Dissimilarity Based Contrastive Divergence for Anomaly Detection

Sahil Manocha ES13B1019@IITH.AC.IN
Adepu Ravi Sankar CS14RESCH11001@IITH.AC.IN
Vineeth N Balasubramanian VINEETHNB@IITH.AC.IN
Indian Institute of Technology Hyderabad, Sangareddy, Telangana, India 502285

Abstract

This paper describes training of a Restricted Boltzmann Machine(RBM) using dissimilarity-based contrastive divergence to obtain an anomaly detector. We go over the merits of the method over other approaches and describe the method's usefulness to obtain a generative model.

1. Introduction

Over the millennia, our senses have been optimized to detect anomalies often manifested as a foreboding or feeling of apprehension. Anomalies, which are inconsistent patterns in data, must not be confused with noise. Anomaly detection refers to abnormal data points that offer actionable insights. For example, an abnormality in satellite imagery data may help national security efforts. In this paper, a method for anomaly detection using semi-supervised training is presented. We use Dissimilarity-based Contrastive Divergence (Diss-CD), a variant of the Contrastive Divergence algorithm, to train a RBM.

In the past, both generative (such as Gaussian Mixture Models) as well as discriminative (such as one-class Support Vector Machines) methods have been proposed for anomaly detection applications. A purely discriminative model that tells anomalies apart from normal data is, by principle, inadequate as the whole spectrum of anomalies cannot possibly be provided to identify the ideal discriminant. Even a large enough dataset of anomalous training examples is hard to obtain due to the very nature of an anomaly. Thus, a generative model that models the distribution of the normal data should ideally give better and more generalizable results for anomaly detection, motivating us to use Restricted Boltzmann Machines (RBMs) in this work.

Furthermore, anomaly detection is shown to work better (Song et al., 2007) when domain knowledge is incorporated in the model. Thus, a semi-supervised training method should be even more suitable for our objective. While discriminative RBMs have been used before for anomaly detection (Fiore, 2013), our approach uses the RBM's generative capability, in a semi-supervised manner. The choice of dissimilar training data affords supervision in the proposed method.

Additionally, as stated by Lee *et al.* in (Lee, 2011), use of dissimilar data for untraining has potential in real-world applications such as spam classification where it is found that anomalous data is statistically more correlated when compared to normal data. Therefore, unlearning on some token dissimilar data should capture a wide variety of possible anomalies.

The proposed method for anomaly detection uses a Diss-CD training method for RBMs which results in high reconstruction error for abnormal data. Using this reconstruction error, one can successfully identify anomalies.

2. Background

2.1. RBMs

Energy-Based Models (EBMs) (Hinton, 2002) capture dependencies between variables by associating a scalar energy to each configuration of the variables. The objective is to minimize the scalar energy function for observed configurations. In a stochastic model, this is equivalently stated as maximizing the likelihood of the occurrence of a configuration; if the energy is high, then the likelihood must be low and vice versa. The likelihood measure defined as such over all configurations needs to be normalized to obtain a probability distribution. Computations required to determine the normalizing constant are often in-

Appearing in Proceedings of the 2^{nd} Indian Workshop on Machine Learning, IIT Kanpur, India, 2016.

tractable in an ordinary setting. An approximate method called Contrastive Divergence works well in practice to approximate normalized measures in polynomial time. Several variations of the CD (contrastive divergence) method like PCD and Tempered MCMC (Guillaume Desjardins, 2010) have been introduced to get around some of the issues that invariably crop up in training using the CD method.

2.1.1. Inference in an RBM

An energy function is defined in an RBM that is minimized in the training of the neural network:

$$E(\mathbf{v}, \mathbf{h}) = -(\mathbf{b}^T\mathbf{v} + \mathbf{c}^T\mathbf{h} + \mathbf{h}^T W\mathbf{v})$$

where the probability of a particular training example $\mathbf{v}$ is

$$P(\mathbf{v}) = \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}}{Z}, \qquad where\ Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$$

Although the marginal probability of $\mathbf{v}$ is intractable due to the normalization constant, the conditional probability of the neurons of a particular layer is simple to compute:

$$P(\mathbf{h}|\mathbf{v}) = \prod_{h_i} P(h_i|\mathbf{v})$$
$$P(h_i = 1|\mathbf{v}) = \sigma(c_i + \sum_j W_{ij} v_j)$$

Thus, if we clamp down the values of the visible layer, then we can perform Gibbs sampling to obtain the values of the hidden layer. The visible layer can be computed similarly.

2.1.2. Training an RBM

The objective of training an RBM is to maximize the log likelihood over the training set. A stochastic gradient descent is performed to reduce the negative log likelihood over the training example. The gradient of the log likelihood is

$$\frac{\partial \ln P(\mathbf{v}^{(k)})}{\partial\theta} = -E_{p(\mathbf{h}|\mathbf{v}^{(k)})}[\frac{\partial E(\mathbf{v}^{(k)}, \mathbf{h})}{\partial\theta}] + E_{p(\mathbf{h},\mathbf{v})}[\frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial\theta}]$$

The CD method approximates the expectation of the gradient as the gradient at *fantasy particle* $\mathbf{v}'$.

$$E_{p(\mathbf{h},\mathbf{v})}[\frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial\theta}] \approx \frac{\partial E(\mathbf{v}', \mathbf{h}')}{\partial\theta}$$

2.2. Anomaly Detection

A good anomaly detector ideally behaves differently for aberrant inputs and identifies unusual trends or events. The major challenge in training a model for anomaly detection is that the form of an anomaly is unknown, since there is no specific way to characterize the features of an anomaly. For an exhaustive treatment of anomaly detection methods, refer to (Chandola, 2009). Use of neural networks in anomaly detection, especially in images, is not as widely explored as in other domains (especially temporal data). Even so, examples of image based anomaly detection can be found in the literature (Park et al., 2012) (Saleh, 2013). To the best of our knowledge, this is one of the first efforts to use RBMs for anomaly detection, from a generative standpoint.

Algorithm 1 Dissimilar Contrastive Divergence Algorithm

Input: RBM(W, b, c), Training data S, Dissimilar data $\bar{S}$, Number of Gibbs cycles K, Number of hidden units n, Number of visible units m
Output: DissCD trained RBM
Initialize $W \sim \left[-\frac{\sqrt{6}}{\sqrt{n+m}}, \frac{\sqrt{6}}{\sqrt{n+m}}\right]$, $b = 0$, $c = 0$
for all $\mathbf{pos_v} \in S, \mathbf{neg_v} \in \bar{S}$ **do**
  $V^{(0)} \leftarrow \mathbf{pos_v}$, $V \leftarrow \mathbf{neg_v}$
  $H \leftarrow$ SampleHGivenV(V)
  for $j = 1$ **to** K **do**
    $V \leftarrow$ SampleVGivenH(H)
    $H \leftarrow$ SampleHGivenV(V)
  end for
  $V' \leftarrow V$
  $w_{ij} = w_{ij} + p(h_i = 1|V^{(0)})v_j^{(0)} - p(h_i = 1|V')v'_j$
  $b_j = b_j + v_j^{(0)} - v'_j$
  $c_i = c_i + p(h_i = 1|V^{(0)}) - p(h_i = 1|V')$
end for

3. Diss-CD for Anomaly Detection

In this paper, we initialize the visible layer to a dissimilar data point to obtain the fantasy particle during Contrastive Divergence. This method of Dissimilarity-based Contrastive Divergence (Sankar & N, 2015), an earlier method proposed by us, aims to improve the probability distributions so that our model better approximates the source distribution of the data. This algorithm is described in Algorithm 1. Each epoch of training takes on the order of $O(d + bK)$ time where d and b are the number of batches in the training and dissimilar-training set respectively. K is the number of Gibbs samplings to obtain the fantasy particle.

Our anomaly detection algorithm works by computing the reconstruction error between an initialization x of the visible layer of the RBM and its reconstruction $\bar{x}$ after a sequence of K gibbs sampling. If this reconstruction error is greater than a threshold ε we classify x as an anomaly. The proposed methodology

Algorithm 2 Anomaly Detection Algorithm

Input: Diss-CD trained RBM(W, b, c), Input S, Threshold ε, Gibbs Sampling Steps K
Initialize visible layer $V \leftarrow S$
for $j = 1$ **to** K **do**
  $H \leftarrow$ SampleHGivenV(V)
  $V \leftarrow$ SampleVGivenH(H)
end for
$E \leftarrow \text{MSE}(|V - S|^2)$
if $E \geq \varepsilon$ **then**
  return 1
else
  return 0
end if
Output: 1 if S is anomalous, 0 otherwise

Table 1. Datasets used for training of RBM.

Data set	Description	Size
MNIST	Images of digits (0 to 9)	60000
CIFAR10	10 classes of 32x32 color images	60000
Rectangles	Grayscale Rectangle Contours of 28x28. Area of each rectangle $\geq$ 300.	10000
Triangles	Grayscale Triangle Contours of 28x28. Area of each triangle $\geq$ 240	10000

is described in Algorithm 2. We use the DissCD RBM wherein the reconstructions overwhelmingly resemble the training data which results in high reconstruction error in case of anomalies.

4. Experiments and Results

We conducted experiments to verify the following claims:

1. DissCD-RBM approximates the source distribution of the data better than PCD-RBM.
2. DissCD-RBM gives better performance for anomaly detection using the described method as compared to PCD-RBM.
3. Choice of dissimilar data gives significant difference in performance of the anomaly detector.

The datasets used are described in Table 1. Unfortunately, no public dataset for testing anomaly detection methods on images exists at the moment, a fact which is also highlighted by (Saleh, 2013). For each model, the expected input comes from the source distribution of MNIST digits. Data points which have no resemblance to digits must be classified as anomalies. Two synthetic datasets, *Rectangles* and *Triangles*, both of which are 28×28 images of shape contours (rectangles and triangles respectively) are used to test the effectiveness of different methods of training. They are chosen due to their component edge orientations which match that of digit components. For all experiments, we choose a batch size of 20, learning rate of 0.1, 500 artificial neurons in hidden layer and trained for 15 epochs. The reconstruction error between the input and the corresponding network output is the mean squared error between the image vectors. The error is sampled for i reconstructions where i ranges from 1 to 20. Each reconstruction is computed using a Gibbs sampling cycle starting at the previous reconstruction. Table 2 presents reconstruction error difference between PCD trained RBM and DissCD trained RBM with Rectangles as dissimilar data. Table 3 compares the reconstruction error of anomalies based on the choice of dissimilar data.

Table 2. Reconstruction Error comparison between PCD and DissCD trained RBMs on different datasets (Higher is better for anomaly detection)

	PCD	DissCD
Silhouttes	367.7	465.6
CIFAR10	199.7	251.2
Triangles	85.7	113.6
MNIST	56.6	48.1

5. Discussion and Conclusions

When CIFAR10 and Silhouttes datasets are used as dissimilar inputs to a RBM trained on MNIST (both these datasets are significantly different from MNIST), both PCD and DISS-CD give high reconstruction errors. Although the reconstruction error in case of Diss-CD method is relatively higher than that in PCD-trained RBM, the effectiveness of the anomaly detection methodology is not evident for such vastly different datasets. In any real scenario, such anomalies are unlikely to occur and any results as such are of little practical significance.

Table 3. Comparison of reconstruction error on triangle dataset for different dissimilar training data (Rectangles & CIFAR10 resp.) and PCD (Higher the better for anomaly detection)

	DissCD *(Rec)*	DissCD *(C10)*	PCD
Triangles	113.6	88.6	85.7

The real challenge occurs when anomalies are similar to the training dataset. In our paper, the synthetic rectangle and triangle datasets exhibit this attribute when the training set is MNIST digits. In our experiments, RBM trained using the Diss-CD method gave higher reconstruction error than PCD-trained RBM for data drawn from rectangle or triangle datasets. The observations of reconstruction errors betrays a few

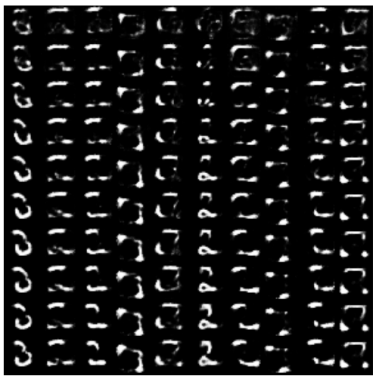

Figure 1. Visible layer initialized to 10 examples from triangle dataset. Sampled after every 2 Gibbs cycles on PCD trained net.

Figure 2. Visible layer initialized to 10 examples from triangle dataset. Sampled after every 2 Gibbs cycles on Diss-CD trained net.

key features of the underlying probability distribution modeled by RBMs. Firstly, the PCD method is unable to properly model the source probability of the MNIST digits data. Thus, the reconstruction may sometimes resemble an anomalous data point. This behavior is not exhibited by the Diss-CD trained networks. In fact, the reconstruction error is slightly lower in case of Diss-CD trained net for the MNIST data which only bolsters its merits as a good generative model. Figure 1 and 2 supports this assertion.

One question that arises is: is the performance of the Diss-CD RBM dependent on the choice of dissimilar dataset? Our experiments demonstrate that this is indeed true. A carefully chosen dissimilar dataset yields a better performing net than an arbitrarily chosen dataset. We tested this hypothesis by untraining on CIFAR10 and Rectangle dataset. The reconstruction error for the Triangle dataset when Rectangle dataset is used for untraining in the Diss-CD RBM is higher. Therefore, the choice of dissimilar data can improve the performance of the anomaly detector.

In summary, Diss-CD provides a novel method to train effective anomaly detectors without departing too much from the traditional method of training RBMs. Furthermore, these networks show themselves to be better generative models than their PCD trained counterparts. As such, the source distribution of the training data is better modeled by the Diss-CD method as compared to the PCD method. The use of such networks as generative models should be explored in future. In our observation, the number of Gibbs samplings for gradient calculation while training must be high (≥ 10) as lower values lead to poor performance on datasets other than the training and untraining datasets. In addition, we found that increase in the number of hidden units did not translate to corresponding increase in performance of the anomaly detector.

References

Chandola, et al. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009. ISSN 0360-0300.

Fiore, et al. Network anomaly detection with the restricted boltzmann machine. *Neurocomput.*, 122:13–23, December 2013. ISSN 0925-2312.

Guillaume Desjardins, et al. Adaptive parallel tempering for stochastic maximum likelihood learning of rbms. *CoRR*, abs/1012.3476, 2010.

Hinton, Geoffrey E. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, August 2002.

Lee, et al. Self-similar properties of spam. IMIS '11, pp. 347–352, Washington, DC, USA, 2011. IEEE Computer Society.

Park, Sangdon, Kim, Wonsik, and Lee, Kyoung Mu. Abnormal object detection by canonical scene-based contextual model. In *(ECCV)*, 2012.

Saleh, et al. Object-centric anomaly detection by attribute-based reasoning. In *(CVPR)*. IEEE, 2013.

Sankar, Adepu Ravi and N, Vineeth. Similarity-based contrastive divergence methods for energy-based deep learning models. In *Proceedings of The 7th ACML*, pp. 391–406, 2015.

Song, Xiuyao, Wu, Mingxi, Jermaine, Christopher, and Ranka, Sanjay. Conditional anomaly detection. *IEEE Trans. on KDD.*, 19(5):631–645, May 2007.