D-Confidence: an active learning strategy which efficiently identifies small classes

Nuno Escudeiro
Instituto Superior de Engenharia do Porto
Rua Dr António Bernardino de Almeida, 431
Porto, P-4200-072 PORTO, Portugal
LIAAD-INESC PORTO L.A.
Rua de Ceuta, 118 - 6
Porto, P-4050-190 PORTO, Portugal
nfe@isep.ipp.pt

Alípio Jorge
LIAAD-INESC PORTO L.A.
Rua de Ceuta, 118 - 6
Porto, P-4050-190 PORTO, Portugal
amjorge@fc.up.pt

Abstract

In some classification tasks, such as those related to the automatic building and maintenance of text corpora, it is expensive to obtain labeled examples to train a classifier. In such circumstances it is common to have massive corpora where a few examples are labeled (typically a minority) while others are not. Semi-supervised learning techniques try to leverage the intrinsic information in unlabeled examples to improve classification models. However, these techniques assume that the labeled examples cover all the classes to learn which might not stand. In the presence of an imbalanced class distribution getting labeled examples from minority classes might be very costly if queries are randomly selected. Active learning allows asking an oracle to label new examples, that are criteriously selected, and does not assume a previous knowledge of all classes. D-Confidence is an active learning approach that is effective when in presence of imbalanced training sets. In this paper we discuss the performance of d-Confidence over text corpora. We show empirically that d-Confidence reduces the number of queries required to identify examples from all classes to learn when compared to confidence, a common active learning criterion.

1 Introduction

Classification tasks require a number of previously labeled cases. A major bottleneck is that case labeling is a laborious task requiring significant human effort. This effort is particularly high in the case of text documents, web pages and other unstructured objects.

The effort required to retrieve representative labeled examples to learn a classification model is not only related to the number of distinct classes (Adami et al., 2005); it is also related to class distribution in the available pool of examples. On a highly imbalanced class distribution, it is particularly demanding to identify examples from minority classes. These, however, may be important in terms of representativeness. Failing to identify cases from underrepresented classes may have costs. Minority classes may correspond to specific information needs which are relevant for specific subgroups of users. In many situations, such as fraud detection, clinical diagnosis, news (Ribeiro and Escudeiro, 2008) and Web resources (Escudeiro and Jorge, 2006), we face the problem of imbalanced class distributions.

The aim of our current work is to get a classification model that is able to fully recognize the target concept, including all the classes to learn no mater how frequent or rare they are.

Our main goal is to identify representative examples for each class in the absence of previous descriptions of some or all the classes. Furthermore, this must be achieved with a reduced number of labeled examples in order to reduce the labeling effort.

There are several learning schemes available for classification. The supervised setting allows users to specify arbitrary concepts. However, it requires a fully labeled training set, which is prohibitive when the labeling cost is high and, besides that, it requires labeled cases from all classes. Semi-supervised learning allows users to state specific needs without

Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing, pages 18–26,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

requiring extensive labeling (Chapelle et al, 2006) but still requires that labeled examples fully cover the target concept. Unsupervised learning does not require any labeling but users have no chance to tailor clusters to their specific needs and there is no guarantee that the induced clusters are aligned with the classes to learn. In active learning, that seems more adequate to our goals, the learner is allowed to ask an oracle (typically a human) to label examples – these requests are called *queries*. The most informative queries are selected by the learning algorithm instead of being randomly selected as is the case in supervised learning.

In this paper we evaluate the performance of *d-Confidence* (Escudeiro and Jorge, 2009) on text corpora. D-Confidence is an active learning approach that tends to explore unseen regions in case space, thus selecting cases from unseen classes faster – with fewer queries – than traditional active learning approaches. D-Confidence selects queries based on a criterion that aggregates the posterior classifier confidence – a traditional active learning criterion – and the distance between queries and known classes. This criterion is biased towards cases that do not belong to known classes (low confidence) and that are located in unseen areas in case space (high distance to known classes). D-confidence is more effective than confidence alone in achieving an homogeneous coverage of target classes.

In the rest of this paper we start by reviewing active learning, in section 2. Section 3 describes d-Confidence. The evaluation process is presented in section 4 and we state our conclusions and expectations for future work in section 5.

2 Active Learning

Active learning approaches (Angluin, 1988; Cohn et al., 1994; Muslea et al., 2006) reduce label complexity – the number of queries that are necessary and sufficient to learn a concept – by analyzing unlabeled cases and selecting the most useful ones once labeled. Queries may be artificially generated (Baum, 1991) – the *query construction* paradigm – or selected from a pool (Cohn et al., 1990) or a stream of data – the *query filtering* paradigm. Our current work is developed under the query filtering approach.

The general idea in active learning is to estimate the value of labeling one unlabeled case. Query-By-Committee (Seung et al., 1992), for example, uses a set of classifiers – the committee – to identify the case with the highest disagreement. Schohn et al. (2000) worked on active learning for Support Vector Machines (SVM) selecting queries – cases to be labeled – by their proximity to the dividing hyperplane. Their results are, in some cases, better than if all available data is used to train. Cohn et al. (1996) describe an optimal solution for pool-based active learning that selects the case that, once labeled and added to the training set, produces the minimum expected error. This approach, however, requires high computational effort. Previous active learning approaches (providing non-optimal solutions) aim at reducing uncertainty by selecting the next query as the unlabeled example on which the classifier is less confident.

Batch mode active learning – selecting a batch of queries instead of a single one before retraining – is useful when computational time for training is critical. Brinker (2003) proposes a selection strategy, tailored for SVM, that combines closeness to the dividing hyperplane – assuring a reduction in the version space close to one half – with diversity among selected cases – assuring that newly added examples provide additional reduction of version space. Hoi et al. (2006) suggest a new batch mode active learning relying on the Fisher information matrix to ensure small redundancy among selected cases. Li et al. (2006) compute diversity within selected cases from their conditional error.

Dasgupta (2005) defines theoretical bounds showing that active learning has exponentially smaller label complexity than supervised learning under some particular and restrictive constraints. This work is extended in Kaariainen (2006) by relaxing some of these constraints. An important conclusion of this work is that the gains of active learning are much more evident in the initial phase of the learning process, after which these gains degrade and the speed of learning drops to that of passive learning. Agnostic Active learning (Balcan et al., 2006), A^2, achieves an exponential improvement over the usual sample complexity of supervised learning in the presence of arbitrary forms of noise. This model is studied by Hanneke (2007) setting general bounds on label complexity.

All these approaches assume that we have an initial labeled set covering all the classes of interest.

Clustering has also been explored to provide an initial structure to data or to suggest valuable queries. Adami et al. (2005) merge clustering and oracle labeling to bootstrap a predefined hierarchy of classes. Although the original clusters provide some structure to the input, this approach still demands for a high validation effort, especially when these clusters are not aligned with class labels. Dasgupta et al. (2008) propose a cluster-based method that consistently improves label complexity over supervised learning. Their method detects and exploits clusters that are loosely aligned with class labels.

Among other paradigms, it is common that active learning methods select the queries which are closest to the decision boundary of the current classifier. These methods focus on improving the decision functions for the classes that are already known, i.e., those having labeled cases present in the training set. The work presented in this paper diverges classifier attention to other regions increasing the chances of finding new labels.

3 D-Confidence Active Learning

Given a target concept with an arbitrary number of classes together with a sample of unlabeled examples from the target space (the working set), our purpose is to identify representative cases covering all classes while posing as few queries as possible, where a query consists of requesting a label to a specific case. The working set is assumed to be representative of the class space – the representativeness assumption (Liu and Motoda, 2001).

Active learners commonly search for queries in the neighborhood of the decision boundary, where class uncertainty is higher. Limiting case selection to the uncertainty region seems adequate when we have at least one labeled case from each class. This class representativeness is assumed by all active learning methods. In such a scenario, selecting queries from the uncertainty region is very effective in reducing version space.

Nevertheless, our focus is on text corpora where only few labeled examples exist and when we are still looking for exemplary cases to qualify the concept to learn. Under these circumstances – while we do not have labeled cases covering all classes – the uncertainty region, as perceived by the active learner, is just a subset of the real uncertainty region. Being limited to this partial view of the concept, the learner is more likely to waste queries. The amount of the uncertainty region that the learner misses is related to the number of classes to learn that have not yet been identified as well as to the class distribution in the training set.

The intuition behinf d-Confidence is that query selection should be based not only on classifier confidence but also on distance to previously labeled cases. In the presence of two cases with equally low confidence d-Confidence selects the one that is farther apart from what is already know, i.e., from previously labeled cases.

3.1 D-Confidence

Common active learning approaches rely on classifier confidence to select queries (Angluin, 1988) and assume that the pre-labeled set covers all the labels to learn – this assumption does not hold in our scenario. These approaches use the current classification model at each iteration to compute the posterior confidence on each known class for each unlabeled case. Then, they select, as the next query, the unlabeled case with the lowest confidence.

D-Confidence, weighs the confidence of the classifier with the inverse of the distance between the case at hand and previously known classes.

This bias is expected to favor a faster coverage of case space, exhibiting a tendency to explore unknown areas. As a consequence, it provides faster convergence than confidence alone. This drift towards unexplored regions and unknown classes is achieved by selecting the case with the lowest d-Confidence as the next query. Lowest d-Confidence is achieved by combining low confidence – probably indicating cases from unknown classes – with high distance to known classes – pointing to unseen regions in the case space. This effect produces significant differences in the behavior of the learning process. Common active learners focus on the uncertainty region asking queries that are expected to narrow it down. The issue is that the uncertainty region is determined by the labels we known at a given iteration. Focusing our search for queries exclusively

Table 1: d-Confidence algorithm.

```
(1)  given $W$; $L_1$ and $K$
(2)  compute distance among cases in $W$
(3)  $i = 1$
(4)  while (not stopping criteria) {
(5)    $U_i = W - L_i$
(6)    learn $h_i$ from $L_i$
(7)    apply $hi$ to $Ui$ generating $conf_i(u_j, c_k)$
(8)    for($u_j in U_i$){
(9)      $dist_i(u_j, c_k) = aggrIndivDistk(u_i, c_k)$
(10)     $dconf_i(u_j, c_k) = \frac{conf_i(u_j,c_k)}{dist_i(u_j,c_k)}$
(11)     $dC_i(u_j) = agConf_k(dconf_i(u_j, c_k))$
(12)   }
(13)   $q_i = u_j : dC_i(u_j) = min_u(dC_i(u))$
(14)   $L_{i+1} = L_i \cup < q_i, label(q_i) >$
(15)   $i + +$
(16) }
```

on this region, while we are still looking for exemplary cases on some labels that are not yet known, is not effective. Unknown classes hardly come by unless they are represented in the current uncertainty region.

In Table 1 we present the d-Confidence algorithm – an active learning proposal specially tailored to achieve a class representative coverage fast.

W is the working set, a representative sample of cases from the problem space. L_i is a subset of W. Members of L_i are the cases in W whose labels are known at iteration i. U, a subset of W, is the set of unlabeled examples. At iteration i, U_i is the (set) difference between W and L_i; K is the number of target concept classes, c_k; h_i represents the classifier learned at iteration i; q_i is the query at iteration i; C_i is the set of classes known at iteration i – that is the set of distinct classes from all L_i elements; $conf_i(u_j, c_k)$ is the posterior confidence on class c_k given case u_j, at iteration i.

D-Confidence for unlabeled cases is computed at steps (8) to (12) in Table 1 as explained below. In (13) the case with the minimum d-Confidence is selected as the next query. This query is added to the labeled set (14), and removed from the unlabeled pool, and the whole process iterates.

Computing d-Confidence d-Confidence is obtained as the ratio between confidence and distance among cases and known classes (Equation 1).

$$\arg\max_k \left(\frac{conf\ (c_k|u)}{median_j\ (dist\ (u, Xlab_{j,k}))} \right) \qquad (1)$$

For a given unlabeled case, u, the classifier generates the posterior confidence w.r.t. known classes (7). Confidence is then divided by an indicator of the distance, $dist()$, between unlabeled case u and all labeled cases belonging to class c_k, $Xlab_{j,k}$ (9). This distance indicator is the $median$ of the distances between case u and all cases in $Xlab_{j,k}$. The median is expected to soften the effect of outliers. At step (10) we compute $dconf_i(u, c_k)$ – the d-Confidence for each known class, c_k, given the case u – by dividing class confidence for a given case by aggregated distance to that class.

Finally, d-Confidence of the case is computed, $dC_i(u)$, as the maximum d-Confidence on individual classes, $agConf_k(conf_i(u, c_k))$, at step (11).

4 Evaluation

D-Confidence was evaluated on two text corpora. We have selected a stratified sample from the 20 Newsgroups (NG) – with 500 documents – and another one from the R52 set of the Reuters-21578 collection (R52) – with 1000 documents. The NG dataset has documents from 20 distinct classes while the R52 dataset has documents from 52 distinct classes. These samples have been selected because they have distinct class distributions.

The class distribution of NG is fairly balanced (Figure 1) with a maximum frequency of 35 and a minimum frequency of 20.

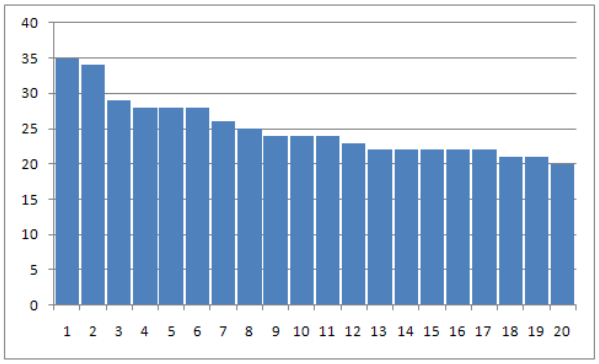

Figure 1: Class distribution in NG dataset

On the other hand, the R52 dataset presents an highly imbalanced class distribution (Figure 2).

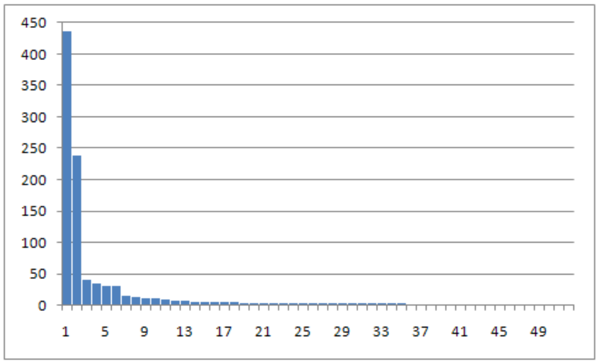

Figure 2: Class distribution in R52 dataset

The most frequent class in R52 has a frequency of 435 while the least frequent has only 2 examples in the dataset. This dataset has 42 classes, out of 52, with a fequency below 10.

4.1 Experimental Setting

We have used Support Vector Machine classifiers (SVM) with linear kernels in all experiments.

In all the experiments we have compared the performance of d-Confidence against confidence – a common active learning setting where query selection is based on low posterior confidence of the current classifier. This comparison is important to evaluate our proposal since d-Confidence is derived from confidence by means of an aggregation with distance in case space. Comparing both these criteria, one against the other, will provide evidence on the performance gains, or losses, of d-Confidence on text when compared to confidence, its baseline.

We have performed 10-fold cross validation on all datasets for standard confidence and d-Confidence active learning. The labels in the training set are hidden from the classifier. In each iteration, the active learning algorithm asks for the label of a single case. For the initial iteration in each fold we give two labeled cases – from two distinct classes – to the classifier. The two initial classes are chosen for each fold, so that different class combinations occur in different folds. Given an initial class to be present in L_1, the specific cases to include in L_1 are randomly sampled from the set of cases on that class. Given the fold, the same L_1 is used for all experiments.

4.2 Results

Our experiments assess the ability of d-Confidence to reduce the labeling effort when compared to confidence.

We have recorded, for each dataset, the number of distinct labels already identified and the progress of the error on the test set for each iteration (generalization error). From these, we have computed, for each dataset, the mean number of known classes and mean generalization error in each iteration over all the cross validation folds (Figures 3 and 4).

The chart legends use *c* for confidence, *dc* for d-Confidence, *e* for generalization error and *kc* for the number of known classes. For convenience of representation the number of classes that are known at each iteration has been normalized to the total number of classes in the dataset thus being transformed into the percentage of known classes instead of the absolute number of known classes. This way the number of known classes and generalization error are both bounded in the same range (between 0 and 1) and we can conveniently represented them in the same chart.

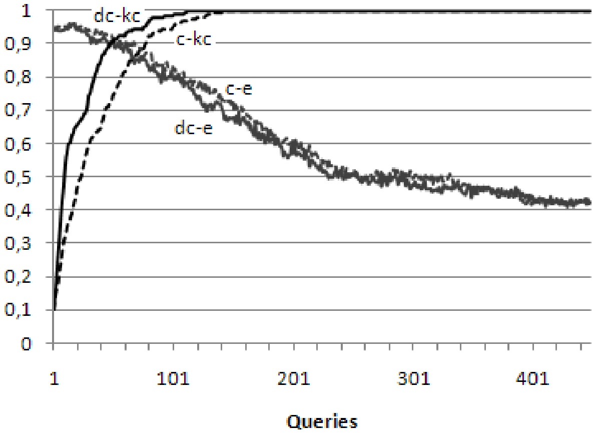

Figure 3: Known classes and error in NG dataset

Means are micro-averages – all the cases are equally weighted – over all iterations for a given dataset and a given selection criterion (confidence or d-Confidence). Besides the overall number of queries required to retrieve labels from all classes and generalization error, we have also observed the mean number of queries that are required to retrieve the first case for each class (Tables 2 to 4) – referred to as *first hit*.

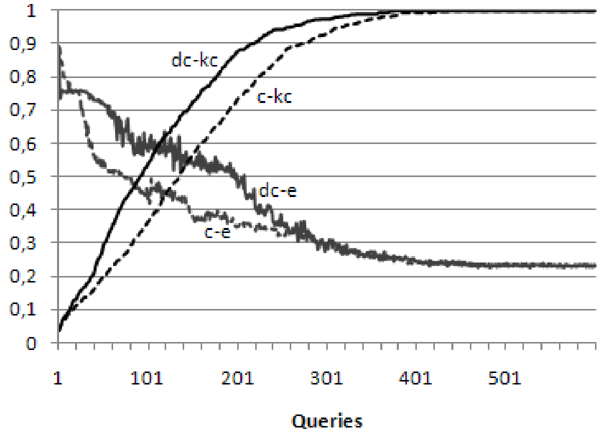

Figure 4: Known classes and error in R52 dataset

We have performed significance tests, t-tests, for the differences of the means observed when using confidence and d-Confidence. Statistically different means, at a significance level of 5%, are bold faced.

When computing first hit for a given class we have omitted the experiments where the labeled set for the first iteration contains cases from that class. Figures 5 and 6 give an overview of the number of queries that are required in each setting to first hit a given number of distinct classes.

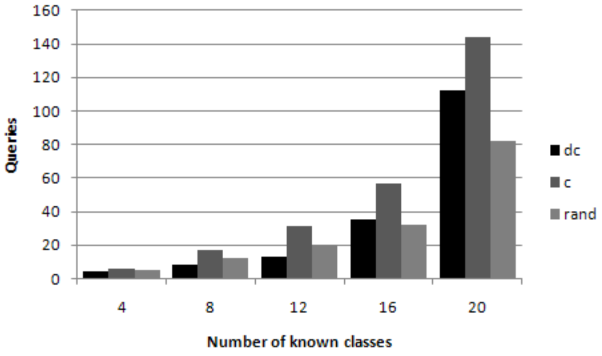

Figure 5: Queries required to identify bunches of distinct classes in NG dataset

A benchmark based on random selection is also provided – averaged over 10 random samples. We have recorded the number of queries required to identify bunches of distinct classes in multiples of 10 for R52 and multiples of 4 in NG.

Table 2: Class distribution (freq) and first hit (c-fh and dc-fh) for the NG dataset.

Class	Freq	c-fh	dc-fh
1	29	36.9	35.7
2	22	41.9	41.1
3	21	**57.3**	76.9
4	34	23.5	**5.9**
5	35	18.9	20.2
6	24	37.1	**15.4**
7	21	53.6	**11.3**
8	24	32.9	**13.1**
9	25	36.3	**9.1**
10	22	**41.1**	48.9
11	22	42.5	**3.5**
12	24	28.6	**4.3**
13	28	18.8	20.4
14	28	25.8	**5.4**
15	22	27.4	**6.2**
16	28	14.9	**2.6**
17	23	**21.4**	27.9
18	26	34.5	**7.7**
19	22	22.2	21.2
20	20	26.7	**6.9**
mean		32.1	19.2

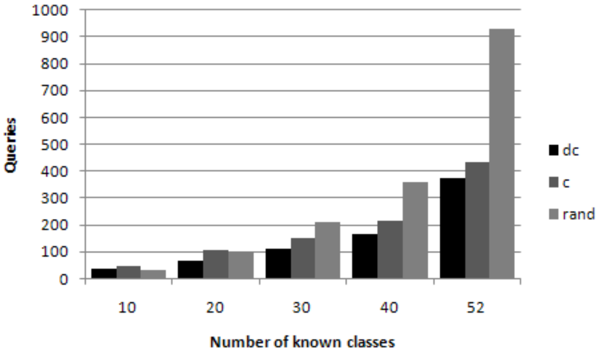

Figure 6: Queries required to identify bunches of distinct classes in R52 dataset

Table 3: Class distribution (Freq) and first hit (c-fh and dc-fh) for the R52 dataset. Only for those classes where d-Confidence outperforms confidence with statistical significance at 5% significance level.

Class	Freq	c-fh	dc-fh
1	239	10.1	**1.6**
2	5	7.2	**1.3**
8	3	103.8	**76.6**
9	7	68.6	**6.6**
10	2	80.0	**10.0**
11	40	83.4	**41.7**
14	2	173.7	**110.6**
15	3	115.6	**64.7**
16	7	96.7	**16.8**
18	5	68.7	**62.9**
22	2	244.4	**197.6**
23	30	153.4	**36.7**
25	4	173.3	**102.9**
26	2	214.1	**123.9**
27	5	206.7	**184.9**
28	2	213.3	**85.2**
29	2	137.6	**44.8**
30	3	159.3	**52.1**
31	2	159.1	**144.8**
32	2	179.7	**123.9**
33	30	160.8	**76.1**
34	15	175.6	**108.7**
36	2	167.4	**107.8**
37	3	118.0	**99.5**
40	2	140.0	**104.7**
43	4	313.1	**256.4**
44	14	216.3	**144.5**
46	12	206	**126.7**
47	2	233.7	**167**
48	3	153.2	**84.1**
49	35	226	**106.9**
50	3	144.3	**75.5**
51	3	148.5	**51.1**
52	2	258.8	**196.5**
mean		156.2	94.0

Table 4: Class distribution (Freq) and first hit (c-fh and dc-fh) for the R52 dataset. Only for those classes where d-Confidence does not outperforms confidence.

Class	Freq	c-fh	dc-fh
3	3	**11.2**	18.0
4	2	**36.4**	72.9
5	6	**23.1**	50.7
6	11	**39.7**	49.7
7	4	**40.1**	89.1
12	2	**128.8**	136.0
13	435	**91.9**	107.8
17	9	**117.0**	135.6
19	2	**123.6**	19.1
20	3	171.7	171.1
21	2	**196.2**	224.0
24	4	**118.6**	178.7
35	4	**146.1**	183.5
38	3	**158.5**	166.4
39	2	152.2	150.4
41	5	**143.6**	154.5
42	3	**188.9**	202.8
45	3	**175.5**	198.7
mean		114.6	128.3

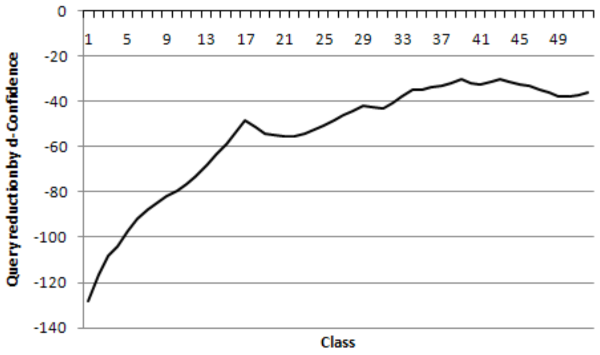

Figure 7: Average gain of d-Confidence to confidence. Classes are sorted by increasing order of their frequency.

4.3 Discussion

The charts in Figures 3 and 4 confirm the results that have been previously reported for standard non-textual datasets (Escudeiro and Jorge, 2009), w.r.t. identification of cases from unknown classes, i.e., d-Confidence reduces the labeling effort that is required to identify examples from all classes. However, the error rate gets worse in the R52 dataset. D-Confidence gets to know more classes from the target concept earlier although less sharply. In the R52 dataset we are exchanging accuracy by representativeness. This might be desirable or not, depending on the specifc task we are dealing with. If we are trying to learn a target concept but we do not know examples from all the classes to learn – for instance if we are in the early stage of a classification problem – this effect might be desirable so we can get a full specification of the target concept with a reduced labeling effort.

It is interesting to notice that d-Confidence outperforms confidence to a greater extent on minority classes. This is obvious in R52 if we compute the cumulative average of the gain in labeling effort that is provided by d-Confidence when compared to confidence (Figure 7).

The gain for each class is defined as the number of queries required by d-Confidence to first hit the class minus the ones that are required by confidence. To compute the moving average, these gains are sorted in increasing order of the class frequency. The average gain starts at -128, for a class with frequency 2, and decreases to the overall average of -36 as class frequency increases up to 435. The bigger gains are observed in the minority classes. Although not as obvious as in R52 this same behaviour is also observed in the NG dataset.

Figures 5 and 6, as well as Tables 2 to 4, show that d-Confidence reduces the labeling effort required to identify unknown classes when compared to confidence. When selecting cases to label randomly, the first bunch of 10 distinct classes is found as fast as with d-Confidence but, from there on, when rare classes come by, d-Confidence takes the lead. The outcome is quite different in the NG dataset. In this dataset d-Confidence still outperforms confidence but it is beaten by random selection of cases after identifying 13.3 classes on average (after 22 queries on average). This observation led us to suspect that when in presence of balanced datasets, d-Confidence identifies new classes faster than random selection in the initial phase of the learning process but selecting cases by chance is better to identify cases in the latest stage of collecting exemplary cases, when few classes remain undetected.

5 Conclusions and Future Work

The evaluation procedure that we have performed provided statistical evidence on the performance of d-Confidence over text corpora when compared to confidence. Although the evaluation has been performed only on two datasets, the conclusions we have reached point out some interesting results.

D-Confidence reduces the labeling effort and identifies exemplary cases for all classes faster that confidence. This gain is bigger for minority classes, which are the ones where the benefits are more relevant.

D-Confidence performs better in imbalanced datasets where it provides significant gains that greatly reduce the labeling effort. For balanced datasets, d-Confidence seems to be valuable in the early stage of the classification task, when few classes are known. In the later stages, random selection of cases seems faster in identifying the few missing classes. However, d-Confidence consistently outperforms confidence.

The main drawback of d-Confidence when applied on imbalanced text corpora is that the reduction in the labeling effort that is achieved in identifying unknown classes is obtained at the cost of increasing error. This increase in error is probably due to the fact that we are diverting the classifier from focusing on the decision function of the majority classes to focus on finding new, minority, classes. As a consequence the classification model generated by d-Confidence is able of identifying more distinct classes faster but gets less sharp in each one of them. This is particularly harmful for accuracy since a more fuzzy decision boundary for majority classes might cause many erroneous guesses with a negative impact on error.

We are now exploring semi-supervised learning to leverage the intrinsic value of unlabeled cases so we can benefit from the reduction in labeling effort provided by d-Confidence without increasing error.

6 References

G. Adami, P. Avesani, and D. Sona. Clustering documents into a web directory for bootstrapping a supervised classification. Data and Knowledge Engineering, 54:301325, 2005.

D. Angluin. Queries and concept learning. Machine Learning, 2:319342, 1988.

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In In ICML, pages 6572. ICML, 2006.

E. Baum. Neural net algorithms that learn in polynomial time from examples and queries. IEEE Transactions in Neural Networks, 2:519, 1991.

K. Brinker. Incorporating diversity in active learning with support vector machines. In Proceedings of the Twentieth International Conference on Machine Learning, 2003.

O. Chapelle, B. Schoelkopf and A. Zien (Eds). Semi-supervised Learning. MIT Press, Cambridge, MA, 2006.

D. Cohn, L. Atlas, and R. Ladner. Training connectionist networks with queries and selective sampling. In Advances in Neural Information Processing Systems, 1990.

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. Machine Learning, (15):201221, 1994.

D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. Journal of Artificial Intelligence Research, 4:129145, 1996.

S. Dasgupta. Coarse sample complexity bonds for active learning. In Advances in Neural Information Processing Systems 18. 2005.

S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In Proceedings of the 25th International Conference on Machine Learning, 2008.

N. Escudeiro and A.M. Jorge. Efficient coverage of case space with active learning. In P. M. L. M. R. Lus Seabra Lopes, Nuno Lau, editor, Progress in Artificial Intelligence, Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA 2009), volume 5816, pages 411422. Springer, 2009.

S. Hanneke. A bound on the label complexity of agnostic active learning. In Proceedings of the 24th International Conference on Machine Learning, 2007.

S. Hoi, R. Jin, and M. Lyu. Large-scale text categorization by batch mode active learning. In Proceedings of the World Wide Web Conference, 2006.

M. Kaariainen. Algorithmic Learning Theory, chapter Active learning in the non-realizable case, pages 63 77. Springer Berlin / Heidelberg, 2006.

M. Li and I. Sethi. Confidence-based active learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28:12511261, 2006.

H. Liu and H. Motoda. Instance Selection and Construction for Data Mining. Kluver Academic Publishers, 2001.

I. Muslea, S. Minton, and C. A. Knoblock. Active learning with multiple views. Journal of Artificial Intelligence Research, 27:203233, 2006.

P. Ribeiro and N. Escudeiro. On-line news ‘a la carte. In Proceedings of the European Conference on the Use of Modern Information and Communication Technologies, 2008.

N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In Proceedings of the International Conference on Machine Learning, 2001.

G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In Proceedings of the International Conference on Machine Learning, 2000.

H. Seung, M. Opper, and H. Sompolinsky. Query by committee. In Proceedings of the 5th Annual Workshop on Computational Learning Theory, 1992.