are the location, scale and shape parameters of the GPD respectively.

2.3 Mixture Models

Mixture models assume that the data consists of sub-populations each generated from a different distribution. It can be used to study the properties of clusters using mixture distributions. In the classic version, when the number of clusters is known, finite mixture models are used with Dirichlet priors. However, when the number of latent clusters is unknown, one can extend finite mixture models to infinite mixture models like Dirichlet process mixture model (DPMM). In DPMM, the mixture distributions being sampled from a Dirichlet process (DP). DP can be viewed as a distribution over a family of distributions, that constitutes a base distribution G_0 which is a prior over the cluster parameters θ and positive scaling parameter α_{DP} . G is a Dirichlet process (denoted as $G \sim DP(G_0, \alpha_{DP})$) if G is a random distribution with same support as the base distribution G_0 and for any measurable finite partition of the support $A_1 \cup A_2 \cup \dots \cup A_k$, we have $(G(A_1), G(A_2), \dots, G(A_k)) \sim \text{Dir}(\alpha_{DP}G_0(A_1), \dots, \alpha_{DP}G_0(A_k))$.

In order to learn the number of clusters from the data, Bayesian non-parametric (BNP) models are used. BNP models like DPMM assume an infinite number of clusters of which only a finite number are populated. It brings forth a finesse in choosing the number of clusters while assuming a prior on the cluster assignments of the data. The prior is given by the Chinese restaurant process (CRP) which is defined analogously to the seating of N customers who sequentially join tables in a Chinese restaurant. Here, the probability of the n^{th} customer joining an existing table is proportional to the table size while the probability the customer forms a new table is always proportional to parameter α , $\forall n \in 1, 2, \dots, N$. This results in a distribution over the set of all partitions of integers $1, 2, \dots, N$. More formally, the distribution can be represented using the following probability function:

$$P(z_n = k | z_{1:n-1}) = \begin{cases} \frac{n_k}{n+\alpha-1} & n_k > 0 \text{ (existing cluster)} \\ \frac{\alpha}{n+\alpha-1} & n_k = 0 \text{ (new cluster)} \end{cases} \quad (3)$$

where z_i is the cluster assignment of the i^{th} data point, n_k is the size of the k^{th} cluster, $\alpha > 0$ is the concentration parameter. Large α values corresponds to an increased tendency of data points to form new clusters⁵.

3 Integrated Clustering and Anomaly Detection (INCAD)

The proposed INCAD model's prior is essentially a modification of a Chinese Restaurant Process (CRP). The seating of a customer at the Chinese restaurant is dependent on an evaluation by a gatekeeper. The gatekeeper decides the ability of the customer to start a new table based on the customer's features relative to existing patrons. If the gatekeeper believes the customer stands out, the customer is assigned a higher concentration parameter α that bumps up their chances of getting a new table. The model was inspired by the work of Blei and Frazier [5] in their distance dependent CRP models. The integrated INCAD model defines a flexible concentration parameter α for each data point. The probabilities are given by:

$$P(z_n | z_{1:n-1}, \mathbf{x}) = \begin{cases} \frac{n_k}{n+\alpha_2-1} & , n_k > 0 \text{ (existing cluster)} \\ \frac{\alpha_2}{n+\alpha_2-1} & , n_k = 0 \text{ (new cluster)} \end{cases}$$

⁵ Our modified model targets this aspect of concentration parameter to generate the desired simultaneous clustering and anomaly detection.

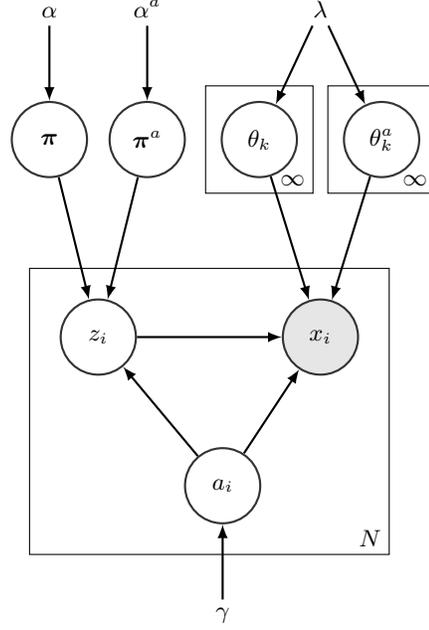


Fig. 1. Graphical representation of the proposed INCAD model.

where z, n, n_k are described earlier and $\alpha_2 = f(\alpha|x_n, \mathbf{x}, \mathbf{z})$ is a function of a base concentration parameter α . The function is chosen such that it is monotonic in $p(x_n)$ where, p is the probability of x_i being in the tail of the mixture distribution. In this paper, f is given by,

$$f(\alpha|x_n, \mathbf{x}, \mathbf{z}) = \begin{cases} \alpha, & \text{if not in tail} \\ \alpha^*, & \text{if in tail} \end{cases}$$

where, $\alpha^* = \frac{100}{1-p_n}$ and

$$p_n = p(x_n) = \begin{cases} \text{Probability of } x_n \text{ being anomalous, } x_n \text{ in tail} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Traditional CRP not only mimics the partition structure of DPMM but also allows flexibility in partitioning differently for different data sets. However, CRP lacks the ability to distinguish anomalies from non-anomalous points. To set differential treatment for anomalous data, the CRP concentration parameter α is modified to be sensitive to anomalous instances and nudge them to form individual clusters. The tail points are clustered using this updated concentration parameter α_2 which is designed to increase with increasing probability of a point being anomalous. This ensures that the probability of tail points forming individual clusters increases as they are further away from the rest of the data.

4 Choice of the Extreme Value Distribution

The choice for G_0^{EV} , the base distribution for the anomalous cluster parameters, is the key for identifying anomalous instances. By choosing G_0^{EV} as the extreme value counterpart of G_0 , the model ensures that the anomalous clusters are statistically “far” from the normal clusters. However, as discussed earlier, not all distributions have a well-defined EVD counterpart. We first describe the inference strategy for the case where G_0^{EV} exists. We then adapt this strategy for the scenario where G^{EV} is not available.

4.1 Inference when G_0^{EV} is available

Inference for traditional DPMMs is computationally expensive, even when conjugate priors are used. MCMC based algorithms [24] and variational techniques [6] have been typically used for inference. Here we adopt the Gibbs sampling based MCMC method for conjugate priors (Algorithm 1 [24]). This algorithm can be thought of as an extension of a Gibbs sampling-based method for a fixed mixture model, such that the cluster indicator z_i can take values between 1 and $K + 1$, where K is the current number of clusters.

Gibbs Sampling Though INCAD is based on DPMM, the model has an additional anomaly classification variable a_i that determines the estimation of the rest of the parameters. In the Gibbs sampling algorithm for INCAD, the data points $\{x_i\}_{i=1}^N$ are observed and the number of clusters K , cluster indicators $\{z_i\}_{i=1}^N$ and anomaly classifiers $\{a_i\}_{i=1}^N$ are latent. Using Markov property and Bayes rule, we derive the posterior probabilities for $z_i \forall i \in 1, 2, \dots, N$ as:

$$\begin{aligned}
 P(z_i = k | x_{\cdot}, z_{-i}, \alpha, \alpha^*, \boldsymbol{\pi}, \boldsymbol{\pi}^a, \lambda, \{\theta_k\}, \{\theta_k^a\}, a_{\cdot}, \gamma) &= P(z_i = k | x_{\cdot}, z_{-i}, \alpha, \alpha^*, \{\theta_k\}, \{\theta_k^a\}, a_i) \\
 &\propto \begin{cases} P(z_i = k | z_{-i}, \alpha, \theta_k) P(x_i | z_i = k, z_{-i}, \theta_k, \alpha) & a_i = 0 \\ P(z_i = k | z_{-i}, \alpha^*, \theta_k^a) P(x_i | z_i = k, z_{-i}, \theta_k^a, \alpha^*) & a_i = 1 \end{cases} \\
 &= \begin{cases} \frac{n_k}{(n + \alpha - 1)} F(x_i | \theta_k) & a_i = 0 \\ \frac{n_k}{(n + \alpha^* - 1)} F(x_i | \theta_k^a) & a_i = 1 \end{cases} \quad (5)
 \end{aligned}$$

where $\alpha^* = \frac{100}{1-p_i}$, p_i is the probability of x_i being anomalous and K is the number of non-empty clusters. Thus, the posterior probability of forming a new cluster denoted by $K + 1$ is given by:

$$\begin{aligned}
 P(z_i = K + 1 | x_{\cdot}, z_{-i}, \alpha, \alpha^*, \boldsymbol{\pi}, \boldsymbol{\pi}^a, \lambda, \{\theta_k\}, \{\theta_k^a\}, a_{\cdot}, \gamma) &= P(z_i = K + 1 | x_i, z_{-i}, \alpha, \alpha^*, \lambda, a_i) \\
 &\propto \begin{cases} P(z_i = K + 1 | z_{-i}, \alpha, \lambda) P(x_i | z_i = K + 1, z_{-i}, \alpha, \lambda, a_i) & a_i = 0 \\ P(z_i = K + 1 | z_{-i}, \alpha^*, \lambda) P(x_i | z_i = K + 1, z_{-i}, \alpha^*, \lambda, a_i) & a_i = 1 \end{cases} \\
 &= \begin{cases} \frac{\alpha}{n + \alpha - 1} \int F(x_i | \theta) G_0(\theta | \lambda) d\theta & a_i = 0 \\ \frac{\alpha^*}{n + \alpha^* - 1} \int F(x_i | \theta^a) G_0^{EV}(\theta^a | \lambda) d\theta^a & a_i = 1 \end{cases} \quad (6)
 \end{aligned}$$

Similarly, the parameters for clusters $k \in \{1, 2, \dots, K\}$ are sampled from:

$$\theta_k \propto G_0(\theta_k|\lambda)\mathcal{L}(\mathbf{x}_k|\theta_k) \text{ if cluster is not anomalous} \quad (7)$$

$$\theta_k^a \propto G_0^{EV}(\theta_k^a|\lambda)\mathcal{L}(\mathbf{x}_k|\theta_k^a) \text{ if cluster is anomalous} \quad (8)$$

where $\mathbf{x}_k = \{x_i|z_i = k\}$ is the set of all points in cluster k . Finally, to identify the anomaly classification of the data, the posterior probability of a_i is given by:

$$\begin{aligned} P(a_i = 1|x_i, z_i, \alpha, \alpha^*, \boldsymbol{\pi}, \boldsymbol{\pi}^a, \lambda, \{\theta_k\}, \{\theta_k^a\}, \gamma) &= P(a_i = 1|x_i, z_i, \alpha^*, \lambda, \{\theta_k^a\}, \gamma) \\ &\propto \sum_{k=1}^{K+1} P(a_i = 1|x_i, z_i = k, z_{-i}, \alpha^*, \lambda, \{\theta_k^a\}, \gamma) * P(z_i = k|x_i, z_{-i}, \alpha^*, \lambda, \{\theta_k^a\}, \gamma) \\ &= \sum_{k=1}^K P(x_i|\theta_k^a)\gamma \frac{n_k}{(n + \alpha^* - 1)} + \left(\int F(x_i|\theta^a)G_0^{EV}(\theta^a|\lambda)d\theta^a \right) \gamma \frac{\alpha^*}{n + \alpha^* - 1} \end{aligned} \quad (9)$$

Similarly,

$$\begin{aligned} &P(a_i = 0|x_i, z_i, \alpha, \lambda, \{\theta_k\}, \gamma) \\ &\propto \sum_{k=1}^K P(x_i|\theta_k)(1 - \gamma) \frac{n_k}{(n + \alpha - 1)} + \left(\int F(x_i|\theta)G_0(\theta|\lambda)d\theta \right) (1 - \gamma) \frac{\alpha}{n + \alpha - 1} \end{aligned} \quad (10)$$

4.2 Inference when G_0^{EV} is not available

The estimation of G_0^{EV} is required on two occasions. Firstly, while sampling the parameters of anomalous clusters when generating the data and estimating the posterior distribution. Secondly, to compute the probability of the point being anomalous when estimating the updated concentration parameter α_2 . When estimating G_0^{EV} is not feasible, the following two modifications to the original model are proposed:

1. Since an approximate G_0^{EV} distribution need not belong to the family of conjugate priors of F , we need a different approach to sample the parameters for anomalous clusters. Thus, we assume $\theta^a \sim G_0$ for sampling the parameters $\{\theta_k^a\}_{k=1}^{\infty}$ for anomalous clusters.
2. To estimate the probability of a point being anomalous, use the approach described by Clifton et al. [9].

The pseudo-Gibbs sampling algorithm, presented in Algorithm 2, has been designed to address the cases when G_0^{EV} is not available. For such cases, the modified f is given by,

$$f(\alpha|x_n, \mathbf{x}, \mathbf{z}) = \begin{cases} \alpha & , \text{if not in tail} \\ \alpha * (1 - ev_prop) + \frac{100}{1-p_n} * ev_prop & , \text{if in tail} \end{cases}$$

where ev_prop determines the effect of a anomalous behavior on the concentration parameter⁶. ev_prop is solely used to speed up the convergence in the Gibbs sampling. Since the estimation of the G^{EV} distribution is not always possible, an alternate/pseudo Gibbs sampling algorithm has been presented in Algorithm 2.

⁶ It can be seen that when the data point x_n has extreme or rare features differentiating it from all existing clusters, the corresponding density $f_y(x_n)$ decreases. This makes it a left tail point in the distribution G_y . The farther away it is in the tail, the lower the probability of its δ -nbd being in the tail and hence a higher $f(\alpha|x_n, \mathbf{x}, \mathbf{z})$.

Algorithm 1: Gibbs Sampling Algorithm when G_0^{EV} is available

Given $z^{(t-1)}, a^{(t-1)}, \{\theta_k^{(t-1)}\}, \{\theta_k^{a(t-1)}\}$ from previous iterations. Let K be the total number of clusters found till last iteration. Sample $z^{(t)}, a^{(t)}, \{\theta_k^{(t)}\}, \{\theta_k^{a(t)}\}$ as follows

1. Set $z_i = z_i^{(t-1)}$ and $a_i = a_i^{(t-1)}$
2. **for** each point $i \rightarrow 1$ **to** N **do**
 - (a) Remove x_i from its cluster z_i .
 - (b) If x_i is the only point in its cluster, $n_{z_i} = 0$ after step (2)(a). Remove the cluster and update K to $K-1$.
 - (c) Rearrange cluster indices to ensure none are empty.
 - (d) Sample z_i from the Multinomial distribution given by Equations 5 and 6
 - (e) If $z_i = K + 1$, then sample new cluster parameters from the following distribution (It must be noted that the above posterior distribution was derived under the assumption of independence and exchangeability of priors for mathematical ease.)

$$\theta \left| x_i, z_i, \{\theta_k^{(t-1)}\}, \{\theta_k^{a(t-1)}\}, a_i^{(t-1)} \right.$$

$$\propto \begin{cases} \alpha G_0(\theta|\lambda) F(x_i|\theta) + \sum_{j \neq i} F(x_i|\theta_{z_j}) \delta(\theta - \theta_{z_j}^{(t-1)}) \delta(a_j^{(t-1)}), & a_i^{(t-1)} = 0 \\ \alpha^* G_0^{EV}(\theta|\lambda) F(x_i|\theta) + \sum_{j \neq i} F(x_i|\theta_{z_j}) \delta(\theta - \theta_{z_j}^{(t-1)}) \delta(a_j^{(t-1)} - 1), & a_i^{(t-1)} = 1 \end{cases}$$

Update $K=K+1$.

- (f) For each cluster $k \in \{1, 2, \dots, K\}$, sample cluster parameters θ_k and θ_k^a using Equations 7 and 8.
 - (g) Sample the anomaly classification a_i from the Binomial distribution given by Equations 9 and 10.
 - (h) Set $z_i^{(t)} = z_i$ and $a_i^{(t)} = a_i$.
-

4.3 Exchangeability

A model is said to be exchangeable when for any permutation S of $\{1, 2, \dots, n\}$, $P(x_1, x_2, \dots, x_n) = P(x_{S(1)}, x_{S(2)}, \dots, x_{S(n)})$. Looking at the joint probability of the cluster assignments for the integrated model, we know,

$$P(z_1, z_2, \dots, z_n | \mathbf{x}) = P(z_1 | \mathbf{x}) P(z_2 | z_1, \mathbf{x}) \dots P(z_n | z_{1:n-1}, \mathbf{x})$$

Without loss of generality, let us assume there are K clusters. Let, for any $k < K$, the joint probability of all the points in cluster k be given by

$$\left(\frac{\alpha * p_{k,1}}{I_{k,1} + \alpha - 1} + \frac{\alpha^* * (1 - p_{k,1})}{I_{k,1} + \alpha^* - 1} \right) \prod_{n_k=2}^{N_k} \left(\frac{(n_k - 1) * p_{k,n_k}}{I_{k,n_k} + \alpha - 1} + \frac{(n_k - 1) * (1 - p_{k,n_k})}{I_{k,n_k} + \alpha^* - 1} \right)$$

Algorithm 2: Gibbs Sampling Algorithm when G_0^{EV} is not available

Given $z^{(t-1)}, a^{(t-1)}, \{\theta_k^{(t-1)}\}, \{\theta_k^{a(t-1)}\}$ from previous iterations. Let K be the total number of clusters found till last iteration. Sample $z^{(t)}, a^{(t)}, \{\theta_k^{(t)}\}, \{\theta_k^{a(t)}\}$ as follows

1. Set $z = z^{(t-1)}$ and $a = a^{(t-1)}$
 2. **for** each point $i \rightarrow 1$ **to** N **do**
 - (a) Steps 2a to 2d in Algorithm 1
 - (b) If $z_i = K + 1$, then set the cluster distribution to be multivariate normal with the new cluster mean as x_i and cluster variance as Σ which is pre-defined.
Update $K=K+1$.
 - (c) For each cluster $k \in \{1, 2, \dots, K\}$, sample cluster parameters θ_k and θ_k^a using Equation 7.
 - (d) Sample the anomaly classification a_i from the Binomial(p_i) where p_i is given by Equation 4. If most of the cluster instances are classified as anomalous, classify all of the cluster's instances as anomalies.
 - (e) Set $z^{(t)} = z$ and $a^{(t)} = a$.
-

where N_k is the size of the cluster k , $I_{k,i}$ is the index of the i^{th} instance joining the k^{th} cluster and $p_{k,i} = p_{I_{k,i}}$. Thus, the joint probability for complete data is then given by

$$\frac{\prod_{k=1}^K \left[(I_{k,1} - 1)p_{k,1}(\alpha - \alpha^*) + \alpha^*(I_{k,1} + \alpha - 1) \prod_{n_k=2}^{N_k} (n_k - 1)(I_{k,n_k} + \alpha - 1 + p_{k,n_k}(\alpha^* - \alpha)) \right]}{\prod_{i=1}^N ((i + \alpha - 1)(i + \alpha^* - 1))}$$

which is dependent on the order of the data. This shows that the model is not exchangeable unless $\alpha = \alpha^*$ or $p_{k,n_k} = 0$ or $p_{k,n_k} = 1$. These conditions effectively reduce the prior distribution to a traditional CRP model. Hence, it can be concluded that the INCAD model cannot be modified to be exchangeable.

Non-exchangeable models in streaming settings Though exchangeability is a reasonable assumption in many situations, the evolution of behavior over time is not captured by traditional exchangeable models. In particular for streaming settings, using non-exchangeable models captures the effect of the order of the data. In such settings, instances that are a result of new evolving behavior should be monitored (as anomalous) until the behavior becomes relatively prevalent. Similarly, relapse of outdated behaviors (either normal or anomalous) should also be subjected to critical evaluation due to extended lag observed between similar instances. Such order driven dependency can be well captured in non-exchangeable models making them ideal for studying streaming data.

4.4 Adaptability to sequential data

One of the best outcomes of having a non-exchangeable prior is its ability to capture the *drift or evolution* in the behavior(s) either locally or globally or a mixture of both. INCAD model serves as a perfect platform to detect these changes and delivers an adaptable classification and clustering. The model has a straightforward extension to sequential settings where the model evolves with

every incoming instance. Rather than updating the model for entire data with each new update, the streaming INCAD model re-evaluates only the tail instances. This enables the model to identify the following evaluations in the data:

1. New trends that are classified as anomalous but can eventually grow to become normal.
2. Previously normal behaviors that have vanished over time but have relapsed and hence become anomalous (eg. disease relapse post complete recovery)

The Gibbs sampling algorithm for the streaming INCAD model is given in Algorithm 3.

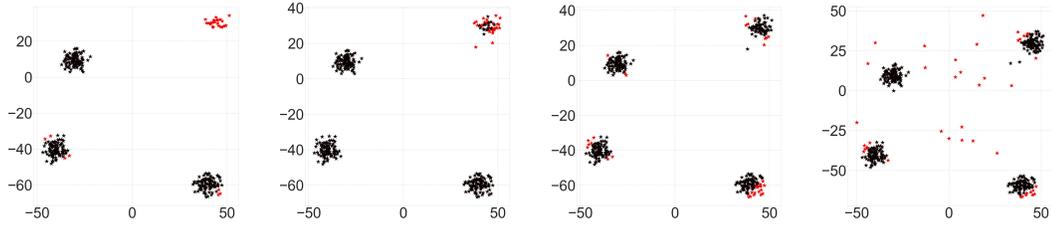


Fig. 2. Evolution of anomaly classification using streaming INCAD: The classification into anomalous and normal instances are represented by ● and ● respectively. Note the evolution of the classification of top right cluster from anomalous to normal with incoming data

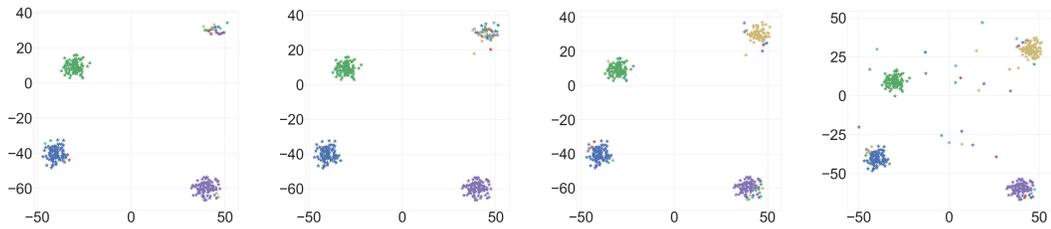


Fig. 3. Evolution of clustering using streaming INCAD: Each cluster is denoted by a different color. Notice the evolution of random points in top right corner into a well formed cluster in the presence of more data

5 Results

In this section, we evaluate the proposed model using benchmark streaming datasets from NUMENTA. The streaming INCAD model’s anomaly detection is compared with SPOT algorithm developed by Siffer et al. [27]. The evolution of clustering and anomaly classification using the streaming INCAD model is visualized using simulated dataset. In addition, the effect of batch vs

Algorithm 3: Algorithm for Streaming Extension

Perform clustering on a small portion of the data (10-20%) using non-streaming model

Set $ev_{prop} = exp^{-0.5}$

for each new data point x_N **do**

1. Compute the mixture proportions m_{para} and the mixture density for all the data.
 Compute $t_1 = q^{th}$ percentile pdf value to identify the tail points
2. For each x_i s.t. $f(x_i) < t_1$ repeat steps 2a→2d of Algorithm 2

If cluster size $\leq 0.05 * N$ then, classify all the cluster points as anomalies.

stream proportion on quality of performance is presented. For Gibbs sampling initialization, the data was assumed to follow a mixture of MVN distributions. 10 clusters were initially assumed with the same initial parameters. The cluster means were set to the sample mean and the covariance matrix as a multiple of the sample covariance matrix, using a scalar constant. The concentration parameter α was always set to 1.

5.1 Simulated Data

For visualizing the model’s clustering and anomaly detection, a 2-dimensional data set of size 400 with 4 normal clusters and 23 anomalies sampled from a normal distribution centered at (0,0) and a large variance was generated for model evaluation. Small clusters and data outliers are regarded as true anomalies. Data from the first 3 normal clusters (300 data points) were first modeled using non-streaming INCAD. The final cluster and the anomalies were then used as updates for the streaming INCAD model. The evolution in the anomaly classification is presented in Figure 2.

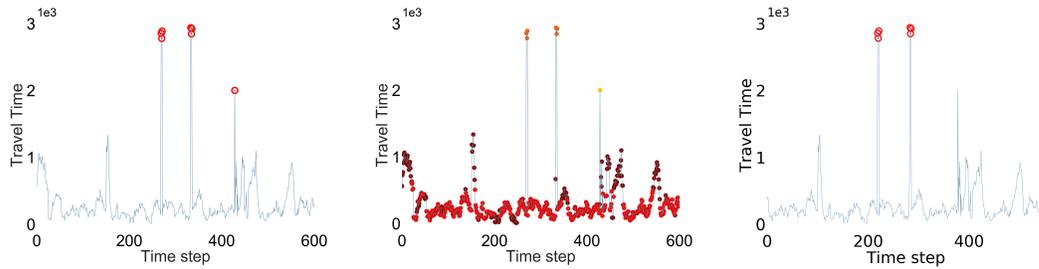


Fig. 4. NUMENTA traffic data: (from left to right) Anomaly detection and clustering using streaming INCAD and anomaly detection using SPOT [27]

5.2 Model Evaluation on NUMENTA Data

Two data sets from NUMENTA [23] namely the real traffic data and the AWS cloud watch data were used as benchmarks for the streaming anomaly detection. The streaming INCAD model was compared with the SPOT algorithm developed by Siffer et al. [27]. Unlike SPOT algorithm, the

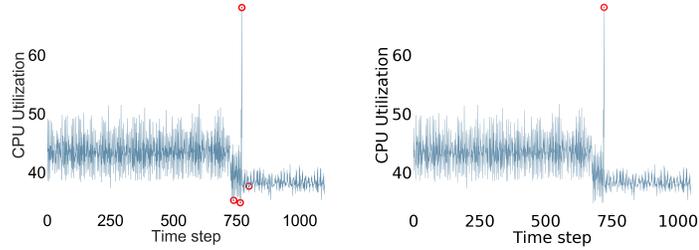


Fig. 5. Anomaly detection on NUMENTA AWS cloud watch data using streaming INCAD(left) and SPOT [27] (right)

streaming INCAD is capable of modeling data with more than one feature. Thus, the data instance, as well as the time of the instance, were used to develop the anomaly detection model. Since the true anomaly labels are not available, the model’s performance with respect to SPOT algorithm was evaluated based on the ability to identify erratic behaviors. The model results on the datasets using streaming INCAD and SPOT have been presented in Figures 4 and 5.

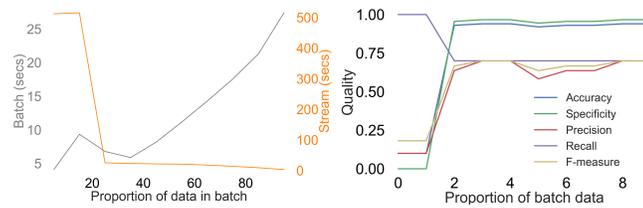


Fig. 6. Proportion of data for batch model vs Quality: Computational time (left) and Proportion of data for the batch model (in tens) vs Quality of Anomaly detection (right)

6 Sensitivity to Batch Proportion

Streaming INCAD model re-evaluates the tail data at each update, the dependency of the model’s performance on the current state must be evaluated. Thus, various metrics were used to study the model’s sensitivity to the initial batch proportion. Figure 6 shows the effect of batch proportion on computational time and performance of anomaly detection. The simulated data defined in Section 5.1 was used for the sensitivity analysis. It can be seen that the computational time is optimal for 25% of data used in to run the non-streaming INCAD model. As anticipated, precision, accuracy, specificity, and f-measure for the anomaly detection were observed to plateau after a significant increase.

7 Conclusion and Future Work

A detailed description of the INCAD algorithm and the motivation behind it has been presented in this paper. The model’s definition of an anomaly and its adaptable interpretation sets the model

apart from the rest of the clustering based anomaly detection algorithms. While past anomaly detection methods lack the ability to simultaneously perform clustering and anomaly detection or to the INCAD model not only defines a new standard for such integrated methods but also breaks into the domain of streaming anomaly detection. The model's ability to identify anomalies and cluster data using a completely data-driven strategy permits it to capture the evolution of multiple behaviors and patterns within the data.

Additionally, the INCAD model can be smoothly transformed into a streaming setting. The model is seen to be robust to the initial proportion of the data subset that was evaluated using the non-streaming INCAD model. Moreover, this sets up the model to be extended to distribution families beyond multivariate normal. Though one of the key shortcomings of the model is its computational complexity in Gibbs sampling in the DPMM clusters, the use of faster methods such as variational inference might prove to be useful.

8 Acknowledgements

The authors would like to acknowledge University at Buffalo Center for Computational Research (<http://www.buffalo.edu/ccr.html>) for its computing resources that were made available for conducting the research reported in this paper. Financial support of the National Science Foundation Grant numbers NSF/OAC 1339765 and NSF/DMS 1621853 is acknowledged.

References

- [1] Husam Al-Behadili, Arne Grumpe, Lubaba Migdadi, and Christian Wöhler. Semi-supervised learning using incremental support vector machine and extreme value theory in gesture data. In *Computer Modelling and Simulation (UKSim)*, pages 184–189. IEEE, 2016.
- [2] Mennatallah Amer and Markus Goldstein. Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer. In *Proc. of the 3rd RCOMM 2012*, pages 1–12, 2012.
- [3] Charles E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, 2(6), 1974.
- [4] Stephen D Bay and Mark Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proc. of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 29–38. ACM, 2003.
- [5] David M Blei and Peter I Frazier. Distance dependent chinese restaurant processes. In *ICML*, pages 87–94, 2010.
- [6] David M. Blei and Michael I. Jordan. Variational methods for the dirichlet process. In *Proc. of the Twenty-first International Conference on Machine Learning*, pages 12–, 2004.
- [7] Philip K Chan, Matthew V Mahoney, and Muhammad H Arshad. A machine learning approach to anomaly detection. Technical report, 2003.
- [8] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.
- [9] David A Clifton, Lei Clifton, Samuel Hugueny, and Lionel Tarassenko. Extending the generalised pareto distribution for novelty detection in high-dimensional spaces. *Journal of Signal Processing Systems*, 74(3):323–339, 2014.
- [10] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pages 77–101. Springer, 2002.

- [11] Joshua French, Piotr Kokoszka, Stilian Stoev, and Lauren Hall. Quantifying the risk of heat waves using extreme value theory and spatio-temporal functional data. *Computational Statistics & Data Analysis*, 131:176–193, 2019.
- [12] Zhouyu Fu, Weiming Hu, and Tieniu Tan. Similarity based vehicle trajectory clustering and anomaly detection. In *ICIP*, volume 2, pages II–602. IEEE, 2005.
- [13] João Gama, Indrè Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):44, 2014.
- [14] Pedro Garcia-Teodoro, Jesus Diaz-Verdejo, Gabriel Maciá-Fernández, and Enrique Vázquez. Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, 28(1-2):18–28, 2009.
- [15] Markus Goldstein and Seiichi Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173, 2016.
- [16] Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research*, 46:235–262, 2013.
- [17] Sreelekha Guggilam, S. M. Arshad Zaidi, Varun Chandola, and Abani Patra. Bayesian anomaly detection using extreme value theory. *arXiv preprint arXiv:1905.12150*, 2019.
- [18] Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650, 2003.
- [19] N. Hjort, C. Holmes, P. Mueller, and S. Walker. *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, Cambridge, UK, 2010.
- [20] Jialiang Jiang, Jessica Castner, Sharon Hewner, and Varun Chandola. Improving quality of care using data science driven methods. In *UNYTE Scientific Session - Hitting the Accelerator: Health Research Innovation through Data Science*, 2015.
- [21] Nan Jiang and Le Gruenwald. Research issues in data stream association rule mining. *ACM Sigmod Record*, 35(1):14–19, 2006.
- [22] Christopher Kruegel and Giovanni Vigna. Anomaly detection of web-based attacks. In *Proc. of the 10th ACM conference on Computer and communications security*, pages 251–261, 2003.
- [23] Alexander Lavin and Subutai Ahmad. Evaluating real-time anomaly detection algorithms—the numenta anomaly benchmark. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 38–44. IEEE, 2015.
- [24] Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.
- [25] Carl Edward Rasmussen. The infinite gaussian mixture model. In *In Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, 2000.
- [26] Matthew S. Shotwell and Elizabeth H. Slate. Bayesian outlier detection with dirichlet process mixtures. *Bayesian Anal.*, 6(4):665–690, 12 2011.
- [27] Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet. Anomaly detection in streams with extreme value theory. In *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1067–1075, 2017. ISBN 978-1-4503-4887-4.
- [28] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
- [29] J. Varadarajan, R. Subramanian, N. Ahuja, P. Moulin, and J. M. Odobez. Active online anomaly detection using dirichlet process mixture model and gaussian process classification. In *2017 IEEE WACV*, pages 615–623, 2017.