Modeling Anomaly Detection with a Deep Boltzmann Machine

Devang Kulshreshtha[1] and Kaushal Kumar Shukla[2]

Indian Institute of Technology, Varanasi - Dept of Computer Science and Engineering
Varanasi - India

Abstract. In this paper, we show how a two-layer Deep Boltzmann Machine (DBM) can be applied directly to anomaly detection. We develop effective inference and learning procedures to model data distribution and effectively separate anomalous from non-anomalous points. We investigate on the decision criterion for outlier detection and propose two models - e-DBM and r-DBM for the same. Using grid search strategy for automatic hyperparameter optimization of DBMs, we show that our method outperform several state-of-the-art detectors on various benchmark datasets.

1 Introduction

Outlier detection models aim to detect samples from data that deviate from the expected patterns of normal data points. A panoply of different models have been proposed for outlier detection. These include linear models [1],distance-based methods [2], and density-based methods [3]. In all these anomaly detection algorithms, the approach is to learn the distribution of normal data, and points deviating from the base distribution can be classified as outliers.

Owing to the recent success of deep neural models in vision and speech tasks have increased their interest in this paradigm also. The first neural network based approach was the Autoencoder Outlier Detection [4] model. [5] uses replicator neural networks which model the data by a multi-layer perceptron. A variety of other neural network based architectures have been proposed [6] since then. Recently, [7] propose an ensemble of autoencoders coupled with adaptive sampling which further improve training time of their algorithm.

However, energy based models have not been systematically explored and developed for anomaly detection. Restricted Boltzmann Machines (RBM) are the most commonly used Energy-Based Models (EBMs) which are the building blocks of deep generative models like Deep Belief Networks [8] and Deep Boltzmann Machines (DBMs) [9]. DBMs have been very popular in a variety of tasks such as document modeling [10] and vision.

In this work, we introduce a two hidden layer DBM model, a new anomaly detection algorithm based on generative probabilistic models. Our DBM model efficiently characterizes the distribution of normal data, which allows accurate separation of anomalous points from normal points. Moreover, due to a large number of hyperparameters involved in the training, we adopt the grid search strategy for hyperparameter optimization of DBMs. This significantly simplifies the training procedure and generalizes our approach to all datasets.

We use 2 criteria to define the outlier score - *energy function* of DBM and *reconstruction error*. We evaluate our criterion under different scenarios and analyze their effectiveness. We perform experiments on several datasets by comparing our method with various state-of-the-art approaches. Using AUC score of the ROC curve as a metric, we show that our approach is able to beat or match competitor approaches for most datasets.

2 Background

2.1 Continuous Restricted Boltzmann Machine (CRBM)

RBM is the basic building block of deep energy based models. The energy function for a continuous input RBM is given by -

$$E(\mathbf{v},\mathbf{h}) = \frac{1}{2}||\mathbf{v} - b||_2^2 - \sum_{i=1}^{H} g(W_i^T\mathbf{v} + b^{'})$$
$$s.t.\ \mathbf{h} = g(W^T\mathbf{v} + b^{'})$$

where $\mathbf{v}$, $\mathbf{h}$ are input and hidden vectors. $W \in R^{d*H}$, $b \in R^d$, and $b^{'} \in R^H$ are the weight matrix, visible bias and hidden bias respectively. $g(x) = log(1+e^x)$ is the softplus function. The goal is to optimize $W, b, b^{'}$ such that $E(\mathbf{v},\mathbf{h})$ summed over all training data is minimized. We use Persistent Contrastive Divergence [11] to learn the weights which works well and is fast and scalable.

2.2 Deep Boltzmann Machine (DBM)

DBM is a deep undirected model with multiple layers of hidden variables. Each layer learns abstract, high-level correlations between neurons of previous layer. High-level representations are thus built from unlabeled inputs. Being an energy-based model, the energy function of a continuous input DBM with one visible layer $\mathbf{v}$ and two hidden layers $\mathbf{h}^{(1)}, \mathbf{h}^{(2)}$ is given by -

$$E(\mathbf{v},\mathbf{h}^{(1)},\mathbf{h}^{(2)}) = \frac{1}{2}||\mathbf{v} - b||_2^2 - \sum_{i=1}^{H1} g(\mathbf{h}_i^{(1)}) - \sum_{i=1}^{H2} g(\mathbf{h}_i^{(2)})$$
$$s.t.\ \mathbf{h}^{(1)} = g(W^{(1)T}\mathbf{v} + b^{'})\ ,\ \mathbf{h}^{(2)} = g(W^{(2)T}\mathbf{h}^{(1)} + b^{''})$$

where $W^{(1)}$ and $W^{(2)}$ are weight matrices and $b, b^{'}, b^{''}$ are the biases. We use the method developed in [9] to optimize our weights. The algorithm is composed of two steps - layerwise pre-training and generative fine-tuning.

1. Pre-training: Initialize each DBM layer by training it as RBM using contrastive divergence.

2. Generative fine-tuning: Perform mean-field inference using variational approximation in the positive phase and stochastic approximation procedure (SAP) in the negative phase.

[9] provide a detailed theoretical explanation for the above algorithm.

Dataset	Samples	Features	Percent Outliers (%)	H1	H2	Batch-size
Cardio	1831	21	9.6	16	32	8
Optdigits	5,216	64	2.9	32	64	16
Pendigits	6,870	16	2.2	12	24	16
Waveform	3,509	21	4.7	20	40	16

Table 1: Summary and hyperparameters of the data sets

3 Applying DBMs to Anomaly Detection

Suppose we have a set of N data observations $X = \{X^{(1)}, X^{(2)}, ...X^{(N)}\}$ which contain both normal and outlier samples. The task is to find outliers set $O = \{x_i\}_{i=1}^{n}$ where $n << N$ which deviate from normal data distribution.

Performing outlier detection by DBM is (naturally) based on the assumption that anomalous samples are assigned lower probabilities by the model when clamped on **v**. Then, by choosing an appropriate threshold p_{th} we output a sample X as an outlier if $P(\mathbf{v} = X) <= p_{th}$ and inlier otherwise. Although the exact calculation of $P(\mathbf{v})$ is infeasible due to intractable partition function and inference [9], we consider two alternative decision criteria for the same-

1. **Energy score:** In a DBM, lower $P(\mathbf{x})$ of samples means higher E since probability is proportional to the sum of exponent of negative of energy function over all hidden configurations [9]. Consequently, one can define an energy threshold E_{th} which is compared to the energy of the test sample.
$$x = \begin{cases} outlier, & \text{if } E(\mathbf{v} = x) >= E_{th} \\ inlier, & \text{otherwise} \end{cases} \tag{1}$$
2. **Reconstruction error:** The reconstruction error of a sample x is given by $R(x) = ||x - x_r||_2^2$ where x_r is sampled from hidden layer h during the negative phase. For finding anomalous points, we choose an appropriate threshold R_{th} such that x is an outlier if $R(x) >= R_{th}$. A DBM learns to approximate the original data and it's underlying distribution. Consequently, the reconstruction error of points belonging to base distribution is low, although this doesn't directly correspond to lower energy values.

4 Experimental Evaluation

4.1 Datasets and Evaluation Metric

We evaluate our algorithm on several datasets from the UCI machine learning repository[1]. In most datasets, the *minority* class is labeled as anomalous and *majority* class as non-anomalous. In cases where there is no minority class ($< 10\%$), we downsample a class to make the outlier ratio less than 10%. Table 1 presents a statistical summary of the 4 datasets used in our experiments.

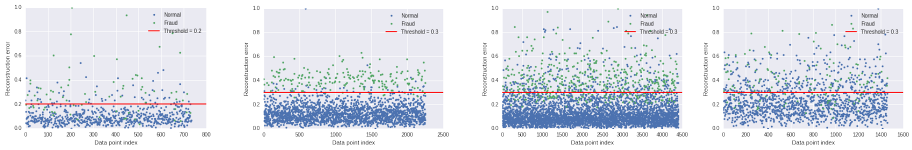
[1] https://archive.ics.uci.edu/ml/datasets.html

The Cardiotocography dataset consists of 3 classes - *normal*, *suspect*, and *pathologic*. We remove the *suspect* class , mark *normal* samples as inliers and *pathologic* samples as outliers. The Optdigits dataset consists of 10 classes (0-9) containing evenly distributed samples. Class 0 samples are marked outliers while the rest as inliers. The Pendigits dataset is from [12]. The Waveform dataset contains 3 types of labels - 0,1,2. Samples belonging to class 0 are downsampled to 10% and marked as outliers, while the remaining 2 classes as inliers. Note that the same methodology for dataset creation has been followed as in [7, 13, 12],

Each dataset is split in ratio of 1:1 for training and testing and only inliers are considered during the training process. The area under ROC curve (AUC) is used as a metric to compare the accuracy of our anomaly detecter against previously published methods.

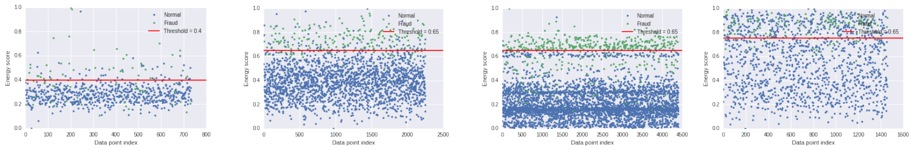
4.2 Baselines

We chose existing competitor methods for comparison and draw useful insights from results on different datasets. Specifically we use ensemble-centric algorithm HiCS [12], conventional full-space LOF outlier detection [3], spectral non-linear dimensionality reduction method [13], neural network based method by Hawkins [5], and finally the state-of-the-art ensemble of autoencoders coupled with adaptive size sample method, also called *RandNet* [7].



(a) Cardio (b) Optdigits (c) Pendigits (d) Waveform

Fig. 1: Outlier detection quality of r-DBM model on various datasets



(a) Cardio (b) Optdigits (c) Pendigits (d) Waveform

Fig. 2: Outlier detection quality of e-DBM model on various datasets

4.3 Details DBM training - Grid Search for Hyperparameter Optimization

For developing an end-to-end training procedure of DBMs without manual interference, we use grid search to automatically fine-tune hidden neurons ($H1$

and $H2$) and batch-size for each dataset. This guarantees fair comparison with other approaches as well as makes the proposed method more generalizable. We evaluate mini-batches of size 8-40 in steps of 8. For $H1$ and $H2$, we try neurons from between $F/2$ to $2 * F$ in steps of 2, where F is the number of features in dataset. The resulting values of $H1, H2$ and batch-size after applying grid search strategy are shown in Table 1.

During pre-training stage, each RBM is trained for 40 epochs using a learning rate of 0.05/batch-size, weightcost of 0.001 and momentum set to 0.5. CD learning is kept fixed at $T = 1$ during the training period. During generative fine-tuning the number of mean field updates is set to 30.

4.4 Accuracy Results

The models developed using the threshold criteria of reconstruction error and energy score are abbreviated as r-DBM and e-DBM respectively. Figure 1 and 2 shows the division of anomalous and normal points according to the reconstruction error/energy score. We manually pick a threshold (different for e-DBM and r-DBM) above which all points will be labeled as outliers. The threshold can be learned more accurately by tuning the parameter on validation dataset for higher precision/recall. Note that the AUC score on test set remains *unaffected* by the threshold value.

The AUC accuracy results are shown in Table 2 and Figure 3. The result show that our model outperforms state-of-the-art methods on all the benchmarks except for the Cardio dataset. In Optdigits dataset, the difference between our method and second best method is as high as 11.92%. The AUC score is very high at 99.55%, which can be realized by looking at Figure 3b. For a threshold of 0.3 (r-DBM) and 0.65 (e-DBM), our model is able to separate out anomalous and non-anomalous points with almost no crossing over of false points. The same however doesn't hold good for the Waveform dataset (Fig 3d). Nevertheless, our model is well above the other competitor approaches by atleast 12.6%.

Dataset	r-DBM	e-DBM	HiCS	LOF	Spectral	Hawkins	RandNet
Cardio	87.70	76.18	92.37	50.63	78.90	92.36	**92.87**
Optdigits	98.17	**99.55**	43.63	67.11	2.66	87.73	87.11
Pendigits	95.45	**97.39**	60.61	54.37	87.88	89.81	93.44
Waveform	81.51	**82.65**	52.94	55.48	62.88	61.57	70.05

Table 2: AUC comparison with state-of-the-art methods. Best performing system highlighted in boldface.

4.4.1 Energy score vs. Reconstruction error

Comparing the decision criteria of our DBM, we see that e-DBM consistently outperforms r-DBM on all except Cardio dataset. This leads us to the conjecture that energy score is a more accurate decision criteria than compared to reconstruction error. This is expected as while optimizing the weights of a DBM, our goal is to maximize the log-likelihood and lower the energy of the training data (or non-anomalous points). The learning algorithm does not minimize the reconstruction error, although it keeps diminishing after every epoch.

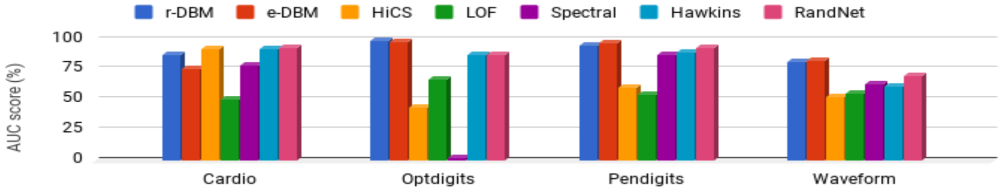


Fig. 3: Results obtained by applying various algorithms on 4 datasets.

References

[1] Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. Technical report, MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING, 2003.

[2] 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.

[3] 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.

[4] Graham Williams, Rohan Baxter, Hongxing He, Simon Hawkins, and Lifang Gu. A comparative study of rnn for outlier detection in data mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 709–712. IEEE, 2002.

[5] Simon Hawkins, Hongxing He, Graham Williams, and Rohan Baxter. Outlier detection using replicator neural networks. In *DaWaK*, volume 2454, pages 170–180. Springer, 2002.

[6] Markos Markou and Sameer Singh. Novelty detection: a review-part 2:: neural network based approaches. *Signal processing*, 83(12):2499–2521, 2003.

[7] Jinghui Chen, Saket Sathe, Charu Aggarwal, and Deepak Turaga. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 90–98. SIAM, 2017.

[8] Geoffrey E Hinton. Deep belief networks. *Scholarpedia*, 4(5):5947, 2009.

[9] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial Intelligence and Statistics*, pages 448–455, 2009.

[10] Nitish Srivastava, Ruslan R Salakhutdinov, and Geoffrey E Hinton. Modeling documents with deep boltzmann machines. *arXiv preprint arXiv:1309.6865*, 2013.

[11] Geoffrey E Hinton. A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer, 2012.

[12] Fabian Keller, Emmanuel Muller, and Klemens Bohm. Hics: High contrast subspaces for density-based outlier ranking. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1037–1048. IEEE, 2012.

[13] Saket Sathe and Charu Aggarwal. Lodes: Local density meets spectral outlier detection. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 171–179. SIAM, 2016.