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

An overarching goal of active learning strategies is to reduce the human effort when labeling datasets and training machine learning methods. In this work, we focus on the analysis of a (theoretical) quasi-optimal, ground-truth-based strategy for labeling instances, which we refer to as the upper limit of performance (ULoP). Our long-term goal is to improve existing active learning strategies and to narrow the gap between current strategies and the outstanding performance of ULoP. In an observational study conducted on five datasets, we leverage visualization methods to better understand how and why ULoP selects instances. Results show that the strategy of ULoP is not constant (as in most state-of-the-art active learning strategies) but changes within the labeling process. We identify three phases that are common to most observed labeling processes, partitioning the labeling process into (1) a Discovery Phase, (2) a Consolidation Phase, and (3) a Fine Tuning Phase.

CCS Concepts

•***Human-centered computing*** → *Information visualization;* •***Theory of computation*** → *Active learning;*

1. Introduction

Labeling refers to the task of assigning additional information to data instances, such as class labels, categories, or relevance scores. Today, labels are essential for the supervised training of machine learning (ML) models. Prominent examples for ML tasks include the recognition of different objects (e.g. people, cars, street signs [KSH12]), or between handwritten digits [LC10]. As such, labeling data instances is a precondition for supervised ML.

Labeling is, however, an expensive and time-consuming task because in practice a large number of labeled instances is required to enable successful training of accurate ML models. Active learning (AL) [Set12] is an incremental learning methodology that aims at reducing labeling costs. AL puts the user into the ML loop and actively selects candidates for labeling (according to a labeling strategy) to improve the ML model in an efficient way. Recently, it has been shown that AL can benefit substantially from the combination with visual analytics approaches in a unified process, referred to as visual-interactive labeling (VIAL) [BZSA18]. VIAL combines the strengths of humans and active ML models in the selection of meaningful candidates for learning.

Recent experiments with different labeling strategies have shown that the selection of useful candidate instances is far from being optimal both in pure AL as well as in combined VIAL approaches. To assess the performance of strategies a quasi-optimal selection strategy was defined that always queries the most useful instance for the learner. The most useful instance is found simply by trying out all possibilities in a greedy fashion. This strategy serves as an upper bound for AL (which we call *Upper Limit of Performance*, ULoP in the following). Experiments revealed a large gap in performance between existing strategies and ULoP, which demonstrates that there is a considerable potential for improvement of strategies [BHZ*17]. The observed performance gap raises interesting research questions regarding how a good selection strategy should work:

- how are instances selected by the ULoP strategy compared to existing AL strategies and VIAL strategies?
- are there certain patterns or rules in the process of selecting instances that can be observed from the ULoP?
- how to select samples for labeling in an optimal way and how to formalize selection strategies that better facilitate learning?

We expect that a closer visual analysis of the ULoP strategy may lead to new insights why the strategy considerably outperforms other selection strategies. Based on these findings it may be possible to create better strategies to facilitate labeling.

The comparative analysis of different AL strategies has a long tradition [RM01], providing insight into performances of different classes of strategies as well as dependencies to data characteristics see, e.g., Burr Settle's survey [Set12]. Seifert and Granitzer [SG10] investigated the performance of user-based picking strategies compared to uncertainty based sampling (AL) [Set09]. Building upon the former, a recent experiment analyzed the performance of formalized user strategies in detail, using the ULoP strategy as a means to anticipate potential space for improvement [BZL*18].

We present the results of an observational study of the ULoP strategy and leverage visualization to obtain deeper insights into the quasi-optimal labeling process. Our contributions are as follows:

- we present the results of the labeling process performed by ULoP, visually observed by five analysts for five datasets
- we condense the observation results and identify commonalities and differences in patterns observed for the five datasets.

DOI: 10.2312/eurovisshort.20181085

- we propose the partition of the labeling process into three principle phases (*Discovery Phase*, *Consolidation Phase*, and *Fine Tuning Phase*), where each of which requires individual labeling strategies to achieve optimal performance.

2. Observation Methodology

Observation Goal and Guiding Questions: We conduct an experiment[†] that seeks to understand what makes the ULoP strategy particularly powerful compared to existing strategies. For this purpose, we use a visual analysis approach that allows the in-depth investigation of each iteration of the ULoP labeling process. We conducted five observation trials for each dataset, each performed by a well-informed data analysis expert from our team (all of them authors of this paper). The procedure was repeated for five different datasets. The experiment time roughly took one minute for every iteration The goal was to derive insights for every dataset. Additionally, we describe the commonalities and differences of findings across the characteristics of individual datasets to draw more general conclusions and formulate guidelines. Overall, we investigate three guiding questions with the overall goal to improve candidate selection in future approaches. Given the assumption that low-dimensional visual representations reveal interesting patterns about ULoP's characteristics, our guiding questions were: Can we infer common denominators for visual patterns? Are there changes in the course of the labeling process? Are there commonalities and differences between datasets?

Formalization of the Upper Limit of Performance: Many labeling strategies have in common that they try to determine the particular instance from the unlabeled dataset for labeling which is most likely to improve the ML model most [Set09]. To simulate a quasi-optimal labeling strategy, our ULoP strategy is modeled by executing a greedy search for instances based on ground truth data. In each labeling iteration, the algorithm evaluates the benefit of each individual unlabeled instance by retraining the classifier with that instance and finally selects the instance which contributes most to classification performance. We use an ensemble of four classifiers including Random Forest [Bre01], NaiveBayes [HKP12], SVM (based on SMO) [Pla98], and a MultilayerPerceptron classifier [RHW86] to reduce the influence of a particular classifier and its learning strategy on the experiment outcome.

Datasets: We employ five different datasets with heterogeneous characteristics to reduce dataset bias in the experiments and to investigate the labeling behavior in different situations. Table 1 provides an overview.

Experimental Setup: A pre-condition for the success of the observational experiment is the ability to identify visual patterns of the ULoPs strategy. The output of the ULoP algorithm is a continuous value between zero and one, reflecting the performance gain of a given instance for the next labeling iteration. For balanced label sets, we use accuracy [FHOM09] and for unbalanced datasets the F1 score [SWY75]. The performance scores are computed on an independent test set for every dataset.

A core benefit of our visual approach is the ability to show not only winning candidates, but the distribution of gains across all candidate instances. In our labeling experiment, we map the accuracy gain of candidates to a continuous univariate color map (gain

[†] we use the term experiment in its general meaning, and not the narrow interpretation as null-hypothesis testing experiments, which is sometimes associated with it in the VIS and HCI communities.

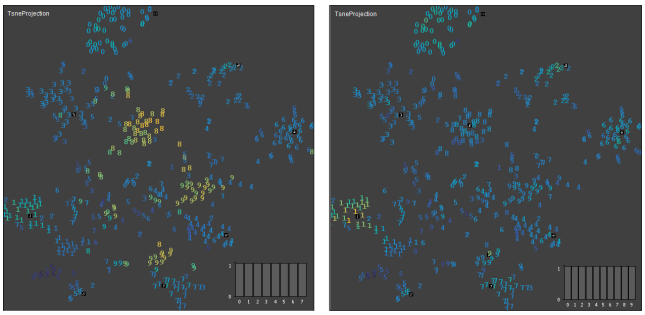

Figure 1: *Visual interface used for the observational study. Two iterations of the labeling process for the MNIST dataset are shown. Dimensionality reduction (here: t-SNE) allows the representation of instances in 2D. Color encodes how much the learning model will benefit from labeling a candidate (orange is best). Left: The Results of ULoP strategy suggest that labeling instances from classes 8 or 9, after all remaining classes (digits) have already been labeled once, is most beneficial. Right: Every class has been labeled exactly once. The Output of ULoP then suggests that choosing digits of class 1, located at the left margin of the manifold would yield the strongest improvement.*

increasing from dark blue to saturated orange). Instances are depicted with a textual information about their true class (unlabeled as well as labeled). Instances already labeled by ULoP are colored black. To represent the high-dimensional instances in 2D, we apply dimensionality reduction. The resulting data representation is depicted in Figure 1. To account for reconstruction errors, the the data experts were informed to switch between four algorithms in the labeling process (PCA [Jol02], t-SNE [vdMH08], non-metric MDS [Kru64], and Sammons Mapping [Sam69]). According to the informal feedback of the experts, t-SNE was predominantly used during observation.

We always start with completely unlabeled data to investigate how ULoP resolves bootstrap problems (e.g., occurring in AL [AP11]). However, we randomize the selection of the very first instance for labeling resulting in unique trials for every analyst. This is important since ULoP is a deterministic process and a randomization is necessary to avoid biases. The observation of a labeling process ends when the performance measure converges (stop criterion). The longest trials were conducted for the ISOLET dataset with 26 classes (criterion reached with approx. 80 iterations).

3. Results

We first summarize the observations of the labeling process for each of the five datasets. The results are presented in a supplemental material document in detail and exemplified in Figures 1, 2, 3, 4,

Dataset	Cls.	Train	Test	Bal.
MNIST [LC10] handwritten digits	10	500	1000	yes
FRAUD [PCJB15] credit card frauds	2	500	2500	no
IRIS [Lic13] classes of iris plants	3	75	75	yes
GENDER [Bec16] female/male voices	2	500	1000	yes
ISOLET [CMF90] spoken letters	26	650	650	yes

Table 1: *Employed datasets. The table summarizes the number of classes and the number of training and testing instances for each dataset, and whether the label distribution is balanced or not.*

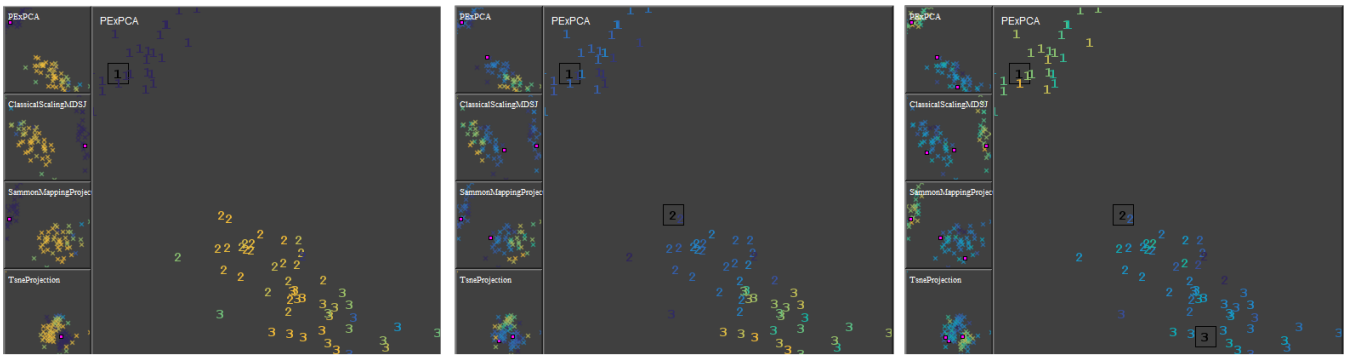

Figure 2: *Visual observation of the IRIS dataset observed with PCA (iterations two to four). Class 1 (setosa) is clearly separated and was labeled first. Second, class 2 is labeled (versicolor) and finally, class 3 (virginica) is labeled with an instance near the class center of gravity. The fourth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

and 5. In the following, we reflect on differences and similarities of these findings, and draw generalizable conclusions.

3.1. Generalization of ULoP Results

During the observations of labeling processes across the five datasets, we have identified different behaviors and labeling patterns. Individual observation results for the different datasets are compiled in the supplementary material due to limited space. Here we have consolidated our observations to derive general behavioral patterns of ULoP that are similar across all datasets. A major insight gained in all five datasets was the existence of three core phases in the labeling process, which we refer to as *Discovery Phase*, *Consolidation Phase*, and *Fine Tuning Phase*. According to these phases, we structure the following result descriptions.

3.1.1. Discovery Phase

The discovery phase starts with the completely unlabeled dataset. In the very first labeling iterations, the ULoP strategy always labeled each class exactly once as soon as possible, and thereby effectively solved the bootstrap problem (5/5 analysts, 5/5 datasets). We have the impression that some classes were preferred by the ULoP strategy. In case of the MINST dataset the classes 0, 1, 6, and 7 were labeled comparatively early, while classes 2, 4, 5, 6, 8, and 9 were addressed later. For the ISOLET dataset, early classes were A, I, Y, X, and C in contrast to the late classes B, V, and P. Overall, we identified the following influencing factors:

- Compact classes are favored. This is most probably because they facilitate the inference of class predictions for a series of instances from labeling a single representative
- Similar to the former, cluster structures are favored for labeling
- Classes with a clear separation from other classes are favored
- Classes located in marginal areas of the manifold are preferred

Additionally, a driving principle of the ULoP strategy during the discovery phase is to obtain a uniform sampling of the (projected) feature space, i.e. ULoP tries to spatially distribute the labels across the entire space. To sum up, the primary goal in the discovery phase is to discover how many classes the dataset is composed of and to solve primarily the bootstrap problem to obtain a first complete classification model. The discovery phase ends with the selection of exactly one representative instance for every class.

3.1.2. Consolidation Phase

In the second phase, we observed that different strategies were applied by the ULoP algorithm.

- We infer that ULoPs results always lead to a balanced number of instances, even though classes are not selected in a strictly alternating way.
- Dense structures in the data (i.e. clusters) were preferred. In particular, compact and well separated classes were favored (similar to our observations in the discovery phase). We were able to clearly observe this behavior for all datasets except the ISOLET dataset for which the 2D projection is particularly dense and the large number of classes impedes a clear clustering.
- Inside of clusters, it is not necessarily the (visible) centroid which is considered as most important. We cannot decide if this is a particular characteristic of the ULoP strategy or a visual artifact. In fact, there is a tendency to label centroid-near instances for datasets with low number of classes (GENDER voice, FRAUD, IRIS), while in datasets with many classes inter-class dependencies may introduce additional constraints.
- The ULoP strategy favors instances within clusters, which have already been labeled once. We assume this pattern to be beneficial for classification because it may help to strengthen the trust of the classifier in a cluster and thereby in its class model.
- In cases where a class is split into two clusters, both clusters are covered before one cluster is further consolidated. This shows that in the consolidation phase ULoP's results (mostly) capture the coarse structure (topology) of the classes.

The result of the consolidation phase is a label set that (i) is balanced, (ii) captures the coarse shape of the class distributions, and (iii) samples instances in areas of overlapping classes.

3.1.3. Fine Tuning Phase

After the consolidation phase, different strategies are addressed by the ULoP strategy to further improve the labeling. At a glance, we observed a transition from synoptic to elementary behavior, i.e., instances such as outliers are selected which are not necessarily representative for larger structures (e.g. clusters). The most frequently observed labeling strategies include:

- Examination of local unexplored structures in the data
- Refinement of not yet well separated classes
- Focus on areas with multiple overlapping classes
- Labeling of outlier instances

At the start of the fine-tuning phase, the classifier's performance usually had already achieved a high level for the respective dataset. Further labeling merely caused small changes in the classifier performance. The major purpose of the fine tuning phase seems to be a refinement of the classes with respect to special cases (e.g. outliers)

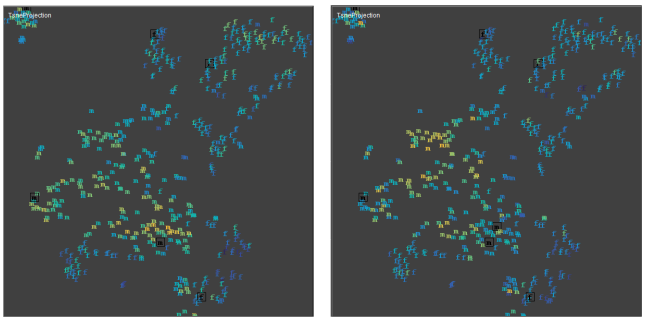

Figure 3: *Visual observation of the GENDER voice dataset after six iterations (three male, three female labels, see the six black instances with rectangles). A pattern can be seen in iteration seven and eight: two male instances are chosen to consolidate the large male cluster at the center. With that,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

which are difficult to model statistically. This may also be a reason why performance decreases are sometimes observed in this phase, i.e. the trained statistical models get biased by outliers.

3.2. Discussion

ULoP The greedy ULoP strategy is limited (and thus only quasi-optimal) by the fact that it looks exactly one iteration into the future. Especially in the IRIS experiment with its similar-looking classes 2 and 3, it became apparent that the shortsighted nature of ULoP causes problems. In IRIS, ULoP labeled 31 instances of class 1 and only three instances of class 2 and 3 each during the labeling process. The limited search range of ULoP yields a strongly imbalanced distribution of class labels which may impede further labeling iterations.

Allowing to query two or three additional instance labels in such cases could solve this imbalance and could yield a much better accuracy in the medium term. Such an extension of ULoP raises further questions like: how many instances should we take into consideration? And how do we deal with combinatorial explosion of possible solutions? Looking only two instances into the future in a dataset of n instances would yield a computational complexity of $O(n^2)$, and would in practice require the classifiers to be trained millions of times in every iteration.

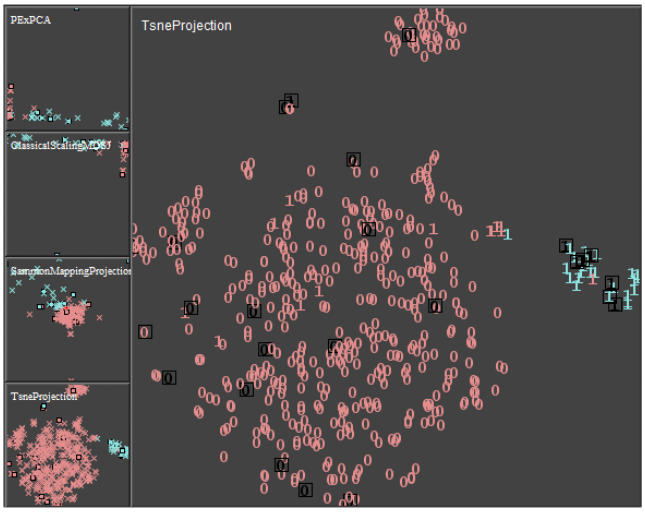

Figure 4: *Visual observation of the FRAUD dataset after 25 iterations using a class-based color coding. Frauds (1-labels) build a separate cluster with all dimensionality reduction techniques. In t-SNE it can be seen that few fraud instances in the large 0 cluster are still classified as no fraud requiring 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.*

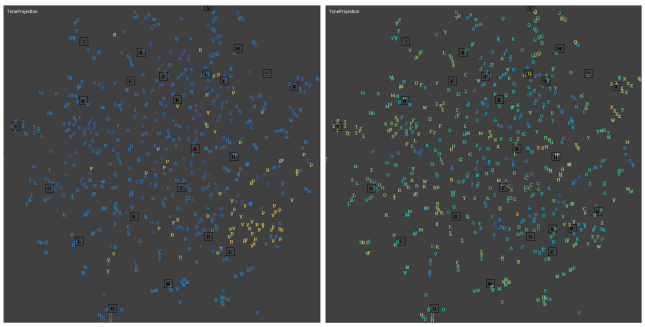

Figure 5: *Visual observation of the ISOLET dataset in iteration 24 and 27. In iteration 24 only the labels B, P, and V are still missing. Interestingly, those letters are phonetically similar and mostly located in a distinct region of the manifold. In iteration 27 every label has been seen exactly once. The preferences for the next label have a slight tendency towards compact clusters in border regions.*

Dimensionality Reduction Dimensionality reduction has shown to be useful for the analysis of labeling strategies in many cases. However, especially for the ISOLET dataset (with 26 classes) we observed that dimensionality reduction also poses some challenges. This became most apparent for datasets with a large number of classes where we observed many class confusions in the embedding manifold. Further research into visual interfaces may be useful to support dimensionality reduction with additional views.

4. Conclusion

We presented the results of an observational study of a quasi-optimal strategy for labeling data, conducted on five datasets. The major conclusion of our investigation is that the labeling process can be partitioned into different phases where each phase follows different instance selection strategies. This is an interesting finding, especially due to the fact that in today's active learning approaches usually one strategy is employed over the entire labeling process. In the *Discovery Phase* every class is labeled exactly one time, in the *Consolidation Phase* the coarse class structures are sampled with additional labels, and in the *Fine Tuning Phase* class boundaries and outliers are refined. With the results, we made one step towards better understanding existing potentials and mechanisms to improve future labeling strategies. 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 unknown class cardinality.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG), Project No. I 2850 (-N31), Lead Agency Procedure (D-A-CH) "Visual Segmentation and Labeling of Multivariate Time Series (VISSECT)".

References

[AP11] ATTENBERG J., PROVOST F.: Inactive learning?: Difficulties employing active learning in practice. *ACM SIGKDD Explorations Newsletter 12*, 2 (Mar. 2011), 36–41. doi:10.1145/1964897.1964906. 2

[Bec16] BECKER K.: Gender recognition by voice – identify a voice as male or female, 2016. Accessed: 2017-12-05. 2

[BHZ*17] BERNARD J., HUTTER M., ZEPPELZAUER M., FELLNER D., SEDLMAIR M.: Comparing visual-interactive labeling with active learning: An experimental study. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 24*, 1 (2017). accepted for TVCG. doi:10.1109/TVCG.2017.2744818. 1

[Bre01] BREIMAN L.: Random forests. *Machine Learning 45*, 1 (2001), 5–32. doi:10.1023/A:1010933404324. 2

[BZL*18] BERNARD J., ZEPPELZAUER M., LEHMANN M., MÜLLER M., SEDLMAIR M.: Towards user-centered active learning algorithms. *Computer Graphics Forum (CGF)* (2018). to appear. 1

[BZSA18] BERNARD J., ZEPPELZAUER M., SEDLMAIR M., AIGNER W.: VIAL: a unified process for visual interactive labeling. *The Visual Computer* (Mar 2018). doi:10.1007/s00371-018-1500-3. 1

[CMF90] COLE R., MUTHUSAMY Y., FANTY M.: The ISOLET spoken letter database. *Tect. Rep* (1990), 90–004. 2

[FHOM09] FERRI C., HERNÁNDEZ-ORALLO J., MODROIU R.: An experimental comparison of performance measures for classification. *Pattern Recognition Letters 30*, 1 (2009), 27–38. doi:10.1016/j.patrec.2008.08.010. 2

[HKP12] HAN J., KAMBER M., PEI J.: *Data Mining – Concepts and Techniques*. Morgan Kaufmann Publishers, 2012. 3rd edition. 2

[Jol02] JOLLIFFE I. T.: *Principal Component Analysis*, 3rd ed. Springer, 2002. 2

[Kru64] KRUSKAL J. B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1 (1964), 1–27. 2

[KSH12] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105. 1

[LC10] LECUN Y., CORTES C.: MNIST handwritten digit database. URL: http://yann.lecun.com/exdb/mnist/. 1, 2

[Lic13] LICHMAN M.: UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml. 2

[PCJB15] POZZOLO A. D., CAELEN O., JOHNSON R. A., BONTEMPI G.: Calibrating probability with undersampling for unbalanced classification. In *Computational Intelligence and Data Mining (CIDM)* (2015), IEEE, pp. 159–166. doi:10.1109/SSCI.2015.33. 2

[Pla98] PLATT J.: Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning*, Schoelkopf B., Burges C., Smola A., (Eds.). MIT Press, 1998. URL: http://research.microsoft.com/~jplatt/smo.html. 2

[RHW86] RUMELHART D. E., HINTON G. E., WILLIAMS R. J.: Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. MIT Press, Cambridge, MA, USA, 1986, ch. Learning Internal Representations by Error Propagation, pp. 318–362. URL: http://dl.acm.org/citation.cfm?id=104279.104293. 2

[RM01] ROY N., MCCALLUM A.: Toward optimal active learning through monte carlo estimation of error reduction. *International Conference on Machine Learning (ICML)* (2001), 441–448. 1

[Sam69] SAMMON J. W.: A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers 18*, 5 (1969), 401–409. doi:10.1109/T-C.1969.222678. 2

[Set09] SETTLES B.: *Active Learning Literature Survey*. Tech. Report 1648, Univ. of Wisconsin–Madison, 2009. 1, 2

[Set12] SETTLES B.: Active learning. *Synthesis Lectures on Artif. Intell. and Machine Learning 6*, 1 (2012), 1–114. 1

[SG10] SEIFERT C., GRANITZER M.: User-based active learning. In *IEEE Conference on Data Mining Workshops (ICDMW)* (2010), pp. 418–425. doi:10.1109/ICDMW.2010.181. 1

[SWY75] SALTON G., WONG A., YANG C. S.: A vector space model for automatic indexing. *Communications of the ACM 18*, 11 (Nov. 1975), 613–620. doi:10.1145/361219.361220. 2

[vdMH08] VAN DER MAATEN L., HINTON G. E.: Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research (JMLR) 9* (2008), 2579–2605. 2