How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?

Nicolas Goix

How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?

Nicolas Goix NICOLAS.GOIX@TELECOM-PARISTECH.FR
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France

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 unsupervised 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.

1. Introduction

When labels are available, classical ways to evaluate the quality of an anomaly scoring function are the ROC and PR curves. Unfortunately, most of the time, data come without any label. In lots of industrial setups, labeling datasets calls for costly human expertise, while more and more unlabeled data are available. A huge practical challenge is therefore to have access to criteria able to discriminate between unsupervised algorithms without using any labels. In this paper, we formalize and justify the use of two such criteria designed for unsupervised anomaly detection (AD), and adapt them to large dimensional data. Strong empirical performance demonstrates the relevance of our approach.

The common underlying assumption behind AD is that anomalies occur in low probability regions of the data generating process. This formulation motivates many statistical AD methods. Classical parametric techniques (Barnett & Lewis, 1994; Eskin, 2000) assume that the normal data are generated by a distribution belonging to some specific and *a priori* known parametric model. The most popular non-parametric approaches include algorithms based on density (level set) estimation (Schölkopf et al., 2001; Scott & Nowak, 2006; Breunig et al., 2000), on dimensionality reduction (Shyu et al., 2003; Aggarwal & Yu, 2001) or on decision trees (Liu et al., 2008). One may refer to (Hodge & Austin, 2004; Chandola et al., 2009; Patcha & Park, 2007; Markou & Singh, 2003) for overviews of current research on AD. It turns out that the overwhelming majority of AD algorithms return more than a binary label, normal/abnormal. They first compute a *scoring function*, which is converted to a binary prediction, typically by imposing some threshold based on its statistical distribution.

What is a scoring function? As anomalies are very rare, their structure cannot be observed in the data, in particular their distribution. It is common and convenient to assume that anomalies occur in the tail of F the distribution of normal data, so that the goal is to estimate density level sets of F. This setup is typically the one of the One-Class Support Vector Machine (OneClassSVM) algorithm developped in (Schölkopf et al., 2001), which extends the SVM methodology (Shawe-Taylor & Cristianini, 2004) to handle training using only positive information. The underlying assumption is that we observe data in $\mathbb{R}^d$ from the normal class only, with underlying distribution F and underlying density $f : \mathbb{R}^d \to \mathbb{R}$. The goal is to estimate density level sets $(\{\mathbf{x}, f(\mathbf{x}) > t\})_{t>0}$ with t close to 0. Such estimates are encompassed into a *scoring function*: any measurable function $s : \mathbb{R}^d \to \mathbb{R}_+$ integrable *w.r.t.* the Lebesgue measure $\mathrm{Leb}(.)$, whose level sets are estimates of the level sets of the density. Any scoring function defines a preorder on $\mathbb{R}^d$ and thus a ranking on a set of new observations. This ranking can be interpreted as a degree of abnormality, the lower $s(x)$, the more abnormal x.

How to know if a scoring function is good? How can we know if the preorder induced by a scoring function s is 'close' to that of f, or equivalently if these induced level sets are close to those of f? The problem is to define this notion of proximity into a criterion $\mathcal{C}$, optimal scoring functions s^* being then defined as those optimizing $\mathcal{C}$. It turns out that for any strictly increasing transform

Presented at ICML2016 Anomaly Detection Workshop, New York, NY, USA, 2016.

$T : \text{Im(f)} \to \mathbb{R}$, the level sets of $T \circ f$ are exactly those of f. Here and hereafter, Im(f) denotes the image of the mapping f. For instance, $2f$ or f^2 are perfect scoring functions, just as f. Thus, we cannot simply consider a criterion based on the distance of s to the true density, *e.g.* $\mathcal{C}(s) = \|s - f\|$. We seek for a similar criterion which is invariant by increasing transformation of the output s. In other words, the criterion should be defined in such a way that the collection of level sets of an optimal scoring function $s^*(x)$ coincides with that related to f. Moreover, any increasing transform of the density should be optimal regarding $\mathcal{C}$.

In the litterature, two functional criteria admissible *w.r.t.* these requirements have been introduced: the Mass-Volume (MV) (Clémençon & Jakubowicz, 2013) and the Excess-Mass (EM) (Goix et al., 2015) curves. Formally, it allows to consider $\mathcal{C}^{\Phi}(s) = \|\Phi(s) - \Phi(f)\|$ (instead of $\|s - f\|$) with $\Phi : \mathbb{R} \to \mathbb{R}_+$ verifying $\Phi(T \circ s) = \Phi(s)$ for any scoring function s and increasing transform T. Here $\Phi(s)$ denotes either the mass-volume curve MV_s of s or its excess-mass curve EM_s, which are defined in the next section. While such quantities have originally been introduced to build scoring functions *via* Empirical Risk Minimization (ERM), the MV-curve has been used recently for the calibration of the One-Class SVM (Thomas et al., 2015). When used to attest the quality of some scoring function, the volumes induced become unknown and must be estimated, which is challenging in large dimension.

In this paper, we define two numerical performance criteria based on MV and EM curves, which are tested *w.r.t.*three classical AD algorithms. A wide range on real labeled datasets are used in the benchmark. In addition, we propose a method based on feature sub-sampling and aggregating. It allows to scale this methodology to high-dimensional data which we use on the higher-dimensional datasets. We compare the results to ROC and PR criteria, which use the data labels hidden to MV and EM curves.

This paper is structured as follows. Section 2 introduces EM and MV curves and defines associated numerical criteria. In Section 3, the feature sub-sampling based methodology to extend their use to high dimension is described. Finally, experiments on a wide range of real datasets are provided in Section 4.

2. Mass-Volume and Excess-Mass based criteria

We place ourselves in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We observe n $i.i.d.$ realizations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of a random variable $\mathbf{X} : \Omega \to \mathbb{R}^d$ representing the normal behavior, with *c.d.f.* F and density f *w.r.t.* the Lebesgue measure on $\mathbb{R}^d$. We denote by $\mathcal{S}$ the set of all scoring functions, namely any measurable function $s : \mathbb{R}^d \to \mathbb{R}_+$ integrable *w.r.t.* the Lebesgue measure. We work under the assumptions that the density f has no flat parts and is bounded. Excess-Mass and Mass-Volume curves are here introduced in a different way they originally were in (Clémençon & Jakubowicz, 2013; Goix et al., 2015). We use equivalent definitions for them since the original definitions were more adapted to the ERM paradigm than to the issues adressed here.

Preliminaries. Let $s \in \mathcal{S}$ be a scoring function. In this context (Clémençon & Jakubowicz, 2013; Goix et al., 2015), the MV and EM curves of s can be written as

$$MV_s(\alpha) = \inf_{u \geq 0} \ \text{Leb}(s \geq u) \ \ s.t. \ \mathbb{P}(s(\mathbf{X}) \geq u) \geq \alpha \quad (1)$$

$$EM_s(t) = \sup_{u \geq 0} \ \mathbb{P}(s(\mathbf{X}) \geq u) \ - \ t\text{Leb}(s \geq u) \quad (2)$$

for any $\alpha \in (0, 1)$ and $t > 0$. The optimal curves are $MV^* = MV_f = MV_{T \circ f}$ and $EM^* = EM_f = EM_{T \circ f}$ for any increasing transform $T : \text{Im(f)} \to \mathbb{R}$. It can be proven (Clémençon & Jakubowicz, 2013; Goix et al., 2015) that for any scoring function s, $MV^*(\alpha) \leq MV_s(\alpha)$ for all $\alpha \in (0, 1)$ and $EM^*(t) \geq EM_s(t)$ for all $t > 0$.

Numerical unsupervised criteria. The main advantage of EM compared to MV is that the area under its curve (AUC) is finite, even if the support of the distribution F is not. As curves cannot be trivially compared, consider the L^1-norm $\|.\|_{L^1(I)}$ with $I \subset \mathbb{R}$ an interval. As $MV^* = MV_f$ is below MV_s pointwise, $\arg\min_s \|MV_s - MV^*\|_{L^1(I)} = \arg\min \|MV_s\|_{L^1(I)}$. We thus define $\mathcal{C}^{MV}(s) = \|MV_s\|_{L^1(I^{MV})}$, which is equivalent to consider $\|MV_s - MV^*\|_{L^1(I^{MV})}$ as mentioned in the introduction. As we are interested in evaluating accuracy on large density level-sets, one natural interval I^{MV} would be for instance $[0.9, 1]$. However, MV diverges in 1 when the support is infinite, so that we arbitrarily take $I^{MV} = [0.9, 0.999]$. The smaller is $\mathcal{C}^{MV}(s)$, the better is the scoring function s. Similarly, we consider $\mathcal{C}^{EM}(s) = \|EM_s\|_{L^1(I^{EM})}$, this time considering $I^{EM} = [0, EM^{-1}(0.9)]$, with $EM_s^{-1}(0.9) := \inf\{t \geq 0, \ EM_s(t) \leq 0.9\}$, as $EM_s(0)$ is finite (equal to 1). We point out that such small values of t correspond to large level-sets. Also, we have observed that $EM_s^{-1}(0.9)$ (as well as $EM_f^{-1}(0.9)$) varies significantly depending on the dataset. Generally, for datasets in large dimension, it can be very small (in the experiments, smallest values are of order 10^{-7}) as it is of the same order of magnitude as the inverse of the total support volume.

Estimation. As the distribution F of the normal data is generally unknown, MV and EM curves must be estimated. Let $s \in \mathcal{S}$ and $\mathbf{X}_1, \ \ldots, \ \mathbf{X}_n$ be an i.i.d. sample with common distribution F and set $\mathbb{P}_n(s \geq t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{s(\mathbf{X}_i) \geq t}$. The empirical MV and EM curves of s are then simply defined as empirical version of (1) and (2),

$$\widehat{MV}_s(\alpha) = \inf_{u \geq 0} \{\text{Leb}(s \geq u) \ \ s.t. \ \mathbb{P}_n(s \geq u) \geq \alpha\} \quad (3)$$

$$\widehat{EM}_s(t) = \sup_{u \geq 0} \mathbb{P}_n(s \geq u) \; - \; t\text{Leb}(s \geq u) \qquad (4)$$

Note that in practice, the volume $\text{Leb}(s \geq u)$ is estimated using Monte-Carlo approximation, which only applies to small dimensions. Finally, we obtain the empirical EM and MV based performance criteria:

$$\widehat{\mathcal{C}}^{EM}(s) = \|\widehat{EM}_s\|_{L^1(I^{EM})} \quad I^{EM} = [0, \widehat{EM}^{-1}(0.9)], \qquad (5)$$

$$\widehat{\mathcal{C}}^{MV}(s) = \|\widehat{MV}_s\|_{L^1(I^{MV})} \quad I^{MV} = [0.9, 0.999]. \qquad (6)$$

3. Scaling with dimension

In this section we propose a methodology to scale the use of the EM and MV criteria to large dimensional data. It consists in sub-sampling training *and* testing data along features, thanks to a parameter d' controlling the number of features randomly chosen for computing the (EM or MV) score. Replacement is done after each draw of features $F_1, \ldots, F_m$. A partial score $\widehat{\mathcal{C}}_k^{MV}$ (resp. $\widehat{\mathcal{C}}_k^{EM}$) is computed for each draw F_k using (5) (resp. (6)). The final performance criteria are obtained by averaging these partial criteria along the different draws of features. This methodology is described in Algorithm 1. A drawback from this approach is that we do not evaluate combinations of more than d' features within the dependence structure. However, according to our experiments, this is enough in most of the cases. Besides, we solve two major drawbacks inherent to MV or EM criteria, which come from the Lebesgue reference measure: **1)** EM or MV performance criteria cannot be estimated in large dimension, **2)** EM or MV performance criteria cannot be compared when produced from spaces of different dimensions.

Algorithm 1 Evaluate AD algo. on high-dimensional data

Inputs: AD algorithm $\mathcal{A}$, data set $X = (x_i^j)_{1 \leq i \leq n, 1 \leq j \leq d}$, feature sub-sampling size d', number of draws m.

for $k = 1, \ldots, m$ **do**

randomly select a sub-group F_k of d' features

compute the associated scoring function $\widehat{s}_k = \mathcal{A}\big((x_i^j)_{1 \leq i \leq n,\, j \in F_k}\big)$

compute $\widehat{\mathcal{C}}_k^{EM} = \|\widehat{EM}_{\widehat{s}_k}\|_{L^1(I^{EM})}$ using (5) or $\widehat{\mathcal{C}}_k^{MV} = \|\widehat{MV}_{\widehat{s}_k}\|_{L^1(I^{MV})}$ using (6)

end for

Return performance criteria:

$$\widehat{\mathcal{C}}^{EM}_{high\_dim}(\mathcal{A}) = \frac{1}{m} \sum_{k=1}^{m} \widehat{\mathcal{C}}_k^{EM} \quad \text{(idem for MV)}$$

Remark 1 *(*FEATURE IMPORTANCES*) With standard MV and EM curves, the benefit of using or not some feature j in training* cannot *be evaluated, since reference measures of $\mathbb{R}^d$ and $\mathbb{R}^{d+1}$ cannot be compared. Solving the second drawback precisely allows to evaluate the importance of features. By sub-sampling features, we can compare accuracies with or without using feature j: when computing $\widehat{\mathcal{C}}^{MV}_{high\_dim}$ or $\widehat{\mathcal{C}}^{EM}_{high\_dim}$ using Algorithm 1, this is reflected in the fact that j can (resp. cannot) be drawn.*

Remarks on theoretical grounds and default parameters are provided in supplementary material.

4. Benchmarks

Does performance in term of EM/MV correspond to performance in term of ROC/PR? Can we recover, on some fixed dataset and without using any labels, which algorithm is better than the others (according to ROC/PR criteria)? In this section we study four different empirical evaluations (ROC, PR, EM, MV) of three classical state-of-the-art AD algorithms, One-Class SVM (Schölkopf et al., 2001), Isolation Forest (Liu et al., 2008), and Local Outlier Factor (LOF) algorithm (Breunig et al., 2000), on 12 well-known AD datasets. Two criteria use labels (ROC and PR based criteria) and two do not (EM and MV based criteria). For ROC and PR curves, we consider the area under the (full) curve (AUC). For the excess-mass curve $EM(t)$ (resp. mass-volume curve), we consider the area under the curve on the interval $[0, EM^{-1}(0.9)]$ (resp. $[0.9, 0.999]$) as described in Section 2. A full description of the datasets is available in supplementary material. The experiments are performed both in a novelty detection framework (also named semi-supervised framework, the training set consisting of normal data only) and in an unsupervised framework (the training set is polluted by anormal data). In the former case, we simply removed anomalies from the training data, and EM and PR criteria are estimated using only normal data. In the latter case, the anomaly rate is arbitrarily bounded to 10% max, and EM and PR criteria are estimated with the same test data used for ROC and PR curves, without using their labels. Recall that standards EM and MV performance criteria refering on the Lebesgue measure, they require volume estimation. They only apply to continuous datasets, with small dimension ($d \leq 8$). The datasets verifying these requirements are *http*, *smtp*, *pima*, *wilt* and *adult*. For the other datasets, we use the performance criteria $\widehat{\mathcal{C}}^{MV}_{high\_dim}$ and $\widehat{\mathcal{C}}^{EM}_{high\_dim}$ computed with Algorithm 1. We arbitrarily chose $m = 50$ and $d' = 5$, which means that 50 draws of 5 features, with replacement after each draw, are done. Other parameters have also been tested but are not presented here. This default parameters are a compromise between computational time and performance, in particular on the largest dimensional datasets. The latter require a relatively large product $m \times d'$, which is the maximal number of different features that can be drawn.

Table 1. Results for the novelty detection setting. One can see that ROC, PR, EM, MV often do agree on which algorithm is the best (in bold), which algorithm is the worse (underlined) on some fixed datasets. When they do not agree, it is often because ROC and PR themselves do not, meaning that the ranking is not clear.

Dataset	iForest				OCSVM				LOF			
	ROC	PR	EM	MV	ROC	PR	EM	MV	ROC	PR	EM	MV
adult	**0.661**	**0.277**	**1.0e-04**	**7.5e01**	0.642	0.206	2.9e-05	4.3e02	0.618	0.187	1.7e-05	9.0e02
http	0.994	0.192	1.3e-03	9.0	**0.999**	**0.970**	**6.0e-03**	**2.6**	0.946	0.035	8.0e-05	3.9e02
pima	0.727	0.182	5.0e-07	**1.2e04**	**0.760**	**0.229**	**5.2e-07**	1.3e04	0.705	0.155	3.2e-07	2.1e04
smtp	0.907	0.005	1.8e-04	9.4e01	0.852	**0.522**	**1.2e-03**	8.2	**0.922**	0.189	1.1e-03	**5.8**
wilt	0.491	0.045	4.7e-05	2.1e03	0.325	0.037	**5.9e-05**	**4.5e02**	**0.698**	**0.088**	2.1e-05	1.6e03
annthyroid	**0.913**	**0.456**	**2.0e-04**	2.6e02	0.699	0.237	6.3e-05	**2.2e02**	0.823	0.432	6.3e-05	1.5e03
arrhythmia	**0.763**	**0.487**	**1.6e-04**	**9.4e01**	0.736	0.449	1.1e-04	1.0e02	0.730	0.413	8.3e-05	1.6e02
forestcov.	0.863	0.046	3.9e-05	2.0e02	0.958	0.110	5.2e-05	1.2e02	**0.990**	**0.792**	**3.5e-04**	**3.9e01**
ionosphere	0.902	0.529	9.6e-05	7.5e01	**0.977**	**0.898**	**1.3e-04**	**5.4e01**	0.971	0.895	1.0e-04	7.0e01
pendigits	0.811	0.197	2.8e-04	2.6e01	0.606	0.112	2.7e-04	2.7e01	**0.983**	**0.829**	**4.6e-04**	**1.7e01**
shuttle	0.996	0.973	1.8e-05	5.7e03	0.992	0.924	**3.2e-05**	**2.0e01**	**0.999**	**0.994**	7.9e-06	2.0e06
spambase	**0.824**	**0.371**	**9.5e-04**	**4.5e01**	0.729	0.230	4.9e-04	1.1e03	0.754	0.173	2.2e-04	4.1e04

EM, MV, ROC and PR curves AUCs are presented in Table 1 for the novelty detection framework. Additional figures and results for the unsupervised framework are available in supplementary material. Results from Table 1 can be summarized as follows. Consider the 36 possible pairwise comparisons between the three algorithms over the twelve datasets

$$\big\{\big(A_1 \text{ on } \mathcal{D}, A_2 \text{ on } \mathcal{D}\big),\, A_{12} \in \{\text{iForest, LOF, OCSVM}\},\\ \mathcal{D} \in \{\text{adult}, \ldots, \text{spambase}\}\big\}. \quad (7)$$

For each dataset $\mathcal{D}$, there are three possible pairs (iForest on $\mathcal{D}$, LOF on $\mathcal{D}$), (OCSVM on $\mathcal{D}$, LOF on $\mathcal{D}$) and (OCSVM on $\mathcal{D}$, iForest on $\mathcal{D}$). Then the EM-score discriminates 28 of them (78%) as ROC score does, and 29 (81%) of them as PR score does. Intuitively this can be interprated as follows. Choose randomly a dataset $\mathcal{D}$ among the twelve available, and two algorithms A_1, A_2 among the three available. This amounts to choose at random a pairwise comparison (A_1 on $\mathcal{D}$, A_2 on $\mathcal{D}$) among the 36 available. Suppose that according to ROC criterion, A_1 is better than A_2 on dataset $\mathcal{D}$, *i.e.* (A_1 on $\mathcal{D}$) $\succ$ (A_2 on $\mathcal{D}$). Then the EM-score discriminates A_1 and A_2 on dataset $\mathcal{D}$ in the same way, *i.e.* also finds A_1 to be better than A_2 on dataset $\mathcal{D}$, this with 78 percent chance.

Besides, let us consider pairs (A_1 on $\mathcal{D}$, A_2 on $\mathcal{D}$) which are similarly ordered by ROC and PR criteria, namely *s.t.* A_1 is better than A_2 (or the reverse) on dataset $\mathcal{D}$ according to both EM and PR. According to Table 1, this represents every pairs but one in *spambase* and two in *smtp*. Then, one achieves $27/33 = 82\%$ of similarly discriminated pairs (*w.r.t.* to ROC and PR criteria). Moreover, EM is able to recover the exact (*w.r.t.* ROC and PR criteria) ranking of (A_1 on $\mathcal{D}$, A_2 on $\mathcal{D}$, A_3 on $\mathcal{D}$) on every datasets $\mathcal{D}$ excepting *wilt* and *shuttle*. For *shuttle*, note that ROC scores are very close to each other (0.996, 0.992, 0.999) and thus not clearly discriminates algorithms. The only significant error committed by EM is for the *wilt* dataset (on which no feature sub-sampling is done due to the low dimension). This may come from anomalies not being far enough in the tail of the normal distribution, *e.g.* forming a cluster near the support of the latter distribution.

Same conclusions and similar accuracies hold for MV-score, which only makes one additional error on the pair (iForest on *pima*, OCSVM on *pima*). Considering all the 36 pairs (7), one observes 75% of good comparisons *w.r.t.* ROC-score, and 72% *w.r.t.* PR score. Considering the pairs which are similarly ordered by ROC and PR criteria, this rate increases to $25/33 = 76\%$. The errors are essentially made on *shuttle*, *wild* and *annthyroid* datasets.

To conclude, when one algorithm has better performance than another on some fixed dataset, according to both ROC and PR AUCs, one can expect to recover it without using labels with an accuracy of 82% in the novelty detection framework (and 77% in the unsupervised framework, cf. supplementary material).

5. Conclusion

We (almost) do not need labels to evaluate anomaly detection algorithms (on continuous data). According to our benchmarks, the EM and MV based numerical criteria introduced in this paper are (in approximately 80 percent of the cases) able to recover which algorithm is better than the other on some dataset (with potentially large dimensionality), without using labels. High-dimensional datasets are dealt with using a method based on feature sub-sampling. This method also brings flexibility to EM and MV criteria, allowing for instance to evaluate the importance of features.

References

Aggarwal, C.C. and Yu, P.S. Outlier detection for high dimensional data. In *ACM Sigmod Record*, 2001.

Barnett, V. and Lewis, T. *Outliers in statistical data*, volume 3. Wiley New York, 1994.

Breunig, M.M., Kriegel, H.P., Ng, R.T., and Sander, J. LOF: identifying density-based local outliers. In *ACM sigmod record*, 2000.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.*, 2009.

Clémençon, S. and Jakubowicz, J. Scoring anomalies: a M-estimation approach. In *AISTATS*, 2013.

Clémençon, S. and Robbiano, S. Anomaly Ranking as Supervised Bipartite Ranking. In *ICML*, 2014.

Eskin, E. Anomaly detection over noisy data using learned probability distributions. In *ICML*, 2000.

Goix, N., Sabourin, A., and Clémençon, S. On Anomaly Ranking and Excess-Mass Curves. In *AISTATS*, 2015.

Hodge, V.J. and Austin, J. A survey of outlier detection methodologies. *Artif. Intel. Review*, 2004.

KDDCup. The third international knowledge discovery and data mining tools competition dataset. 1999.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Liu, F.T., Ting, K.M., and Zhou, Z.H. Isolation Forest. In *ICDM*, 2008.

Markou, M. and Singh, S. Novelty detection: a review part 1: statistical approaches. *Signal proc.*, 2003.

Patcha, A. and Park, J.M. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 2007.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in Python. *JMLR*, 2011.

Polonik, W. Measuring Mass Concentrations and Estimating Density Contour Cluster-An excess Mass Approach. *The Annals of Statistics*, 1995.

Polonik, W. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 1997.

Schölkopf, B., Platt, J.C, Shawe-Taylor, J., Smola, A.J, and Williamson, R.C. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7): 1443–1471, 2001.

Schubert, E., Wojdanowski, R., Zimek, A., and Kriegel, H.-P. On Evaluation of Outlier Rankings and Outlier Scores. In *SDM*, pp. 1047–1058. SIAM, 2012.

Scott, C.D and Nowak, R.D. Learning minimum volume sets. *The Journal of Machine Learning Research*, 7:665–704, 2006.

Shawe-Taylor, J. and Cristianini, N. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Shyu, M.L., Chen, S.C., Sarinnapakorn, K., and Chang, L. A novel anomaly detection scheme based on principal component classifier. Technical report, DTIC Document, 2003.

Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A.A. A detailed analysis of the kdd cup 99 data set. In *IEEE CISDA*, 2009.

Thomas, A., Feuillard, V., and Gramfort, A. Calibration of One-Class SVM for MV set estimation. In *DSAA*, 2015.

Yamanishi, K., Takeuchi, J.I., Williams, G., and Milne, P. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *KDD*, 2000.

Supplementary Material

5.1. additional intuition behind EM/MV

Note that $MV^*(\alpha)$ is the optimal value of the constrained minimization problem

$$\min_{\Gamma \text{ borelian}} \text{Leb}(\Gamma) \quad s.t. \ \mathbb{P}(\mathbf{X} \in \Gamma) \geq \alpha. \tag{8}$$

The minimization problem (8) has a unique solution Γ^*_α of mass α exactly, referred to as *minimum volume set* (Polonik, 1997): $MV^*(\alpha) = \text{Leb}(\Gamma^*_\alpha)$ and $\mathbb{P}(\mathbf{X} \in \Gamma^*_\alpha) = \alpha$.

Similarly, the optimal EM curve is linked with the notion of density excess-mass (as introduced in the seminal contribution (Polonik, 1995)). The main idea is to consider a Lagrangian formulation of the constrained minimization problem obtained by exchanging constraint and objective in (8),

$$EM^*(t) := \max_{\Omega \text{ borelian}} \{\mathbb{P}(\mathbf{X} \in \Omega) - t\text{Leb}(\Omega)\}. \tag{9}$$

Figure 1 compares the mass-volume and excess-mass approaches.

Remark 2 *(*LINK WITH ROC CURVE*) To evaluate unsupervised algorithms, it is common to generate uniform outliers and then use the ROC curve approach. Up to identify the Lebesgue measure of a set to its empirical version (*i.e. *the proportion of uniform point inside), this approach is equivalent to using the mass-volume curve (Clémençon & Robbiano, 2014). However, in the former approach, the volume estimation does not appear directly, so that the (potentially huge) amount of uniform points needed to provide a good estimate of a volume is often not respected, yielding optimistic performances.*

5.2. Remarks on the feature sub-sampling based Algorithm 1.

Remark 3 *(*THEORETICAL GROUNDS*) Criteria $\widehat{\mathcal{C}}^{MV}_{high\_dim}$ or $\widehat{\mathcal{C}}^{EM}_{high\_dim}$ do not evaluate a specific scoring function s produced by some algorithm (on some dataset), but the algorithm itself* w.r.t. *the dataset at stake. Indeed, these criteria proceed with the average of partial scoring functions on sub-space of $\mathbb{R}^d$. We have no theoretical guaranties that the final score does correspond to some scoring function defined on $\mathbb{R}^d$. In this paper, we only show that from a practical point of view, it is a useful and accurate methodology to compare algorithms performance on large dimensional datasets.*

Remark 4 *(*DEFAULT PARAMETERS*) In our experiments, we arbitrarily chose $m = 50$ and $d' = 5$. This means that 50 draws of 5 features (with replacement after each draw) have been done. Volume in spaces of dimension 5 have thus to be estimated (which is feasible with Monte-Carlo), and 50 scoring functions (on random subspaces of dimension 5) have to be computed by the algorithm we want to evaluate. The next section shows (empirically) that these parameters achieve a good accuracy on the collection of datasets studied, the largest dimension considered being 164.*

5.3. Datasets description

The characteristics of these reference datasets are summarized in Table 2. They are all available on the UCI repository (Lichman, 2013) and the preprocessing is done in a classical way. We removed all non-continuous attributes as well as attributes taking less than 10 differents values. The *http* and *smtp* datasets belong to the KDD Cup '99 dataset (KDDCup, 1999; Tavallaee et al., 2009), which consists of a wide variety of hand-injected attacks (anomalies) in a closed network (normal background). They are classicaly obtained as described in (Yamanishi et al., 2000). These datasets are available on the *scikit-learn* library (Pedregosa et al., 2011). The *shuttle* dataset is the fusion of the training and testing datasets available in the UCI repository. As in (Liu et al., 2008), we use instances from all different classes but class 4. In the *forestcover* data, the normal data are the instances from class 2 while instances from class 4 are anomalies (as in (Liu et al., 2008)). The *ionosphere* dataset differentiates 'good' from 'bad' radars, considered here as abnormal. A 'good' radar shows evidence of some type of structure in the ionosphere. A 'bad' radar does not, its signal passing through the ionosphere. The *spambase* dataset consists of spam or non-spam emails. The former constitute the abnomal class. The *annthyroid* medical dataset on hypothyroidism contains one normal class and two abnormal ones, which form the outlier set. The *arrhythmia* dataset reflects the presence and absence (class 1) of cardiac arrhythmia. The number of attributes being large considering the sample size, we removed attributes containing missing data. The *pendigits* dataset contains 10 classes corresponding to the digits from 0 to 9, examples being handwriting samples. As in (Schubert et al., 2012), the abnormal data are chosen to be those from class 4. The *pima* dataset consists of medical data on diabetes. Patients suffering from diabetes (positive class) were considered outliers. The *wild* dataset involves detecting diseased trees in Quickbird imagery. Diseased trees (class 'w') is the abnormal class. In the *adult* dataset, the goal is to predict whether income exceeds $ 50K/year based on census data. Only the 6 continuous attributes are kept.

5.4. complementary results

Results from the unsupervised framework (training and testing data are polluted by outliers) are similar for both

Figure 1. Comparison between $MV^*(\alpha)$ and $EM^*(t)$

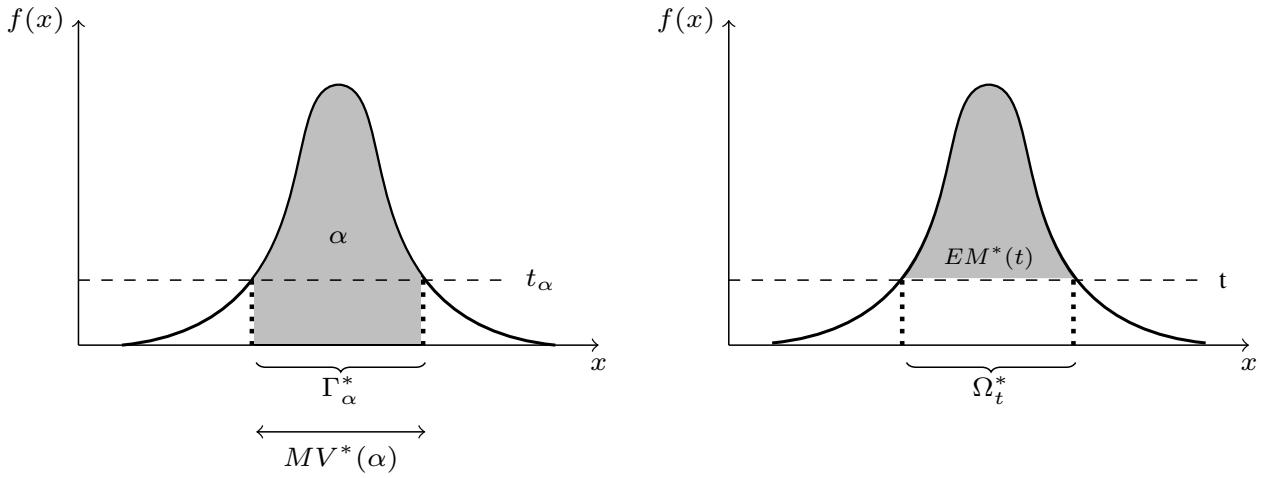

Table 2. Original Datasets characteristics

	nb of samples	nb of features	anomaly class	
adult	48842	6	class '$> 50K$'	(23.9%)
http	567498	3	attack	(0.39%)
pima	768	8	pos (class 1)	(34.9%)
smtp	95156	3	attack	(0.03%)
wilt	4839	5	class 'w' (diseased trees)	(5.39%)
annthyroid	7200	6	classes $\neq 3$	(7.42%)
arrhythmia	452	164	classes $\neq 1$ (features 10-14 removed)	(45.8%)
forestcover	286048	10	class 4 (vs. class 2)	(0.96%)
ionosphere	351	32	bad	(35.9%)
pendigits	10992	16	class 4	(10.4%)
shuttle	85849	9	classes $\neq 1$ (class 4 removed)	(7.17%)
spambase	4601	57	spam	(39.4%)

EM and MV criteria. We just observe a slight decrease in accuracy. Considering all the pairs, one observes $26/36 = 72\%$ (resp. $27/36 = 75\%$) of good comparisons *w.r.t.* ROC-score (resp. *w.r.t.* PR score) for EM, and 75% (resp. 78%) of good comparisons *w.r.t.* ROC-score (resp. *w.r.t.* PR score) for MV. Considering the pairs which are similarly ordered by ROC and PR criteria, the rate for EM as for MV increases to $24/31 = 77\%$. Figure 5.4 shows excess-mass and mass-volume curves on the adult dataset in a novelty detection setting. Corresponding figures for the other datasets follow.

Table 3. Results for the unsupervised setting still remains good: one can see that ROC, PR, EM, MV often do agree on which algorithm is the best (in bold), which algorithm is the worse (underlined) on some fixed datasets. When they do not agree, it is often because ROC and PR themselves do not, meaning that the ranking is not clear.

Dataset	iForest				OCSVM				LOF			
	ROC	PR	EM	MV	ROC	PR	EM	MV	ROC	PR	EM	MV
adult	**0.644**	**0.234**	**6.6e-05**	**2.7e02**	0.627	0.184	1.8e-05	5.6e02	0.545	0.098	7.4e-06	1.9e03
http	**0.999**	**0.686**	1.4e-03	2.2e01	0.994	0.207	**5.7e-03**	**3.3**	0.354	0.019	9.8e-05	3.9e02
pima	**0.747**	0.205	**1.2e-06**	**1.2e04**	0.742	**0.211**	6.0e-07	1.9e04	0.686	0.143	6.0e-07	3.2e04
smtp	0.902	0.004	2.7e-04	8.6e01	0.852	**0.365**	**1.4e-03**	7.7	**0.912**	0.057	1.1e-03	**7.0**
wilt	0.443	0.044	3.7e-05	2.2e03	0.318	0.036	**3.9e-05**	**4.3e02**	**0.620**	**0.066**	2.0e-05	8.9e02
annthyroid	**0.820**	**0.309**	**6.9e-05**	7.7e02	0.682	0.187	4.1e-05	**3.1e02**	0.724	0.175	1.6e-05	4.1e03
arrhythmia	**0.740**	0.416	**8.4e-05**	**1.1e02**	0.729	**0.447**	6.8e-05	1.2e02	0.729	0.409	5.6e-05	1.5e02
forestcov.	0.882	0.062	3.2e-05	2.3e02	**0.951**	**0.095**	4.4e-05	1.4e02	0.542	0.016	**2.4e-04**	**4.6e01**
ionosphere	0.895	0.543	7.4e-05	9.3e01	**0.977**	**0.903**	**8.7e-05**	**7.7e01**	0.969	0.884	6.9e-05	1.0e02
pendigits	0.463	0.077	2.7e-04	2.5e01	0.366	0.067	2.6e-04	2.8e01	**0.504**	**0.089**	**4.5e-04**	**1.6e01**
shuttle	**0.997**	**0.979**	7.1e-07	1.2e05	0.992	0.904	**5.8e-06**	**1.7e02**	0.526	0.116	7.1e-07	1.7e07
spambase	**0.799**	**0.303**	**2.2e-04**	**3.5e01**	0.714	0.214	1.5e-04	2.9e02	0.670	0.129	3.7e-05	2.7e04

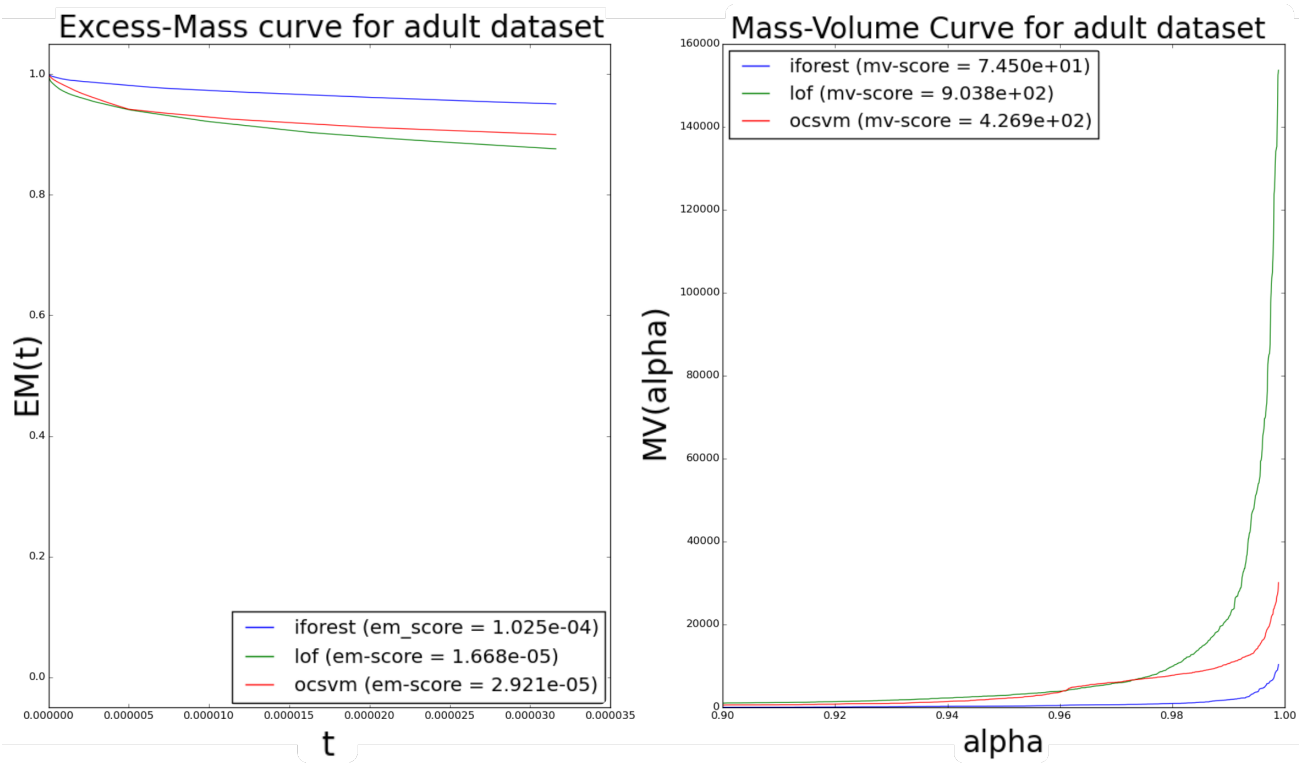

Figure 2. MV and EM curves for adult dataset (novelty detection framework). We can see that both in terms of EM and MV curves, iForest is found to perform better than OCSVM, which is itself found to perform better than LOF. Comparing to Table 1, ROC and PR AUCs give the same ranking (iForest on adult $\succ$ OCSVM on adult $\succ$ LOF on adult). The 3 pairwise comparisons (iForest on adult, LOF on adult), (OCSVM on adult, LOF on adult) and (OCSVM on adult, iForest on adult) are then similarly ordered by EM, PR, MV and EM criteria.

Figure 3. MV and EM curves for http dataset (novelty detection framework)

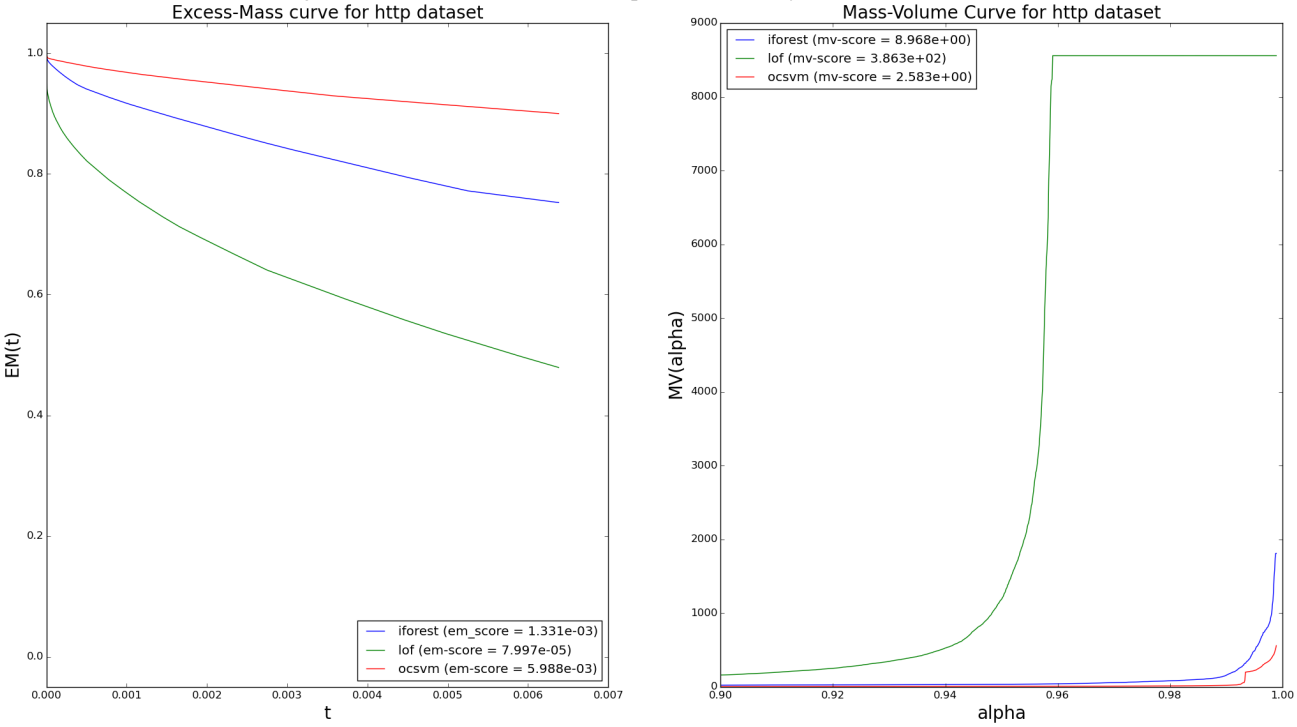

Figure 4. MV and EM curves for http dataset (unsupervised framework)

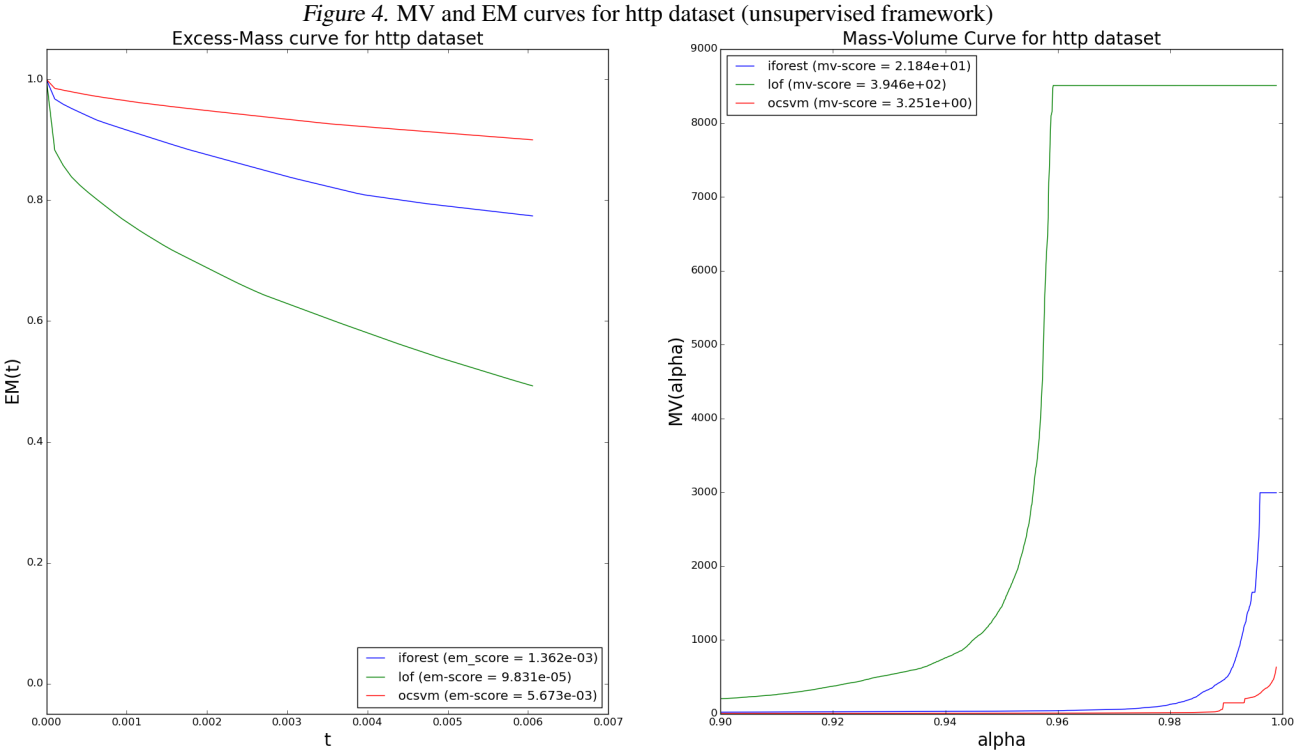

Figure 5. MV and EM curves for pima dataset (novelty detection framework)

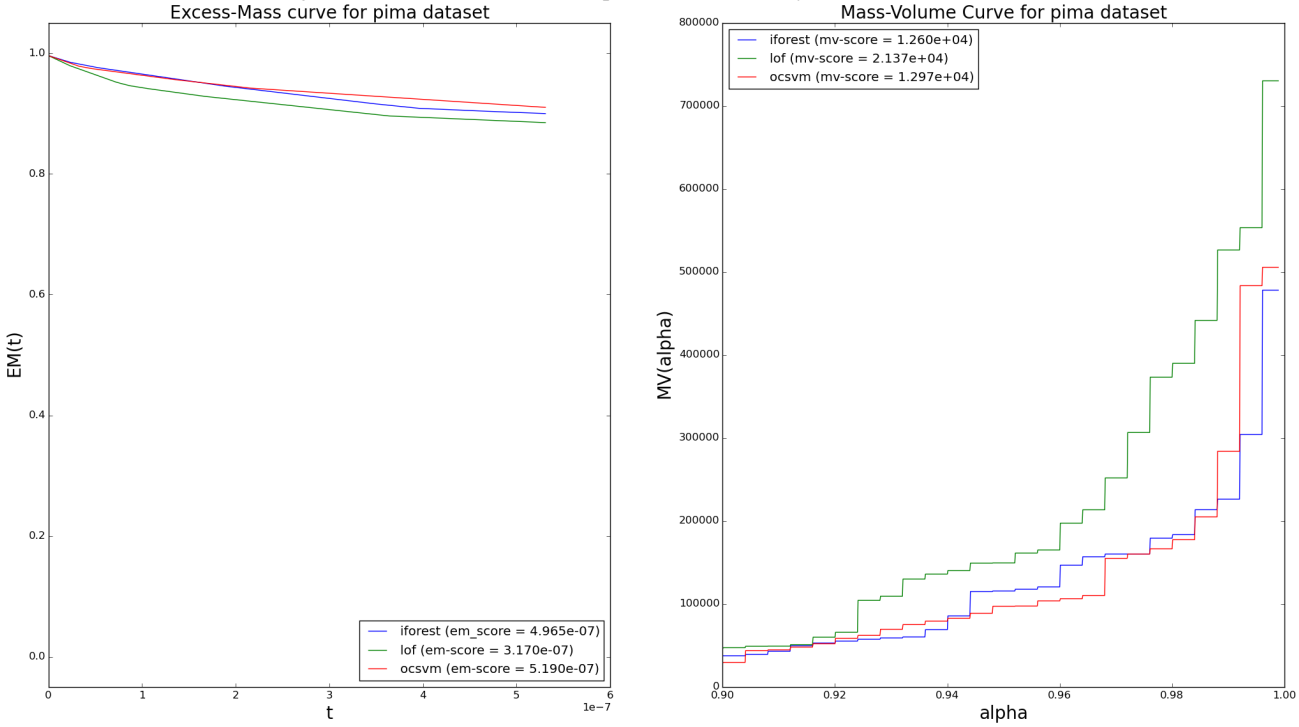

Figure 6. MV and EM curves for pima dataset (unsupervised framework)

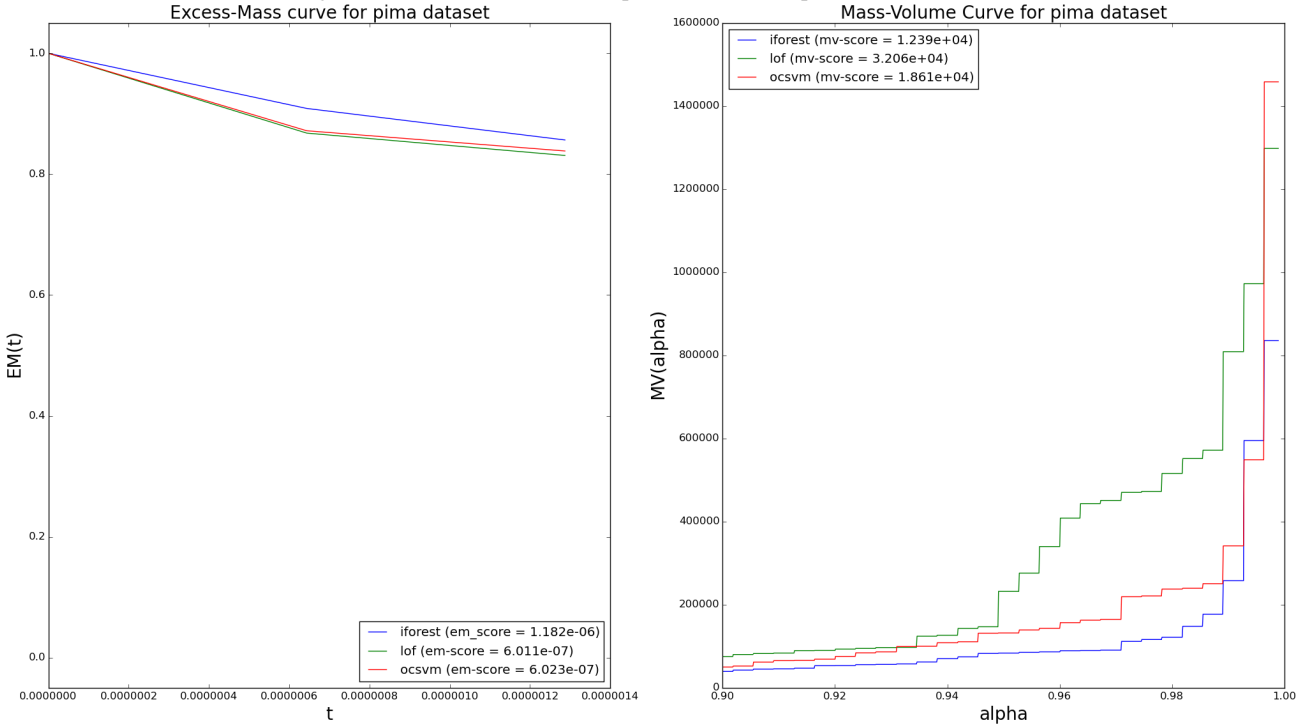

Figure 7. MV and EM curves for smtp dataset (novelty detection framework)

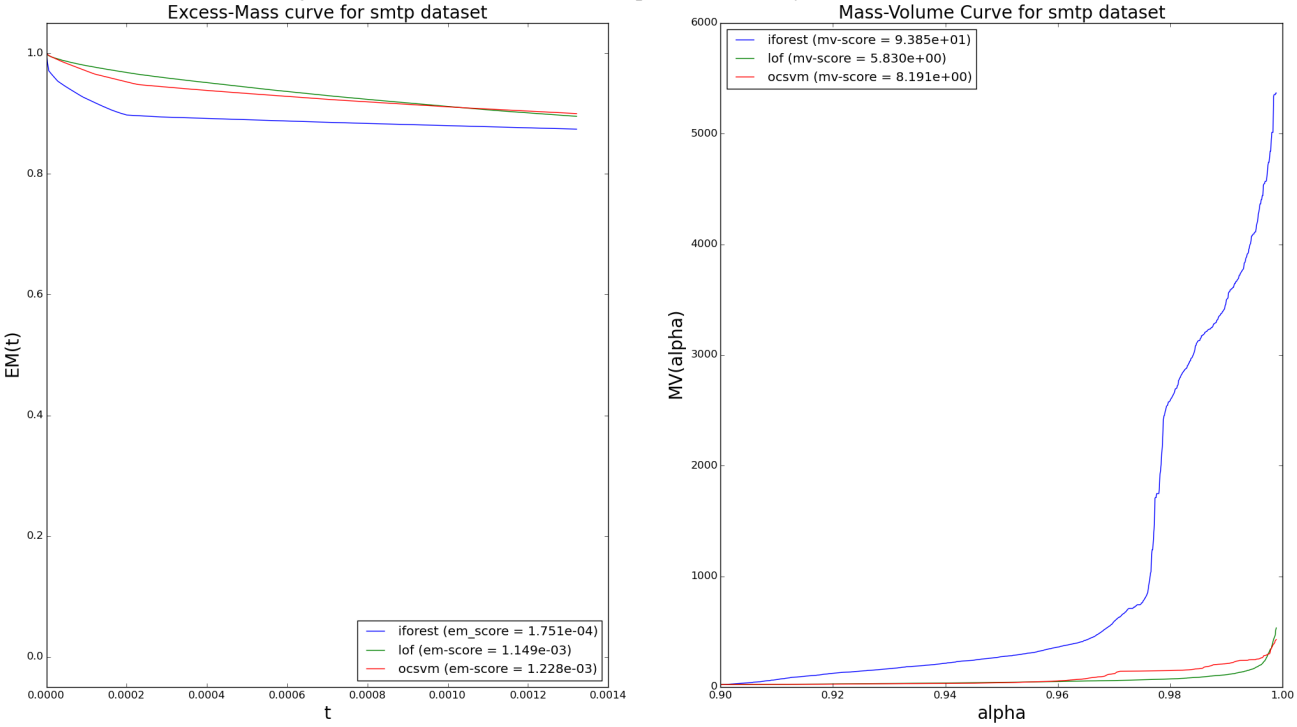

Figure 8. MV and EM curves for smtp dataset (unsupervised framework)

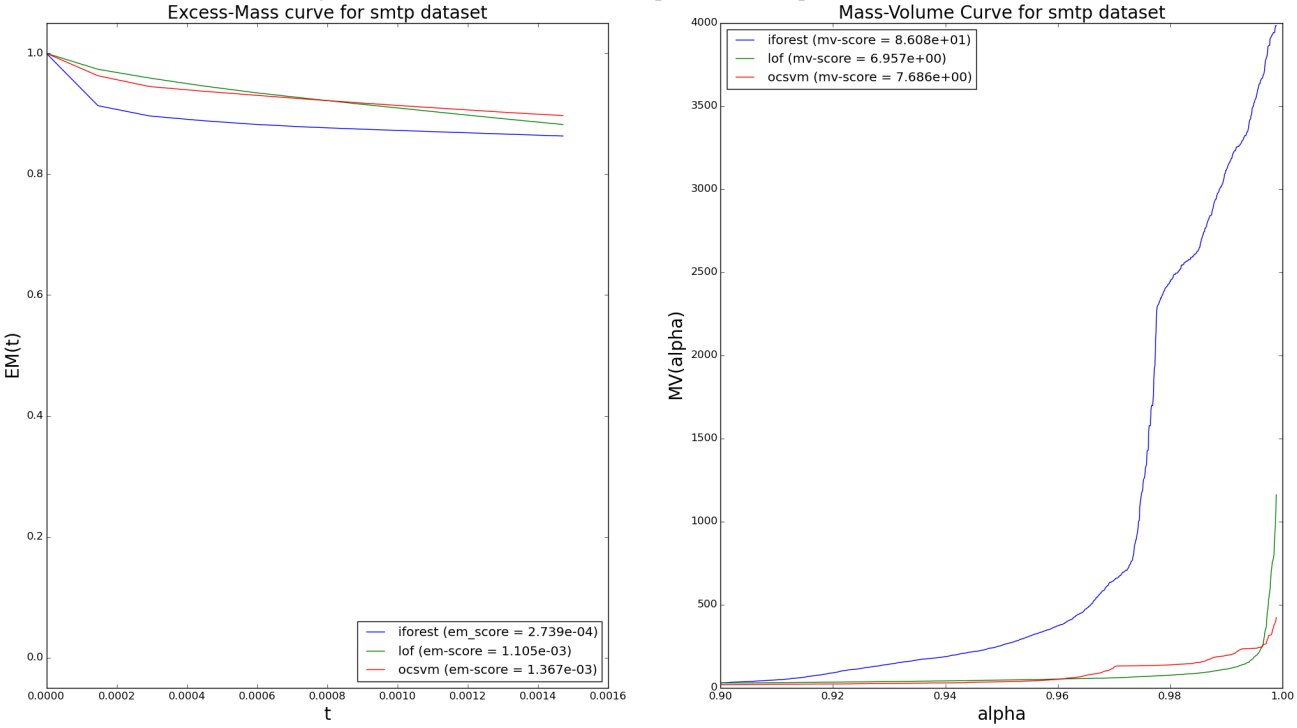

Figure 9. MV and EM curves for wilt dataset (novelty detection framework)

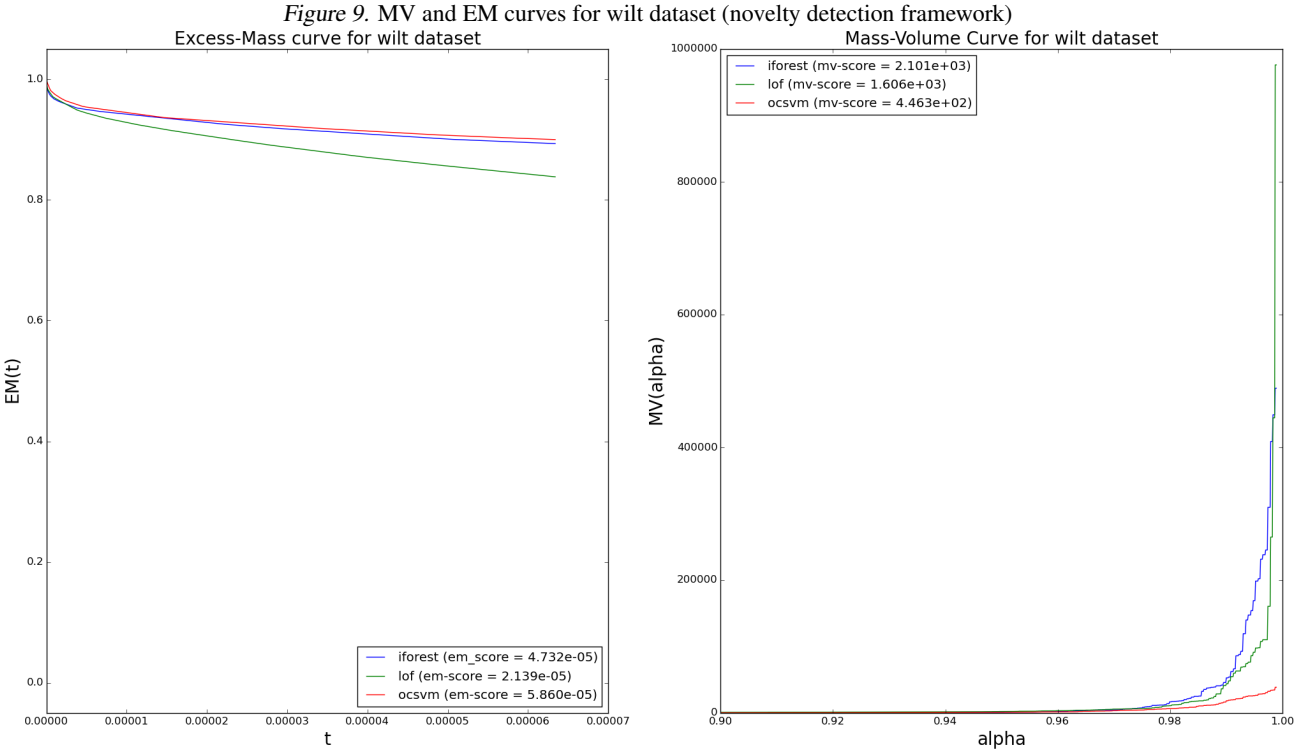

Figure 10. MV and EM curves for wilt dataset (unsupervised framework)

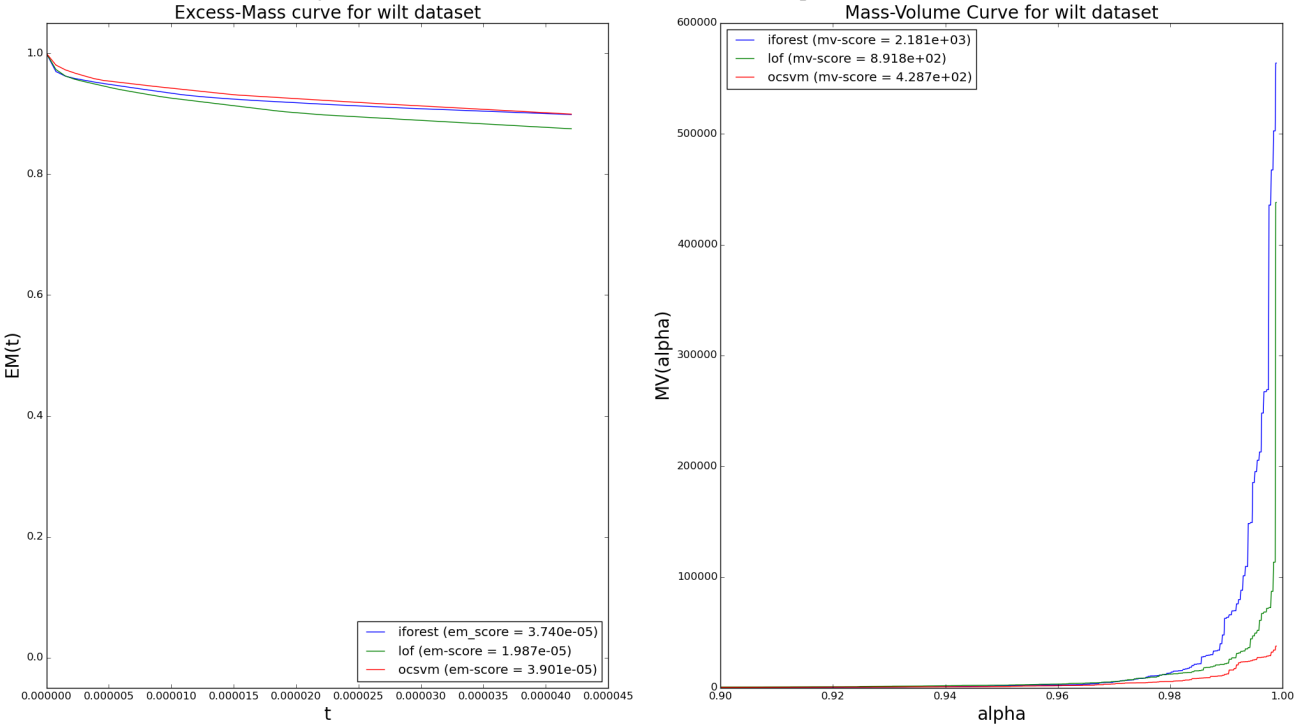

Figure 11. MV and EM curves for adult dataset (novelty detection framework).

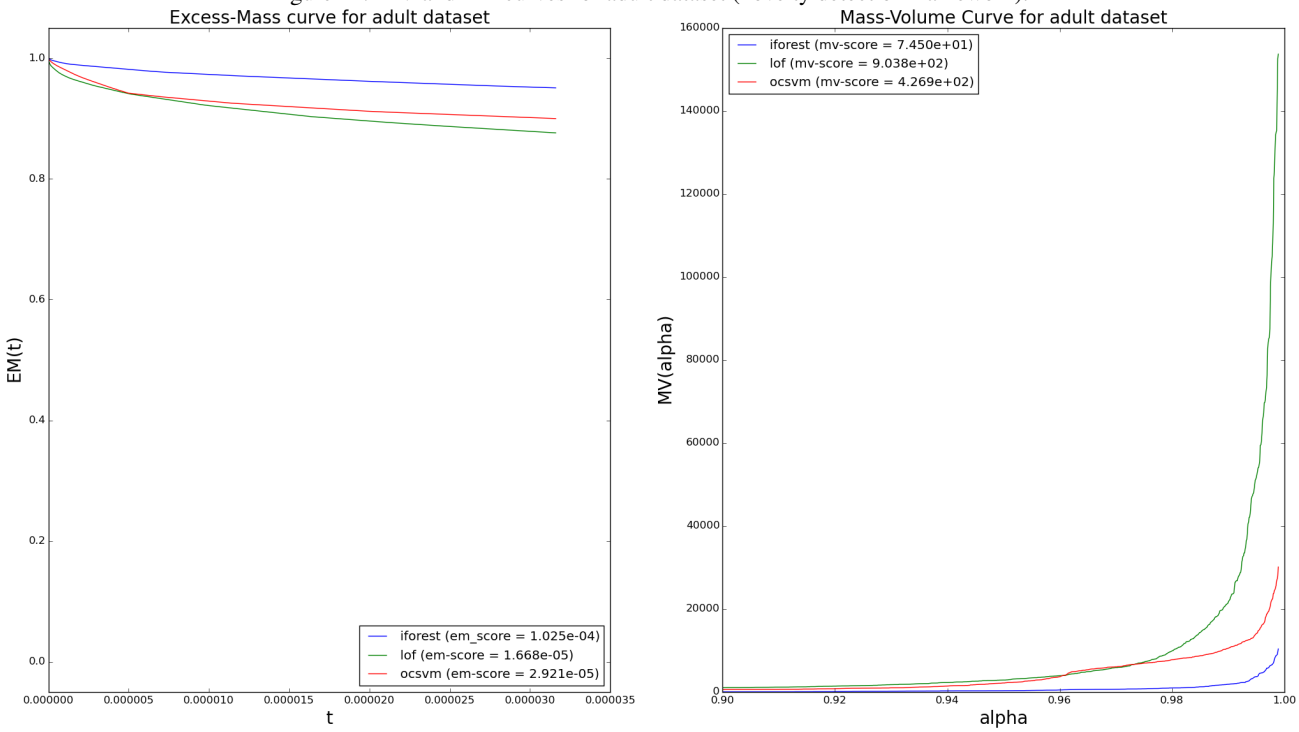

Figure 12. MV and EM curves for adult dataset (unsupervised framework)