Eurographics Conference on Visualization (EuroVis) 2018
J. Heer, H. Leitte, and T. Ropinski
(Guest Editors)

Volume 37 (2018), Number 3

Towards User-Centered Active Learning Algorithms – Supplemental Materials –

SUBMISSION ID 1249

Abstract

This document contains supplemental material to the work presented in the main manuscript. Overall, we provide four sections. First, we present in-depth details about the formalization of the building blocks used to create the user strategies in Section 1. Second, we provide details about the characteristics of the four data sets used in the performance analysis experiment in Section 2. In Section 3 and Section 4, we show detailed information about observations made in the two experiment parts for every data set, leading to eight figures and paragraphs of results.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation

1. Formalization of Building Blocks

We identified 11 low-level building blocks that are repeatedly used to create the higher-level user strategies. While the paper contains an introduction to the building blocks as well as descriptions and abstract formalizations, this document provides in-depth information about formalization details and default parameters. The formalizations cover techniques from data mining, machine learning, statistics, and information retrieval. Let V_c be the set of all candidate feature vectors.

Let V_t be the set of all training feature vectors. $V = V_c \cup V_t$. $d : V \times V \rightarrow \mathbb{R}$ is called distance function. In all of the building blocks $T \subseteq V_c$, $S \subseteq V$ and $x \in V_c$ are arbitrarily chosen.

Nearest Spatial Neighbors (NSN) Algorithm that retrieves instances in the vicinity of a focused instance. Can be used to assess characteristics of local structures in the data set. Can be implemented based on k (number of NN, default) or based on an orbit epsilon.

Let $v_1, ..., v_n$ be an ordering of all $v \in S$ with $x \notin S$, such that $d(x, v_1) \leq ... \leq d(x, v_n)$. $kNN(x, S, k) = \{v_1, ..., v_k\}$

Spatial Balancing (SPB) Component that tries to balance the distribution of instances across the entire data set emphasizing undiscovered areas.

$B(x) = \min_{t \in V_t} d(x,t)$

$SPB(V_c) = \arg\max_{x \in V_c} B(x)$

Clustering (CLU) Assigns similar instances to groups. Can be used to select instances in the center of dense areas, border areas, or areas of intersecting clusters.

$CLU(V_c) = \arg\min f_{cost}(\{C_1, ..., C_n\})$ such that

$\forall x \in V_c : \exists C_k : x \in C_k$ and

$\forall x \in V_c : (x \in C_k \wedge x \in C_l) \implies k = l$

Assume k-means is used as clustering algorithm. Then we have

$$f_{cost}(\{C_1, ..., C_n\}) = \sum_{i=1}^{n} \sum_{x \in C_i} ||x, m_i||_2$$

$$m_i = \frac{\sum_{x \in C_i} x}{|C_i|}$$

Density Estimation (DEN) Identifies dense areas in the data set. Can be used to select common rather than special or unique observations.

$DEN(x) = score_{DEN}(x)$

An exemplary implementation of $score_{DEN}(x)$ is

$score_{DEN}(x) = -\sum_{v \in kNN(x, V_c, k)} \frac{d(x,v)^2}{k}$

Outlier Detection (OUT) Identifies instances in sparsely populated regions. Can be used to select instances with special or even unique characteristics.

$OUT(T) = \{score_{OUT}(v, T) \geq t : v \in T\}$, $t \in \mathbb{R}$ is a threshold.

One possible definition of $score_{OUT}(v, T)$ is

$score_{OUT}(v, T) = -DEN(v)$

Compactness Estimation (CE) Identifies the compactness of groups of instances (clusters/classes). Can be used to prefer either compact or diverse distributions of instances in clusters/classes. $CE(T) = score_{CE}(T)$

Assume the Variance is used to rate the compactness of Clusters. Then we have

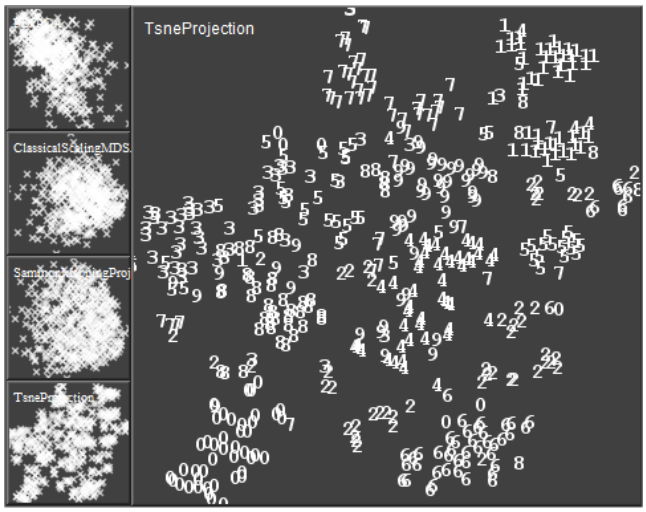

Figure 1: *MNIST data set represented with four different dimensionality reduction techniques. t-SNE is enlarged showing that the true labels of the instances build a series of well-separated cluster patterns while at the center many classes are intersecting. Looking at the PCA result (top left) reveals that MNIST hardly contains outliers.*

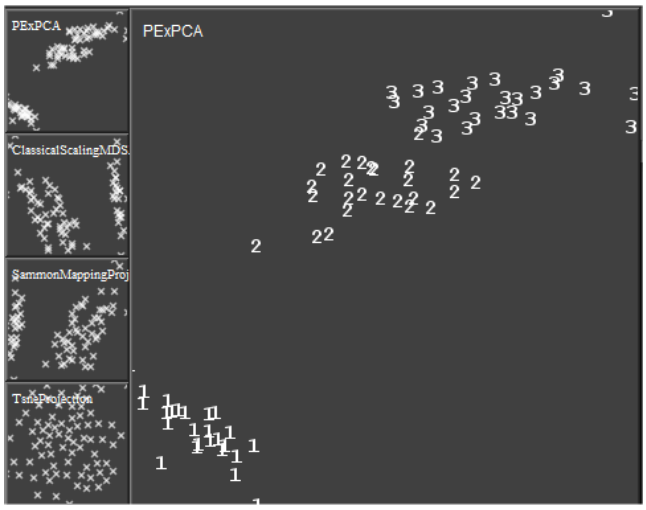

Figure 2: *IRIS data set represented with four different dimensionality reduction techniques. PCA depicts this well-known data set in a familiar way as seen in a series of images and related works. Class 1 is well separated from the intersecting classes 2 and 3.*

$score_{CE}(T) = \frac{1}{|T|}\sum_{x\in T}(d(x,m_i))^2$, where m_i is the average of the cluster.

Ideal Instance Identification (III) Pre-defined set of instances with perfectly exposed characteristics for a given label. Highly associated to semantical characteristics of user preferences.

$$U(x) = \begin{cases} 1, & \text{if } x \text{ is considered ideal by the user} \\ 0, & \text{otherwise} \end{cases}$$

$III(x) = \min_{u\in\{v\in V:U(v)=1\}} d(x,u)$

Class Likelihood (CL) represents the likelihoods l provided by a given (pre-trained) classifier f for an unlabeled instance x as:

$CL(x) = f(x) = l$, with $l \in \mathbb{R}^{|Y|}$ such that

$0 \le l_i \le 1$ for all $i = 1,...,|Y|$ and

$\sum_{i=1}^{|Y|} l_i = 1$

Class Prediction (CP) Prediction of a classifier applied on every candidate instance, based on the class likelihood. To achieve robust predictions we employ the results of a probabilistic ensemble classifier. Classifiers included in the ensemble are Naive Bayes [DHS*73], Random Forest [Bre01] (RF), Multilayer Perceptron [HSW89] (MP), Support Vector Machine [CV95] (SVM).

$CP(x) = y' = \arg\max(CL(x)), y' \in Y$

Local Class Diversity (LCD) Assessment of the diversity of a predicted class distribution in the vicinity of an instance.

$LCD(x) = div(p)$, $p \in \mathbb{R}^{|Y|}$ such that $p_i = p(x,y_i)$

$p(x,y_i) = \frac{|\{v\in kNN(x,V_c,k):CP(v)=y_i\}|}{k}$

Assume the entropy is used as a diversity measure. The we have

$div(p) = \sum_{i=1}^{|Y|} -p_i \log(p_i)$

Local Class Separation (LCS) Assessment of the presence of classes in the vicinity of an instance (assuming that classes are not necessarily disjoint). Inspired by cluster validity measures. Measure of separation for a set of nearest neighbors of every class. Can be used to identify regions with high class uncertainties.

$LCS(x) = f_{sep}(C_1,...,C_n)$ with

$C_i = kNN(x,\{v \in V : CP(v) = y_i\},k)$ for $i = 1,...,|Y|$

where $f_{sep}(C_1,...,C_n)$ is a separation scoring function. Such a scoring function could be based on Dunn-like indices [**?**], the Davies-Bouldin Index [**?**], or Silhouettes [**?**], for example.

2. Data Set Characteristics

The concrete results of the experiment depend on the data set of choice. Thus, we need to compare the strategies on the basis of different data sets. In our case, we selected data sets according to the following considerations: First, the data set should have numerical data/features and contain a class variable (binary or categorical). Second, the size of the data should include at least several thousand instances. This allows using hundreds of instances for candidates and thousands of instances for testing. Third, the data set should be publicly available. Fourth, the data sets should be intuitive, well-known, or even heavily applied in practice. Finally, we aimed at covering a broad range of data-centered characteristics, such as (1) binary classification versus multi-class classification, (2) equally-balanced label distribution versus unbalanced distribution, (3) proof of concept versus real-world complexity, as well as (4) being marked with few versus many outliers.

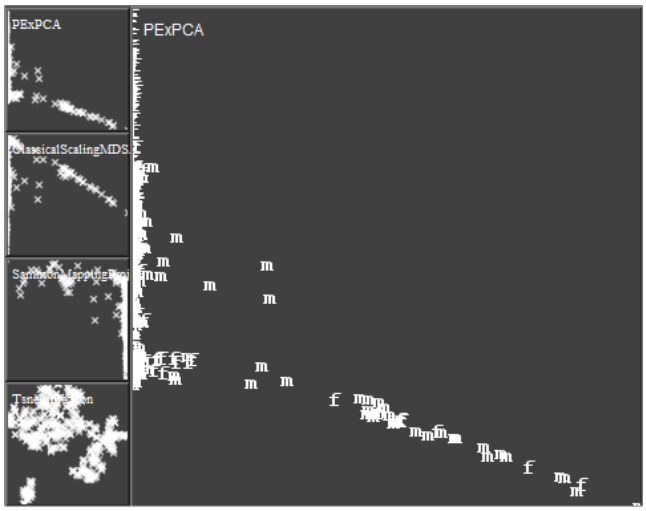

Figure 3: *GENDER voice data set represented with four different dimensionality reduction techniques. PCA is enlarged depicting a series of outliers of both classes (f, m) aligned along two common axis (west, on towards north one towards south-east). Outliers may hamper the results of classifiers, building blocks, and user strategies as well. In addition, it is difficult to identify patterns such as clusters or well separated class areas. In fact, t-SNE reveals cluster structures.*

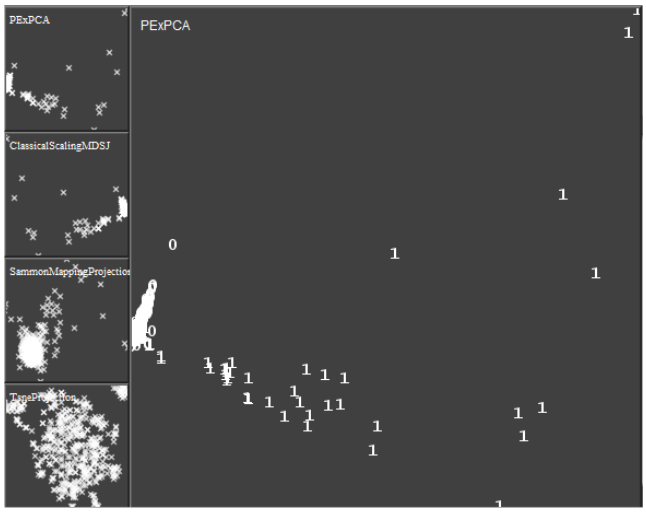

Figure 4: *FRAUD data set represented with four different dimensionality reduction techniques. PCA is enlarged depicting a series of outliers associated to class 1 (fraud) sparse distributed in the 2D embedding. Outliers may hamper the results of classifiers, building blocks, and user strategies as well. In turn, class 0 (no fraud) seems to build a compact cluster.*

2.1. MNIST Data Set

The *MNIST* data set [LBBH98] perfectly meets all requirements to the data. It consists of thousands of images showing handwritten digits. Each raw digit is represented by a $28x28$ image yielding a 784 dimensional vector in the original space. For faster classification, we use a descriptor that extracts slices in horizontal, vertical, and diagonal direction. Overall, the feature vector contains 42 numerical dimensions.

Visual Analysis of the structure of the MNIST data set is provided in Figure 1. We assess a series of visual patterns as well as a low degree of severe outliers.

2.2. IRIS Data Set

The *IRIS* data set [Lic13] does not fulfill the criterion of thousands of instances, rather it consists of three classes with 50 instances each. However, we consider *IRIS* valuable as an intuitive, frequently used data set to proof the concept. The four dimensions are used as numerical features.

Visual Analysis of the structure of the IRIS data set is provided in Figure 2. PCA reveals the typical appearance of IRIS as presented in many works before. The data set is very small, contains two intersected classes, and virtually no outliers.

2.3. GENDER VOICE Data Set

The *Gender Recognition by Voice* data set [Bec16] contains acoustic properties of the voice and speech to identify the gender of speaker. This data set consists of 3,168 instances which are preprocessed by acoustic analysis with an analyzed frequency range of 0-280 Hz. The outcome is a 21 dimensional numerical vector of acoustic properties for each instance. The data set contains a considerable amount of outliers.

Visual Analysis of the structure of the GENDER voice data set is provided in Figure 3. PCA reveals that the data set contains a series of outliers that may have an influence on the performance of user strategies.

2.4. FRAUD Detection Data Set

The Credit Card *FRAUD* data set [PCJB15] contains transactions of credit cards recorded in Europe in two days in September 2013. Overall, 492 frauds are included in the 284,807 transactions (0.172%). As such, the data set can be used to assess an unbalanced binary classification problem. The data consists of 28 numerical features as a result of a PCA [Jol02] processing step. The data set contains a considerable amount of outliers.

We shed light on the unbalanced label distribution as an important characteristic for the performance of data-centered strategies. Figure 4 provides an overview of the data set. Instances that contain label 0 (no fraud) occur ten times as often than labels with 1 (fraud). Figures 5 and 6 provide evidence that the performance of clustering for the labeling process is harmed by an unbalanced label distribution.

Visual Analysis of the structure of the FRAUD data set is provided in Figure 3. PCA reveals that the data set contains a series of outliers of class 1 (fraud) that may have an influence on the per-

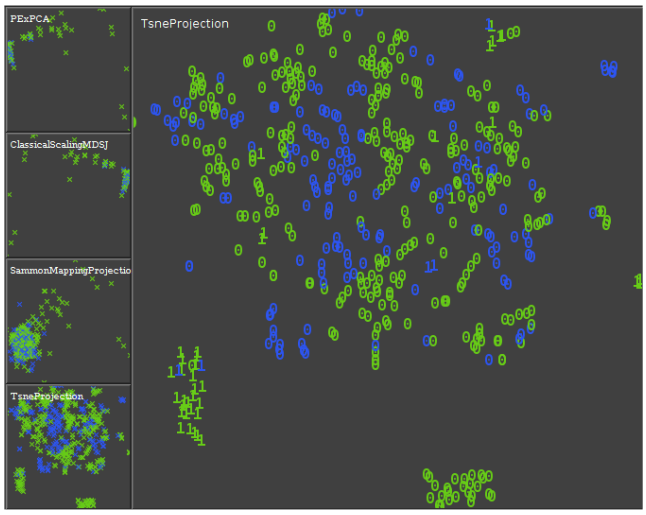

Figure 5: *FRAUD data set represented with t-SNE. The result of building block Clustering is shown, the two clusters are depicted with blue and green colors. By looking at the true class labels (0 and 1), it becomes apparent why this result does not contribute to the labeling process: the clustering result does not reflect the distribution of classes. Unbalanced data sets seem to hamper the performance of Centroids First and other density-centered strategies.*

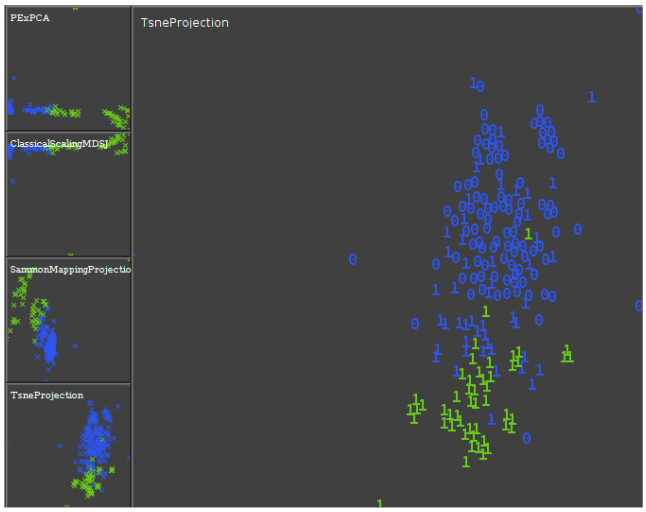

Figure 6: *Repetition of the analysis presented in Figure 5. In contrast to the latter analysis, we use a balanced set of instances (for hypothesis testing; not part of the experiments in the main paper). Now, the two clusters are well-separated and reflect the true class distribution. We infer that unbalanced label distributions harm the performance of data-centered building blocks such as Clustering and thus, of the Centroids First strategy.*

formance of user strategies. Most instances of class 0 are allocated west in a compact way.

3. Results for RQ_1: Analysis of the Bootstrap Problem

The very first phase of the labeling phase is often associated with bootstrap problems of model-based techniques. In the following, we list a series of observations for every data set with an emphasis on the bootstrap problem and strategies that may be able to trackle the problem. Generalizable insights are also presented in the main paper.

3.1. MNIST Data Set

1. As expected, ULoP performs best. Interestingly, its performance is decreasing after 10 iterations (when the bootstrap phase is finished). On average, it takes 7 more iterations to surpass the accuracy level reached after 10 iterations. This might be caused by the fact that some classes are labeled twice after 11 to 17 iterations, while others are only labeled only once. As such, this over-representation of some classes seems to decrease the performance. This observation is generalizable for our ULoP implementation: we executed ULoP trials as well.
2. Model-based strategies perform worse than Random.
3. Ideal Labels First outperforms all other strategies up to the 20th iteration. Considering the accuracy integral over all iterations, this strategy provides the best performance. Ideal Labels First is also the strategy which needs the fewest iterations for visiting labeling all classes (see the boxplot visualization).
4. Two ALs overtake the user strategies after 40 iteration. Thus, we identify sort of a sweet spot between data-centered strategies and ALs.

In general, data-based strategies perform best, followed by ALs and model-based strategies.

3.2. IRIS Data Set

1. The performance chart of most strategies follows a similar pattern: A steep increase in the beginning, later stagnation on a high level, respectively.
2. Ideal Labels First is dominating in the beginning, up to the 8th iteration. After that, Outliers and Cluster Borders provide the best performance.
3. The performance of model-based strategies varies a lot. While the Class Borders Refinement strategy has a very poor performance, the Class Outlier Strategy works quite well.

In summary, some data-based strategies have a particularly good start, are then overtaken by outlier- and model-based strategies, which are finally overtaken by most ALs. The small size of the data set may add to the observation that outlier strategies performed well (which was exceptional with this data set).

3.3. GENDER VOICE Data Set

1. ALs performance lies in a narrow corridor. Relative to user strategies and compared to their performance on other data sets, ALs perform well on this data set.
2. Centroids First provides the best performance up to the break-even point at the 15th iteration, afterwards it gets surpassed by most of the ALs.

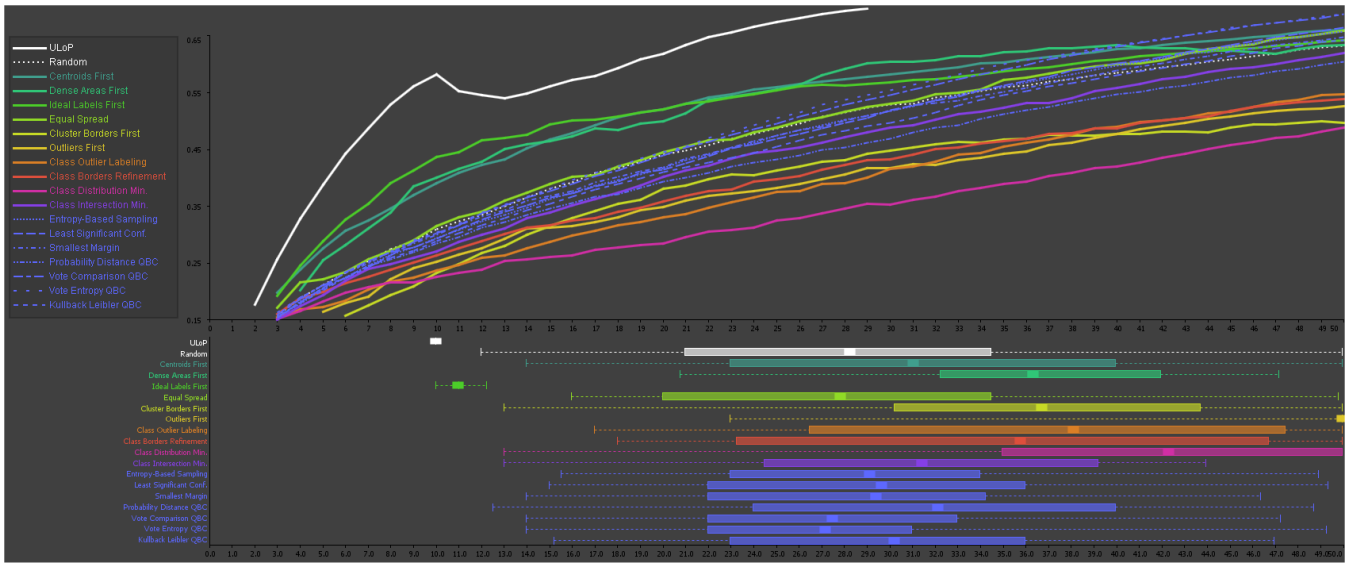

Figure 7: *Average performance of strategies in the very first 50 iterations of the labeling process (MNIST data set). Most data-based user strategies perform particularly well in the very first iterations (Ideal Labels First, Centroids First, Dense Areas First). Most model-based user strategies perform below average. Focusing on the distinction between density-based and outlier-based strategies, the density-based perform particularly well while outlier-based strategies perform poor. The performance of ALs is unremarkable at start, but increases for more iterations. Boxplots at the bottom show the iterations when strategies produced at least one label for each class. The Ideal Labels First strategy visits all class labels remarkably early (all 10 labels in 11 iterations on average). Model-based strategies require more iterations to see each label at least one time. As a general rule, strategies with good accuracies also visited every class label earlier. We infer that data-centered strategies are a means to tackle the bootstrap problem.*

3. Up to the 10th iteration, only two strategies beat Random.
4. The performance of outlier strategies on this data set is especially poor.
5. In the beginning, there is a huge performance gap between ULoP and the remaining strategies. This gap becomes smaller with with more iterations.

To sum up, Centroids First shows a good performance in the beginning. Later, ALs are dominating.

3.4. FRAUD DETECTION Data Set

1. Most of the user strategies experience huge accuracy changes between two consecutive iterations. ALs do not show these variations. We assume that the strategies may escape jump between recurring local patterns.
2. Cluster Borders First performs well over the first 5 iterations. However, the performance does not improve anymore after this initial phase. ALs deliver the best performance later on, almost reaching the results of ULoP.
3. Density-centered strategies (Dense Areas First, Centroids First) perform very poor on this data set. One explanation is the unbalance of this data set (ratio 10:1). We expect many dense areas within the dominating class. To prove this, we performed a visual analysis of the data set shown in Section 2.4.

Overall, all ALs provide good results on this data set. Cluster Borders First performs very well early on, but stagnates afterwards. Class Intersection is again the best model-based user strategy.

4. Results for RQ_2: Performance Comparison after Bootstrap

In the following, we list a series of observations for every data set. Generalizable insights are also presented in the main paper.

4.1. MNIST Data Set (Figure 11)

1. ULoP outperforms every strategy by far. After about 20 iterations, the gradient declines a little bit, but the accuracy increases steadily.
2. From the set of strategies, the data-based user strategies yield the best results in the very first 30 iterations, particularly Centroids First, Dense Areas First, and Ideal Labels First. All these strategies belong to the strategies preferring clusters and dense areas. So we may conclude that looking for such patterns in the early iterations leads to a good basis – even after having resolved the bootstrap problem.
3. As opposed to this, AL strategies start quite slow and do not compete with the data-based approaches in the early stages. But they increase very constantly and outperform those strategies since about iteration 35 (interesting sweet spot).
4. Model-based user strategies perform below Random. Only Class Intersection can compete, the remaining model-based strategies come in last.

To sum up, no strategy can keep up with ULoP, but strategies with a bias to centroids in clusters and dense areas in general are the most promising attempts on this data set. MNIST is a multi-class data set, so we may conclude that these approaches are the

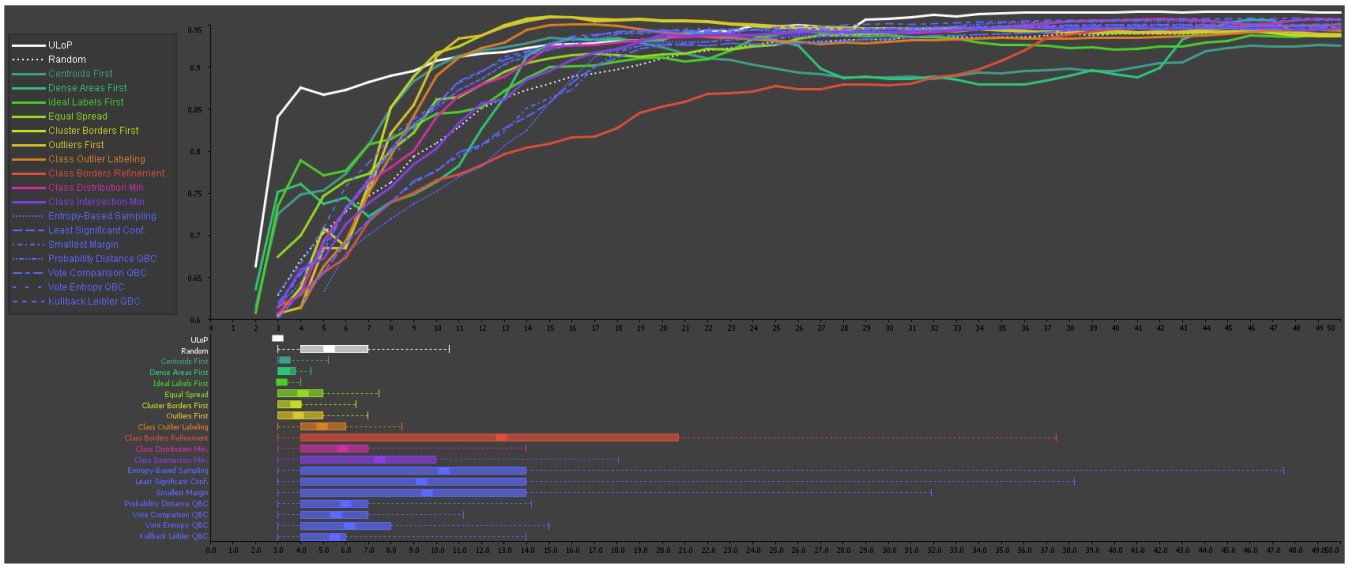

Figure 8: *Average performance of strategies in the very first 50 iterations of the labeling process (IRIS data set). Most data-based user strategies perform well in very early iterations, only Cluster Borders First and Outliers First seem to be less appropriate to tackle the bootstrap problem. Most model-based user strategies perform at the pace of Random and ALs at start. In performance of ALs is unremarkable at start, but increases for more iterations. An interesting finding is the decrease in performance of Centroid First and Dense Areas First, beginning with the 25th iteration; we assume that the very small data set only requires/contains few dense instances per class. Boxplots at the bottom show the iterations when strategies produced at least one label for each class. Data-centered user strategies consistently perform better than Random and ALs. The model-based Class Border strategy performs considerably weak. As a rule, strategies with good accuracies seem to label all classes earlier. We infer that data-centered user strategies are a means to tackle the bootstrap problem.*

best way to handle such data sets. AL strategies need some time to master the bootstrap problem. The typically quite specific queries of AL seem to be worth not before some iterations are passed. Prior, more general approaches win the race.

4.2. IRIS Data Set (Figure 12)

1. ULoP again outperforms the remaining strategies, reaching an accuracy of about 90% already after six iterations.
2. Some data-based strategies perform weak, at start especially Dense Areas First and Equal Spread. In later iterations, Centroids First and Dense Areas First have a considerable low, losing accuracy, far below the remaining strategies.
3. AL strategies start well and reach a high level already after about ten iterations. Model-based user strategies perform on a comparable level.

In summary, all strategies perform quite well and constantly. Data-based strategies, especially those with a bias to centroids and dense areas, can not keep up with the remaining strategies. IRIS is a very small data set, so we may conclude that those more general approaches do not lead to adequate results since anomalies and outliers have a higher impact on the performance in such data sets, as well as every individual false classified instance.

4.3. GENDER VOICE Data Set (Figure 13)

1. ULoP's performance curve is very smooth, continually increasing. The performance curves of AL strategies behave similar to ULoP, though on a considerably lower level.
2. Class Intersection is the only model-based user strategy that can compete with AL strategies. Especially Class Distribution Minimization and Class Outliers Labeling perform very weak and unsteadily.
3. Dense Areas First, Equal Spread, and Equal Spread start strong at first, but are not able to keep pace with AL strategies between 10th and 20th iteration. Again, we identify a sweet spot when AL starts to outperform data-based strategies. Centroids First performs below Random, almost not at all increasing between 10th and 20th iteration.

To conclude, GENDER VOICE is a two-class data set with a considerably increased percentage amount of outliers in relation to the other data sets. This fact seems to cause difficulties in data-based and cluster-based strategies like Centroids First. These strategies benefit from more separable data sets, whereas they struggle with classes consisting of several clusters.

4.4. FRAUD DETECTION Data Set (Figure 14)

1. ULoP again outperforms all remaining strategies, followed by AL strategies, Class Borders, and Class Intersection.
2. Ideal Label strategy starts similarly strong, but can not keep its level after 10 iterations.

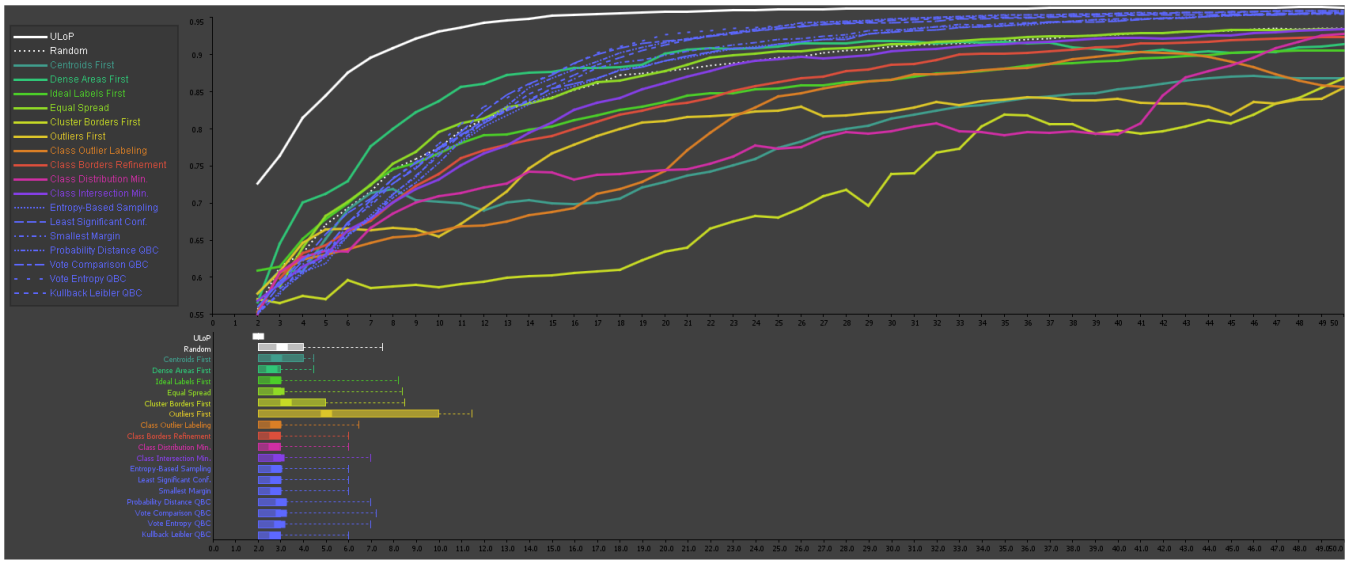

Figure 9: *Average performance of strategies in the very first 50 iterations of the labeling process (GENDER VOICE data set). The Dense Areas First strategy considerably outperforms the remaining strategies in early iterations, followed by Ideal Labels First and Equal Spread. Most model-based user strategies perform unremarkable in early iterations. The Centroids First strategy shows the weakest score; In-depth analysis revealed that the data set consists a reasonable number of outliers which might have caused problems to the Custering building block. The boxplots at the bottom (showing in each iteration every class label was at least seen on time) shows that Cluster Borders and the Outliers Strategy perform worst. Outlier-based strategies seem to be no means to address the bootstrap problem.*

3. Class Distribution makes a jump in the performance after the 23th iteration; one explanation may be the escape from some local pattern. In any case, this indicates that some strategies establish a general order of instance selection, regardless the randomized starting conditions.
4. Equal Spread, Dense Areas First, and Centroids First perform distinctly below Random, suffering some heavy losses during the labeling process (particularly Dense Areas First).
5. The Random baseline on this data set works very weak in relation to previous data sets.

Fraud Detection incorporates an interesting structure, consisting of one highly overweight class and a small class, containing some greatly outlying instances. This special consistency may be the reason for the hampering Random strategy since its working mechanisms are not that complex and elaborated as those of model-based approaches. The set of centroid-oriented strategies seem to have to cope with this problem as well. In Section2.4 we conduct a visual analysis of the unbalanced FRAUD data set. As a result, most dense regions and clusters are located in the over-represented class 0 (no fraud). We conclude that unbalanced class distributions in the data hamper the performance of many data-centered strategies, and may require individual treatment.

References

[Bec16] BECKER K.: Gender recognition by voice – identify a voice as male or female, 2016. Accessed: 2017-12-05. 3

[Bre01] BREIMAN L.: Random forests. *Machine learning 45*, 1 (2001), 5–32. 2

[CV95] CORTES C., VAPNIK V.: Support-vector networks. *Machine learning 20*, 3 (1995), 273–297. 2

[DHS*73] DUDA R. O., HART P. E., STORK D. G., ET AL.: *Pattern classification*, vol. 2. Wiley New York, 1973. 2

[HSW89] HORNIK K., STINCHCOMBE M., WHITE H.: Multilayer feedforward networks are universal approximators. *Neural networks 2*, 5 (1989), 359–366. doi:10.1016/0893-6080(89)90020-8. 2

[Jol02] JOLLIFFE I. T.: *Principal Component Analysis*, 3rd ed. Springer, 2002. doi:10.1007/978-1-4757-1904-8_7. 3

[LBBH98] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324. doi:10.1109/5.726791. 3

[Lic13] LICHMAN M.: UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml. 3

[PCJB15] POZZOLO A. D., CAELEN O., JOHNSON R. A., BONTEMPI G.: Calibrating probability with undersampling for unbalanced classification. In *Symposium Series on Computational Intelligence and Data Mining (CIDM)* (2015), IEEE, pp. 159–166. doi:10.1109/SSCI.2015.33. 3

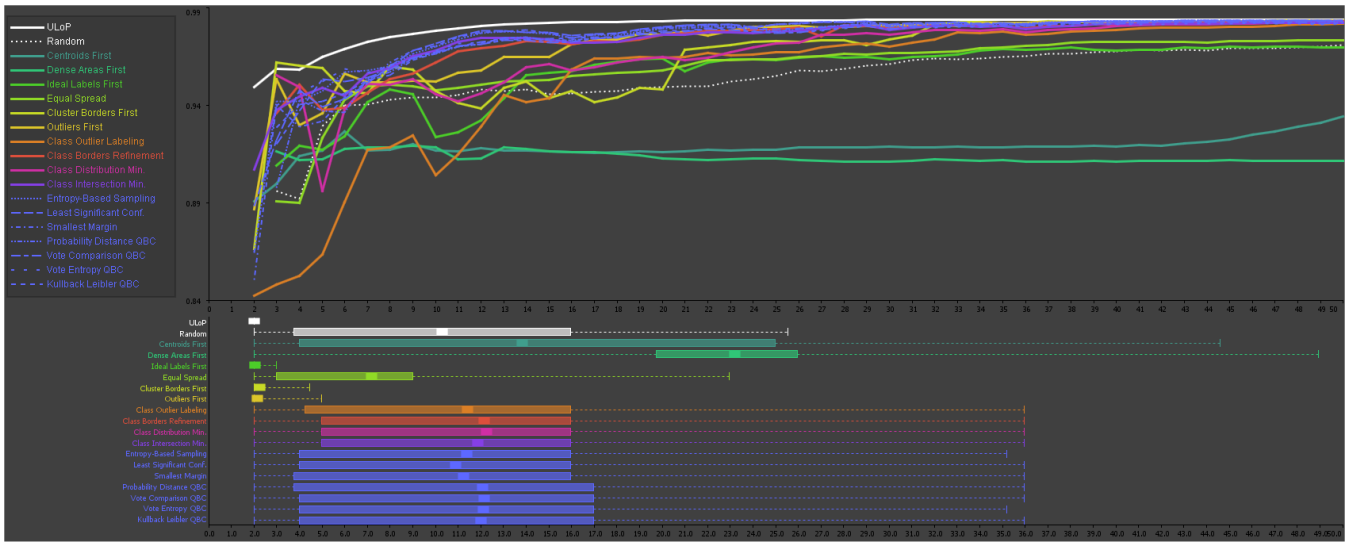

Figure 10: *Average performance of strategies in the very first 50 iterations of the labeling process (FRAUD data set). Most strategies achieve a high performance very fast, the difference to the ULoP is comparatively small. However, Dense Areas First and Centroids First perform particularly weak. Visual analysis of the data set revealed the high number of outliers may have had a considerable influence on the strategy performances. Interestingly, the outlier strategies (Outliers First, Cluster Borders First) performed well for this outlier-prone data set. Boxplots at the bottom show the iterations when strategies produced at least one label for each class. Ideal Labels strategy again performs particularly well, together with the outlier-based strategies (Cluster Borders, Outlier Strategy). The performances of model-based user strategies are comparable to Random and AL.*

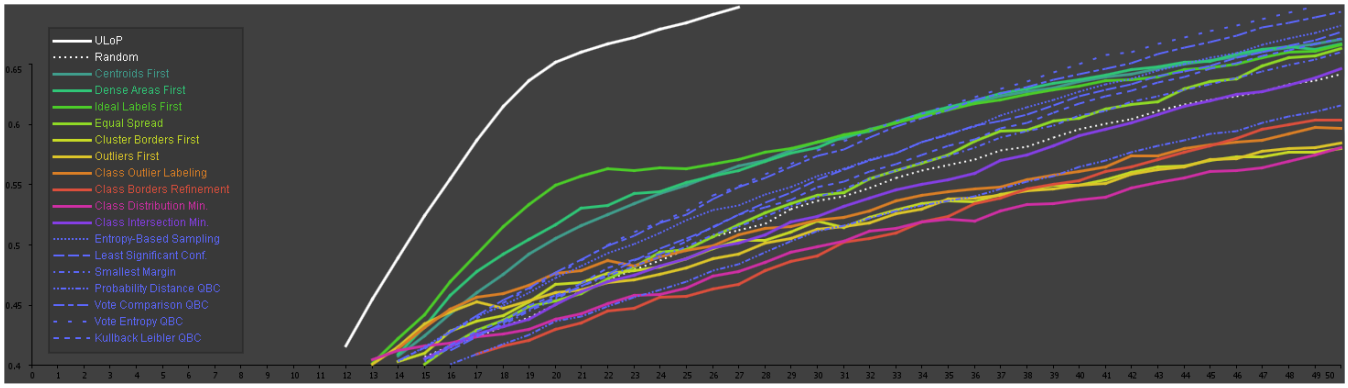

Figure 11: *Average performance of strategies after the initialization with one labeled instance per class (MNIST data set). The ULoP still outperforms remaining strategies significantly. Three data-based user strategies (Centroids First, Dense Areas First, Ideal Labels First) perform considerably better than the remaining strategies. AL strategies start at a moderate level, but achieve particularly well performances in later phases. Using the afromentioned data-centered user strategies and the ALs, we assess a break-even point in the performances at the 35th iteration. Class Intersection can compete with Random, remaining model-centered strategies perform below Random. In general, data-centered strategies with a focus on dense patterns/clusters in the data set perform particularly well while model-based user strategies perform below Random.*

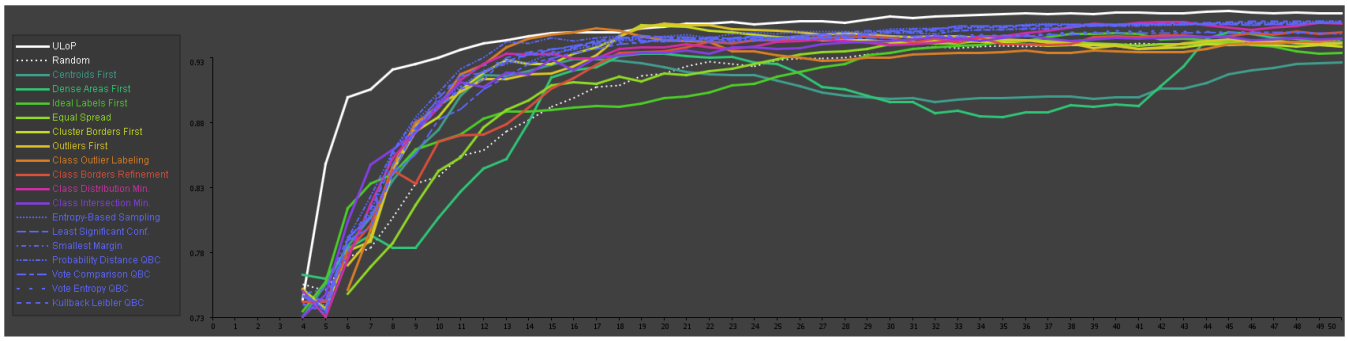

Figure 12: *Average performance of strategies after the initialization with one labeled instance per class (IRIS data set). ULoP (Greedy, green) performs best, followed by a series of strategies including most model-based user strategies. The data-based user strategies Dense Areas First and Equal Spread perform comparatively weak at start. An interesting observation is the performance of Centroids First and Dense Areas First, which decreases after 25 iterations. We assume that phenomenon may be due to the very small data size. The remaining strategies achieve performances over 90%.*

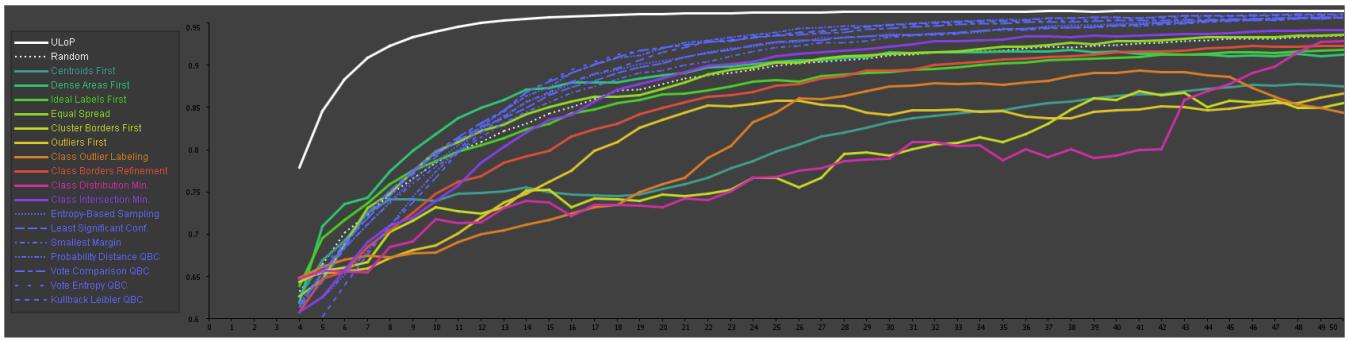

Figure 13: *Average performance of strategies after the initialization with one labeled instance per class (GENDER VOICE data set). ULoP shows the best performance by far. Dense Areas First, Ideal Labels First, and Equal Spread start strong, but outperformed by ALs between the 10th and 20st iteration. Class Intersection Min. is the best model-based strategy that almost competes with ALs. The remaining model-based strategies perform below Random. Centroids First seems to suffer from the existence of outliers showing particularly weak performance.*

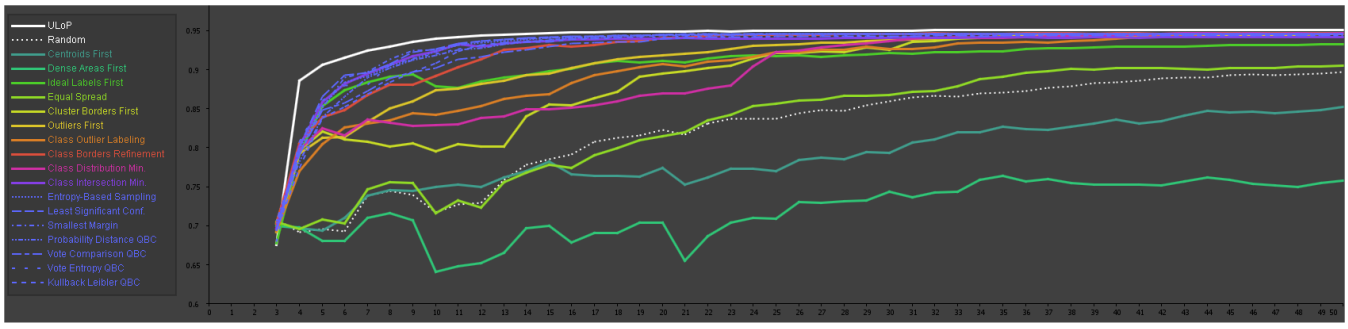

Figure 14: *Average performance of strategies after the initialization with one labeled instance per class (FRAUD detection data set). Class Intersection and most AL strategies have particularly good performances, followed by Ideal Labels First and Class Borders Refinement. Dense Areas First, Equal Spread, and Centroids First perform particularly weak, possibly hampered by the series of existing outliers. An interesting insight is the weak performance of the Random baseline. This may be an indication that the concepts implemented in model-based strategies are highly valuable compared to Random.*