Noname manuscript No.
(will be inserted by the editor)

Energy-Based Anomaly Detection for Mixed Data

Kien Do · Truyen Tran · Svetha Venkatesh

Received: xxx / Revised: xxx / Accepted: xxx

Abstract Anomalies are those deviating significantly from the norm. Thus anomaly detection amount to finding data points located far away from their neighbors, i.e., those lying in low density regions. Classic anomaly detection methods are largely designed for single data type such as continuous or discrete. However, real world data is increasingly heterogeneous, where a data point can have both discrete and continuous attributes. Mixed data poses multiple challenges including (a) capturing the inter-type correlation structures and (b) measuring deviation from the norm under multiple types. These challenges are exaggerated under (c) high-dimensional regimes. In this paper, we propose a new scalable unsupervised anomaly detection method for mixed data based on Mixed-variate Restricted Boltzmann Machine (Mv.RBM). The Mv.RBM is a principled probabilistic method that estimates density of mixed data. We propose to use *free-energy* derived from Mv.RBM as anomaly score as it is identical to data negative log-density up-to an additive constant. We then extend this method to detect anomalies across multiple levels of data abstraction, an effective approach to deal with high-dimensional settings. The extension is dubbed `MIXMAD`, which stands for MIXed data Multilevel Anomaly Detection. In `MIXMAD`, we sequentially constructs an ensemble of mixed-data Deep Belief Nets (DBNs) with varying depths. Each DBN is an energy-based detector at a predefined abstraction level. Predictions across the ensemble are finally combined via a simple rank aggregation method. The proposed methods are evaluated on a comprehensive suit of synthetic and real high-dimensional datasets. The results demonstrate that for anomaly detection, (a) a proper handling mixed-types is necessary, (b) free-energy is a powerful anomaly scoring method, (c) multilevel abstraction of data is important for high-dimensional data, and (d) empirically Mv.RBM and `MIXMAD` are superior to popular unsupervised detection methods for both homogeneous and mixed data.

Keywords mixed data; mixed-variate restricted Boltzmann machine; deep belief net; multi-level anomaly detection

K. Do, T. Tran, S. Venkatesh
Center for Pattern Recognition and Data Analytics (PRaDA),
Deakin University, Australia.
E-mail: {dkdo,truyen.tran,svetha.venkatesh}@deakin.edu.au

1 Introduction

A vital skill for living organism is detecting large deviations from the norm. Except for few deadly instances, we learn to detect anomalies by observing and exploring, without supervision. Unsupervised anomaly detection does not assume any domain knowledge about abnormality, and hence is cheap and pervasive. A disciplined approach is to identify instances lying in low density regions [11]. However, estimating density in realistic settings is difficult, especially on mixed and high-dimensional data.

Mixed-data is a pervasive but rarely addressed phenomenon: a data attribute can be any type such as continuous, binary, count or nominal [28]. Most existing anomaly detection methods, however, assume homogeneous data types. Gaussian mixture models (GMMs), for instance, require data to be continuous and normally distributed – a strong assumption rarely met in practice. One approach to mixed-type data is to reuse existing methods. For example, we can transform multiple types into a single type via a process known as *data coding*. For instance, nominal data is often coded as a set of zeros except for one active element. But it leads to information loss because the derived binary variables are considered independent in subsequent analysis. Further, binary variables are not naturally supported by numerical methods such as GMM and PCA. Another way is to modify existing methods to accommodate multiple types. However, the modifications are often heuristic. For distance-based methods such as k-NN [4] we need to define type-specific distances, then combine these distances into a single measure. Because type-specific distances differ in scale and semantics, finding a suitable combination is non-trivial.

Another pervasive challenge is *high dimensionality* [11]. Under this condition, non-parametric methods that define a data cube to estimate relative frequency are likely to fail. It is because the number of cubes grows exponentially with data dimensions, thus a cube with only a few or no observed data points needs not be in a low density region. An alternative is to use distance to k-nearest neighbors, assuming that the larger distance, the less dense the region [4]. But distance in high dimensional space is sensitive to a small change in each dimension, and easily distorted by redundant and irrelevant dimensions. This necessitates *data abstraction* in which data is transformed into a more abstract form.

To sum up, a disciplined approach to mixed-data and high-dimensional anomaly detection demands meeting four criteria: (i) *capturing between-type correlation structure*, (ii) *offering abstraction on raw data and detection on abstracted data*, (iii) *offering an effective way to measure the deviation from the norm*, and importantly (iv) *being efficient to compute*.

To this end, we propose a *new energy-based approach* which models multiple types simultaneously and provides a fast mechanism for identifying low density regions[1]. Under this energy-based framework, anomalies have higher *free-energy* than the rest. To be more precise, we adapt and extend a recent method called Mixed-variate Restricted Boltzmann Machine (Mv.RBM) [44]. Mv.RBM is a generalization of the classic RBM – originally designed for binary data, and is a building block for many deep learning architectures [6, 21] in recent years. Mv.RBM has been applied for representing *regularities* in survey analysis [44], multimedia [32] and healthcare [31], but not for anomaly detection, which searches for *irregularities*. Mv.RBM captures the correlation structure between types through factoring – data types are assumed to be conditionally independent given a generating mechanism.

We contribute to the literature of anomaly detection in several ways. First we extend Mv.RBM to cover *counts*, a type often seen in practice, but not previously modeled in Mv.RBM. We then propose to use *free-energy* as anomaly score to rank mixed-type instances.

[1] A preliminary version of this paper has been published in [16].

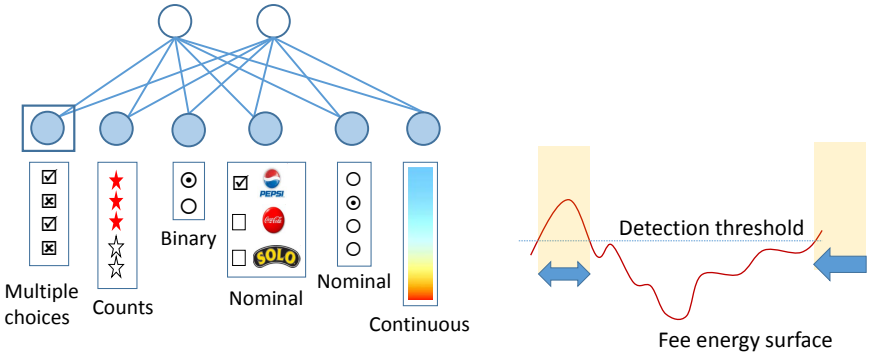


Fig. 1 **Left**: Mix-variate Restricted Boltzmann machine (Mv.RBM) for mixed-type data [44]. Filled circles denote visible inputs, empty circles denote hidden units. Multiple choices are modeled as multiple binaries, denoted by a filled circle in a clear box. **Right**: Mv.RBM's free-energy as anomaly scoring method.

In RBMs, free-energy equals the negative log-density up to an additive constant, and thus offering a principled way for density-based anomaly detection. Importantly, estimation of Mv.RBM is very efficient, and scalable to massive datasets. Likewise, free-energy is computed easily through a single matrix projection. Thus Mv.RBM coupled with free-energy meets three criteria (i,iii, and iv) outlined above for anomaly detection. See Fig. 1 for a graphical illustration.

For criterion (ii) – *data abstraction*, we leverage recent advances in unsupervised deep learning to abstract the data into multilevel low-dimensional representations [6]. While deep learning has revolutionized supervised learning [27], it is rarely used for unsupervised anomaly detection [41,49]. In particular, we adapt Deep Belief Net (DBN) [21], which is a deep generalization of RBM, for mixed data. A DBN is built by successively learning a new RBM based on the output of previous RBMs. A trained DBN is thus a layered model that allows multiple levels of data abstraction. Due to its stepwise construction, we can use the top layer of the DBN as an anomaly detector. The anomaly detection procedure is as follows: First apply multiple layered abstractions to the data, and then estimate the anomalies at each level. Finally, an anomaly score is aggregated across levels. These together constitute a method called `MIXMAD`, which stands for *MIX*ed data *M*ultilevel *A*nomaly *D*etection. While `MIXMAD` bears some similarity with the recent ensemble approaches [2], the key difference is the we rely on multiple data abstractions, not data resampling nor random subspaces which are still on the original data level. In `MIXMAD`, as the depth increases and the data representation is more abstract, the energy landscape gets smoother, and thus it may detect different anomalies. For reaching anomaly consensus across depth-varying DBNs, `MIXMAD` employs a simple yet flexible rank aggregation method based on p-norm.

We validate the proposed approach through an extensive set of synthetic and real experiments against well-known baselines, which include classic single-type methods (PCA, GMM and one-class SVM), as well as state-of-the-art mixed-type methods (ODMAD [26], Beta mixture model (BMM) [8] and GLM-t [28]). The experiments demonstrate that (a) a proper handling mixed-types is necessary, (b) free-energy is a powerful and efficient anomaly scoring method, (c) multilevel abstraction of data is important, and (d) empirically `MIXMAD` is superior to popular unsupervised detection methods for both homogeneous and mixed data.

In summary, we claim the following contributions:

- Introducing a new anomaly detection method for mixed-type data. The method is based on *free-energy* derived from a recent method known as Mixed-variate Restricted Boltzmann Machine (Mv.RBM). The method is theoretically motivated and efficient.
- Extension of Mv.RBM to handle the type of counts.
- Introducing the concept of Multilevel Anomaly Detection (MAD) that argues for reaching consensus across multiple abstractions of data.
- Deriving `MIXMAD`, an efficient MAD algorithm to build a sequence Deep Belief Nets, each of which is an anomaly detector. All detectors are then combined using a flexible p-norm aggregation that allows tuning along the conservative/optimistic axis.
- A comprehensive evaluation of Mv.RBM and `MIXMAD` on realistic high-dimensional single/mixed-type datasets against a large suite of competing methods.

The rest of the paper is organized as follows. Section 2 reviews relevant background on anomaly detection. Section 3 introduces the problem of anomaly detection using density and building blocks for `MIXMAD`. The main contributions of the paper are presented in Section 4 (Mv.RBM) and Section 5 (`MIXMAD`). Section 6 reports a comprehensive suite of experiments on various settings, followed by a section for discussion and conclusion.

2 Background

Anomalies are those characterized by irregular characteristics [11]. A wide range of unsupervised methods have been proposed for *homogeneous data,* for example, distance-based (e.g., k-NN [4]), density-based (e.g., LOF [9], LOCI [35]), cluster-based (e.g., Gaussian mixture model or GMM), projection-based (e.g., PCA) and max margin (One-class SVM). Distance-based and density-based methods model the local behaviors around each data point while cluster-based methods group similar data points together into clusters. Projection-based methods, on the other hand, find a data projection that is sensitive to outliers. These popular methods commonly assume continuous attributes. *Categorical data* demands separate treatments as these existing notions of distance metrics, density or projection are not easily translatable [3, 14, 22, 34, 40]. A popular method is pattern mining, in which co-occurrence statistics of discrete attributes are examined [3, 14, 34].

Given that continuous and categorical attributes demand separate treatments, it is very challenging to address both data types in a unified manner [25, 28, 42, 48, 50]. The work of [19] introduces LOADED, which uses frequent pattern mining to define the score of each data point in the nominal attribute space and links it with a precomputed correlation matrix for each item set in the continuous attribute space. A more memory-efficient method is RELOAD [33], which employs a set of Naïve Bayes classifiers with continuous attributes as inputs to predict abnormality of nominal attributes instead of aggregating over a large number of item sets. Another method is ODMAD [26], a two--step procedure. First it detects anomalies using nominal attributes. Then the remaining of the points are examined over continuous attribute space. In [8], separate scores over nominal data space and numerical data space are calculated for each data point. The list of two dimensional score vectors of data was then modeled by a mixture of bivariate beta distributions. Abnormal objects could be detected as having a small probability of belonging to any components. The work of [50] introduces POD, which stands for Pattern-based Anomaly Detection. A pattern is a subspace formed by a particular nominal field and all continuous fields. A logistic classifier is trained for each subspace pattern, in which continuous and nominal attributes are explanatory and response variables, respectively. The probability returned by the classifier measures the degree to which an instance deviates from a specific pattern. This is called Categorical Anomaly Factor (COF). The collection of

COFs and k-NN distance form the final anomaly score for a data example. Given a nominal attribute, POD models the functional relationship between continuous variables.

For all the methods mentioned above, their common drawback is that they are only able to capture intra-type correlation, but not inter-type correlation. The work of [28] introduces a Generalized Linear Model that uses a latent variable for correlation capturing and an another latent variable following Student-t distribution as an error buffer. The main advantage of this method is that it provides strong a statistical foundation for modeling distribution of different types. However, the inference for detecting outliers is inexact and expensive to compute. We wish to emphasize that we intend to cover more types than just nominal and continuous. In particular, our model choices are capable of modeling count [36], preference, and ordinal data simultaneously [44]. Importantly, the computation complexity is linear in number of variables.

Another challenge is *high-dimensional data* [4,51]. This type is sensitive to irrelevant and redundant attributes, causing failure of low-dimensional techniques [1]. Popular solutions include feature selection, dimensionality reduction (such as using PCA) and subspace analysis [51]. Our approach to high-dimensionality is through multiple levels of abstraction. The abstraction uncovers regularities in data, and consequently helps mitigate noise, redundancy and irregularity.

The recent successes of *deep neural networks* in classification tasks have inspired some work in anomaly detection. A strategy is to use unsupervised deep networks such as Deep Belief Nets and Stacked Autoencoders as feature detectors. The features are then fed into well-established detection algorithms [47]. Another strategy is to use reconstruction errors by Stacked Autoencoders as anomaly scores [5,18,38,39]. A problem with this approach is that the final model still operates on raw data, which can be noisy and high-dimensional. Further, reconstruction error does not necessarily reflect data density [23], making it hard to justify the method. A better approach is to use deep networks to estimate the energy directly [17, 49]. When data is sequential, there have been several attempts to use a recurrent neural net known as the Long Short-Term Memory (LSTM) for anomaly detection [7,12,13,29,41,46]. They do not handle mixed data, however.

3 Preliminaries

3.1 Density-Based Anomaly Detection for Mixed Data

Given a data instance $\boldsymbol{x}$, a principled density-based anomaly detection method is:

$$-\log P(\boldsymbol{x}) \geq \beta \tag{1}$$

for some predefined threshold β. Here $-\log P(\boldsymbol{x})$ serves as the anomaly scoring function. Gaussian mixture models (GMMs), for example, estimate the density using $P(\boldsymbol{x}) = \sum_k \alpha_k \mathcal{N}(\boldsymbol{x}; \mu_k, \Sigma_k)$ where $\{\alpha_k\}$ are mixing coefficients and $\mathcal{N}(\boldsymbol{x}; \mu_k, \Sigma_k)$ is a multivariate normal density.

When data is discrete or mixed-type, estimating $P(\boldsymbol{x})$ is highly challenging for several reasons. For discrete data, the space is exponentially large and there are usually no closed form expressions of density. Mixed data poses further challenges because since we need to model between-type correlation. For example, for two variables of different types x_1 and x_2, we need to specify either $P(x_1, x_2) = P(x_1)P(x_2 \mid x_1)$ or $P(x_1, x_2) = P(x_2)P(x_1 \mid x_2)$.

With this strategy, the number of pairs grows quadratically with the number of types. Most existing methods follow this approach and they are designed for a specific pair such as binary and Gaussian [15]. They neither scale to large-scale problems nor support arbitrary types such as binary, continuous, nominal, and count.

In Section 4 we will present a scalable solution based on Mixed-variate Restricted Boltzmann Machines (Mv.RBM) [44]. But first we briefly review its homogeneous version for binary data – the classic RBM.

3.2 Restricted Boltzmann Machines

Given binary input variables $\boldsymbol{x} \in \{0,1\}^N$ and hidden variables $\boldsymbol{h} \in \{0,1\}^K$, RBM defines the joint distribution [21]:

$$P(\boldsymbol{x}, \boldsymbol{h}) \propto \exp\left(\boldsymbol{a}'\boldsymbol{x} + \boldsymbol{b}'\boldsymbol{h} + \boldsymbol{h}'\boldsymbol{W}\boldsymbol{x}\right) \tag{2}$$

where $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{W}\}$ are model parameters. The posterior $P(\boldsymbol{h} \mid \boldsymbol{x})$ and data generative process $P(\boldsymbol{x} \mid \boldsymbol{h})$ in RBM are factorized as:

$$P(\boldsymbol{h} \mid \boldsymbol{x}) = \prod_k P(h_k \mid \boldsymbol{x}); \qquad P(\boldsymbol{x} \mid \boldsymbol{h}) = \prod_i P(x_i \mid \boldsymbol{h}) \tag{3}$$

Model estimation in RBM amounts to maximize data likelihood with respect to model parameters. It is typically done by n-step Contrastive Divergence (CD-n) [20], which is an approximate but fast method. In particular, for each parameter update, CD-n maintains a very short Mote Carlo Markov chain (MCMC), starting from the data, runs for n steps, then collects the samples to approximate data statistics. The MCMC is efficient because of the factorizations in Eq. (3), that is, we can sample all hidden variables in parallel through $\hat{\boldsymbol{h}} \sim P(\boldsymbol{h} \mid \boldsymbol{x})$ and all visible variables in parallel through $\hat{\boldsymbol{x}} \sim P(\boldsymbol{x} \mid \boldsymbol{h})$. For example, for Gaussian inputs, the parameters are updated as follows:

$$\begin{aligned} b_k &\leftarrow b_k + \eta\left(\bar{h}_{k|\boldsymbol{x}} - \bar{h}_{k|\hat{\boldsymbol{x}}}\right) \\ a_i &\leftarrow a_i + \eta\left(x_i - \hat{x_i}\right) \\ W_{ik} &\leftarrow W_{ik} + \eta\left(x_i \bar{h}_{k|\boldsymbol{x}} - \hat{x}_i \bar{h}_{k|\hat{\boldsymbol{x}}}\right) \end{aligned}$$

where $\bar{h}_{k|\boldsymbol{x}} = P(h_k = 1 \mid \boldsymbol{x})$ and $\eta > 0$ is the learning rate. This learning procedure scales linearly with n and data size.

4 Energy-Based Anomaly Detection

In this section, we present a scalable method for mixed-data anomaly detection based on Mixed-variate Restricted Boltzmann Machine (Mv.RBM) [44]. Mv.RBM estimates the data density of all types simultaneously. It bypasses the problems with detailed specifications and quadratic complexity by using latent binary variables, as outlined in Sec. 3.1. Here correlation between types is not modeled directly but is factored into indirect correlation with latent variables. As such we need only to model the correlation between a type and the latent binary. This scales linearly with the number of types.

Mv.RBM was primarily designed for data representation which transforms mixed data into a homogeneous representation, which serves as input for the next analysis stage. Our adaptation, on the other hand, proposes to use Mv.RBM as outlier detector directly, without going through the representation stage.

Func.	Binary	Gaussian	Nominal	Count
$E_i(x_i)$	$-a_i x_i$	$\frac{x_i^2}{2} - a_i x_i$	$-\sum_c a_{ic}\delta(x_i, c)$	$\log x_i! - a_i x_i$
$G_{ik}(x_i)$	$-W_{ik}x_i$	$-W_{ik}x_i$	$-\sum_c W_{ikc}\delta(x_i, c)$	$-W_{ik}x_i$

Table 1 Type-specific energy sub-functions. Here $\delta(x_i, c)$ is the identity function, that is, $\delta(x_i, c) = 1$ if $x_i = c$, and $\delta(x_i, c) = 0$ otherwise. For Gaussian, we assume data has unit variance. Multiple choices are modeled as multiple binaries.

4.1 Multi-Variate Restricted Boltzmann Machine

We first review Mv.RBM for a mixture of binary, Gaussian and nominal types, then extend to cover counts. See Fig. 1 for a graphical illustration. Mv.RBM is an extension of RBM for multiple data types. Let us start with classic RBM for binary data. We rewrite the joint distribution of RBM in Eq. (2) as follows:

$$P(\boldsymbol{x}, \boldsymbol{h}) \propto \exp\left(-E(\boldsymbol{x}, \boldsymbol{h})\right)$$

where $E(\boldsymbol{x}, \boldsymbol{h})$ is energy function of the following form:

$$E(\boldsymbol{x}, \boldsymbol{h}) = \sum_{i=1}^{N} E_i(x_i) + \sum_{k=1}^{K} \left(-b_k + \sum_{i=1}^{N} G_{ik}(x_i)\right) h_k \tag{4}$$

where $E_i(x_i) = -a_i x_i$ and $G_{ik}(x_i) = -W_{ik}x_i$.

Mv.RBM extends RBM by redefining the energy function to fit multiple data types. See Fig. 1 for a graphical illustration. The energy function of Mv.RBM differs from that of RBM by the using multiple type-specific energy sub-functions $E_i(x_i)$ and $G_{ik}(x_i)$ as listed[2] in Table 1. Here we extend the work of [44] to support counts by using Poisson distributions [36]:

$$E_i(x_i) = \log x_i! - a_i x_i; \qquad G_{ik}(x_i) = -W_{ik}x_i \tag{5}$$

The posterior $P(h_k \mid \boldsymbol{x})$ has the same form across types, that is, the activation probability $P(h_k = 1 \mid \boldsymbol{x})$ is sigmoid $\left(b_k - \sum_i G_{ik}(x_i)\right)$. On the other hand, the generative process is type-specific. For example, for binary data, the activation probability $P(x_i = 1 \mid \boldsymbol{h})$ is sigmoid $\left(a_i + \sum_k W_{ik}h_k\right)$; and for Gaussian data, the conditional density $P(x_i \mid \boldsymbol{h})$ is $\mathcal{N}\left(a_i + \sum_k W_{ik}h_k; \mathbf{1}\right)$.

Learning in Mv.RBM is almost identical to that of RBM, as described in Sec. 3.2. The only difference is the generative distribution $P(\boldsymbol{x} \mid \boldsymbol{h})$ is the product of mixed-type distributions, i.e., $P(\boldsymbol{x} \mid \boldsymbol{h}) = \prod_i P_i(x_i \mid \boldsymbol{h})$ for $P_i(x_i \mid \boldsymbol{h})$ is type-specific.

4.2 Mv.RBM for Anomaly Detection

In Mv.RBM, types are not correlated directly but through the common hidden layer. Here types are conditionally independent given $\boldsymbol{h}$, but since $\boldsymbol{h}$ are hidden, types are dependent as in $P(\boldsymbol{x}) = \sum_{\boldsymbol{h}} P(\boldsymbol{x}, \boldsymbol{h})$. Since $\boldsymbol{h}$ is discrete, Mv.RBM can be considered as a mixture model of 2^K components that shared the same parameter. This suggests that Mv.RBM can be used for outlier detection in the same way that GMM does (e.g., see Sec. 3.1), but it may fit data much better. Data density can be rewritten as:

[2] The original Mv.RBM also covers rank, but we do not consider in this paper.

$$P(\boldsymbol{x}) \propto \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{x}, \boldsymbol{h})\right) = \exp\left(-F(\boldsymbol{x})\right)$$

Here $F(\boldsymbol{x}) = -\log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{x}, \boldsymbol{h})\right)$ is known as *free-energy*, which can be estimated as

$$F(\boldsymbol{x}) = \sum_i E_i(x_i) - \sum_k \log\left(1 + \exp\left(b_k - \sum_i G_{ik}(x_i)\right)\right) \tag{6}$$

Notice that the free-energy equals the negative log-density up to an additive constant:

$$F(\boldsymbol{x}) = -\log P(\boldsymbol{x}) + \text{constant}$$

Thus *we can use the free-energy as the anomaly score* to rank data instances, following the detection rule in Eq. (1). Importantly, the free-energy can be computed in linear time.

4.2.1 Remark

To sum up, Mv.RBM, coupled with free-energy, offers a disciplined approach to mixed-type anomaly detection that meet three desirable criteria:

- capturing correlation structure between types by factoring via a binary hidden layer,
- measuring deviation from the norm using free-energy in Eq. (6), and
- anomaly scores are efficient to compute with just linear complexity in Eq. (6).

A major challenge of unsupervised outlier detection is the phenomenon of *swamping effect*, where an inlier is misclassified as outlier, possibly due a large number of true outliers in the data [37]. When data models are highly expressive – such as large RBMs and Mv.RBMs – outliers are included by the models as if they have patterns themselves, even if these patterns are weak and differ significantly from the regularities of the inliers. One way to control the model expressiveness is to limit the number of hidden layers K (hence the number of mixing components 2^K).

5 Detecting Anomalies Across Data Abstraction Levels

Real world data may be high-dimensional, noisy and redundant. Low level data may also hide anomalies. For example, two images may have similar intensity histograms at the pixel level, but have very different semantics. These challenges suggest *data abstraction* as a preprocessing step. Indeed, abstraction is a powerful tool for several reasons. First, abstraction may uncover the inherent low-dimensional manifold, and thus eliminating redundancy and reducing noises in the raw data. Second, abstraction may reveal regular patterns, making it easier to single out irregularities.

In the previous sections, we have shown that the RBM family models the mechanism to generate data through the hidden layer. The posterior $P(\boldsymbol{h} \mid \boldsymbol{x})$ serves as data abstraction. We can abstract the data further by using samples from the posterior $\hat{\boldsymbol{h}} \sim P(\boldsymbol{h} \mid \boldsymbol{x})$ as input for the next RBM. This process can be repeated to uncover multilevels of data abstraction. This is essentially the procedure of producing Deep Belief Networks (DBNs) [21].

With Mv.RBM at the first level to model mixed data, mixed-variate DBNs (Mv.DBNs) can be built in the same way. In this section, we present a solution of using Mv.DBNs

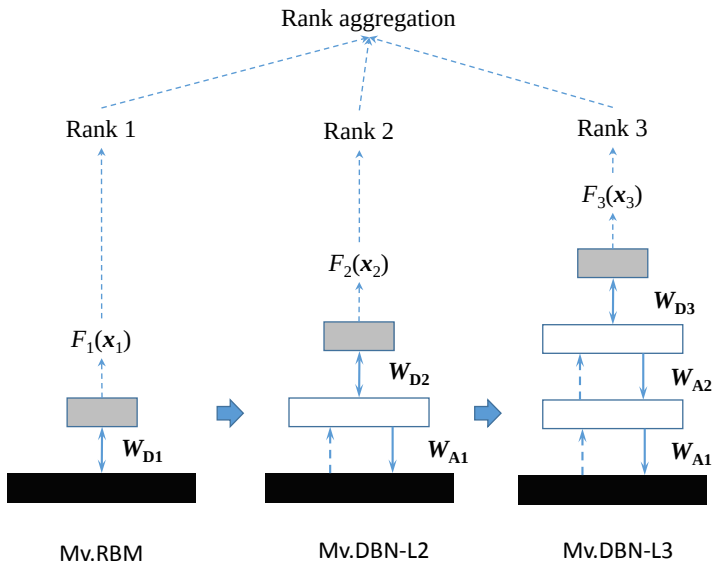


Fig. 2 `MIXMAD`: MIXed data Multilevel Anomaly Detection based on Mixed-variate RBM (Mv.RBM) and successive Mix-variate DBNs (Mv.DBN). Mv.DBNs are "grown" sequentially (left to right), with abstraction layer inserted. Filled boxes represent data input, empty boxes represent abstraction layers, and shaded boxes represent the hidden layer of the detection Mv.RBM/RBM.

for anomaly detection called `MIXMAD`, which stands for **MIX**ed data **M**ultilevel **A**nomaly **D**etection. `MIXMAD` consists of **(a)** Mv.RBM [44], and **(b)** several Mixed-variate DBNs of different depths built on top of Mv.RBM. The Mv.RBM converts mixed data into homogeneous representation of binary variables. The subsequent RBMs operate on binary variables. Subsequent Mv.DBNs reuse parameters learnt by the previous Mv.DBNs. The Mv.RBM and Mv.DBNs are all density-based anomaly detectors following Eq. (1). Each detector assigns an anomaly score for each data instance. The score is used to rank data instances. All rank lists are then aggregated to produce a single rank list of anomalies. See Fig. 2 for a graphical illustration. For each Mv.DBN, lower Mv.RBMs/RBMs are feature extractors and the top RBM is used as anomaly detector. In subsequent subsections, we describe the components of `MIXMAD` in more detail.

5.1 Multilevel Detection Procedure With Mv.RBM/DBN Ensemble

The main idea is to recognize that the RBM at the top of the Mv.DBN operates on data abstraction $\boldsymbol{x}_L$, and the RBM's prior density $P_L(\boldsymbol{x}_L)$ can replace $P(\boldsymbol{x})$ in Eq. (1). Recall that the input $\boldsymbol{x}_l$ to the intermediate RBM at level l is the abstraction of the lower level data as:

$$\boldsymbol{x}_l \sim \text{Bernoulli}\left(\sigma\left(\boldsymbol{b}_{l-1} + \boldsymbol{W}_{l-1}\boldsymbol{x}_{l-1}\right)\right) \tag{7}$$

The prior density $P_L(\boldsymbol{x}_L)$ can be rewritten as:

$$P(\boldsymbol{x}_L) \propto \exp(-F_L(\boldsymbol{x}_L)) \tag{8}$$

where

$$F_L(\boldsymbol{x}_L) = -\boldsymbol{b}_L^{'}\boldsymbol{x}_L - \sum_k (1 + \log(\boldsymbol{a}_{Lk} + \boldsymbol{W}_{Lk}\boldsymbol{x}_L)) \tag{9}$$

This free energy, like the case of Mv.RBM in Eq. (6) can also be used as an anomaly score of abstracted data, and the anomaly region is defined as:

$$\mathcal{R} \in \{\boldsymbol{x} \mid F_L(\boldsymbol{x}_L) \geq \beta\}$$

Once the Mv.DBN has been trained, the free-energy can be approximated by a deterministic function, where the intermediate input $\boldsymbol{x}_l$ in Eq. (7) is replaced by:

$$\boldsymbol{x}_l = \sigma(\boldsymbol{b}_{l-1} + \boldsymbol{W}_{l-1}\boldsymbol{x}_{l-1}) \tag{10}$$

Given that we can have multiple free-energies across levels of the Mv.DBNs, there are multiple anomaly scores per data instance. Each score reflects the nature of abnormality at the corresponding level of abstraction. This suggests that we can combine anomaly detection across levels. This appears to resemble the idea of ensemble approach [2], but differing from how the ensemble is constructed. Since free-energies differ across levels, direct combination of free-energies is not possible. A sensible approach is through rank aggregation, that is, the free-energies at each level are first used to rank instances from the lowest to the highest energy. The rank now serves as an anomaly score. The decision threshold is determined by the percentile $\alpha \in (0, 100)$, which is user-defined.

5.1.1 p-norm Rank Aggregation

Ideally optimal rank aggregation minimizes the disagreement with all ranks. The minimization requires searching through a permutation space of size $n!$ for n instances, which is intractable. However, we need not to care about optimal rank of all instances. In our application, we pay more attention to the a small portion of data at the top. Denote by $s_{li} \geq 0$ the anomaly score of instance i at level l. We propose to use the following p-norm aggregation:

$$\bar{s}_i(p) = \left(\sum_{l=1}^{L} s_{li}^p\right)^{1/p} \tag{11}$$

where $p > 0$ is a tuning parameter.

There are two detection regimes under this aggregation scheme. The detection at $p < 1$ is *conservative*, that is, individual high outlier scores are suppressed in favor of a consensus. The other regime is *optimistic* at $p > 1$, where the top anomaly scores tend to dominate the aggregation. This aggregation subsumes several methods as special cases: $p = 1$ reduces to the classic Borda count when s_{li} is rank position; $p = \infty$ reduces to the max: $\lim_{p\to\infty} \bar{s}_i(p) = \max_l \{s_{li}\}$.

Input: data $\mathcal{D} = \{\boldsymbol{x}\}$; **Output**: Anomaly rank.
User-defined parameters: depth L, hidden sizes $\{K_1, K_2, ..K_L\}$, and p.

1. Set $\boldsymbol{x}_1 \leftarrow \boldsymbol{x}$.
2. For each level $l = 1, 2, ..., L$:
   (a) Train a *detection RBM* (or *Mv.RBM* if $l = 1$) on $\boldsymbol{x}_l$ with K_L hidden units;
   (b) Estimate free-energy $F_l(\boldsymbol{x}_l)$ using Eqs. (9,10);
   (c) Rank data according to $F_l(\boldsymbol{x}_l)$;
   (d) If $l < L$
      i. Train an *abstraction RBM* on $\boldsymbol{x}_l$ with K_l hidden units;
      ii. Abstracting data using Eq. (7) to generate $\boldsymbol{x}_{l+1}$;
3. Aggregate ranks using p-norm in Eq. (11).

Fig. 3 Multilevel anomaly detection algorithm.

5.1.2 Separation of Abstraction and Detection

Recall from that we use RBMs for both abstraction (Eq. (7)) and anomaly detection (Eq. (9)). Note that data abstraction and anomaly detection are different goals – abstraction typically requires more bits to adequately disentangle multiple factors of variation [6], while detection may require less bits to estimate a rank score. Fig. 3 presents the multilevel anomaly detection algorithm. It trains one RBM and $(L-1)$ Mv.DBNs of increasing depths – from 2 to L – with time complexity linear in L. They produces L rank lists, which are then aggregated using Eq. (11).

5.2 Complexity analysis

`MIXMAD` offers not only a principled method for anomaly detection but also computation advantage. Recall that in the case of single-layer Mv.RBM described in Sec. 3.1, the computational complexity is linear in number of dimensions N, number of hidden units K, and number of data points n, i.e., $\mathcal{O}(nNK)$. The `MIXMAD` is a stack of Mv.RBM at the bottom and several binary RBMs at the upper layers. In practice, we choose $K_1 \ll N$ (to prevent the swamping effect as discussed in Sec. 4.2.1) for the bottom layer, and $K_{l+1} \leq K_l$ for upper layers $l = 1, .., L$; and thus adding more layers only introduces a small multiplicative constant in computational complexity.

Training typically stops at 100 epochs, regardless of the data model and data size. To estimate the cut-off threshold, the rank operations add $\mathcal{O}(n \log n)$ computation steps. These are scalable to high dimensional settings and large datasets. For example, the training over KDD data matrix of size $75,669 \times 41$ costs about 0.3s/epoch for Mv.RBM and about 4.2s/epoch for `MIXMAD` with $L = 3$ on one GPU GeForce GTX 980Ti 12GB. Testing on $32,417$ data points only take several seconds.

6 Experiments

This section reports experiments and results of the proposed energy-based methods on a comprehensive suite of datasets. We first present the cases for single data type in Section 6.1, then for mixed data in Section 6.2.

	Dims	#train	#test	%anomaly
MNIST	784	3,000	1,023	4.9
InternetAds	174	1,682	1,682	5.0
Preterm (37wks)	369	3,000	5,104	10.9
Preterm (34wks)	369	3,000	5,104	6.5

Table 2 Data statistics.

6.1 Homogeneous Data

6.1.1 Data

We use three *high-dimensional* real-world datasets with very different characteristics: *handwritten digits (MNIST)*, *Internet ads* and *clinical records of birth episodes.*

- The *MNIST* has $60,000$ gray images of size 28×28 for training and $10,000$ images for testing[3]. The raw pixels are used as features (784 dimensions). Due to ease of visualization and complex data topology, this is an excellent data for testing anomaly detection algorithms. We use digit '8' as normal and a small portion (˜5%) of other digits as anomalies. This proves to be a challenging digit compared to other digits – see Fig. 4 (left) for failure of pixel-based k-nearest neighbor methods. We randomly pick 3,000 training images and keep all the test set.
- The second dataset is *InternetAds* with 5% anomaly injection as described in [10]. As the data size is moderate (1,682 instances, 174 features), no train/test splitting is used.
- The third dataset consists of *birth episodes* collected from an urban hospital in Sydney, Australia in the period of 2011–2015 [45]. Preterm births are anomalies that have a critical impact on the survival and development of the babies. In general, births occurring within 37 weeks of gestation are considered preterm. We are also interested in early preterm births, e.g., those occurring with 34 weeks of gestation. This is because the earlier the birth, the more severe the case, leading to more intensive care. Features include 369 clinically relevant facts collected in the first few visits to hospital before 25 weeks of gestation. The data is randomly split into a training set of 3,000 cases, and a test set of 5,104 cases.

All data are normalized into the range [0,1], which is known to work best in [10]. Data statistics are reported in Table 2.

6.1.2 Baselines

We compare the proposed method against four popular unsupervised baselines – *k-NN*, *PCA*, *Gaussian mixture model (GMM)*, and *one-class SVM (OCSVM)* [11]. (a) The k-NN uses the mean distance from a test case to the k nearest instances as anomaly score [4]. We set $k = 10$ with Euclidean distance. (b) For PCA, $\alpha\%$ total energy is discarded, where α is the estimated anomaly rate in training data. The reconstruction error using the remaining eigenvectors is used as the anomaly score. (c) The GMMs have four clusters and are regularized to work with high dimensional data. The negative log-likelihood serves as anomaly score. (d) The OCVSMs have RBF kernels with automatic scaling. We also consider RBM [17] as baseline, which is a special case of Mv.RBM when data is all binary.

[3] http://yann.lecun.com/exdb/mnist/

Param.	*MNIST*	*InternetAds*	*Preterm*
K_D	10	10	10
K_A	70	50	70
N	784	174	369

Table 3 Settings of the MAD. K_D is the number of hidden units in the detection RBM, K_A in the abstraction RBMs, and N is data dimensions..

We use the following evaluation measures: *Area Under ROC Curve (AUC)*, and *NDCG@T*. The AUC reflects the average discrimination power across the entire dataset, while the NDCG@T places more emphasis on the top retrieved cases.

6.1.3 MAD Implementation

Abstraction RBMs have the same number of hidden units while detection RBM usually have smaller number of hidden units. All RBMs are trained using CD-1 [20] with batch size of 64, learning rate of 0.3 and 50 epochs. Table 3 list model parameters used in experimentation.

6.1.4 Results

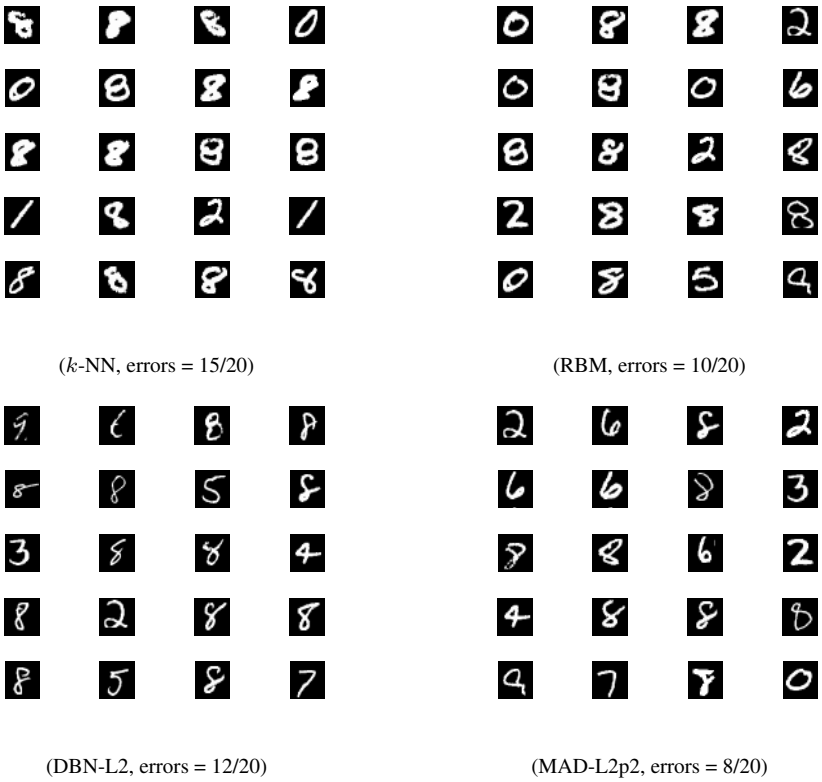
To see how MAD (Multilevel Anomaly Detection) combines evidences from detectors in the ensemble, we run the algorithms (RBM, the DBN with 2 layers, and the MAD that combines RBM and DBN results). Fig. 4 plots detection by the RBM/DBN/MAD against the classic k-NN on the MNIST dataset. k-NN fails 15 out of 20 cases, mostly due to the variation in stroke thickness, which is expected for matching based on raw pixels. RBM and DBN have different errors, confirming that anomalies differ among abstraction levels. Finally, the ensemble of RBM/DBN, then MAD improves the detection significantly. The error is mostly due to the high variation in styles (e.g., and 8 with open loops).

Table 4 reports the Area Under the ROC Curve (AUC) for all methods and datasets. Overall MAD with 2 or 3 hidden layers works well. The difference between the baselines and MAD is amplified in the NDCG measure, as shown in Table 5. One possible explanation is that the MAD is an ensemble – an anomaly is considered anomaly if it is detected by all detectors at different abstraction levels. One exception is the max-aggregation (where $p \to \infty$ in Eq. (11)), where the detection is over-optimistic.

6.2 Mixed Data

We now present experiments on synthetic and real-world mixed data. For comparison, we implement well-known single-type anomaly detection methods including Gaussian mixture model (GMM), Probabilistic Principal Component Analysis (PPCA) and one-class SVM (OCSVM). The number of components of PPCA model is set so that the discarded energy is the same as the anomaly rate in training data. For OCSVM, we use radial basis kernel with $\nu = 0.7$. GMM and PPCA are probabilistic, and thus data log-likelihood can be computed for anomaly detection.

Since all of these single-type methods assume numerical data, we code nominal types using dummy binaries. For example, a *A* in the nominal set {*A,B,C*} is coded as (1,0,0) and *B* as (0,1,0). This coding causes some nominal information loss, since the coding does not ensure that only one value is allowed in nominal variables. For all methods, the detection



(k-NN, errors = 15/20) (RBM, errors = 10/20)

(DBN-L2, errors = 12/20) (MAD-L2p2, errors = 8/20)

Fig. 4 Anomaly detection on MNIST test set for the top 20 digits. Normal digit is "8". MAD stands for Multilevel Anomaly Detection.

Method	*MNIST*	*InternetAds*	*Preterm (37wks)*	*Preterm (34wks)*
k-NN	0.804	0.573	0.596	0.624
PCA	0.809	0.664	0.641	0.673
GMM	0.839	0.725	0.636	0.658
OCSVM	0.838	0.667	0.646	0.676
RBM	0.789	0.712	0.648	0.677
MAD-L2p.5	**0.867**	**0.829**	0.627	**0.729**
MAD-L2p1	**0.880**	**0.827**	**0.645**	**0.748**
MAD-L2p2	**0.897**	**0.816**	**0.661**	**0.761**
MAD-L2p∞	**0.892**	**0.765**	**0.660**	**0.745**
MAD-L3p.5	0.787	**0.789**	**0.674**	**0.757**
MAD-L3p1	0.814	**0.775**	**0.689**	**0.765**
MAD-L3p2	**0.847**	**0.758**	**0.685**	**0.759**
MAD-L3p∞	**0.876**	**0.734**	**0.668**	**0.742**

Table 4 The Area Under the ROC Curve (AUC). L is the number of hidden layers, p is the aggregation parameter in Eq. (11), bold indicate better performance than baselines. Note that RBM is the limiting case of MAD with $L = 1$.

Method	*MNIST*	*InternetAds*	*Preterm (37wks)*	*Preterm (34wks)*
k-NN	0.218	0.413	0.362	0.188
PCA	0.488	0.225	0.505	0.356
GMM	0.458	0.415	0.438	0.223
OCSVM	0.423	0.094	0.471	0.172
RBM	0.498	0.421	0.429	0.216
MAD-L2p.5	**0.666**	**0.859**	**0.945**	**0.831**
MAD-L2p1	**0.667**	**0.859**	**0.945**	**0.831**
MAD-L2p2	**0.666**	**0.859**	**0.945**	**0.831**
MAD-L2p∞	**0.536**	0.271	**0.741**	**0.576**
MAD-L3p.5	**0.732**	**0.908**	**0.798**	**0.625**
MAD-L3p1	**0.732**	**0.908**	**0.798**	**0.626**
MAD-L3p2	**0.732**	**0.902**	**0.769**	**0.597**
MAD-L3p∞	0.360	**0.598**	0.370	0.113

Table 5 The NDCG@20. L is the number of hidden layers, p is the aggregation parameter in Eq. (11), bold indicate better performance than baselines. Note that RBM is the limiting case of MAD with $L = 1$.

threshold is based on the α percentile of the training anomaly scores. Whenever possible, we also include results from other recent mixed-type papers, ODMAD [26], Beta mixture model (BMM) [8] and GLM-t [28]. We followed the same mechanism they used to generate anomalies.

6.2.1 Synthetic Data

We first evaluate the behaviors of Mv.RBM on synthetic data with controllable complexity. We simulate mixed-type data using a generalized Thurstonian theory, where Gaussians serve as underlying latent variables for observed discrete values. Readers are referred to [43] for a complete account of the theory. For this study, the underlying data is generated from a GMM of 3 mixing components with equal mixing probability. Each component is a multivariate Gaussian distributions of 15 dimensions with random mean and positive-definite covariance matrix. From each distribution, we simulate 1,000 samples, creating a data set size 3,000. To generate anomalies, we randomly pick 5% of data, and add uniform noise to each dimension, i.e., $x_i \leftarrow x_i + e_i$ where $e_i \sim \mathcal{U}$. For visualization, we use t-SNE to reduce the dimensionality to 2 and plot the data in Fig. 5.

Out of 15 variables, 3 are kept as Gaussian and the rest are used to create mixed-type variables. More specifically, 3 variables are transformed into binaries using random thresholds, i.e., $\tilde{x}_i = \delta(x_i \geq \theta_i)$. The other 9 variables are used to generate 3 nominal variables of size 3 using the rule: $\tilde{x}_i = \arg\max(x_{i1}, x_{i2}, x_{i3})$.

Models are trained on 70% data and tested on the remaining 30%. This testing scheme is to validate the generalizability of models on unseen data. The learning curves of Mv.RBM are plotted in Fig. 6. With the learning rate of 0.05, learning converges after 10 epochs. No overfitting occurs.

The decision threshold β in Eq. (1) is set at 5 percentile of the training set. Fig. 7 plots the anomaly detection performance of Mv.RBM (in F-score) on test data as a function of model size (number of hidden units). To account for random initialization, we run Mv.RBM 10 times and average the F-scores. It is apparent that the performance of Mv.RBM is competitive against that of GMM. The best F-score achieved by GMM is only about 0.35, lower than the worst F-score by Mv.RBM, which is 0.50. The PCA performs poorly, with F-score of 0.11, possibly because the anomalies does not conform to the notion of residual subspace assumed by PCA.

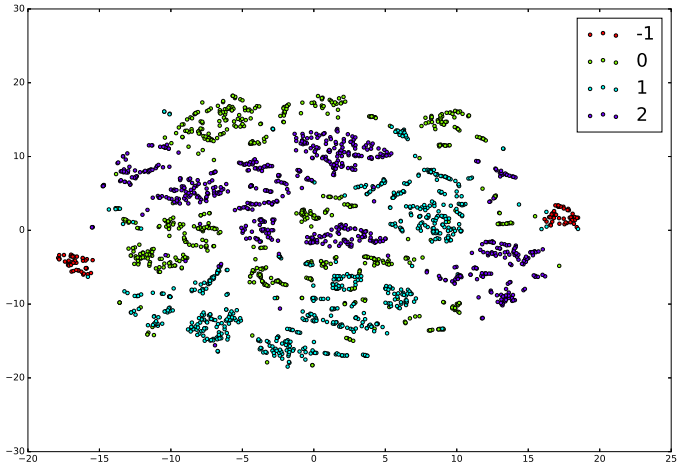


Fig. 5 Synthetic data with 3 normal clusters (cluster IDs 0,1,2) and 1 set of scattered anomalies (ID: -1, colored in red). Best viewed in color.

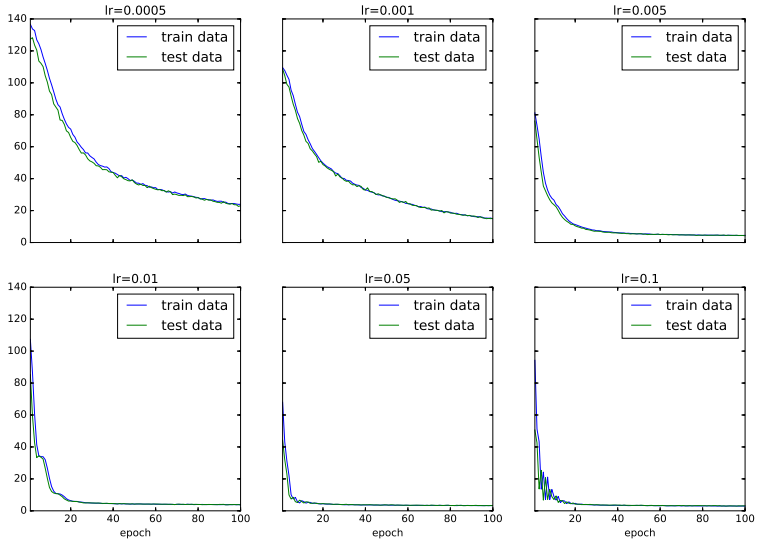


Fig. 6 Learning curves of Mv.RBM (50 hidden units) on synthetic data for different learning rates. The training and test curves almost overlap, suggesting no overfitting. Best viewed in color.

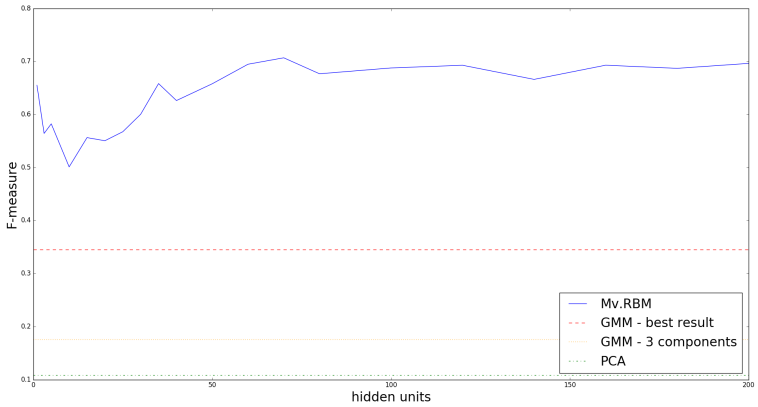


Fig. 7 Performance of Mv.RBM in F-score on synthetic data as a function of number of hidden units. Horizontal lines are performance measures of PCA (in green) and GMM (best result in red; and with 3 components in yellow). Best viewed in color.

The performance difference between Mv.RBM and GMM is significant considering the fact that the underlying data distribution is drawn from a GMM. It suggests that when the correlation between mixed attributes is complex like this case, using GMM even with the same number of mixture components cannot learn well. Meanwhile, Mv.RBM can handle the mixed-type properly, *without knowing the underlying data assumption*. Importantly, varying the number of hidden units does not affect the result much, suggesting the stability of the model and it can free users from carefully crafting this hyper-parameter.

6.2.2 Real Data

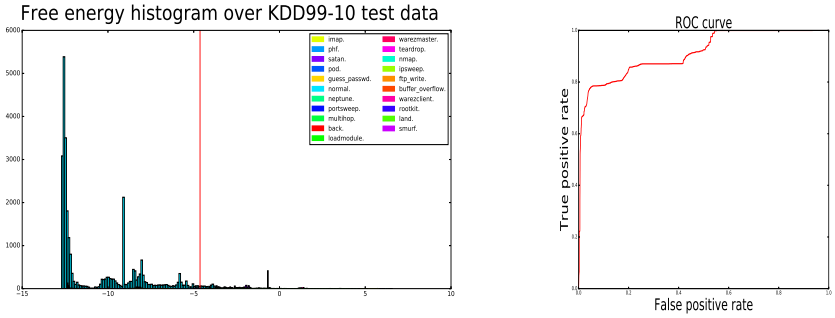
For real-world applications, we use a wide range of mixed-type datasets. From the UCI repository[4], we select 7 datasets which were previously used as benchmarks for mixed-type anomaly detection [8, 19, 28]. Data statistics are reported in Table 6. We generate anomalies by either using rare classes whenever possible, or by randomly injecting a small proportion of anomalies, as follows:

- **Using rare classes**: For the KDD99 *10 percent* dataset (KDD99-10), intrusions (anomalies) account for 70% of all data, and thus it is not possible to use full data because anomalies will be treated as normal in unsupervised learning. Thus, we consider all normal instances from the original data as inliers, which accounts for 90% of the new data. The remaining 10% anomalies are randomly selected from the original intrusions.
- **Anomalies injection**: For the other datasets, we treat data points as normal objects and generate anomalies based on a contamination procedure described in [8, 28]. anomalies are created by randomly selecting 10% of instances and modifying their default values. For numerical attributes (Gaussian, Poisson), values are shifted by 2.0 to 3.0 times standard deviation. For discrete attributes (binary, categorical), the values are switched to alternatives.

[4] https://archive.ics.uci.edu/ml/datasets.html

Dataset	No. Instances		No. Attributes				
	Train	Test	Bin.	Gauss.	Nominal	Poisson	Total
KDD99-10	75,669	32,417	4	15	3	19	41
Australian Credit	533	266	3	6	5	0	14
German Credit	770	330	2	7	11	0	20
Heart	208	89	3	6	4	0	13
Thoracic Surgery	362	155	10	3	3	0	16
Auto MPG	303	128	0	5	3	0	8
Contraceptive	1136	484	3	0	4	1	8

Table 6 Characteristics of mixed-type datasets. The proportion of anomalies are 10%.



(a) Free-energy histogram (b) ROC curve

Fig. 8 anomaly detection on the KDD99-10 dataset. (a) Histogram of free-energies. The vertical line separates data classified as inliers (left) from those classified as anomalies (right). The color of majority (light blue) is inlier. Best viewed in color. (b) ROC curve (AUC = 0.914).

Numerical attributes are standardized to zero means and unit variance. For evaluation, we randomly select 30% data for testing, and 70% data for training. Note that since learning is unsupervised, anomalies must also be detected in the training set since there are no ground-truths. The anomalies in the test set is to test the generalizability of the models to unseen data.

6.2.3 Models setup

The number of hidden units in Mv.RBM is set to $K = 2$ for the KDD99-10 dataset, and to $K = 5$ for other datasets. The parameters of Mv. RBM are updated using stochastic gradient descent, that is, update occurs after every mini-batch of data points. For small datasets, the mini-batch size is equal to the size of the entire datasets while for KDD99-10, the mini-batch size is set to 500. The learning rate is set to 0.01 for all small datasets, and to 0.001 for KDD99-10. Small datasets are trained using momentum of 0.8. For KDD99-10, we use Adam [24], with $\beta_1 = 0.85$ and $\beta_2 = 0.995$. For small datasets, the number of mixture components in GMM is chosen using grid search in the range from 1 to 30 with a step size of 5. For KDD99-10, the number of mixture components is set to 4.

	KDD	AuCredit	GeCredit	Heart	ThSurgery	AMPG	Contra.
GMM (*)	0.42	0.74	0.86	0.89	0.71	1.00	0.62
OCSVM (*)	0.54	0.84	0.86	0.76	0.71	1.00	0.84
PPCA (*)	0.55	0.38	0.02	0.64	0.70	0.67	0.02
BMM	–	0.97	0.93	0.87	0.94	0.62	0.67
ODMAD	–	0.94	0.81	0.63	0.88	0.57	0.52
GLM-t	–	–	–	0.72	–	0.64	–
Mv.RBM	0.71	0.90	0.95	0.94	0.90	**1.00**	0.91
MIXMAD-L2p0.5	0.72	0.93	**0.97**	0.94	**0.97**	**1.00**	**0.95**
MIXMAD-L2p1	0.72	0.93	0.95	0.94	**0.97**	**1.00**	**0.95**
MIXMAD-L2p2	0.69	0.93	**0.97**	0.94	**0.97**	**1.00**	**0.95**
MIXMAD-L2p∞	0.69	0.73	**0.97**	**1.00**	**0.97**	**1.00**	**0.95**
MIXMAD-L3p0.5	**0.73**	**0.98**	**0.97**	0.94	**0.97**	0.70	**0.95**
MIXMAD-L3p1	0.72	**0.98**	**0.97**	0.94	**0.97**	0.70	**0.95**
MIXMAD-L3p2	0.71	**0.98**	**0.97**	0.94	**0.97**	0.70	**0.95**
MIXMAD-L3p∞	0.50	0.78	**0.97**	0.94	**0.97**	0.57	**0.95**

Table 7 Anomaly detection F-score on mixed data. (*) baseline single-type methods worked on coded data.

6.2.4 Results

Fig. 8(a) shows a histogram of free-energies computed using Eq. (6) on the KDD99-10 dataset. The inliers/anomalies are well-separated into the low/high energy regions, respectively. This is also reflected in an Area Under the ROC Curve (AUC) of 0.914 (see Fig. 8(b)).

The detection performance in term of F-score on test data is reported in Tables 7. The mean of all single type scores is 0.66, of all competing mixed-type scores is 0.77, and of Mv.RBM scores is 0.91. These demonstrate that (a) a proper handling of mixed-types is required, and (b) Mv.RBM is highly competitive against other mixed-type methods for anomaly detection. Point (a) can also be strengthened by looking deeper: On average, the best competing mixed-type method (BMM) is better than the best single-type method (OCSVM). For point (b), the gaps between Mv.RBM and other methods are significant: On average, Mv.RBM is better than the best competing method – the BMM (mixed-type) – by 8.3%, and better than the worst method – the PPCA (single type), by 111.6%. On the largest dataset – the KDD99-10 – Mv.RBM exhibits a significant improvement of 29.1% over the best single type method (PPCA).

7 Discussion

As an evidence to the argument in Section 4.2.1 about separating the abstraction and detection RBMs, we found that the sizes of the RBMs that work well on the MNIST do not resemble those often found in the literature (e.g., see [21]). For example, typical numbers of hidden units range from 500 to 1,000 for a good generative model of digits. However, we observe that 10 to 20 units for detection RBMs and 50-100 units for abstraction RBMs work well in our experiments, regardless of the training size. This suggests that the number of bits required for data generation is higher than those required for anomaly detection. This is plausible since accurate data generation model needs to account for all factors of variation and a huge number of density modes. On the other hand, anomaly detection model needs only to identify *low density regions* regardless of density modes. An alternative explanation is that since the CD-1 procedure used to train RBMs (see Section 3.2) creates *deep energy wells around each*

data points, an expressive model may lead to more false alarms. Thus, the smoothness of the energy surface may play an important role in anomaly detection. Our `MIXMAD` algorithm offers a consensus among multiple energy surface, and thus can be considered as a way to mitigate the energy wells issue.

The construction procedure of DBN has been proved to be equivalent to the variational renomarlization groups in physics [30]. In particular, with layerwise construction, the data is rescaled – the higher layer operates on a coarser data representation. This agrees with our initial motivation for the `MIXMAD`. Finally, although not implemented here, the `MIXMAD` lends itself naturally to detecting anomalies in *multimodal data* with diverse modal semantics. For example, an image can be equipped with high-level tags and several visual representations. Each data representation can be modelled as a Mv.RBM at the right level of abstraction. The upper RBMs then integrate all information into coherent representations.

Conclusion

This paper has introduced a new energy-based method for mixed-data anomaly detection based on Mixed-variate Restricted Boltzmann Machine (Mv.RBM) and Deep Belief Networks. Mv.RBM avoids direct modeling of correlation between types by using binary latent variables, and in effect, models the correlation between each type and the binary type. We derived free-energy, which equals the negative log of density up-to a constant, and use it as the anomaly score. We then generalized the idea to detecting anomalies via multilevel abstraction. We introduced `MIXMAD`, a procedure to train a sequence of Deep Belief Networks, each of which provides a ranking of anomalies. All rankings are then aggregated through a simple p-norm trick. Overall, the method is highly scalable – the computational complexity grows linearly with number of types. Our experiments on mixed-type datasets of various types and characteristics demonstrate that the proposed method is competitive against the well-known baselines designed for single types, and recent models designed for mixed-types. These results (a) support the hypothesis that in mixed-data, proper modeling of types should be in place for anomaly detection, and (b) show Mv.RBM is a powerful density-based anomaly detection method, and (c) learning data representation through multilevel abstraction is a sensible strategy for high-dimensional settings; and (d) `MIXMAD` is a competitive method.

Acknowledgments

This work is partially supported by the Telstra-Deakin Centre of Excellence in Big Data and Machine Learning.

References

1. Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pages 420–434. Springer, 2001.
2. Charu C Aggarwal and Saket Sathe. Theoretical foundations and algorithms for outlier ensembles. *ACM SIGKDD Explorations Newsletter*, 17(1):24–47, 2015.
3. Leman Akoglu, Hanghang Tong, Jilles Vreeken, and Christos Faloutsos. Fast and reliable anomaly detection in categorical data. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 415–424. ACM, 2012.

4. Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 15–27. Springer, 2002.
5. John Becker, Timothy C Havens, Anthony Pinar, and Timothy J Schulz. Deep belief networks for false alarm rejection in forward-looking ground-penetrating radar. In *SPIE Defense+ Security*, pages 94540W–94540W. International Society for Optics and Photonics, 2015.
6. Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.
7. Loïc Bontemps, James McDermott, Nhien-An Le-Khac, et al. Collective anomaly detection based on long short-term memory recurrent neural networks. In *International Conference on Future Data and Security Engineering*, pages 141–152. Springer, 2016.
8. Mohamed Bouguessa. A practical outlier detection approach for mixed-attribute data. *Expert Systems with Applications*, 42(22):8637–8649, 2015.
9. Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.
10. Guilherme O Campos, Arthur Zimek, Jörg Sander, Ricardo JGB Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, pages 1–37, 2015.
11. Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.
12. Sucheta Chauhan and Lovekesh Vig. Anomaly detection in ecg time signals via deep long short-term memory networks. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–7. IEEE, 2015.
13. Min Cheng, Qian Xu, Jianming Lv, Wenyin Liu, Qing Li, and Jianping Wang. Ms-lstm: A multi-scale lstm model for bgp anomaly detection. In *Network Protocols (ICNP), 2016 IEEE 24th International Conference on*, pages 1–6. IEEE, 2016.
14. Kaustav Das, Jeff Schneider, and Daniel B Neill. Anomaly pattern detection in categorical datasets. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–176. ACM, 2008.
15. Alexander R De Leon and Keumhee Carrière Chough. *Analysis of Mixed Data: Methods & Applications*. CRC Press, 2013.
16. Kien Do, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Outlier detection on mixed-type data: An energy-based approach. *International Conference on Advanced Data Mining and Applications (ADMA 2016)*, 2016.
17. Ugo Fiore, Francesco Palmieri, Aniello Castiglione, and Alfredo De Santis. Network anomaly detection with the restricted Boltzmann machine. *Neurocomputing*, 122:13–23, 2013.
18. Ni Gao, Ling Gao, Quanli Gao, and Hai Wang. An intrusion detection model based on deep belief networks. In *Advanced Cloud and Big Data (CBD), 2014 Second International Conference on*, pages 247–252. IEEE, 2014.
19. Amol Ghoting, Matthew Eric Otey, and Srinivasan Parthasarathy. Loaded: Link-based outlier and anomaly detection in evolving data sets. In *ICDM*, pages 387–390, 2004.
20. G.E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.
21. G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
22. Dino Ienco, Ruggero G Pensa, and Rosa Meo. A semisupervised approach to the detection and characterization of outliers in categorical data. *IEEE transactions on neural networks and learning systems*, 2016.
23. Hanna Kamyshanska and Roland Memisevic. The potential energy of an autoencoder. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(6):1261–1273, 2015.
24. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
25. Anna Koufakou and Michael Georgiopoulos. A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. *Data Mining and Knowledge Discovery*, 20(2):259–289, 2010.
26. Anna Koufakou, Michael Georgiopoulos, and Georgios C Anagnostopoulos. Detecting outliers in high-dimensional datasets with mixed attributes. In *DMIN*, pages 427–433. Citeseer, 2008.
27. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
28. Yen-Cheng Lu, Feng Chen, Yating Wang, and Chang-Tien Lu. Discovering anomalies on mixed-type data using a generalized student-t based approach. *IEEE Transactions on Knowledge and Data Engineering, DOI:10.1109/TKDE.2016.2583429*, 2016.

29. Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *Proceedings of ESANN*, pages 89–94. Presses universitaires de Louvain, 2015.
30. Pankaj Mehta and David J Schwab. An exact mapping between the variational renormalization group and deep learning. *arXiv preprint arXiv:1410.3831*, 2014.
31. T.D. Nguyen, T. Tran, D. Phung, and S. Venkatesh. Latent patient profile modelling and applications with mixed-variaterestricted Boltzmann machine. In *Proc. of Pacific-Asia Conference on Knowledge Discovery and DataMining (PAKDD)*, Gold Coast, Queensland, Australia, April 2013.
32. T.D. Nguyen, T. Tran, D. Phung, and S. Venkatesh. Learning sparse latent representation and distance metric for imageretrieval. In *Proc. of IEEE International Conference on Multimedia & Expo*, California, USA, July 15-19 2013.
33. Matthew Eric Otey, Srinivasan Parthasarathy, and Amol Ghoting. Fast lightweight outlier detection in mixed-attribute data. *Techincal Report, OSU–CISRC–6/05–TR43*, 2005.
34. Hao-Ting Pai, Fan Wu, and Pei-Yun S Sabrina Hsueh. A relative patterns discovery for enhancing outlier detection in categorical data. *Decision Support Systems*, 67:90–99, 2014.
35. Spiros Papadimitriou, Hiroyuki Kitagawa, Phillip B Gibbons, and Christos Faloutsos. Loci: Fast outlier detection using the local correlation integral. In *Data Engineering, 2003. Proceedings. 19th International Conference on*, pages 315–326. IEEE, 2003.
36. R. Salakhutdinov and G. Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 2009.
37. Robert Serfling and Shanshan Wang. General foundations for studying masking and swamping robustness of outlier identifiers. *Statistical Methodology*, 20:79–90, 2014.
38. Jianwen Sun, Reto Wyss, Alexander Steinecker, and Philipp Glocker. Automated fault detection using deep belief networks for the quality inspection of electromotors. *tm-Technisches Messen*, 81(5):255–263, 2014.
39. Takaaki Tagawa, Yukihiro Tadokoro, and Takehisa Yairi. Structured denoising autoencoder for fault detection and analysis. In *ACML*, 2014.
40. Guanting Tang, Jian Pei, James Bailey, and Guozhu Dong. Mining multidimensional contextual outliers from categorical relational data. *Intelligent Data Analysis*, 19(5):1171–1192, 2015.
41. Adrian Taylor, Sylvain Leblanc, and Nathalie Japkowicz. Anomaly detection in automobile control network data with long short-term memory networks. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pages 130–139. IEEE, 2016.
42. Khoi-Nguyen Tran and Huidong Jin. Detecting network anomalies in mixed-attribute data sets. In *Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on*, pages 383–386. IEEE, 2010.
43. T. Tran, D. Phung, and S. Venkatesh. Thurstonian Boltzmann Machines: Learning from Multiple Inequalities. In *International Conference on Machine Learning (ICML)*, Atlanta, USA, June 16-21 2013.
44. T. Tran, D.Q. Phung, and S. Venkatesh. Mixed-variate restricted Boltzmann machines. In *Proc. of 3rd Asian Conference on Machine Learning (ACML)*, Taoyuan, Taiwan, 2011.
45. Truyen Tran, Wei Luo, Dinh Phung, Jonathan Morris, Kristen Rickard, and Svetha Venkatesh. Preterm birth prediction: Deriving stable and interpretable rules from high dimensional data. *Conference on Machine Learning in Healthcare, LA, USA*, 2016.
46. Aaron Tuor, Samuel Kaplan, Brian Hutchinson, Nicole Nichols, and Sean Robinson. Deep learning for unsupervised insider threat detection in structured cybersecurity data streams. 2017.
47. Yao Wang, Wan-dong Cai, and Peng-cheng Wei. A deep learning approach for detecting malicious JavaScript code. *Security and Communication Networks*, 2016.
48. Mao Ye, Xue Li, and Maria E Orlowska. Projected outlier detection in high-dimensional mixed-attributes data set. *Expert Systems with Applications*, 36(3):7104–7113, 2009.
49. Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. *arXiv preprint arXiv:1605.07717*, 2016.
50. Ke Zhang and Huidong Jin. An effective pattern based outlier detection approach for mixed attribute data. In *Australasian Joint Conference on Artificial Intelligence*, pages 122–131. Springer, 2010.
51. Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387, 2012.