Master-Seminar 1: Data Analytics

Instructor : Eya Boumaiza

Tuesday 14:00 - 16:00
Room: B 026

Anomaly detection (or outlier detection) is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data. Typically, anomalous data can be connected to some kind of problem or rare event such as e.g. bank fraud, medical problems, structural defects etc. This connection makes it more challenging to decide which data points can be considered anomalies especially if no labels are provided (Unsupervised Learning). Possibly, some normal or anomalous examples can be labeled, in addition to a large set of unlabeled data (Semi-supervised Learning). This seminar focuses on discussing state-of-the-art research publications dedicated for Unsupervised and Semi-supervised Anomaly detection.

Reading List

#	Authors	Title & Abstract	Venue & citations	Year
1	S. Ramaswamy et al.,	**Title:** Efficient Algorithms for Mining Outliers from Large Data Sets **Abstract:** In this paper, we propose a novel formulation for distance-based outliers that is based on the distance of a point from its k^{th} nearest neighbor. We rank each point on the basis of its distance to its k^{th} nearest neighbor and declare the top n points in this ranking to be outliers. In addition to developing relatively straightforward solutions to finding such outliers based on the classical nested loop join and index join algorithms, we develop a highly efficient partition-based algorithm for mining outliers. This algorithm first partitions the input data set into disjoint subsets, and then prunes entire partitions as soon as it is determined that they cannot contain outliers. This results in substantial savings in computation. We present the results of an extensive experimental study on real-life and synthetic data sets. The results from a real-life NBA database highlight and reveal several expected and unexpected aspects of the database. The results from a study on synthetic data sets demonstrate that the partition-based algorithm scales well with respect to both data set size and data set dimensionality.	ACM SIGMOD **Cit:** 2001	2000
2	E. Eskin et al.,	**Title:** A geometric framework for unsupervised anomaly detection: Detecting Intrusions in Unlabeled Data **Abstract:** Most current intrusion detection systems employ signature-based methods or data mining-based methods which rely on labeled training data. This training data is typically expensive to produce. We present a new geometric framework for unsupervised anomaly detection, which are algorithms that are designed to process unlabeled data. In our framework, data elements are mapped to a feature space which is typically a vector space $\mathcal{R}^d$. Anomalies are detected by determining which points lies in sparse regions of the feature space. We present two feature maps for mapping data elements to a feature space. Our first map is a data-dependent normalization feature map which we apply to network connections. Our second feature map is a spectrum kernel which we apply to system call traces. We present three algorithms for detecting which points lie in sparse regions of the feature space. We evaluate our methods by performing experiments over network records from the KDD CUP 1999 data set and system call traces from the 1999 Lincoln Labs DARPA evaluation.	Applications of Data Mining in Computer Security **Cit:** 661	2002

3	Z. He et al.,	**Title:** Discovering cluster-based local outliers **Abstract:** In this paper, we present a new definition for outlier: cluster-based local outlier, which is meaningful and provides importance to the local data behavior. A measure for identifying the physical significance of an outlier is designed, which is called cluster-based local outlier factor (CBLOF). We also propose the FindCBLOF algorithm for discovering outliers. The experimental results show that our approach outperformed the existing methods on identifying meaningful and interesting outliers.	Pattern Recognition Letters **Cit:** 623	2003
4	W. Fan et al.,	**Title:** Using artificial anomalies to detect unknown and known network intrusions **Abstract:** Intrusion detection systems (IDSs) must be capable of detecting new and unknown attacks, or anomalies. We study the problem of building detection models for both pure anomaly detection and combined misuse and anomaly detection (i.e., detection of both known and unknown intrusions) . We propose an algorithm to generate artificial anomalies to coerce the inductive learner into discovering an accurate boundary between known classes (normal connections and known intrusions) and anomalies. Empirical studies show that our pure anomaly detection model trained using normal and artificial anomalies is capable of detecting more than 77% of all unknown intrusion classes with more than 50% accuracy per intrusion class. The combined misuse and anomaly detection models are as accurate as a pure misuse detection model in detecting known intrusions and are capable of detecting at least 50% of unknown intrusion classes with accuracy measurements between 75% and 100% per class.	Knowledge and Information Systems **Cit:** 266	2004
5	N. Abe et al.,	**Title:** Outlier Detection by Active Learning **Abstract:** Most existing approaches to outlier detection are based on density estimation methods. There are two notable issues with these methods: one is the lack of explanation for outlier flagging decisions, and the other is the relatively high computational requirement. In this paper, we present a novel approach to outlier detection based on classification, in an attempt to address both of these issues. Our approach isbased on two key ideas. First, we present a simple reduction of outlier detection to classification, via a procedure that involves applying classification to a labeled data set containing artificially generated examples that play the role of potential outliers. Once the task has been reduced to classification, we then invoke a selective sampling mechanism based on active learning to the reduced classification problem. We empirically evaluate the proposed approach using a number of data sets, and find that our method is superior to other methods based on the same reduction to classification, but using standard classification methods. We also show that it is competitive to the state-of-the-art outlier detection methods in the literature based on density estimation, while significantly improving the computational complexity and explanatory power.	ACM SIGKDD **Cit:** 274	2006

6	F. Liu et al.,	**Title:** Isolation Forest **Abstract:** Most existing model-based approaches to anomaly detection construct a profile of normal instances, then identify instances that do not conform to the normal profile as anomalies. This paper proposes a fundamentally different model-based method that explicitly isolates anomalies instead of profiles normal points. To our best knowledge, the concept of isolation has not been explored in current literature. The use of isolation enables the proposed method, iForest, to exploit sub-sampling to an extent that is not feasible in existing methods, creating an algorithm which has a linear time complexity with a low constant and a low memory requirement. Our empirical evaluation shows that iForest performs favourably to ORCA, a near-linear time complexity distance-based method, LOF and random forests in terms of AUC and processing time, and especially in large data sets. iForest also works well in high dimensional problems which have a large number of irrelevant attributes, and in situations where training set does not contain any anomalies.	IEEE **Cit:** 776	2008
7	F.T. Liu et al.,	**Title:** Isolation-based Anomaly Detection **Abstract:** Anomalies are data points that are few and different. As a result of these properties, we show that, anomalies are susceptible to a mechanism called isolation. This article proposes a method called Isolation Forest (iForest), which detects anomalies purely based on the concept of isolation without employing any distance or density measure—fundamentally different from all existing methods. As a result, iForest is able to exploit subsampling (i) to achieve a low linear time-complexity and a small memory-requirement and (ii) to deal with the effects of swamping and masking effectively. Our empirical evaluation shows that iForest outperforms ORCA, one-class SVM, LOF and Random Forests in terms of AUC, processing time, and it is robust against masking and swamping effects. iForest also works well in high dimensional problems containing a large number of irrelevant attributes, and when anomalies are not available in training sample.	TKDD **Cit:** 324	2012
8	W. Liu et al.,	**Title:** Unsupervised One-Class Learning for Automatic Outlier Removal **Abstract:** Outliers are pervasive in many computer vision and pattern recognition problems. Automatically eliminating outliers scattering among practical data collections becomes increasingly important, especially for Internet inspired vision applications. In this paper, we propose a novel one-class learning approach which is robust to contamination of input training data and able to discover the outliers that corrupt one class of data source. Our approach works under a fully unsupervised manner, differing from traditional one-class learning supervised by known positive labels. By design, our approach optimizes a kernel-based max-margin objective which jointly learns a large margin one-class classifier and a soft label assignment for inliers and outliers. An alternating optimization algorithm is then designed to iteratively refine the classifier and the labeling, achieving a provably convergent solution in only a few iterations. Extensive experiments conducted on four image datasets in the presence of artificial and real-world outliers demonstrate that the proposed approach is considerably superior to the state-of-the-arts in obliterating outliers from contaminated one class of images, exhibiting strong robustness at a high outlier proportion up to 60%.	IEEE **Cit:** 73	2014

9	Y. Xia et al.,	**Title:** Learning Discriminative Reconstructions for Unsupervised Outlier Removal **Abstract:** We study the problem of automatically removing outliers from noisy data, with application for removing outlier images from an image collection. We address this problem by utilizing the reconstruction errors of an autoencoder. We observe that when data are reconstructed from low-dimensional representations, the inliers and the outliers can be well separated according to their reconstruction errors. Based on this basic observation, we gradually inject discriminative information in the learning process of an autoencoder to make the inliers and the outliers more separable. Experiments on a variety of image datasets validate our approach.	ICCV **Cit:** 51	2015
10	N. Goix et al.,	**Title:** How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms? **Abstract:** When sufficient labeled data are available, classical criteria based on Receiver Operating Characteristic (ROC) or Precision-Recall (PR) curves can be used to compare the performance of un-supervised anomaly detection algorithms. However , in many situations, few or no data are labeled. This calls for alternative criteria one can compute on non-labeled data. In this paper, two criteria that do not require labels are empirically shown to discriminate accurately (w.r.t. ROC or PR based criteria) between algorithms. These criteria are based on existing Excess-Mass (EM) and Mass-Volume (MV) curves, which generally cannot be well estimated in large dimension. A methodology based on feature sub-sampling and aggregating is also described and tested, extending the use of these criteria to high-dimensional datasets and solving major drawbacks inherent to standard EM and MV curves.	ICML **Cit:** 19	2016
11	T. Ergen et al.,	**Title:** Unsupervised and Semi-supervised Anomaly Detection with LSTM Neural Networks **Abstract:** We investigate anomaly detection in an unsupervised framework and introduce Long Short Term Memory (LSTM) neural network based algorithms. In particular, given variable length data sequences, we first pass these sequences through our LSTM based structure and obtain fixed length sequences. We then find a decision function for our anomaly detectors based on the One Class Support Vector Machines (OC-SVM) and Support Vector Data Description (SVDD) algorithms. As the first time in the literature, we jointly train and optimize the parameters of the LSTM architecture and the OC-SVM (or SVDD) algorithm using highly effective gradient and quadratic programming based training methods. To apply the gradient based training method, we modify the original objective criteria of the OC-SVM and SVDD algorithms, where we prove the convergence of the modified objective criteria to the original criteria. We also provide extensions of our unsupervised formulation to the semi-supervised and fully supervised frameworks. Thus, we obtain anomaly detection algorithms that can process variable length data sequences while providing high performance, especially for time series data. Our approach is generic so that we also apply this approach to the Gated Recurrent Unit (GRU) architecture by directly replacing our LSTM based structure with the GRU based structure. In our experiments, we illustrate significant performance gains achieved by our algorithms with respect to the conventional methods.	SP **Cit:** 7	2017

12	T. Schlegl et al.,	**Title:** Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery **Abstract:** Obtaining models that capture imaging markers relevant for disease progression and treatment monitoring is challenging. Models are typically based on large amounts of data with annotated examples of known markers aiming at automating detection. High annotation effort and the limitation to a vocabulary of known markers limit the power of such approaches. Here, we perform unsupervised learning to identify anomalies in imaging data as candidates for markers. We propose AnoGAN, a deep convolutional generative adversarial network to learn a manifold of normal anatomical variability, accompanying a novel anomaly scoring scheme based on the mapping from image space to a latent space. Applied to new data, the model labels anomalies, and scores image patches indicating their fit into the learned distribution. Results on optical coherence tomography images of the retina demonstrate that the approach correctly identifies anomalous images, such as images containing retinal fluid or hyperreflective foci.	IPMI **Cit:** 284	2017
13	L. Deecke et al.,	**Title:** Image Anomaly Detection with Generative Adversarial Networks **Abstract:** Many anomaly detection methods exist that perform well on low-dimensional problems however there is a notable lack of effective methods for high-dimensional spaces, such as images. Inspired by recent successes in deep learning we propose a novel approach to anomaly detection using generative adversarial networks. Given a sample under consideration, our method is based on searching for a good representation of that sample in the latent space of the generator; if such a representation is not found, the sample is deemed anomalous. We achieve state-of-the-art performance on standard image benchmark datasets and visual inspection of the most anomalous samples reveals that our method does indeed return anomalies.	ECML **Cit:** 7	2018
14	I. Golan et al.,	**Title:** Deep Anomaly Detection Using Geometric Transformations **Abstract:** We consider the problem of anomaly detection in images, and present a new detection technique. Given a sample of images, all known to belong to a normal” class (e.g., dogs), we show how to train a deep neural model that can detect out-of-distribution images (i.e., non-dog objects). The main idea behind our scheme is to train a multi-class model to discriminate between dozens of geometric transformations applied on all the given images. The auxiliary expertise learned by the model generates feature detectors that effectively identify, at test time, anomalous images based on the softmax activation statistics of the model when applied on transformed images. We present extensive experiments using the proposed detector, which indicate that our algorithm improves state-of-the-art methods by a wide margin.	NIPS **Cit:** 13	2018

15	L. Ruff et al.,	**Title:** Deep One-Class Classification **Abstract:** Despite the great advances made by deep learning in many machine learning problems, there is a relative dearth of deep learning approaches for anomaly detection. Those approaches which do exist involve networks trained to perform a task other than anomaly detection, namely generative models or compression, which are in turn adapted for use in anomaly detection; they are not trained on an anomaly detection based objective. In this paper we introduce a new anomaly detection method—Deep Support Vector Data Description—, which is trained on an anomaly detection based objective. The adaptation to the deep regime necessitates that our neural network and training procedure satisfy certain properties, which we demonstrate theoretically. We show the effectiveness of our method on MNIST and CIFAR-10 image benchmark datasets as well as on the detection of adversarial examples of GTSRB stop signs.	ICML **Cit:** 62	2018
16	L. Ruff et al.,	**Title:** Deep Support Vector Data Description for Unsupervised and Semi-Supervised Anomaly Detection **Abstract:** Deep approaches to anomaly detection have recently shown promising results over shallow approaches on high-dimensional data. Typically anomaly detection is treated as an unsupervised learning problem. In practice however, one may have— in addition to a large set of unlabeled samples—access to a small pool of labeled samples, e.g. a subset verified by some domain expert as being normal or anomalous. Semi-supervised approaches to anomaly detection make use of such labeled data to improve detection performance. Few deep semi-supervised approaches to anomaly detection have been proposed so far and those that exist are domain-specific. In this work, we present Deep SAD, an end-to-end methodology for deep semisupervised anomaly detection. Using an information-theoretic perspective on anomaly detection, we derive a loss motivated by the idea that the entropy for the latent distribution of normal data should be lower than the entropy of the anomalous distribution. We demonstrate in extensive experiments on MNIST, Fashion-MNIST, and CIFAR-10 along with other anomaly detection benchmark datasets that our approach is on par or outperforms shallow, hybrid, and deep competitors, even when provided with only few labeled training data.	ICML **Cit:** 1	2019
17	L. Ruff et al.,	**Title:** Deep Semi-Supervised Anomaly Detection **Abstract:** Deep approaches to anomaly detection have recently shown promising results over shallow approaches on high-dimensional data. Typically anomaly detection is treated as an unsupervised learning problem. In practice however, one may have—in addition to a large set of unlabeled samples—access to a small pool of labeled samples, e.g. a subset verified by some domain expert as being normal or anomalous. Semi-supervised approaches to anomaly detection make use of such labeled data to improve detection performance. Few deep semi-supervised approaches to anomaly detection have been proposed so far and those that exist are domain-specific. In this work, we present Deep SAD, an end-to-end methodology for deep semi-supervised anomaly detection. Using an information-theoretic perspective on anomaly detection, we derive a loss motivated by the idea that the entropy for the latent distribution of normal data should be lower than the entropy of the anomalous distribution. We demonstrate in extensive experiments on MNIST, Fashion-MNIST, and CIFAR-10 along with other anomaly detection benchmark datasets that our approach is on par or outperforms shallow, hybrid, and deep competitors, even when provided with only few labeled training data.	(Under Review) **Cit:** 0	2019

18	L. Ruff et al.,	**Title:** Self-Attentive, Multi-Context One-Class Classification for Unsupervised Anomaly Detection on Text **Abstract:** There exist few text-specific methods for unsupervised anomaly detection, and for those that do exist, none utilize pre-trained models for distributed vector representations of words. In this paper we introduce a new anomaly detection method—Context Vector Data Description (CVDD)—which builds upon word embedding models to learn multiple sentence representations that capture multiple semantic contexts via the self-attention mechanism. Modeling multiple contexts enables us to perform contextual anomaly detection of sentences and phrases with respect to the multiple themes and concepts present in an unlabeled text corpus. These contexts in combination with the self-attention weights make our method highly interpretable. We demonstrate the effectiveness of CVDD quantitatively as well as qualitatively on the well-known Reuters, 20 Newsgroups, and IMDB Movie Reviews datasets.	ACL **Cit:** 0	2019
19	G. Pang et al.,	**Title:** Deep Anomaly Detection with Deviation Networks **Abstract:** Although deep learning has been applied to successfully address many data mining problems, relatively limited work has been done on deep learning for anomaly detection. Existing deep anomaly detection methods, which focus on learning new feature representations to enable downstream anomaly detection methods, perform indirect optimization of anomaly scores, leading to data-inefficient learning and suboptimal anomaly scoring. Also, they are typically designed as unsupervised learning due to the lack of large-scale labeled anomaly data. As a result, they are difficult to leverage prior knowledge (e.g., a few labeled anomalies) when such information is available as in many real-world anomaly detection applications. This paper introduces a novel anomaly detection framework and its instantiation to address these problems. Instead of representation learning, our method fulfills an end-to-end learning of anomaly scores by a neural deviation learning, in which we leverage a few (e.g., multiple to dozens) labeled anomalies and a prior probability to enforce statistically significant deviations of the anomaly scores of anomalies from that of normal data objects in the upper tail. Extensive results show that our method can be trained substantially more data-efficiently and achieves significantly better anomaly scoring than state-of-the-art competing methods.	ACM SIGKDD **Cit:** 0	2019