Selectively inhibiting learning bias for active sampling

Brazilian Conference on Intelligent Systems, IV, 2015, Natal.
http://www.producao.usp.br/handle/BDPI/49974

Selectively inhibiting learning bias for active sampling

Davi P. dos Santos
and André C. P. L. F. de Carvalho
Universidade de São Paulo (USP)
Instituto de Ciências Matemáticas e Computação (ICMC)
São Carlos - SP, Brazil
Email: davips@icmc.usp.br, andre@icmc.usp.br

***Abstract*—Efficient training of machine learning algorithms requires a reliable labeled set from the application domain. Usually, data labeling is a costly process. Therefore, a selective approach is desirable.**

Active learning has been successfully used to reduce the labeling effort, due to its parsimonious process of querying the labeler. Nevertheless, many active learning strategies are dependent on early predictions made by learning algorithms. This might be a major problem when the learner is still unable to provide reliable information. In this context, agnostic strategies can be convenient, since they spare internal learners - usually favoring exploratory queries. On the other hand, prospective queries could benefit from a learning bias.

In this article, we highlight the advantages of the agnostic approach and propose how to explore some of them without foregoing prospection. A simple hybrid strategy and a visualization tool called *ranking curves*, are proposed as a proof of concept. The tool allowed to see clearly when the presence of a learner was possibly detrimental. Finally, the hybrid strategy was successfully compared to its counterpart in the literature, to pure agnostic strategies and to the usual baseline of the field.

***Keywords*—active learning; machine learning; agnostic**

I. Introduction

Several machine learning algorithms have been proposed to induce models able to deal with a variety of application tasks. Such algorithms need a reliable sample of the application domain for training, which, in the case of classification, is a set of labeled instances. Usually, labeling requires expensive human supervision effort. To reduce this effort, active learning is a useful approach due to its parsimonious process of querying the supervisor [1]. However, many active learning strategies are dependent on learning algorithms to evolve its internal model - called *learner*. This is a major problem when neither the domain peculiarities nor the user expertise are enough to define the proper algorithm in advance. Without a proper algorithm, its predictions are unreliable, inadequate to be used by the active learning strategy. A possibility to determine the best learning algorithm, would be to rely on the initial training set to select it by cross-validation, but the number of labeled instances is frequently scarce or inexistent. In this context, agnostic strategies are convenient because they do not need internal learners. Indeed, agnostic sampling still allows improvements in label complexity over traditional passive learning [2]. On the other hand, agnostic approaches lack the prospective capability of a learning bias that could speed up the learning process regarding the number of queries. Ideally, both, resilience to a still roughly trained/inadequate learner, and prospection capability are desirable properties of an active learning strategy.

In this study, we empirically demonstrate that agnostic strategies are more effective than gnostic approaches in the first exploratory steps, while the learner bias is important later in the learning process. As a result, we propose a hybrid approach able to identify the moments when relying on the learner is detrimental. It achieved better performance than the baseline and directly related strategies.

This article is organized as follows. Background information concerning active learning and related work is provided in Section II, including a review of pertinent active learning strategies; the proposed method is presented in Section III; experiments and results are detailed in Section IV; and, finally, conclusions are drawn in Section V.

II. Related Work

An effective way to label data selectively is the employment of active learning [1]. It is reasonable to only acquire labels for the most important part of the data, due to acquisition costs. The focus of this work is the *pool-based query*, when the learner is given the freedom to choose the most informative instance among several others in a pool [3]. Depending on the learning algorithm, there are several successful strategies for the pool-based setting [4]. Three strategies are relevant for this work. They are explained in sections II-A, II-B and II-C, respectively: uncertainty sampling, density-weighted sampling and cluster-based sampling. Convenient abbreviations are provided between parentheses after the name of each strategy.

A. Uncertainty Sampling

A purely prospective strategy is Uncertainty Sampling (Unc). It focuses on exploitation while neglecting exploration. The criterion to select an instance is based on the maximum posterior probability. The posterior probability is given by a probabilistic model [3], which provides how probable is the class y given an instance $\boldsymbol{x}$. The maximum posterior probability is presented in Equation 1:

$$P_{max}(\boldsymbol{x}) = \max_{y} P(y|\boldsymbol{x}) \quad (1)$$

978-1-5090-0016-6/15 $31.00 © 2015 IEEE
DOI 10.1109/BRACIS.2015.17

where $\boldsymbol{x}$ is sampled from the pool $\mathcal{U}$, and $P(y|\boldsymbol{x})$ is the posterior conditional probability of $\boldsymbol{x}$ being of the class y. The output of a probability-based model, for instance, can be adopted as $P(y|\boldsymbol{x})$. The uncertainty sampling strategy consists of querying the most informative instance, i.e. the instance with the lowest $P_{max}(\boldsymbol{x})$, to explore the decision boundary in the attribute space - or parameter space, depending on the classifier. Although this measure depends on a probabilistic model, probability distributions can be estimated for other families of learning algorithms.

This strategy requires only a single training on the labeled instances to be able to test all remaining candidates. Margin and entropy measures are alternative variants to uncertainty for multiclass problems.

B. Density-weighted Sampling

While Unc exploits only the uncertainty of the model, density weighted (DW) strategies exploit the data distribution in addition to the uncertainty measure. The *information density* measure [5] is given by Equation 2:

$$ID(\boldsymbol{x}) = H(\boldsymbol{x})\frac{1}{|\mathcal{U}|}\sum_{\boldsymbol{u}\in\mathcal{U}} sim(\boldsymbol{x},\boldsymbol{u}) \tag{2}$$

or the *training utility* (TU) [6], measure adopted in this work, shown in Eq. 3:

$$TU(\boldsymbol{x}) = ID(\boldsymbol{x})(\sum_{\boldsymbol{l}\in\mathcal{L}} sim(\boldsymbol{x},\boldsymbol{l}))^{-1} \tag{3}$$

where $\mathcal{L}$ is the set of labeled instances. Any similarity $sim(\boldsymbol{x},\boldsymbol{u})$ and informativity measures $H(\boldsymbol{x})$ can be adopted. In this work, the Euclidean distance was adopted and transformed into a similarity measure by the formula in Eq. 4:

$$sim(\boldsymbol{x},\boldsymbol{u}) = \frac{1}{1+d(\boldsymbol{x},\boldsymbol{u})} \tag{4}$$

The uncertainty $P_{max}(\boldsymbol{x})$ was adopted as the informativity measure in the experiments.

The complexity order is $\mathcal{O}(1)$, but $|\mathcal{U}|^2$ distance calculations are needed for each query; they should be cached in fast access memory before the active sampling process is started.

C. Cluster-based Sampling

The active sampling process can exploit natural clusters in the pool, instead of performing queries that focus on decision boundaries. One such approach is the Hierarchical Sampling (HS) [7]. It is based on hierarchical clustering [8]. Hierarchical clustering organizes the instances according to a hierarchy, usually represented by a tree. Each *leaf node* represents an instance and each *parent node* represents a similarity/proximity relation between its children. A child is a leaf or a parent of other children. Any node can be seen as a single group of all its descendant leaves. Different pruning choices can be made, leading to different partitioning schemes. Once the instances are partitioned, each group is a cluster.

In the active learning context, instances are queried with higher probability from the most impure and representative clusters. Purity is the proportion between the most frequent class and the others. Representativeness is the amount of instances contained in the cluster. Both measures are combined, and the pruning is defined, according to statistical bounds; details in [7]. It is guaranteed to not perform worse than random sampling. This strategy and random sampling are examples of agnostic approaches. In the experiments, the original implementation provided by the authors was adopted. The clustering algorithm employed was Ward's average linkage method [8].

III. Proposal

We propose two strategies to assess the effect of the learner in the strategy during learning. One agnostic, to track the effects of the learning bias absence; and, a hybrid agnostic-gnostic/exploratory-prospective approach.

A. Agnostic

The usefulness of taking an exploratory step before making any assumptions about the data is intuitive. When one needs to investigate a problem in more depth, data obtained from the exploration step can allow more confident prospection. One form of prospection is to resort to a learned model. However, any model is biased towards assumptions about the data distribution [9]. Therefore, there is a trade-off in the balance of exploration and exploitation which can lead to a confident but inefficient, or to an efficient but highly biased search of the instance space. For purposes of comparison of both cases, we defined a strategy to show the effect of the absence of a learning bias. It can be seen as the agnostic version of TU, called Density-weighted Agnostic Sampling, ATU, to simplify future references in the text.

Agnostic approaches have some advantages. They can query all the instance space instead of only the areas near a learner decision boundary. The absence of a learner to spend training time makes agnostic strategies faster than gnostic ones. Also, it avoids the problem of the choice of a learning algorithm in advance. Another advantage is the independence among queries, which allows e.g. to have simultaneous multiple oracles and to answer the queries in any order[1]. However, an agnostic approach like random sampling have a non uniform exploratory nature. Redundant instances or outliers can be preferred by chance over more representative instances.

We propose that even without a learner, density weighting can be employed to ensure densely and unlabeled areas in the instance space are queried. The new informativity measure is presented in Eq. 5.

$$ATU(\boldsymbol{x}) = \frac{1}{|\mathcal{U}|}\sum_{\boldsymbol{u}\in\mathcal{U}} sim(\boldsymbol{x},\boldsymbol{u})(\sum_{\boldsymbol{l}\in\mathcal{L}} sim(\boldsymbol{x},\boldsymbol{l}))^{-1} \tag{5}$$

The equation of this illustrative strategy balances proximity to dense unlabeled groups of instances and departure from already labeled areas of the instance space.

[1]However, the order is relevant when the application requires prediction capability during the active sampling process.

B. Hybrid

The lack of a decision boundary can imply in no exploitation. This is not the case with HS, because, despite having no learner, it is still able to exploit impure clusters. Therefore, HS has an implicit bias, when it takes into account the available labels. Beyond questioning its real agnostic nature, when employing HS strategy, the desirable independence between queries does not hold. Nonetheless, ATU is also devoid of a learner, but unlike HS, it is able to not depend on labels. The main weakness of ATU is that it cannot optimize towards a target, like improving directly the model predictions or attacking impure clusters. An alternative is to adopt a learner late in the learning process.

A pure gnostic strategy like Unc, can optimize the search for the final decision boundary. Its successive partitioning of the instance space in two parts is analogous to a binary search [10], which in ideal conditions, can lead to an exponential reduction in labeling costs [11]. Although, in adverse conditions, the budget can be wasted on a few instances and its vicinity, before all the extent of a useful boundary is unveiled. Therefore, a combination between exploitation and exploration is desirable. This is already done by TU, where both behaviors are components of the same measure. However, none of them can act in their pure form. This results in permanent influence of the learner, even if its learning stage is not yet adequate to make predictions. There is also no distinction between early and late queries. We propose a new technique to circumvent the disadvantages of both, *presence* and *absence* of learner, called Hybrid Density-based Training Utility Active Learning (HTU).

HTU consists of selective use of TU and ATU when needed and Unc to do internal comparisons. The goal of HTU is to resort to the learner only when exploration is not making relevant contributions anymore. This can be estimated comparing how TU and Unc would sort the available instances according to their relevance. If the two sorted pools were identical, then only Unc would suffice to provide the next query, i.e., agnosticism would have become unneeded. However, this coincidence, or any result near it, is highly improbable most of the time. Therefore, the correlation between the outcome from the two strategies should be quantified, avoiding the comparison of instance positions in two ordered lists. Since both strategies provide informativity measures, it is straightforward to generate two lists of values for the entire pool and compare them via Pearson correlation [12]. Values too close to unit indicates a small contribution of the agnostic part of the TU informativity measure (Eq. 3). This is an indicator that suggests to switch from ATU to TU (or from ATU to Unc) - which is equivalent to embed a learner. An arbitrary value representing a very strong correlation, such as higher than 0.8, can be adopted as threshold [13]. In this work, the value 0.999 was chosen based on datasets that were not included in the experiments[2] as follows. Figure 1 shows the mean ranking position for values of different orders of magnitude in the interval $[0.9; 0.999999]$. 10-fold cross-validation was used to estimate performance of HTU according to the κ kappa measure after 100 queries. The kappa measure [15] is given by Equation 6.

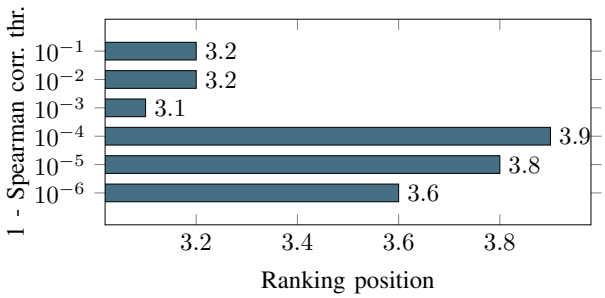

Fig. 1. Ranking position for different Spearman correlation thresholds.

$$\kappa = \frac{a - \frac{\boldsymbol{v}_{pred} \cdot \boldsymbol{v}_{exp}}{n}}{1 - \frac{\boldsymbol{v}_{pred} \cdot \boldsymbol{v}_{exp}}{n}} \tag{6}$$

where $\boldsymbol{v}_{pred}$ is a vetor with the total number of instances predicted by class, $\boldsymbol{v}_{exp}$ is a vetor with the total number of instances expected by class, a is the ordinary accuracy and n is the size of the test set.

Finally, a disadvantage of HTU is its intrinsic need of a learner and the corresponding training time, incurring in additional computational costs when compared to ATU. This can be relevant in applications where the querying time is critical.

IV. Experiments

The proposed method was empirically evaluated according to the current methodology in the areas of active learning and machine learning. Additionally, an innovative tool to visualize the learning curve for multiple datasets is presented.

A. Methodology

For each new query, a new model is built/updated and tested against unknown instances previously set apart. In the literature, an arbitrary number of queries (50, 100, 200, $|\mathcal{U}|$ etc.) have been used [16], [17]. We adopted a budget of 200 queries. Five runs of 5-fold cross-validation were applied. Training folds were used as the pool of unlabeled instances - as adopted by [18]. Duplicate instances were removed.

In the experiments, it was assumed that the label of one instance per class was known before the start of the active sampling process[3]. One or more than one instance per class have been used in the literature [17].

The performance indicator is the *Area under the Learning Curve* (ALC) [19] for κ [20]. In strict terms, the indicator is the mean of all κ values, instead of a real area *under* the curve, since there are also negative values. Kappa (κ) was chosen due to the presence of imbalanced datasets.

[2]Datasets from the UCI repository [14]: artificial charac., statlog vehicle sil., connect. vowel red., robot nav. sensor, vertebra column 3c, volcanoes a3, user knowl., page blocks, waveform v2, turkiye stud., car eval., heart disease clev., cardiotoc. 3, abalone 3, connect. vowel, molec. splice junc., wine quality white, statlog landsat sa., cardiotoc. 1., flare, first order theor., leaf, systhetic control, semeion, balance scale, autoUniv au7 300, volcanoes e2, thyroid sick, statlog image segm., autoUniv au7 700, yeast 4, volcanoes b2, mfeat fourier, thyroid ann, volcanoes d1, movement libras. They were limited to 1000 instances.

[3]Except for the Cluster-based strategy.

B. Datasets and algorithms

The evaluation was performed with 195 different settings resulting from the combination of five learning algorithms and third-nine binary labeled datasets from the UCI repository [14]. The employed classifiers were: 5NN, C4.5w[4], NB[5], SVM with RBF[6] and CIELM [21]–[25]. When not stated otherwise, all parameters used the default values from the Weka library [26]. C4.5w was adjusted to predict the probability distribution with Laplacian smoothing and to keep the number of instances within the limit of ten instances. CIELM was built with additive sigmoid nodes, one for each training instance. SVM was of type C-SVC [27] with the following parameters: $\gamma = 0,5$, $C = 1$, cache 200MB and $eps = 0.001$. 5NN was weighted by the complement of the distance.

Datasets are detailed in Table I. They were binarized and standardized or discretized when needed, i.e. for distance calculations or training some of the algorithms (NB, CIELM, SVM and 5NN). There were no missing values.

TABLE I. DATASET DETAILS: NUMBER OF INSTANCES IN THE POOL, NUMBER OF ATTRIBUTES, NUMBER OF NOMINAL ATTRIBUTES AND PERCENTAGE OF EXAMPLES FROM THE MAJORITARY CLASS.

Dataset	#Inst.	#Attrib.	#Nomin.	%Maj. class
1-autoUniv au1 1000	798	20	0	74
2-banana	4233	2	0	55
3-banknote authentic...	1078	4	0	55
4-bupa	273	6	0	58
5-climate simulation...	432	20	0	91
6-habermans survival	226	3	0	72
7-heart disease hung...	234	13	0	64
8-hill valley withou...	970	100	0	50
9-horse colic surgic...	240	27	14	64
10-indian liver patie...	456	10	1	71
11-ionosphere	280	33	0	64
12-kr vs kp	2557	36	36	52
13-mammographic mass	514	5	0	52
14-monks1	346	6	0	50
15-monks2	346	6	6	67
16-monks3	346	6	0	53
17-mushroom	6499	21	21	52
18-ozone eighthr	2021	72	0	94
19-ozone onehr	2022	72	0	97
20-phoneme	4316	5	0	71
21-pima indians diabe...	614	8	0	65
22-qsar biodegradatio...	842	41	0	66
23-ringnorm	5920	20	0	50
24-saheart	370	9	1	65
25-spambase	3366	57	0	60
26-spectf heart	214	44	0	79
27-statlog australian...	552	14	6	56
28-statlog german cre...	800	24	0	70
29-statlog heart	216	13	0	56
30-steel plates fault...	1553	33	0	65
31-thyroid hypothyroi...	2468	25	18	95
32-thyroid sick euthy...	2468	25	18	91
33-tic tac toe	766	9	9	65
34-twonorm	5920	20	0	50
35-vertebra column 2c	248	6	0	68
36-voting	223	16	16	67
37-wdbc	455	30	0	63
38-wholesale channel	352	7	0	68
39-wilt	3855	5	0	95

C. Experimental results

All learning curves have logarithmic shape, as can be seen in Figure 2. As these curves are averaged over all datasets,

[4]Weka version for C4.5 decision tree, called J48.
[5]*Naive Bayes*
[6]*Support Vector Machines with Radial Basis Function*

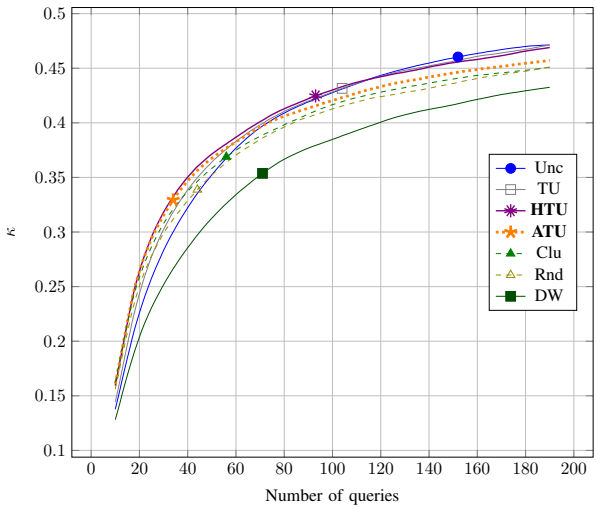

Fig. 2. Average learning curves for all datasets.

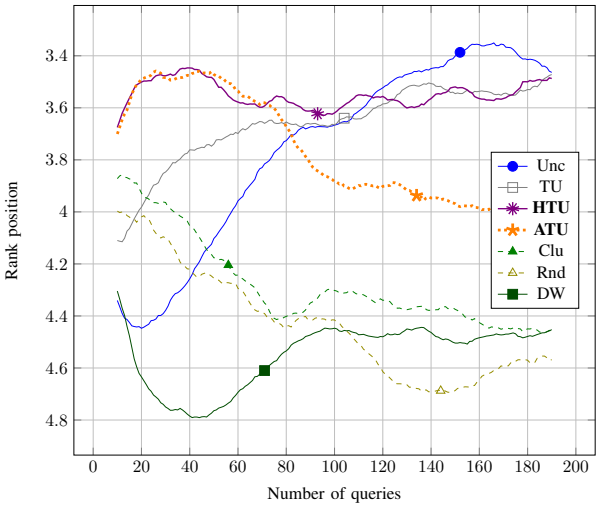

Fig. 3. Average ranking curves for all datasets.

strategies should not be compared by them. The effective range of κ is problem-dependent, causing averages to be biased towards easier datasets. Therefore, a more adequate curve would take into account the relative position among strategies. This can be done by ranking curves. At each time step, the ranking of strategies is averaged for all datasets. The result is the mean position per time step. The plot of all mean positions is the ranking curve presented in Figure 3 for each strategy. Until 80 queries, the two proposed methods are clearly superior to the others. As expected, agnosticism was advantageous in the beginning for both and, as the model evolved, the purely agnostic ATU and the purely gnostic Unc switched roles. This is confirmed by the fact that even the baseline Rnd and HS, both agnostic, were better than Unc early in the process of learning. Only after 100 queries, the most frequent limit employed in the literature, Unc prevails.

These findings were verified statistically by the Friedman-Nemenyi test [28] applied to the ALC-κ values for: all queries;

the first 50; the first 100; and, the last 100 queries. The resulting tables are respectively: II, III, IV and V. All pairs of strategies are compared by their ranking position in each dataset. Each symbol $s_{r,c}$ in a cell at row r and column c indicates that the strategy r is better than strategy c.

In the first 50 queries, the two proposed methods were better than Rnd, Unc and DW with $p = 0.01$ (Table II). The purely prospective Unc was worse than four of five strategies with some degree of exploration ($p \leq 0.10$). The advantage decreases with 100 queries (Table III), but the overall superiority of ATU and HTU remains. The same can be noted with 200 queries (Table IV), except by the overall equivalence between TU and ATU. Finally, with the last 100 queries (Table V), HTU, Unc and TU achieve the same overall performance, with little advantage to Unc, which was slightly better than ATU ($p = 0.10$) - according to the expected and already graphically noticed gnostic-agnostic switch.

TABLE II. ONE VERSUS ONE FOR THE FIRST 50 QUERIES. *Each symbol indicates a p-value: * (0.01), + (0.05) and . (0.10).*

	1	2	3	4	5	6	7
1 - Rnd	-						
2 - HS		-			.		*
3 - **ATU**	*		-		*		*
4 - **HTU**	*			-	*		*
5 - Unc					-		
6 - TU					.	-	*
7 - DW							-

TABLE III. ONE VERSUS ONE FOR THE FIRST 100 QUERIES. *Details in Table II.*

	1	2	3	4	5	6	7
1 - Rnd	-						
2 - HS		-					
3 - **ATU**	*	.	-		.		*
4 - **HTU**	*	+		-	+		*
5 - Unc					-		
6 - TU	+					-	*
7 - DW							-

TABLE IV. ONE VERSUS ONE FOR ALL 200 QUERIES. *Details in Table II.*

	1	2	3	4	5	6	7
1 - Rnd	-						
2 - HS		-					
3 - **ATU**	*	+	-				*
4 - **HTU**	*	*		-			*
5 - Unc	*				-		*
6 - TU	*	+				-	*
7 - DW							-

TABLE V. ONE VERSUS ONE FOR THE LAST 100 QUERIES. *Details in Table II.*

	1	2	3	4	5	6	7
1 - Rnd	-						
2 - HS		-					
3 - **ATU**	*	.	-				
4 - **HTU**	*	*		-			*
5 - Unc	*	*	.		-		*
6 - TU	*	*				-	*
7 - DW							-

V. CONCLUSION

The experiments have clearly demonstrated that agnosticism plays an important role in the beginning of the active learning. We demonstrate experimentally that a density-weighted method can be made agnostic and still achieve good performance within the usual budget adopted in the literature. Additionally, we proposed a proof of concept to demonstrate how to automatically explore such findings. It performed statistically better in the critical part part of the learning curve than the baseline and most relevant contenders. The inclusion of a pure gnostic step is intended as future work. A heuristic, or a meta-learning, approach to objectively define the correlation threshold value, experimentally fixed in 0.999, is also a relevant topic for more in-depth research. Another topic, suggested by a reviewer, is to turn the uncertainty measures into probabilistic measures to balance between exploration and exploitation. All implemented code is available at [29].

ACKNOWLEDGMENT

This research was supported by CAPES, CNPq and FAPESP. The authors would like to thank the reviewers for the valuable suggestions.

REFERENCES

[1] B. Settles, *Active Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2012.

[2] A. Beygelzimer, D. Hsu, J. Langford, and T. Z. 0001, "Agnostic active learning without constraints," *CoRR*, vol. abs/1006.2588, 2010.

[3] D. D. Lewis, "A sequential algorithm for training text classifiers: Corrigendum and additional data," *SIGIR Forum*, vol. 29, no. 2, pp. 13–19, 1995.

[4] D. P. Santos and A. C. P. L. F. Carvalho, "Comparison of active learning strategies and proposal of a multiclass hypothesis space search," in *HAIS*, ser. LNCS, vol. 8480. Springer, 2014, pp. 618–629.

[5] B. Settles, "Curious machines: active learning with structured instances," Ph.D. dissertation, University of Madison Wisconsin, 2008.

[6] A. Fujii, K. Inui, T. Tokunaga, and H. Tanaka, "Selective sampling for example-based word sense disambiguation," *Computational Linguistics*, vol. 24, no. 4, pp. 573–597, 1998.

[7] S. Dasgupta, "Two faces of active learning," *Theoretical Computer Science*, vol. 412, no. 19, pp. 1767–1781, 2011.

[8] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *Comput. J*, vol. 26, no. 4, pp. 354–359, 1983.

[9] T. M. Mitchell, "The need for biases in learning generalizations," in *Readings in Machine Learning*. Morgan Kauffman, 1980, pp. 184–191, book published in 1990.

[10] C. Zhang and K. Chaudhuri, "Beyond disagreement-based agnostic active learning," *CoRR*, vol. abs/1407.2657, 2014.

[11] S. Dasgupta, "Coarse sample complexity bounds for active learning," in *NIPS*, 2005.

[12] S. M. Ross, *Introduction to probability and statistics for engineers and scientists (2. ed.)*. Academic Press, 2000.

[13] J. D. Evans, *Straightforward statistics for the behavioral sciences*. Brooks/Cole, 1996.

[14] K. Bache and M. Lichman, "UCI repository of machine learning databases," University of California, Department of Information and CS, Irvine, CA, Tech. Rep., 2013.

[15] M. Shah, "Generalized agreement statistics over fixed group of experts," in *ECML/PKDD (3)*, ser. LNCS, vol. 6913. Springer, 2011, pp. 191–206.

[16] H. Raghavan, O. Madani, and R. Jones, "Active learning with feedback on features and instances," *JMLR*, vol. 7, pp. 1655–1686, 2006.

[17] Y. Guo and D. Schuurmans, "Discriminative batch mode active learning," in *NIPS*. Curran Associates, Inc, 2007.

[18] P. Melville and R. J. Mooney, "Diverse ensembles for active learning," in *Proceedings of the Twenty-first ICML*. New York, NY, USA: ACM, 2004, pp. 74–.

[19] I. Guyon, G. C. Cawley, G. Dror, and V. Lemaire, "Results of the active learning challenge," in *Active Learning and Experimental Design @ AISTATS*, vol. 16. JMLR.org, 2011, pp. 19–45.

[20] B. D. Eugenio and M. Glass, "The kappa statistic: A second look," *Computational Linguistics*, vol. 30, no. 1, pp. 95–101, 2004.

[21] P. E. Hart, "The condensed nearest neighbor rule (corresp.)," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 515–516, 1968.

[22] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[23] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *ECML*, ser. LNCS, vol. 1398. Springer, 1998, pp. 4–15.

[24] M. A. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998.

[25] G.-B. Huang and L. Chen, "Convex incremental extreme learning machine," *Neurocomputing*, vol. 70, no. 16-18, pp. 3056–3062, 2007.

[26] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "WEKA data mining software: an update," *SIGKDD Explorat.*, vol. 11, no. 1, pp. 10–18, 2009.

[27] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[28] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *JMLR*, vol. 7, pp. 1–30, 2006.

[29] D. P. Santos and A. C. P. L. F. Carvalho, "An Active Learning Library for Scala," *GitHub Software Repository*, Jan 2015, DOI:10.5281/zenodo.13733.