Reversing uncertainty sampling to improve active learning schemes

Cristian Cardellino, Milagro Teruel and Laura Alonso i Alemany

Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba
Argentina

Abstract. Active learning provides promising methods to optimize the cost of manually annotating a dataset. However, practitioners in many areas do not massively resort to such methods because they present technical difficulties and do not provide a guarantee of good performance, especially in skewed distributions with scarcely populated minority classes and an undefined, catch-all majority class, which are very common in human-related phenomena like natural language.
In this paper we present a comparison of the simplest active learning technique, pool-based uncertainty sampling, and its opposite, which we call reversed uncertainty sampling. We show that both obtain results comparable to the random, arguing for a more insightful approach to active learning.

1 Introduction and Motivation

Active learning has been a promising area of machine learning since the mid-nineties [3]. However, unlike other areas of machine learning, active learning has not been so widely adopted in applied areas. A reason for that is put forward by [1]: practicioners do not have any solid ground to take practical decisions as to which configurations are most adequate for their purposes.

We propose to compare the performances of the most used method in active learning, uncertainty sampling [9], and its exact oposite, what we call *reversed uncertainty* sampling. Uncertainty sampling is appealing to the nonexpert because it is a wrapper method that can be applied to any base learner, and even very easily, if we measure uncertainty as the uncertainty provided by the learner itself, without further calculations.

The rest of the paper is organized as follows. In the next Section we review relevant work on the methods that we are putting into practice in our approach. Then, we detail our approach in Section 3. In Section 4 describe the experimental settings and analyze results in Section 5.

2 Relevant Work

In the active learning challenge organized on 2011 by [7], the most used active learning method was uncertainty sampling [9,12,16]. Uncertainty sampling consists in choosing, from a large set of unlabeled instances, those where a classifier

is most uncertain. It performs well in general, and it is very simple to implement, being a wrapper method over any base learner. Uncertainty can be calculated by sophisticated measures, but the practitioner can also resort to the measures of uncertainty provided by an off-the-shelf learner.

It is widely known in the active learning community that discriminative methods (like uncertainty sampling or query by committee [6]) perform well only if a big labeled dataset is available for training, while density estimation methods (like [14]) have a good performance with very few labeled instances [4,17]. This is due to the fact that density estimation is specially good at establishing an initial decision boundary that adequately separates most of the population, even if it is unlabeled. This is so because these methods sample from regions with maximal density and few or no labels at all. In contrast, discriminative methods sample from the regions where the classifier shows least certainty, and as a result provide more certainty in those regions specifically. This helps to refine the decision boundary, but does not take into account the distribution of the unlabeled data.

Despite the benefits of density-based approaches, discriminative methods tend to be easier to implement and more intuitive. That is why we have approached the density-based intuition by a simple approach: reversing simple discriminative methods. By reversing a discriminative method we mean selecting instances where the model has most certainty or lowest entropy, instead of most uncertainty or highest entropy. These can be assumed to be closest to the generative center that the current model assumes correct. Then, having them labeled by an oracle should be a good approximation to testing the current generative centers. If the oracle confirms what the model believes, then the current generative center is consolidated by adding one more instance to the training set. If the oracle rejects what the model believes, the generative center is relocated.

3 Reversing discriminative rankings

The principle behind discriminative approaches to active learning is to choose those instances or features that accumulate the most uncertainty, assuming that labelling those will help decision making information in a region of the instance space that is currently unexplored. This principle implies that providing labels for regions where there are already labels is the same as providing redundant information.

However, when there are very few labeled instances in the training set, the principle behind discriminative approaches does not succeed in improving performance. This is so because discriminative methods are well-suited to *refine* decision boundaries, not to *set* them. To place decision boundaries in the instance space, methods that try to characterize the space, like density estimation methods, are better suited. Instead of refining the decision frontier, these methods aim to locate the generative centers of the data.

We propose to use discriminative methods in a reverse manner; that is, to help locate the generative centers of the data. As described in Section 2, it is well known that discriminative methods do not have the ability to characterize

the instance space and locate generative centers based on such characterization. However, we can use them to check whether our current assumptions about the generative centers are right. The instances where the model has most certainty can be assumed to be closest to the generative center that the current model assumes correct. Then, having them labeled by an oracle is the same as testing the current generative centers. If the oracle confirms what the model believes, then the current generative center is consolidated by adding one more instance to the training set. If the oracle rejects what the model believes, the generative center is relocated, as in a search procedure.

We apply the most popular discriminative method, uncertainty sampling, to try to set the initial decision boundary more adequately. We reverse the ranking criterion of uncertainty sampling: instead of selecting those instances where the classifier is most uncertain, we select those where the classifier is *most certain*. The intuition behind our approach is also at the basis of bootstrapping as semi-supervised learning, but to our knowledge it has not been systematically studied in an active learning context.

4 Experimental Setting

As a base learner we used a Multinomial Naïve Bayes (MNB) classifier, which is commonly employed in active learning because parametrizations and adaptations can be introduced in a principled way without much complication. We exploit the MNB classifier provided by Weka [13] to implement uncertainty sampling to choose instances.

4.1 Data

We used the 20 Newsgroups dataset, a standard dataset for evaluation of active learning [7]. The 20 Newsgroups dataset is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. As features, we use the standard bag-of-words model for the representation of a document, together with bag-of-bigrams and -trigrams.

Since the dataset, represented as a matrix of features occurrences, was too large to deal with in our current hardware setting, we decided to make experiments with smaller samples of 1000 randomly selected documents of the 20000. Features that occurred less than 10 times in the sampled dataset were discarded.

4.2 Experiments

To evaluate the different active learning approaches we simulated the *learning – automatic labelling – manual labelling – retraining* loop as follows. For each instance sampling strategy (uncertainty sampling, reverse uncertainty sampling or random sampling) we run experiments with 10 samples with 1000 randomly selected instances each. Each sample is divided in 800 instances for training and 200 for testing. From the training instances, we obtain 10 random instances and

we train the initial model with MNB. The remaining training dataset is used as our "untagged" dataset, where we apply either uncertainty sampling, reverse uncertainty sampling or random sampling. The 10 instances ranking highest are selected and assigned the class provided by the labeled dataset. Then these newly labeled instances are added to the training dataset and we proceed to infer the new MNB model. This loop continues until all instances are labeled.

To assess the performance of each strategy with skewed distributions, where we had a majority class and some minority classes, for each random sample, we found the 3 classes that had obtained most instances and established them as minority classes, then collapsed the remaining 17 as a single catchall majority class.

In each iteration of the loop, we evaluate the obtained model in the test dataset. As metrics we use accuracy and True Positive Rate (TPR), also known as *sensitivity*, that is, the weighted mean of the rate of true positives identified in each class. This metric is very close to the standard *accuracy* metric [10], but it allows us to obtain a per-class analysis, which is useful to assess performance in the minority classes. Therefore, this is the metric we use for skewed distributions.

5 Analysis of Results

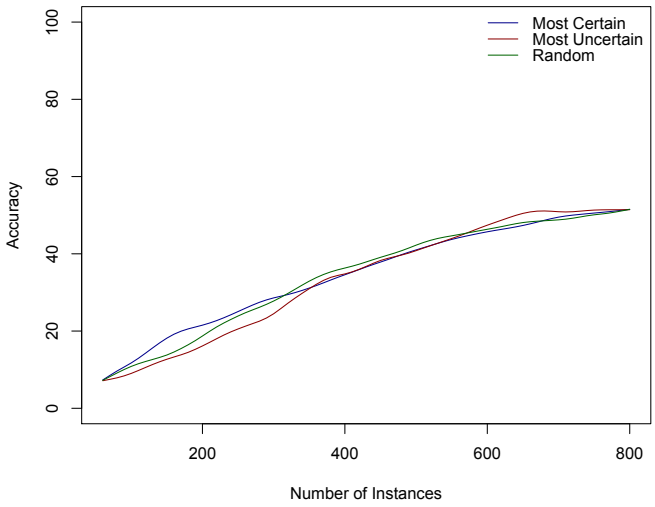

Fig. 1. Mean accuracy across 10 samples of the 20 Newsgroups dataset.

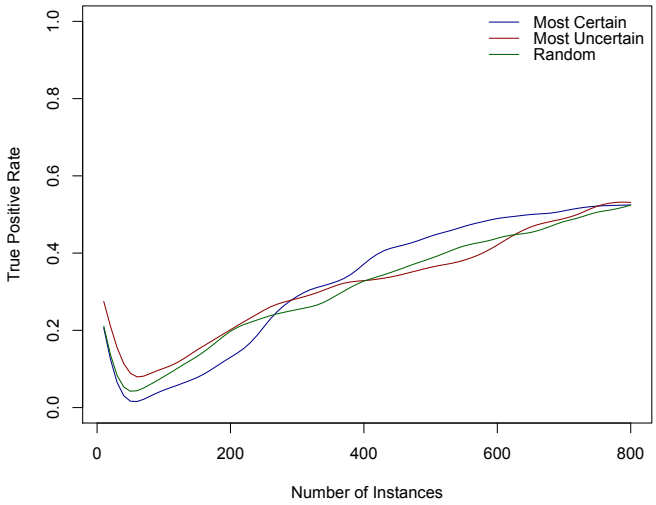

Fig. 2. Mean TPR on 10 experiments on the 20 Newsgroup dataset with 3 minority classes and the remaining 17 classes collapsed into a single catchall majority class.

In Figure 1 we can see the mean accuracy obtained accross 10 1000-instance samples of the balanced dataset. Accuracy figures are comparable to those obtained by other active learning approaches to this dataset, like [16]. We can see a slight tendency for the reversed uncertainty sampling approach to obtain better performances when the training data are scarce, and the three methods to perform indistinguishably as the amount of labeled instances increases. Uncertainty sampling tends to perform worse than the other two methods when there are few examples.

Conversely, in the scenario with 3 minority classes and a catchall majority class, seen in Figure 2, we can see that reversed uncertainty sampling performs worse than the other two strategies when very few instances are available for training, and performs best when the training dataset is bigger, as opposed to what we have seen in the balanced dataset.

To assess the stability of these results, we applied Student's t-test to test against the null hypothesis that results were indistinguishable. As displayed in Table 1, p-values obtained for the test are very high, thus the results cannot be considered different. Thus differences are not significant if the whole range of iterations is taken into account, but significance begins to emerge if we take into account only some parts of the training process. We can see that the most certain and most uncertain strategies are well differentiated both when the dataset has very few or many instances, both in balanced and in skewed distributions.

We can also see that the most certain strategy is more differentiated from random, in some cases by performing worse and in some others by performing better. These differences call for a closer study in different datasets, and describing the contribution of different feature configurations to the behaviour of different instance selection strategies.

	balanced dataset						
	whole	1.half	2.half	1.quarter	2.quarter	3.quarter	4.quarter
certain vs. uncertain	.68	.15	.37	.01	.27	.62	<.001
certain vs. random	.92	.67	.65	.18	.65	.31	.79
uncertain vs. random	.75	.35	.60	.23	.17	.68	<.001
	skewed dataset						
	whole	1.half	2.half	1.quarter	2.quarter	3.quarter	4.quarter
certain vs. uncertain	.86	.10	<.001	<.001	.58	<.001	.03
certain vs. random	.59	.56	<.001	.08	.55	<.001	<.001
uncertain vs. random	.67	.22	.68	.15	.08	.03	.29

Table 1. Significance levels (p-values obtained by applying Student's t-test) to assess the difference between results obtained by different instance selection strategies.

6 Conclusions and Future Work

We have presented a straightforward approach to integrate some of the benefits of insightful, density estimation active learning methods into an active learning approach for non-experts. We have shown that simply reversing the ranking criterion of uncertainty sampling produces results that are not clearly differentiable in terms of performance, but show some tendency comparable to density estimation methods in the scenario where very few labeled instances are available for training, while classical uncertainty sampling performs better if there are more labeled instances. We have shown this is the case for a standard text classification problem, with standard features. This characterization should be useful to reduce the costs of development of labeled datasets from scratch. In addition to characterizing the performance of instance selection methods in a standard dataset, with balanced classes, we have also assessed the performance when we have imbalanced classes. Although results have been tested in a single dataset, we expect that they are somewhat extrapolable to other datasets.

It must be noted that the results obtained so far are preliminary, and require more experimentation on bigger and more diverse datasets, and actual comparison with density estimation methods in the same experimental conditions.

Uncertainty sampling has the benefit of being a wrapper method over virtually any base learner, which facilitates the implementation of such approach in an established machine learning workflow. A very useful addition to improve the usability of this approach would be an accessible way to determine which

selection strategy is more useful at which point in the learning process. In the literature this has been done by complex methods of error estimation [4], which are contradictory with the simple spirit of the approach presented here. We will be working on simpler methods to approach this question, most notably, the p-value obtained by applying hypothesis testing to the difference between results obtained by different strategies.

Once we have established that these different strategies produce significantly different results, we need to study the contribution of different feature configurations to the behaviour of different instance selection strategies.

References

1. Attenberg, J., Provost, F.: Inactive learning?: Difficulties employing active learning in practice. SIGKDD Explor. Newsl. 12(2), 36–41 (Mar 2011), http://doi.acm.org/10.1145/1964897.1964906
2. Bloodgood, M., Vijay-Shanker, K.: Taking into account the differences between actively and passively acquired data: The case of active learning with support vector machines for imbalanced datasets. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. pp. 137–140. Association for Computational Linguistics (2009)
3. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Mach. Learn. 15(2), 201–221 (May 1994), http://dx.doi.org/10.1023/A:1022673506211
4. Donmez, P., Carbonell, J.G., Bennett, P.N.: Dual strategy active learning. In: Kok, J.N., Koronacki, J., de Mántaras, R.L., Matwin, S., Mladenic, D., Skowron, A. (eds.) ECML. Lecture Notes in Computer Science, vol. 4701, pp. 116–127. Springer (2007), http://dblp.uni-trier.de/db/conf/ecml/ecml2007.html#DonmezCB07
5. Ertekin, S.E., Huang, J., Bottou, L., Giles, C.L.: Learning on the border: Active learning in imbalanced data classification. In: In Proc. ACM Conf. on Information and Knowledge Management (CIKM '07 (2007)
6. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Mach. Learn. 28(2-3), 133–168 (Sep 1997), http://dx.doi.org/10.1023/A:1007330508534
7. Guyon, I., Cawley, G.C., Dror, G., Lemaire, V.: Results of the active learning challenge. In: Guyon, I., Cawley, G.C., Dror, G., Lemaire, V., Statnikov, A.R. (eds.) Active Learning and Experimental Design @ AISTATS. JMLR Proceedings, vol. 16, pp. 19–45. JMLR.org (2011), http://dblp.uni-trier.de/db/journals/jmlr/jmlrp16.html#GuyonCDL11
8. He, J., Carbonell, J.G.: Nearest-neighbor-based active learning for rare category detection. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S.T. (eds.) NIPS. Curran Associates, Inc. (2007), http://dblp.uni-trier.de/db/conf/nips/nips2007.html#HeC07
9. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3–12. SIGIR '94, Springer-Verlag New York, Inc., New York, NY, USA (1994), http://dl.acm.org/citation.cfm?id=188490.188495

10. Powers, D.M.W.: Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. Journal of Machine Learning Technologies 2(1), 37:63 (2011), school of Informatics and Engineering, Flinders University, Adelaide, Australia, TR SIE-07-001
11. Tomanek, K., Hahn, U.: Reducing class imbalance during active learning for named entity annotation. In: Proceedings of the Fifth International Conference on Knowledge Capture. pp. 105–112. K-CAP '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1597735.1597754
12. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 2, 45–66 (Mar 2002), http://dx.doi.org/10.1162/153244302760185243
13. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)
14. Xu, Z., Yu, K., Tresp, V., Xu, X., Wang, J.: Representative sampling for text classification using support vector machines. In: Proceedings of the 25th European Conference on Information Retrieval Research (ECIR 2003). LNCS, vol. 2633, pp. 393–407. Springer, Pisa, Italy (April 2003)
15. Zhu, J., Hovy, E.H.: Active learning for word sense disambiguation with methods for addressing the class imbalance problem. EMNLP-CoNLL 7, 783–790 (2007)
16. Zhu, J., Ma, M.: Uncertainty-based active learning with instability estimation for text classification. ACM Trans. Speech Lang. Process. 8(4), 5:1–5:21 (Feb 2012), http://doi.acm.org/10.1145/2093153.2093154
17. Zhu, J., Wang, H., Yao, T., Tsou, B.K.: Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. pp. 1137–1144. Association for Computational Linguistics (2008)
18. Zipf, G.: Human Behavior and the Principle of Least Effort. Addison–Wesley, Cambridge, MA (1949)