Learning from the Best – Visual Analysis of a Quasi-Optimal Data Labeling Strategy

Jürgen Bernard[1,2], Marco Hutter[1], Markus Lehmann[1], Martin Müller[1], Matthias Zeppelzauer[3], and Michael Sedlmair[4]

[1]TU Darmstadt, Germany
[2]Fraunhofer IGD, Germany
[3]St. Pölten University of Applied Sciences, St. Pölten, Austria
[4]Jacobs University Bremen, Germany

Abstract

The supplemental materials document contains information about the observational study in more detail. In particular, we report the results of the observation of five individual data sets, each observed by five analysts.

CCS Concepts

•Human-centered computing → ***Information visualization;*** *HCI design and evaluation methods;* ***•Theory of computation*** → *Active learning;* ***•Computing methodologies*** → *Machine learning;*

1. MNIST

Dimensionality reduction of the MNIST data set reveals clusters for different digits, partially compact and separated (see Figure 6). This characteristic is revealed best with t-SNE. However, at the center of the manifold, we identify several spatial class conflicts, predominantly between the classes 2, 4, 5, 6, 8, and 9.

In the first ten labeling iterations, all ten classes are covered by ULoP by labeling exactly one instance of each class (5/5 analysts). We infer that ULoP can be used to tackle bootstrap problems known from AL in a highly effective way. The very first classes were mostly the same in each trial (classes 0, 1 and 7). The commonality of these classes is (i) a clear separation from other classes, (ii) a compact alignment in the embedding space (t-SNE), as well as (iii) the location in margin areas of the projection manifold (see Figure 6). These classes seem to be easier to model and thus classifiers benefit most from them. A formalization of ULoP's strategy in the first ten iterations may be as follows: *choose one instance from each class, starting with compact and well-separated classes*.

After 20 iterations, the label distribution was still balanced almost equally over all classes. In the further course, the distribution became more unbalanced; easily predictable classes (like class 0 and 7) were selected more frequently than classes that are difficult to predict. *ULoP tries to balance the distribution at this stage, but with a slight bias to reduce to safe instances*. This is to some degree surprising, as one may draw the hypothesis that difficult classes may require more labels.

In later stages, when the label distribution has become unbalanced, the strategy re-balances the distribution. We assume that there is a moment when those easily predictable classes are already predicted well, so it does not make sense to further explore them. Then, ULoP deliberately tends to label wrongly classified instances to improve the accuracy. *Structures consisting of uncertain instances are preferred at this stage of the labeling process.*

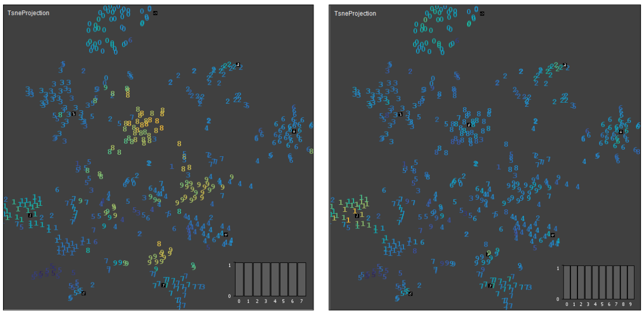

Figure 1: *Visual interface used for the observational study. The screenshots show two iterations of the labeling process for the MNIST dataset. Dimensionality reduction (here: t-SNE) allows the representation of instances in 2D. Color is used to encode how much the learning model will benefit from labeling a candidate (orange is best, here). Left: The ULoP strategy suggests labeling instances from classes 8 or 9, after all remaining digits have already been labeled once. Right: Every label has been seen exactly once. The ULoP then suggests to choose digits of class 1, located at the left margin of the manifold.*

Summary. We draw the conclusion that *the output of the ULoP strategy is related to several different strategies in the course of the process* - and the distinction of these (and their combinations) is not trivial. Some clusters are labeled late in the labeling process, others are labeled quite early, as well as outliers. A possible formulation of ULoP's overall strategy is as follows: *The more compact and separated a cluster, the earlier it is labeled in the process and the*

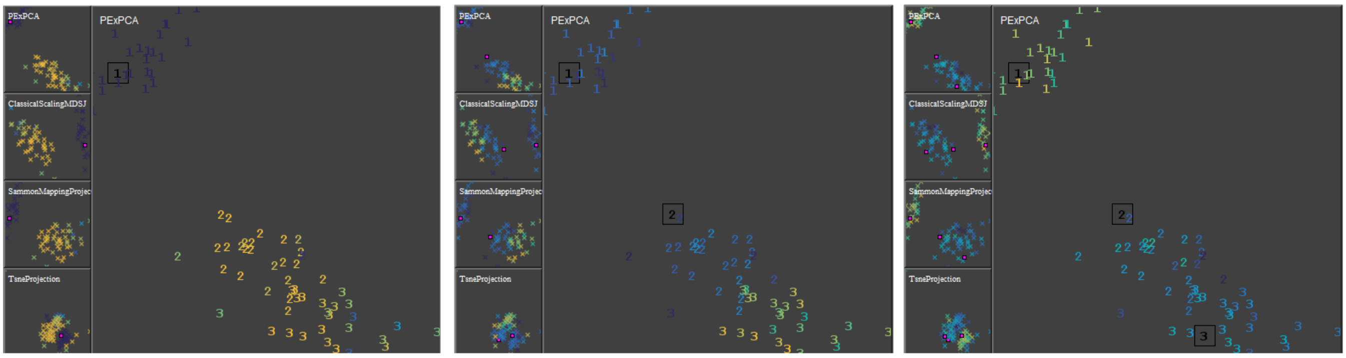

Figure 2: *Visual observation of the IRIS dataset observed with PCA (iterations two to four). Class 1 (setosa) is clearly separated and was labeled first. Second, class 2 is labeled (versicolor) and finally, class 3 (virginica) is labeled with an instance near the class center of gravity. The fourth label is again of class 1 (not shown). Thus, ULoP accessed one instance of each class first, as it did for every other data set.*

higher is the gain of accuracy when choosing one of the cluster's instances. We identified that the instance selected within a cluster was not necessarily the centroid (observed with t-SNE). We assume that neighborship relations to other classes/clusters have a strong influence on the selection of both representative instances and instances that support the separation of classes. In fact, we observed that the output of ULoP often aimed at defining central supporting points of classes/clusters.

2. IRIS

The visual representation of the IRIS data set (see Figure 7) forms a clearly separated cluster (class 1) and a super-cluster consisting of the classes 2 and 3 (partially conflicting).

In the first three iterations of ULoP, each class is covered exactly once (5/5 analysts). Again, we infer that the bootstrap problem can be solved almost perfectly by using ULoP. ULoP's strategy in this first stage was to balance the label distribution, slightly preferring centroid-close instances.

In the subsequent phase, we observed an unexpected behavior. Instances from class 1, (clearly separated), have been selected considerably more frequently than instances from the intersecting classes 2 and 3. In addition, choosing class 2 or 3 led to an intermediate decrease of accuracy (3/5 analysts). Such a selection was often followed by an instance of the opposed class (3 or 2) which again produced an accuracy gain (5/5 analysts). One explanation for these phenomena may be that the implementation of the ULoP algorithm only looks one iteration into the future. When there is no instance that yields a gain in accuracy, the instance which is highest ranked by ULoP leads to a small decrease in accuracy. This problem may be resolved if the ULoP implementation would look more than one iteration into the future, coupled with a massive increase of computational costs. An interesting side aspect regards human cognition: *humans would easily identify the necessity to assign labels to both classes (class 2 and 3)* to achieve a gain in accuracy.

One side-note was the observation that t-SNE performed worst in the separation of class 1 and the two remaining classes. While all analysts made good experiences for the remaining data sets, t-SNE was less useful for the very small IRIS data set. The observation of IRIS revealed one point that may be remarkable: *there are situations in the labeling process where the accuracy will decrease* regardless which instance will be added next. It will be interesting to assess the effect of such local "valleys" in the accuracy progression.

3. GENDER Voice

Dimensionality reduction applied on the GENDER voice dataset reveals several cluster structures for both classes (m and f). With t-SNE, we observe that instances of the two classes are well-separated, see Figure 8.

In the first two iterations of ULoP, both classes are covered (5/5). As such, ULoP can be used to efficiently resolve the bootstrap problem. In early labeling iterations, we identified a strong prioritization of centroid-near instances. *Labeling centroid-like instances in the beginning may contribute to a compact and representative set of initial labels.*

After this first phase, the output of the ULoP strategy changed considerably. We made the observation that many selected instances were close to already labeled instances, rather than located in more uncertain regions. One interpretation of this phenomenon may be that *the repeated labeling of relevant areas in the data set helps to consolidate the classifier.* This strategy was complemented with the selection of instances that tend to separate classes. One analyst used the metaphor of a Dirichlet tessellation [LD50] / Voronoi diagram [Vor08] to describe the distribution of points and resulting class areas in 2D.

After the consolidation of the class centroids, the following selected instances are situated at the border of their visual clusters and also at the border of their respective class. The regions the classifier was most uncertain about were uncovered during the labeling process, similar to an important active learning principle (uncertainty sampling). The observational trials ended after about ten labels per class, when the classifier already achieved a very high accuracy with only marginal space for improvement. In general, ULoP's strategy on GENDER voice can be referred to a balancing approach, *first selecting instances near the cluster/class centroids, then exploring the border areas*.

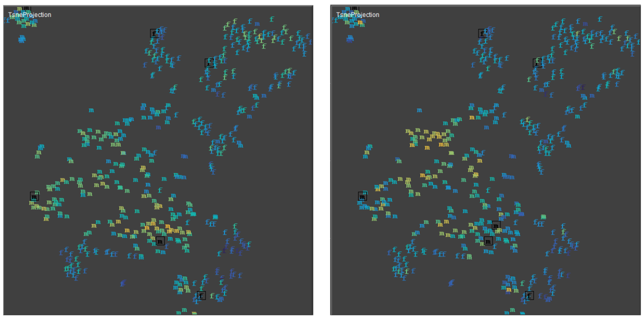

Figure 3: *Visual observation of the GENDER voice dataset after six iterations (three male, three female labels, see the six black instances with rectangles). A pattern can be seen in iteration seven and eight: two male instances are chosen to consolidate the large male cluster at the center. With that, the accuracy increases to 0.88.*

4. FRAUD Detection

FRAUD is a data set with an unbalanced label distribution, including ten times as many regular cases (0) than frauds (1). Dimensionality reduction (t-SNE) reveals a super-cluster, as well as some structures in margin areas, one reflecting frauds (see Figure 9).

In the beginning, both label classes are selected equally. All observed trials started with a 0 class label (5/5). Then, an instance of the 1 class was recommended, i.e., after two iterations both classes were equally covered. Again, we ascertain the particularly good performance to address the bootstrap problem.

Despite balancing, the instances were not chosen in an actual alternating order (2/5). We note that *the amount of labeled instances in early iterations was balanced, even for an unbalanced dataset*. In the further course, the proportional amount of instances of class 1 decreased. However, we did not observe (0/5) that it decreased towards its natural balance (1 : 10).

It should be noted that the labeled instances of class 0 are not equally spread - they are mostly located in the vicinity of class 1 instances. Cluster regions that only contained class 0 instances required only few labels. Opposed to this, clusters of class 1 instances were labeled often, and quite early in the labeling process. Also, those clusters were predicted correctly after just a few iterations and are mostly the only instances in the data set which are predicted as class 1. There are some instances of class 1 in the data set which are assessed with a low performance gain in the entire labeling process. Probably, classifiers lack of complexity to reliably distinguish those special instances from class 1 and would as a consequence mis-classify many instances from class 0 after having modeled these outliers explicitly.

The general strategy which was observed during ULoP's labeling processing can be summarized as follows: *first, label both classes exactly once. Then, label both classes with a bias towards class consolidation. Finally, focus on instances that strengthen local class separation.*

5. ISOLET Spoken Letters

The visualization of the ISOLET data set shows several separated clusters in peripheral areas (see Figure 10). However, moving closer to the center, we identify considerably less cluster/class separation. Part of the super-cluster at the center are the classes B, C, D, E, G, P, T, and V, which also have phonetical similarity.

During the initial phase of the labeling process, *all 26 classes are covered exactly once*. The bootstrap problem is solved effectively. Some classes (A, I, Y, X, and C) are always chosen very early (5/5), while some other classes (B, V, P) tend to be chosen rather at the end of the initial phase. *There is a slight trend towards selecting peripheral classes earlier*. For the ISOLET data set, however, further experiments are required for validation. With the large number of classes, it becomes more difficult to assess whether or not centroid-near instances are preferred over instances at the borderline of classes in the selection process. Switching between different dimensionality reduction helps with the investigation, but does not reveal a clear trend. What can be observed very clearly is the even spatial distribution of selected labels, spanning the entire 2D space (see Figure 10). *Sampling the space uniformly seems to be a key element in the early ULoP's strategy.*

In the phase of consolidation, three analysts again observed the selection of exactly one instance per class. The two remaining analysts identified some instances which were selected a third time before every instance was addressed twice. In summary, *ULoP managed to balance the label distribution also in the consolidation phase*. Similar to the initial labeling phase, instances with a peripheral location in the projection plane tend to be selected earlier than classes in the center of the embedding space.

During the last observed phase beginning with the 53rd iteration (3/5), the accuracy started to stagnate. All analysts reported that the third phase conducted with the ISOLET data set with its 26 classes was the most challenging observation task because keeping the overview was difficult. Similarly, the observation of the labeling process consumed considerably more time with ISOLET. For

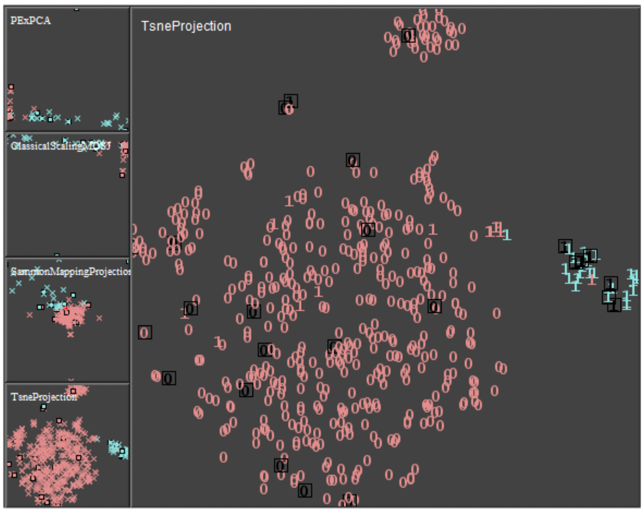

Figure 4: *Visual observation of the FRAUD dataset after 25 iterations using a class-based color coding. Frauds (1-labels) build a separate cluster with all dimensionality reduction techniques. In t-SNE it can be seen that few fraud instances in the large 0 cluster are still classified as no fraud requiring individual treatment (fine tuning). After iteration 10 the average F1 score remained at about 0.96. F1 was chosen to account for the unbalanced class priors.*

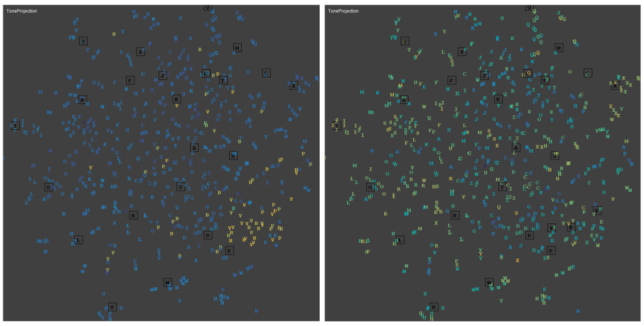

Figure 5: *Visual observation of the ISOLET dataset in iteration 24 and 27. In iteration 24 only the labels B, P, and V have not been seen yet. Interestingly, those letters are phonetically similar and mostly located in a distinct region of the manifold. In iteration 27 every label has been seen exactly once. The preferences for the next label have a slight tendency towards compact clusters in border regions.*

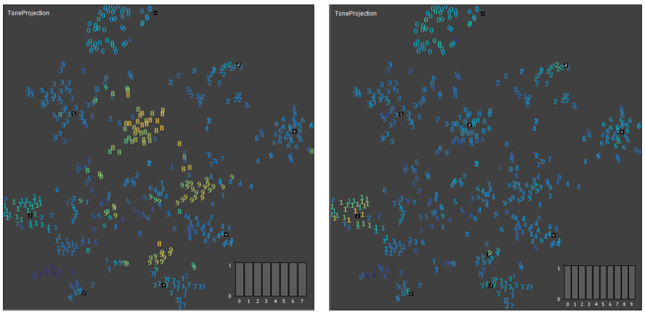

Figure 6: *Visual interface used for the observational study. The screenshots show two iterations of the labeling process for the MNIST dataset. Dimensionality reduction (here: t-SNE) allows the representation of instances in 2D. Color is used to encode how much the learning model will benefit from labeling a candidate (orange is best, here). Left: The Results of ULoP strategy suggest that an instance from class 8 or 9 should be labeled, after all remaining digits have already been labeled once. Right: Every label has been seen exactly once. Afterwards, digits of class 1 located at the left margin of the manifold are highly scored by ULoP.*

multi-class labeling, it may be beneficial to research for other visualization types.

6. MNIST

Dimensionality reduction of the MNIST data set reveals clusters for different digits, partially compact and separated (see Figure 6). This characteristic is revealed best with t-SNE. However, at the center of the manifold, we identify several spatial class conflicts, predominantly between the classes 2, 4, 5, 6, 8, and 9.

In the first ten labeling iterations, ULoP covers all ten classes by labeling exactly one instance of each class (5/5 analysts). We infer that ULoP tackles bootstrap problems known from AL in a highly effective way. The very first classes were mostly the same in each trial (classes 0, 1 and 7). The commonality of these classes is (i) a clear separation from other classes, (ii) a compact alignment in the embedding space (t-SNE), as well as (iii) the location in margin areas of the projection manifold (see Figure 6). These classes seem to be easier to model and thus classifiers benefit most from them. A formalization of ULoP's strategy in the first ten iterations may be as follows: *choose one instance from each class, starting with compact and well-separated classes*.

After 20 iterations, the label distribution was still balanced almost equally over all classes. In the further course, the distribution became more unbalanced; easily predictable classes (like class 0 and 7) were selected more frequently than classes that are difficult to predict. *ULoPs results lead to a balanced distribution at this stage, but with a slight bias to reduce to safe instances*. This is to some degree surprising, as one may draw the hypothesis that difficult classes may require more labels.

In later stages, when the label distribution has become unbalanced, the strategy re-balances the distribution. We assume that there is a moment when those easily predictable classes are already predicted well, so it does not make sense to further explore them. Then, ULoP deliberately tends to label wrongly classified instances to improve the accuracy. *Structures consisting of uncertain instances are preferred at this stage of the labeling process*.

Summary. We draw the conclusion that *ULoP seems to apply several different strategies in the course of the process* - and the distinction of these (and their combinations) is not trivial. Some clusters are labeled late in the labeling process, others are labeled quite early, as well as outliers. A possible formulation of ULoP's overall strategy is as follows: *The more compact and separated a cluster, the earlier it is labeled in the process and the higher is the gain of accuracy when choosing one of the cluster's instances.* We identified that the instance selected within a cluster was not necessarily the centroid (observed with t-SNE). We assume that neighborship relations to other classes/clusters have a strong influence on the selection of both representative instances and instances that support the separation of classes. In fact, we observed that ULoPs output often defines central supporting points of classes/clusters.

7. IRIS

The visual representation of the IRIS data set (see Figure 7) forms a clearly separated cluster (class 1) and a super-cluster consisting of the classes 2 and 3 (partially conflicting).

In the first three iterations of ULoP, each class is covered exactly once (5/5 analysts). Again, we infer that the bootstrap problem is tackled by ULoP almost perfectly. ULoP's strategy in this first stage was to balance the label distribution, slightly preferring centroid-close instances.

In the subsequent phase, we observed an unexpected behavior. Instances from class 1, (clearly separated), have been selected considerably more frequently than instances from the intersecting classes 2 and 3. In addition, choosing class 2 or 3 led to an intermediate decrease of accuracy (3/5 analysts). Such a selection was often followed by an instance of the opposed class (3 or 2) which again produced an accuracy gain (5/5 analysts). One explanation for these phenomena may be that the ULoP algorithm only looks one iteration into the future. When there is no instance that yields a gain in accuracy an instance for which the decrease in accuracy is minimal is ranked highest. This problem may be resolved if the ULoP implementation would look more than one iteration into the future, coupled with a massive increase of computational

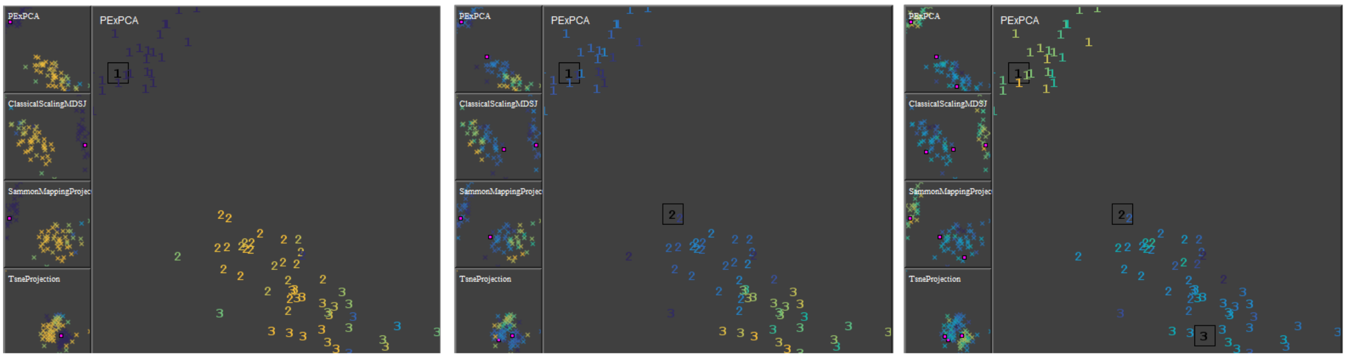

Figure 7: *Visual observation of the IRIS dataset observed with PCA (iterations two to four). Class 1 (setosa) is clearly separated and was labeled first. Second, class 2 is labeled (versicolor) and finally, class 3 (virginica) is labeled with an instance near the class center of gravity. The fourth label is again of class 1 (not shown). Thus, ULoP accessed one instance of each class first, as it did for every other data set.*

costs. An interesting side aspect regards human cognition: *humans would easily identify the necessity to assign labels to both classes (class 2 and 3)* to achieve a gain in accuracy.

One side-note was the observation that t-SNE performed worst in the separation of class 1 and the two remaining classes. While all analysts made good experiences for the remaining data sets, t-SNE was less useful for the very small IRIS data set. The observation of IRIS revealed one point that may be remarkable: *there are situations in the labeling process where the accuracy will decrease* regardless which instance will be added next. It will be interesting to assess the effect of such local "valleys" in the accuracy progression.

8. GENDER Voice

Dimensionality reduction applied on the GENDER voice dataset reveals several cluster structures for both classes (m and f). With t-SNE, we observe that instances of the two classes are well-separated, see Figure 8.

In the first two iterations of ULoP, both classes are covered (5/5). As such, ULoP efficiently resolves the bootstrap problem. In early labeling iterations, we identified a strong prioritization of centroid-near instances. *Labeling centroid-like instances in the beginning may contribute to a compact and representative set of initial labels.*

After this first phase, the strategy of ULoP changed. We made the observation that many selected instances were close to already labeled instances, rather than located in more uncertain regions. One interpretation of this phenomenon may be that *the repeated labeling of relevant areas in the data set helps to consolidate the classifier.* This strategy was complemented with the selection of instances that tend to separate classes. One analyst used the metaphor of a Dirichlet tessellation [LD50] / Voronoi diagram [Vor08] to describe the distribution of points and resulting class areas in 2D.

After having consolidated the class centroids, ULoP started to select instances situated at the border of their visual clusters and also at the border of their respective class. ULoP uncovered the regions the classifier was most uncertain about, similar to an important active learning principle (uncertainty sampling). The observational trials ended after about ten labels per class, when the classifier already achieved a very high accuracy with only marginal space for improvement. In general, ULoP's strategy on GENDER voice can be referred to a balancing approach, *first selecting instances near the cluster/class centroids, then exploring the border areas.*

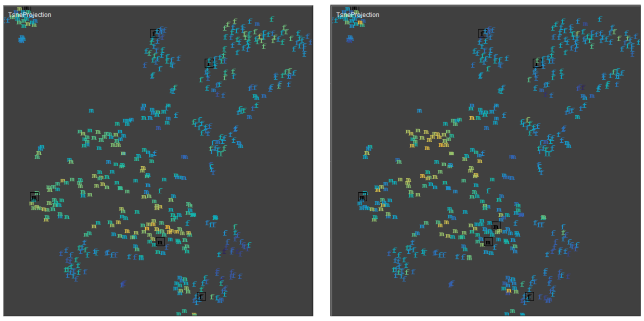

Figure 8: *Visual observation of the GENDER voice dataset after six iterations (three male, three female labels, see the six black instances with rectangles). A pattern can be seen in iteration seven and eight: two male instances are chosen to consolidate the large male cluster at the center. With that, the accuracy increases to 0.88.*

9. FRAUD Detection

FRAUD is a data set with an unbalanced label distribution, including ten times as many regular cases (0) than frauds (1). Dimensionality reduction (t-SNE) reveals a super-cluster, as well as some structures in margin areas, one reflecting frauds (see Figure 9).

In the beginning, both label classes are addressed by ULoP equally. All observed trials started with a 0 class label (5/5). Then, an instance of the 1 class was recommended, i.e., after two iterations both classes were equally covered. Again, we ascertain the particularly good performance to address the bootstrap problem.

Despite balancing, the instances were not chosen in an actual alternating order (2/5). We note that *ULoP balanced the amount of labeled instances in early iterations, even for an unbalanced dataset.* In the further course, the proportional amount of instances of class 1 decreased. However, we did not observe (0/5) that it decreased towards its natural balance (1 : 10).

It should be noted that the labeled instances of class 0 are not equally spread - they are mostly located in the vicinity of class 1 instances. Cluster regions that only contained class 0 instances required only few labels. Opposed to this, clusters of class 1 instances were labeled often, and quite early in the labeling process. Also, those clusters were predicted correctly after just a few iterations and are mostly the only instances in the data set which are predicted as class 1. There are some instances of class 1 in the data set which are assessed with a low performance gain in the entire labeling process. Probably, classifiers lack of complexity to reliably distinguish those special instances from class 1 and would as a consequence mis-classify many instances from class 0 after having modeled these outliers explicitly.

The general strategy of ULoP can be summarized as follows: *first, label both classes exactly once. Then, label both classes with a bias towards class consolidation. Finally, focus on instances that strengthen local class separation.*

10. ISOLET Spoken Letters

The visualization of the ISOLET data set shows several separated clusters in peripheral areas (see Figure 10). However, moving closer to the center, we identify considerably less cluster/class separation. Part of the super-cluster at the center are the classes B, C, D, E, G, P, T, and V, which also have phonetical similarity.

During the initial phase of the labeling process, *ULoP covers all 26 classes exactly once* and thereby effectively solves the bootstrapping problem. Some classes (A, I, Y, X, and C) are always chosen very early (5/5), while some other classes (B, V, P) tend to be chosen rather at the end of the initial phase. *There is a slight trend towards selecting peripheral classes earlier*. For the ISOLET data set, however, further experiments are required for validation. With the large number of classes, it becomes more difficult to assess whether or not centroid-near instances are preferred by ULoP over instances at the border of classes. Switching between different dimensionality reduction helps with the investigation, but does not reveal a clear trend. What can be observed very clearly is the even spatial distribution of selected labels, spanning the entire 2D space (see Figure 10). *Sampling the space uniformly seems to be a key element in the early ULoPs strategy*.

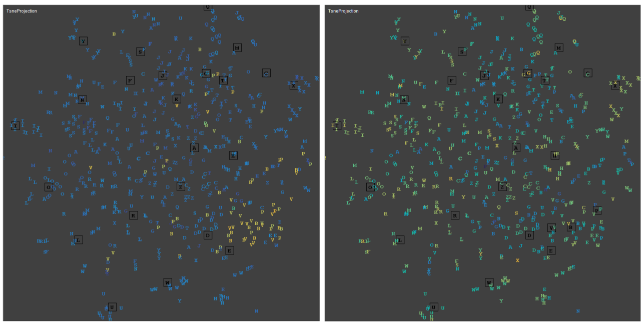

Figure 10: *Visual observation of the ISOLET dataset in iteration 24 and 27. In iteration 24 only the labels B, P, and V have not been seen yet. Interestingly, those letters are phonetically similar and mostly located in a distinct region of the manifold. In iteration 27 every label has been seen exactly once. The preferences for the next label have a slight tendency towards compact clusters in border regions.*

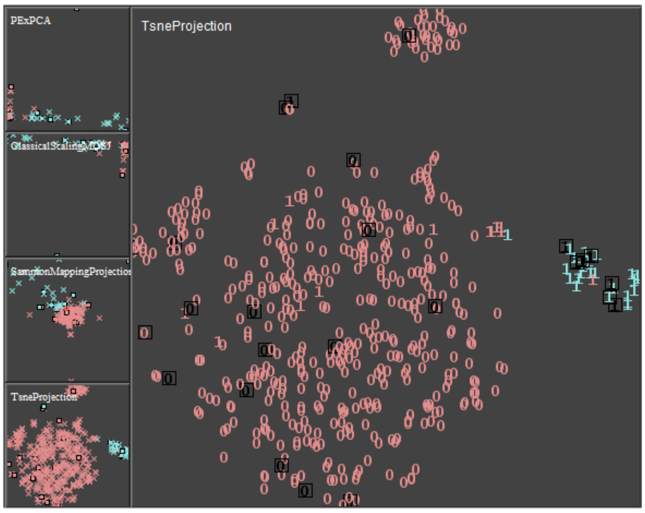

Figure 9: *Visual observation of the FRAUD dataset after 25 iterations using a class-based color coding. Frauds (1-labels) build a separate cluster with all dimensionality reduction techniques. In t-SNE it can be seen that few fraud instances in the large 0 cluster are still classified as no fraud requiring individual treatment (fine tuning). After iteration 10 the average F1 score remained at about 0.96. F1 was chosen to account for the unbalanced class priors.*

In the phase of consolidation, three analysts again observed the selection of exactly one instance per class. The two remaining analysts identified some instances which were selected a third time before every instance was addressed twice. In summary, *ULoPs results lead to a balanced label distribution in the consolidation phase too*. Similar to the initial labeling phase, instances with a peripheral location in the projection plane tend to be selected earlier than classes in the center of the embedding space.

During the last observed phase beginning with the 53rd iteration (3/5), the accuracy started to stagnate. All analysts reported that the third phase conducted with the ISOLET data set with its 26 classes was the most challenging observation task because keeping the overview was difficult. Similarly, the observation of the labeling process consumed considerably more time with ISOLET. For multi-class labeling, it may be beneficial to research for other visualization types.

11. Conclusion

We presented the results of an observational study of the Upper Limit of Performance in labeling, conducted on five data sets. A core insight is the existence of three core phases in the labeling process, i.e. a *discovery phase* where every label is seen (at least) once, a *consolidation phase* where class structures are supported with additional labels and a *fine tuning phase* where class boundaries and outliers are additionally labeled. With the results, we made one step towards the understanding of potentials and mechanisms of future labeling strategies. Other future work includes research and experiments with alternative criteria for upper limits of performance, other visual interfaces for the analysis of labeling strategies (especially for multi-class problems), as well as research into scenarios with an unknown class cardinality.

References

[LD50] LEJEUNE DIRICHLET G.: Über die reduction der positiven quadratischen formen mit drei unbestimmten ganzen zahlen. *Journal für die reine und angewandte Mathematik 40* (1850), 209–227. URL: http://eudml.org/doc/147457. 2, 5

[Vor08] VORONOI G.: Nouvelles applications des paramètres continus à la théorie des formes quadratiques. premier mémoire. sur quelques propriétés des formes quadratiques positives parfaites. *Journal für die reine und angewandte Mathematik 133* (1908), 97–178. URL: http://eudml.org/doc/149276. 2, 5