Anomaly Intrusion Detection Using Incremental Learning of an Infinite Mixture Model with Feature Selection

Wentao Fan[1], Nizar Bouguila[1], and Hassen Sallay[2]

[1] Concordia University, Montreal, QC, Canada
wenta_fa@encs.concordia.ca, nizar.bouguila@concordia.ca
[2] Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia
hmsallay@imamu.edu.sa

Abstract. We propose an incremental nonparametric Bayesian approach for clustering. Our approach is based on a Dirichlet process mixture of generalized Dirichlet (GD) distributions. Unlike classic clustering approaches, our model does not require the number of clusters to be pre-defined. Moreover, an unsupervised feature selection scheme is integrated into the proposed nonparametric framework to improve clustering performance. By learning the proposed model using an incremental variational framework, the number of clusters as well as the features weights can be automatically and simultaneously computed. The effectiveness and merits of the proposed approach are investigated on a challenging application namely anomaly intrusion detection.

Keywords: Mixture models, clustering, Dirichlet process, generalized Dirichlet, feature selection, variational inference, intrusion detection.

1 Introduction

Huge volumes of data are routinely generated by organizations, scientific activities, internet traffic and so on. An important problem is to model these data to improve the process of making automatic decisions [12]. A widely used approach for data modeling and knowledge discovery is clustering. Clustering can be defined as the task of partitioning a given data set $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$ containing N vectors into M homogenous clusters $\mathcal{C}_1, \ldots, \mathcal{C}_M$ such that $\mathcal{C}_j \cap \mathcal{C}_l = \emptyset$, and $\cup_{j=1}^{M} \mathcal{C}_j = \mathcal{X}$. Finite mixture models have been widely applied for clustering during the last two decades [11]. Within finite mixture modeling, selecting the number of components that best describes the underlying data without over- or under-fitting is one of the most challenging problems. This obstacle can be removed by extending finite mixtures to the infinite case through Dirichlet processes [13]. Infinite mixtures allow a natural approach for data clustering. Unlike finite mixtures, the number of clusters does not need to be specified by the practitioner in advance and can be automatically inferred from the dataset. Several approaches have been proposed to learn mixture models. In particular, variational inference has received a lot of attention recently [5,4,1,6]. Variational

P. Lingras et al. (Eds.): RSKT 2013, LNAI 8171, pp. 364–373, 2013.

inference is a deterministic approximation learning technique that only requires a modest amount of computational power in contrast to other well-developed approaches such as Markov chain Monte Carlo (MCMC) techniques, and has a tractable learning process as well. Generally real-world problems involve dynamic data sets where the volume of data continuously grows. Thus, it is crucial to adopt an incremental way to learn the statistical model used for clustering.

In this paper, we adopt an incremental version of variational inference proposed by [7] to learn infinite generalized Dirichlet (GD) mixtures with unsupervised feature selection. The employment of the GD as the basic distribution in our mixture model is motivated by its favorable performance when dealing with non-Gaussian data [2,3]. The advantages of our framework are summarized as following: First, the difficulty of choosing the appropriate number of components is avoided by assuming that there is an infinite number of components. Second, thanks to its incremental nature, it is very efficient when dealing with sequentially arriving data, which is an important factor for real-time applications. Third, within the proposed framework, the model parameters and features saliencies can be estimated simultaneously and automatically. The effectiveness of our approach is illustrated through a challenging task namely anomaly intorsion detection. The rest of this paper is organized as follows. Section 2 reviews briefly the infinite GD mixture model with unsupervised feature selection. In Section 3, we develop an incremental variational inference framework for model learning. Section 4 is devoted to the experimental results. Finally, conclusion follows in Section 5.

2 Infinite GD Mixture Model with Feature Selection

In this section, we review briefly the infinite generalized Dirichlet (GD) mixture model with feature selection, which is constructed using a stick-breaking Dirichlet process framework. If a D-dimensional random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_D)$ is sampled from a mixture of GD distributions with infinite number of components:

$$p(\boldsymbol{Y}|\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{j=1}^{\infty} \pi_j \mathrm{GD}(\boldsymbol{Y}|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) \tag{1}$$

where $\boldsymbol{\pi}$ represents the mixing coefficients with the constraints that are positive and sum to one. Here we adopt the Dirichlet process framework with a stick-breaking representation [15], where the mixing coefficients $\{\pi_j\}$ are constructed by recursively breaking a unit length stick into an infinite number of pieces as $\pi_j = \lambda_j \prod_{k=1}^{j-1}(1-\lambda_k)$. The stick breaking variable λ_j is distributed according to $\lambda_j \sim \mathrm{Beta}(1, \zeta)$, where ζ is a positive real number and is the concentration parameter of the Dirichlet process. In Eq. (1), $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jD})$ and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jD})$ are the positive parameters of the GD distribution $\mathrm{GD}(\boldsymbol{Y}|\boldsymbol{\alpha}_j, \boldsymbol{\beta})$ associated with component j, where $\mathrm{GD}(\boldsymbol{X}|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$ is given by

$$\mathrm{GD}(\boldsymbol{Y}|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) = \prod_{l=1}^{D} \frac{\Gamma(\alpha_{jl}+\beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} Y_l^{\alpha_{jl}-1} \left(1 - \sum_{k=1}^{l} Y_k\right)^{\gamma_{jl}} \tag{2}$$