2010-01-01

EGAL: Exploration Guided Active Learning for TCBR

Rong Hu
Technological University Dublin, rong.hu@tudublin.ie

Sarah Jane Delany
Technological University Dublin, sarahjane.delany@tudublin.ie

Brian Mac Namee
Technological University Dublin, brian.macnamee@tudublin.ie

Recommended Citation

Hu,R. Delany, S.J. & Mac Namee, B. (2010) *EGAL: Exploration guided active learning for TCBR.* Alessandria Italy, 19-22 July, In Proceedings of ICCBR 2010, p156-170, 2010. doi:10.1007/978-3-642-14274-1_13

2010-01-01

EGAL: Exploration Guided Active Learning for TCBR

Rong Hu
Dublin Institute of Technology, rong.hu@dit.ie

Sarah Jane Delany
Dublin Institute of Technology, Sarahjane.Delany@dit.ie

Brian Mac Namee
Dublin Institute of Technology, brian.macnamee@dit.ie

EGAL: Exploration Guided Active Learning for TCBR

Rong Hu, Sarah Jane Delany and Brian Mac Namee

Dublin Institute of Technology, Dublin, Ireland
rong.hu@dit.ie,sarahjane.delany@dit.ie,brian.macnamee@dit.ie

Abstract. The task of building labelled case bases can be approached using active learning (AL), a process which facilitates the labelling of large collections of examples with minimal manual labelling effort. The main challenge in designing AL systems is the development of a selection strategy to choose the most informative examples to manually label. Typical selection strategies use exploitation techniques which attempt to refine uncertain areas of the decision space based on the output of a classifier. Other approaches tend to balance exploitation with exploration, selecting examples from dense and interesting regions of the domain space. In this paper we present a simple but effective exploration-only selection strategy for AL in the textual domain. Our approach is inherently case-based, using only nearest-neighbour-based density and diversity measures. We show how its performance is comparable to the more computationally expensive exploitation-based approaches and that it offers the opportunity to be classifier independent.

1 Introduction

A significant barrier to developing case-based reasoning (CBR) systems in certain domains (particularly textual case-based reasoning (TCBR)) is that labelled case bases can be difficult or expensive to obtain. Active learning (AL) can be used to overcome this problem; building labelled case bases by selecting only the *most informative* examples from a larger unlabelled dataset for labelling by an oracle (normally a human expert) and using these to infer the labels for the remainder of the unlabelled data. The most popular *selection strategy* for choosing these most informative examples is *uncertainty sampling* [12]. Typically in uncertainty sampling a ranking classifier is trained using those examples labelled by the oracle so far and is then used to classify the remaining unlabelled examples. Using the output of the ranking classifier as a measure of classification confidence, those examples for which classifications are least confident are selected for labelling by the oracle. This process is repeated until a stopping criterion is reached - typically a limit on the number of labels given by the oracle.

Uncertainty sampling is considered an *exploitation*-based AL selection strategy which attempts to refine the classification decision boundary in uncertain areas of the feature space and can work well if the initial classification boundary is well shaped. However, with small numbers of labelled examples, it can be

difficult to reliably estimate the boundary, and it has been suggested that exploitation techniques are prone to querying outliers [20]. Exploitation approaches to selection can also suffer from a lack of exploration of the feature space and may not work well in some scenarios - for example XOR-type problems [2].

Other selection strategies have been developed which attempt to balance exploitation with *exploration*, focussing on examples distant from the labelled set with the aim of sampling wider, potentially more interesting areas of the feature space. These multi-faceted approaches have recently become popular. Existing work has combined uncertainty sampling with density information [7, 17, 21]; with diversity information [3, 5, 19, 23]; or with both [22, 29].

However, we believe that by applying an exploration-only approach to AL selection we can create an AL-based labelling system that is inherently case-based (i.e. based only on features of the case base derived from a similarity measure), and does not suffer from the difficulties associated with exploitation-based approaches. Furthermore, using an exploration-only approach is efficient as it does not require the repeated re-training of a classifier and re-classification of the unlabelled case base associated with exploitation-based approaches.

In this paper we present *Exploration Guided Active Learning* (EGAL), a simple, case-based, computationally efficient, exploration-only AL selection strategy that does not use the output of a classifier in its selection decisions. We compare the performance of this new approach to existing exploitation-based and hybrid selection strategies on a selection of text classification datasets.

The rest of the paper is organized as follows: Section 2 discusses the different selection strategies used in active learning, categorising them into exploration- and exploitation-based methods. Approaches that incorporate uncertainty sampling, density sampling, and diversity sampling; and other related work are discussed. We introduce our exploration-based selection strategy, EGAL, in Section 3 showing how it incorporates simple similarity-based measures of density and diversity. Section 4 describes an evaluation of EGAL using seven textual datasets. We conclude in Section 5 discussing how this approach can be adapted for non case-based classification tasks, offering the opportunity for a classifier-independent selection strategy to get over the reusability problem.

2 Review

AL can be used for two purposes: to build a classifier using the smallest number of manually labelled examples; or to build a fully labelled case base using the smallest number of manually labelled examples. While the difference between these two is subtle, and often ignored, it is important. A labelled case base can be useful for many tasks other than simply building a classification model - for example in [30] an AL-labelled case base was used for information retrieval-like search queries.

The advantages of using case-based classifiers in the AL process were appreciated initially by Hasenjager & Ritter [8] who proposed AL algorithms using local learning models; and by Lindenbaun et al. [14] who developed AL strate-

gies for nearest neighbour classifiers. Although any classifier can be used in the exploitation-based AL algorithms, the case-based approach to AL is particularly attractive as confidence scores are easily calculated, and the repeated retraining required in AL is especially efficient - new examples are simply added to the case base. More recent examples of case-based AL include index-driven selection sampling for CBR [26]; developing case retention strategies for CBR [18]; semantic labelling of text [16]; supervised network intrusion detection [13] and building classification systems with a weighted k-nearest neighbour classifier [4]. These applications all tend to use exploitation-based selection strategies.

Previous work which uses the underlying structure of the dataset to include exploration in AL selection strategies can be categorised into three approaches: *density-based sampling*, *diversity-based sampling*, and sampling using a combination of both density and diversity. One technique applied frequently is to identify the underlying structure in the dataset by clustering the unlabelled examples. Approaches that use clustering tend to talk about the *most representative* example [24, 27], which could either use a local inter-cluster measure which could be considered a density approach, or a global intra-cluster measure which could be considered a diversity approach. For clarity we will avoid the term *most representative*, and the remainder of this section will discuss techniques under the distinctions of density-based and diversity-based sampling.

2.1 Using Density in AL

Uncertainty sampling strategies are prone to querying outliers since outliers are likely to have high uncertainty [20]. To overcome this problem, selection strategies which consider density information have been proposed. The intuition is that an example with high density degree is less likely to be an outlier.

Incorporating density information with uncertainty sampling has been shown to boost the performance of AL in various studies [7, 15, 21, 31]. Labelling an example from a highly dense region of the domain space can increase the confidence of the classifications in its neighbourhood. The density of an example is generally calculated as the average similarity of those neighbours of the example within a specified neighbourhood and has been used, for example, to avoid the selection of outliers [31] and to select the most uncertain examples with maximum density [32]. A common approach is to use *density-weighting* where density is defined explicitly and combined as a function of the uncertainty score [17, 21, 31]. Other approaches are more implicit, such as those that cluster the unlabelled examples and use the properties of the clusters to select examples for labelling [27].

Novel uses of density information include He *et al.* [9] who make use of nearest neighbours to compare the local density of each example with that of each of its neighbours and select for labelling the example with the highest difference in density; and Fujii *et al.* [7] who use the neighbours of example x to quantify the increase in the utility score (called training utility) of the remaining unlabelled examples if a label is provided for x. The example which is expected to result in the greatest increase in training utility is selected for labelling.

2.2 Using Diversity in AL

Diversity is used in AL selection strategies mainly in an attempt to overcome the lack of exploration when uncertainty sampling is used. A popular approach to incorporating diversity is to include the *Kernel Farthest First* (KFF) algorithm (which selects those examples that are furthest from the current labelled set) as a member of an ensemble of AL processes [2, 19] (the other members of the ensemble are typically based on uncertainty sampling).

In the information retrieval literature, several AL heuristics which capture the diversity of feedback documents have been proposed [23, 28]. It has been demonstrated in [23] that the performance of traditional relevance feedback (presenting the top k documents according to relevance only) is consistently worse than that of presenting documents with more diversity. Several practical algorithms based on the diversity of the feedback documents have been presented - for example clustering the documents and choosing the cluster centroids to present for labelling [23].

2.3 Using Density & Diversity in AL

Several AL algorithms are proposed in the literature that either explicitly [4, 22, 29] or implicitly [28] combine both density and diversity with uncertainty sampling to select examples for labelling. These ensemble-based approaches have proven to be particularly successful as they have the advantages of all three approaches.

However, to the best of our knowledge, no approach has been described in the literature that combines density sampling and diversity sampling *without* also using uncertainty sampling. Such an exploration-only approach would be especially efficient as it would not require the repeated building of a classifier, or classification of a large set of unlabelled examples. It would also be particularly suited to the task of building labelled case bases as it would be based only on the properties of the case base and an associated similarity measure. The next section will describe our new EGAL algorithm which takes this approach.

3 The Exploration Guided Active Learning Algorithm

This section describes our exploration-only AL selection strategy: Exploration Guided Active Learning (EGAL). We first discuss how we measure density and diversity, and then explain how they are combined. For this discussion, consider a dataset, $\mathcal{D}$, which consists of a pool of unlabelled examples, $\mathcal{U}$, and a case base of labelled examples, $\mathcal{L}$, which grows as examples, x_i, are selected from $\mathcal{U}$ and presented to the oracle for labelling.

Measuring Density: We measure the density of an unlabelled example x_i by considering the similarity to x_i of the examples that are within a pre-defined neighbourhood N_i of x_i, as given in Equation 1. This neighbourhood N_i (see

Equation 2) is set by a similarity threshold α, where $\alpha = \mu - 0.5 \times \delta$; μ and δ being the mean and standard deviation of the pair-wise similarities of all examples in $\mathcal{D}$ respectively.

$$\text{density}(x_i) = \sum_{x_r \in N_i} \text{sim}(x_i, x_r) \tag{1}$$

$$N_i = \{x_r \in \mathcal{D} | \text{sim}(x_i, x_r) \geq \alpha\} \tag{2}$$

Unlike other density measures such as that in [9], we use the sum of the similarities in the neighbourhood N_i instead of the count of the number of neighbours in N_i. The effect of this is to have fewer *ties* in the density-based ranking, which makes for a more straightforward density-based sampling technique. A selection strategy using density alone will select the example(s) with the highest density to present for labelling.

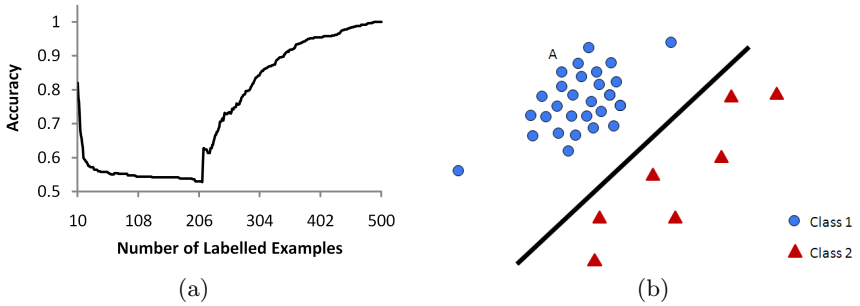
Measuring Diversity: We measure diversity by considering the examples which are most dissimilar to the labelled case base $\mathcal{L}$. Distance being the inverse of similarity, our diversity measure for an example x_i (given in Equation 3) is defined as the distance between x_i and its nearest labelled neighbour. The diversity measure has the advantage of efficient time complexity and it also ensures that the newly selected examples are different from the examples already in $\mathcal{L}$. A selection strategy based on diversity alone would select the example(s) with highest diversity to present for labelling.

$$\text{diversity}(x_i) = \frac{1}{\max_{x_r \in \mathcal{L}} \text{sim}(x_i, x_r)} \tag{3}$$

Combining Density and Diversity: Density and diversity sampling greedily choose examples that optimise locally, which can make them myopic approaches to selection in AL. They can become trapped in local optimums which can result in poor performance globally. An example of density sampling's poor performance is evident in Figure 1(a), which shows the performance of a density-based active learner on a textual dataset of 500 examples starting with 10 initially labelled examples, (details on the selection of the initial case base, the classifier used, and the performance measures used are given in Section 4). This shows a degradation in performance until after 200 or so examples are labelled, at which point performance improves rapidly. Figure 1(b) illustrates how this can happen. With density sampling, examples from class 1 in group A will be repeatedly selected for labelling while examples from class 2 will be ignored, leading to a poorly defined classification boundary during this time. When diversity alone is used, similarly dysfunctional scenarios can arise.

To overcome these problems, we introduce an element of diversity to a density-based sampling approach. Including diversity means that high density examples that are close to labelled examples are not selected for labelling by the oracle.

To determine whether an example should be considered as a candidate for selection, we use a threshold β. If the similarity between an unlabelled example



(a) (b)

Fig. 1. Illustrating how density-based sampling can perform badly

x_i and its nearest neighbour in the labelled case base is greater than β then x_i is not a candidate for selection. We call the set of examples that can be considered for selection the *Candidate Set*, $\mathcal{CS}$, which we define as follows:

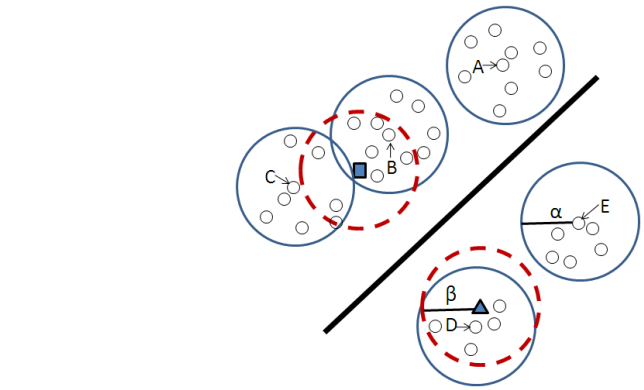
$$\mathcal{CS} = \{\exists x_i \in \mathcal{U} \mid sim(x_i, x_j) \leq \beta, x_j \in \mathcal{L}, \\ sim(x_i, x_j) \geq sim(x_i, x_k), \forall x_k \in \mathcal{L}, j \neq k\}$$

Our EGAL selection strategy ranks the possible candidates for selection (i.e. those in $\mathcal{CS}$) based on their density, and selects those examples with the highest density for labelling first. Thus, examples close to each other in the feature space will not be selected successively for labelling.

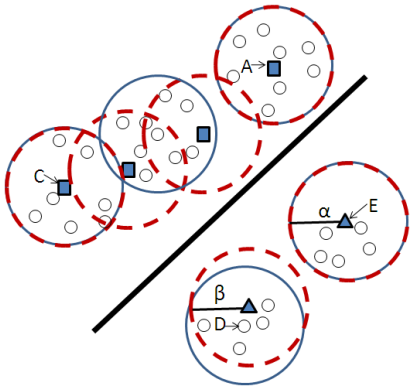
Parameters α and β play an important role in the selection process. α controls the radius of the neighborhood used in the estimation of density, while β controls the radius of the neighbourhood used in the estimation of $\mathcal{CS}$. The values selected for these parameters can significantly impact the overall performance.

Shen et al. [22] use a threshold similar to our β which they set to the average pair-wise similarity of the examples in the whole dataset. Initially, however, we set $\beta = \alpha$ as shown in Figure 2(a), where shaded polygons represent labelled examples in $\mathcal{L}$ and circles represent unlabelled examples in $\mathcal{U}$. The regions defined by α are shown as solid circles for a small number of unlabelled examples (A, B, C, D and E). For clarity of illustration, rather than showing the regions defined by β around every unlabelled example, we show them, as broken circles, around only the labelled examples. The effect, however, is the same: if a labelled example is within the neighbourhood of an unlabelled example defined by β, then the unlabelled example will also be within the neighbourhood of the labelled example defined by β.

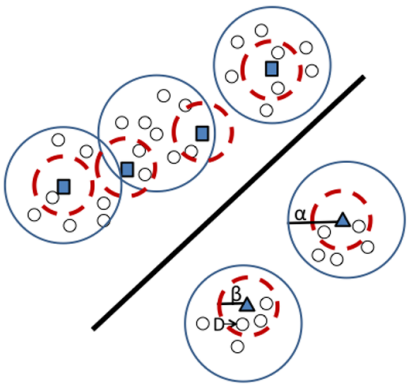
In the example shown in Figure 2(a), since examples B and D have labelled examples in the neighbourhood defined by β, they will not be added to $\mathcal{CS}$. A, C and E, however, will be added. As more examples are labelled, we may reach a stage when there are no examples in the candidate set as there are always labelled examples within the neighbourhood defined by β. This scenario



(a) $\alpha = \beta$ and $\mathcal{CS} \neq \emptyset$



(b) $\alpha = \beta$ and $\mathcal{CS} = \emptyset$



(c) $\alpha \neq \beta$ and $\mathcal{CS} \neq \emptyset$

Fig. 2. The relationship between parameters α and β and the candidate set $\mathcal{CS}$

is shown in Figure 2(b). When this happens we need to increase β to shrink this neighbourhood as shown in Figure 2(c). We update β when we have no examples left in $\mathcal{CS}$ - a unique feature of our approach as far as we are aware.

We use a novel method to update β motivated by a desire to be able to set the size of $\mathcal{CS}$. As the size of the $\mathcal{CS}$ is defined by β, a bigger β value gives us a bigger candidate set. We set β to a value which can give us a candidate set with a size proportional to the number of elements available for labelling (i.e. the size of the unlabelled pool $\mathcal{U}$) as detailed below:

(i) Calculate the similarity between each unlabelled example and its nearest labelled neighbour giving the set S, as follows

$$S = \{s_i = sim(x_i, x_j) \mid x_i \in \mathcal{U}, x_j \in \mathcal{L}, \\ sim(x_i, x_j) \geq sim(x_i, x_k), \forall x_k \in \mathcal{L}, j \neq k\}$$

(ii) Choose the value s_w from S that splits S into two, where

$$S_1 = \{s_i \in S \mid s_i \leq s_w\},$$
$$S_2 = \{s_j \in S \mid s_j > s_w\} \text{ and}$$
$$|S_1| = \lfloor (w \times |S|) \rfloor, 0 \leq w \leq 1$$

(iii) Let $\beta = s_w$, which is the similarity value such that w proportion of unlabelled examples will be in diverse neighbourhoods of the feature space.

The proportion parameter, w, allows us to balance the influence of diversity and density in our selection strategy. When $w = 0$, the EGAL algorithm defaults to pure diversity-based sampling discounting any density information. As w increases, the influence of density increases and the influence of diversity decreases with more examples being added to $\mathcal{CS}$. When $w = 1$ the EGAL algorithm becomes purely a density-based sampling algorithm. We explore the effect of changing the value of the proportion parameter w in Section 4.2.

Our combined strategy can be implemented very efficiently. At the start the pair-wise similarity matrix for the entire dataset and the individual density measure for every example are calculated and cached. At each iteration of the selection algorithm, the updated diversity measure for each example in the unlabelled set, $\mathcal{U}$, is the only calculation necessary. Computationally this is very efficient, especially considering the rebuilding of a classifier and the classification of every unlabelled example required by uncertainty sampling based methods at each iteration of the selection algorithm.

4 Evaluation

To assess the performance of our EGAL algorithm, we performed a comparative evaluation with other AL selection strategies. The objective of our evaluation was firstly to see whether the performance of combining density and diversity information in our EGAL approach was better than density or diversity sampling alone. In addition, we compared EGAL to uncertainty sampling which is the most commonly used AL selection strategy, and density-weighted uncertainty sampling which is the most common approach to combining density and uncertainty. After describing the datasets used, the implementation details of our EGAL approach and the evaluation measures used; this section will describe the results of these experiments.

4.1 Experimental Setup

In our evaluations we used seven balanced text-based classification datasets: a spam dataset [6]; four binary classification datasets derived from the 20-Newsgroup collection[1]; and two binary classification datasets from the Reuters

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/

collection[2]. The properties of each dataset, and the average accuracy achieved in five iterations of 10-fold cross validation using a 5-NN classifier, are shown in Table 1 (accuracies are included as an indication of the difficulty of each classification problem). Each dataset was pre-processed to remove stop-words and stemmed using Porter stemming.

Table 1. Details of datasets used in the evaluation experiments.

Dataset	**Task**	**Examples**	**Feat.**	**Accu.**
20NG-WinXwin	comp.os.ms-windows.misc vs. comp.windows.x	496	8557	91.14%
20NG-Comp	comp.sys.ibm.pc.hardware vs. comp.sys.mac.hardware	500	7044	85.56%
20NG-Talk	talk.religion.misc vs. alt.atheism	500	9000	93.92%
20NG-Vehicle	rec.autos vs. rec.motorcycles	500	8059	92.96%
Reuters-1	acq vs. earn	500	3692	89.56%
Reuters-2	g151 vs. g158	500	6135	95.36%
Spam	spam vs. non-spam	500	18888	96.80%

As the datasets used in our evaluations are fully labelled, the labelling process can be simulated without the need for a human oracle. At each iteration one example from the unlabelled pool, $\mathcal{U}$, is selected for labelling and its label is applied. This process is repeated until the oracle's label budget expires. In order to monitor the performance of the EGAL algorithm, and compare it to other approaches, after each labelling a k-NN classifier is built from the labelled case base, $\mathcal{L}$, and classifications are made for every example remaining in the unlabelled pool, $\mathcal{U}$. These classifications are compared with the actual labels in each dataset and the accuracy of this labelling is used to evaluate the performance of the selection strategy. Accuracy is calculated as $Accuracy = C/|\mathcal{D}|$, where C is the number of correctly labelled examples. Both manually and automatically labelled examples are included in this calculation so as to avoid large fluctuations as new labels are added in the latter stages of the process [10]. Using the accuracy recorded after each manual labelling, a learning curve is constructed to plot the accuracy as a function of the number of labels provided (for example Figure 3(a)). It is important to note that the classifications of the unlabelled pool made after each manual labelling are only for evaluation purposes and are not required by the EGAL algorithm.

In all of the experiments described in this section the same AL process is used. The initial case base contains 10 examples selected for labelling by the oracle using a deterministic clustering approach, as we have found it to be a successful approach to initial case base selection [11]. The same initial case base is used by each AL algorithm for each dataset. When classifiers are used, these

[2] http://www.daviddlewis.com/resources/testcollections/reuters21578/

are 5-NN classifiers using distance weighted voting. Finally, the stopping criteria used by all algorithms is a labelling budget which assumes that the oracle will provide 110 labels for each dataset.

4.2 Exploration of the Effect of the Balancing Parameter w

The density neighbourhood parameter, α, is set to $\mu - 0.5 \times \delta$ (as discussed in Section 3), as preliminary experiments showed it to be a good choice. In order to set the diversity neighbourhood parameter β, a value of w which controls the balance between density and diversity in the EGAL selection process is required. Intuition would suggest that diversity is more important than density, and in order to investigate this experiments were performed with w set to $0.25, 0.50$ and 0.75 on the datasets described previously. Results on two of these datasets are shown in Figure 3. Across the seven datasets it was clear that $w = 0.25$ gave the best results (indicated by the fact that the learning curve for $w = 0.25$ dominates the others) and this value was used in all further experiments. This experiment supports the intuition that diversity is more important than density in the selection process.

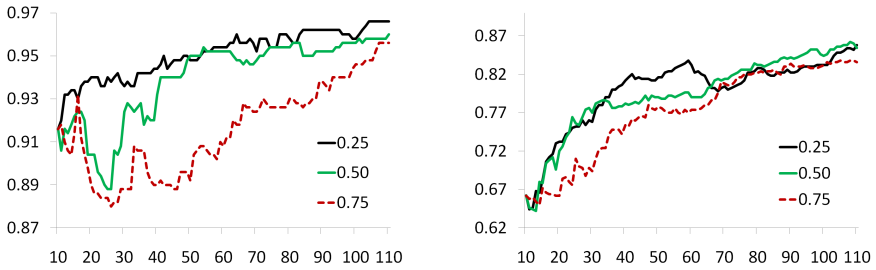


Fig. 3. The effect of the balancing parameter w on the EGAL algorithm

4.3 EGAL Evaluation Results

The results of comparisons between our proposed approach (labelled EGAL), density sampling (labelled Density) and diversity sampling (labelled Diversity) across the seven datasets are summarised in Figure 4. A random sampling strategy (labelled RS), which randomly picks examples for labelling, is also included as a baseline. The results show that density sampling doesn't perform well but that diversity sampling performs consistently better than the baseline random sampling. In addition, incorporating density information with diversity sampling in our EGAL algorithm improves the performance of diversity sampling consistently on all datasets.

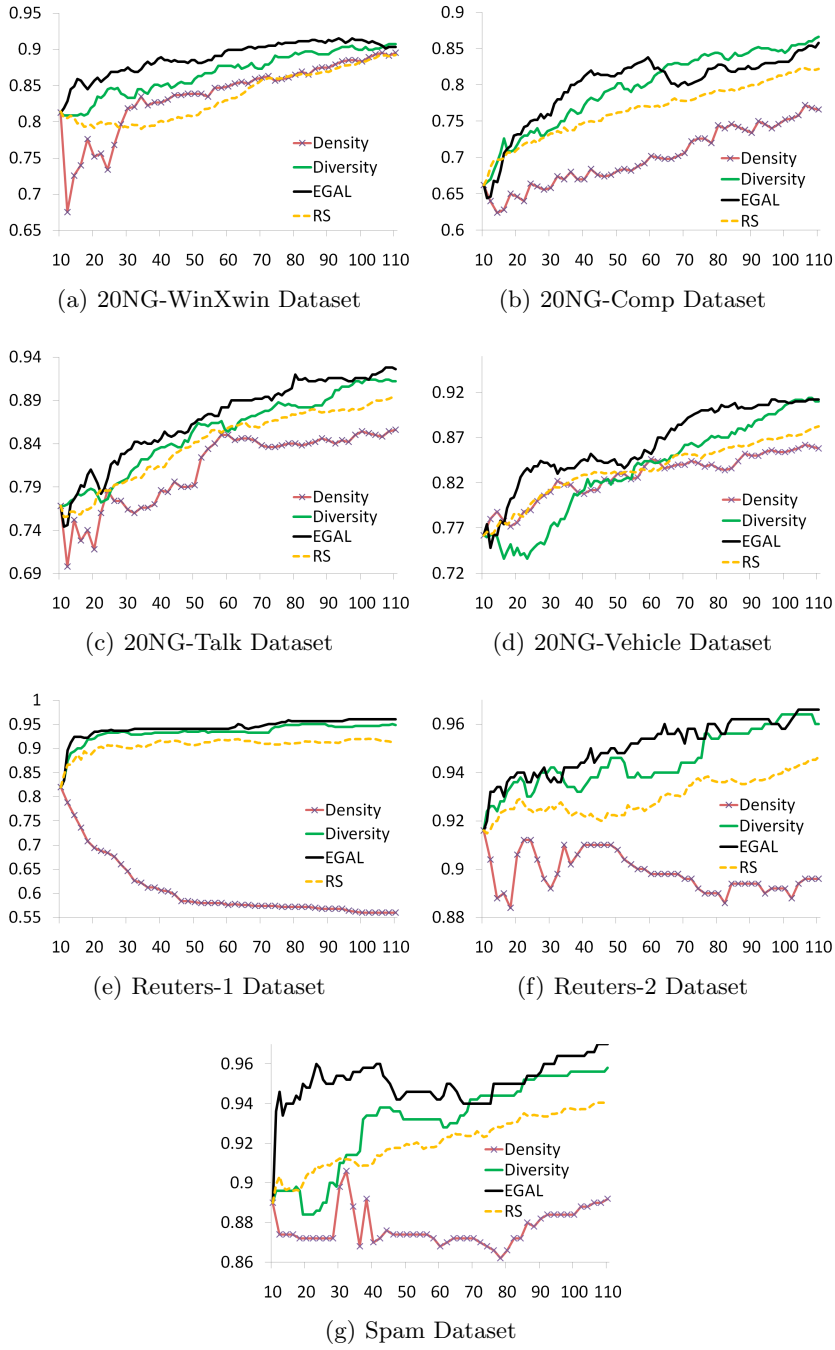


Fig. 4. Comparison of Density, Diversity, EGAL and RS selection strategies

We also compared EGAL to the more frequently used uncertainty sampling (US) using Hu *et al.*'s implementation [10] which is based on a k-NN classifier and density-weighted uncertainty sampling (DWUS) where uncertainty is multiplied with the density measure and examples with the highest resulting ranking score are selected for labelling. The results are shown in Figure 5.

Previous work on density weighted uncertainty sampling has shown an improvement over uncertainty sampling [17, 21]. Interestingly, the results in Figure 5 agree with that conclusion for datasets where density sampling alone also improves performance. However, for datasets where density sampling performs badly (see Figures 5(e), 5(f) and 5(g)) DWUS does not improve performance over US indicating that the density information is having a negative effect on the AL process.

The more interesting benefit of EGAL is in the early stage of the AL process, the first 20 to 30 labellings, where it outperforms both US and DWUS. A detailed analysis of the Area under the Learning Curve (ALC) for learning curves up to a varying number of labels was performed. Illustrative examples of ALC values are given in Table 2. The difference between US and EGAL was found to be significant (at $\alpha = 0.05$) using the Wilcoxon signed-rank test at 30 labels and below. There was no significant difference between DWUS and US at any number of labels.

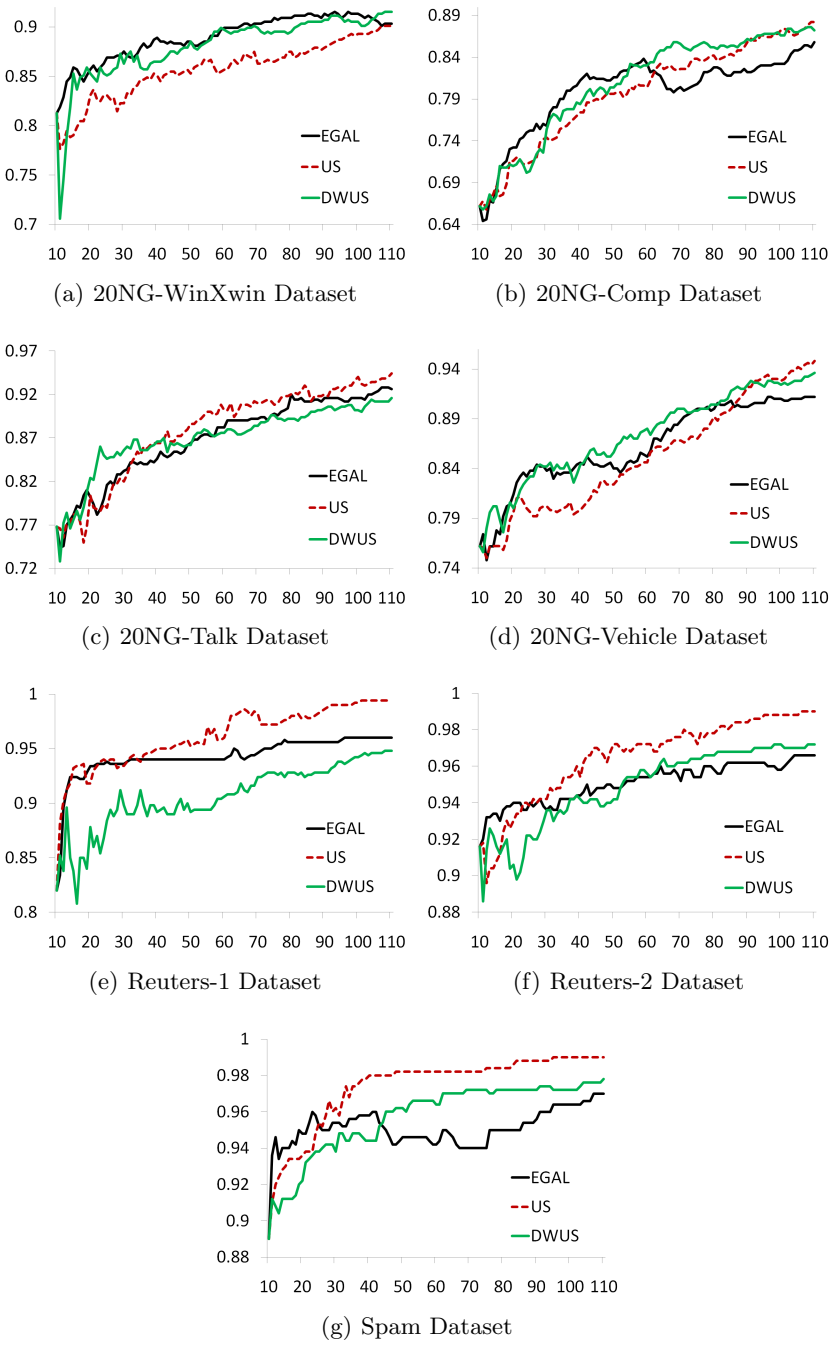
Table 2. Illustrative ALC values for learning curves up to the specified number of labels. The best values across the three approaches are highlighted in bold.

Dataset	30 Labels			60 Labels			110 Labels		
	US	DWUS	EGAL	US	DWUS	EGAL	US	DWUS	EGAL
20NG-WinXwin	16.24	16.69	**17.09**	41.79	42.95	**43.58**	85.72	88.07	**88.97**
20NG-Comp	14.02	14.04	**14.36**	37.47	37.97	**38.69**	79.87	**80.90**	79.89
20NG-Talk	15.74	**16.24**	15.89	41.92	**42.22**	41.56	**87.94**	87.04	86.95
20NG-Vehicle	15.66	**16.17**	16.14	40.20	**41.81**	41.42	85.26	**87.38**	86.35
Reuters-1	**18.49**	17.28	18.42	**47.03**	44.16	46.62	**96.21**	90.63	94.32
Reuters-2	18.53	18.30	**18.71**	**47.42**	46.60	47.11	**96.49**	94.92	95.09
Spam	18.76	18.49	**18.92**	**48.10**	47.15	47.43	**97.41**	95.75	95.12

These results point towards an interesting empirical property of the EGAL algorithm: it can improve the labelling accuracy fastest in the beginning stages of active learning. This would be beneficial in domains where labelling cost is high.

5 Conclusions and Future Work

In this work, we have proposed EGAL, an exploration-only approach to AL-based labelling of case bases. EGAL is inherently case-based as it uses only the notions of density and diversity, based on similarity, in its selection strategy. This



(a) 20NG-WinXwin Dataset

(b) 20NG-Comp Dataset

(c) 20NG-Talk Dataset

(d) 20NG-Vehicle Dataset

(e) Reuters-1 Dataset

(f) Reuters-2 Dataset

(g) Spam Dataset

Fig. 5. Comparison of EGAL, US and DWUS selection strategies

avoids the drawbacks associated with exploitation-based approaches to selection. Furthermore, in contrast to most active learning methods, because EGAL does not use a classifier in its selection strategy it is computationally efficient. We have shown empirical results of EGAL's viability as a useful tool for building labelled case bases, especially in domains where it is desirable to front-load the AL process so that it performs well in the earlier phases - a feature of EGAL demonstrated in our evaluation experiments.

It is on the absence of any particular classifier in EGAL that we intend to focus our future work. AL methods that use a classifier in their selection strategy are tuned to that particular classifier, resulting in poor reusability of the labelled data by other classifiers. This is known as the *reusability problem* in active learning [1, 25]. Tomanek et al. [25] argued that by using a committee-based active learner, the dataset built with one type of classifier can reasonably be reused by another. Another possible solution to the reusability problem is our EGAL algorithm as a classifier-free AL framework. Our future work in this area will check the reusability of the resultant labelled examples from EGAL at training different types of classifier.

Acknowledgments. This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/RFP/CMSF718.

References

1. Baldridge, J., Osborne, M.: Active learning and the total cost of annotation. In: Proc. of EMNLP '04. pp. 9–16 (2004)
2. Baram, Y., El-Yaniv, R., Luz, K.: Online choice of active learning algorithms. Journal of Machine Learning Research 5, 255–291 (2004)
3. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: Proc. of ICML '03. pp. 59–66 (2003)
4. Cebron, N., Berthold, M.R.: Active learning for object classification: from exploration to exploitation. Data Mining and Knowledge Discovery 18(2), 283–299 (2009)
5. Dagli, C.K., Rajaram, S., Huang, T.S.: Combining diversity-based active learning with discriminant analysis in image retrieval. In: Proc. of ICITA '05. pp. 173–178 (2005)
6. Delany, S.J., Cunningham, P., Tsymbal, A., Coyle, L.: A case-based technique for tracking concept drift in spam filtering. Knowledge-Based Systems 18(4–5), 187–195 (2005)
7. Fujii, A., Tokunaga, T., Inui, K., Tanaka, H.: Selective sampling for example-based word sense disambiguation. Computational Linguistics 24(4), 573–597 (1998)
8. Hasenjäger, M., Ritter, H.: Active learning with local models. Neural Processing Letters 7(2), 107–117 (1998)
9. He, J., Carbonell, J.G.: Nearest-neighbor-based active learning for rare category detection. In: Proc. of NIPS '07 (2007)
10. Hu, R., Mac Namee, B., Delany, S.J.: Sweetening the dataset: Using active learning to label unlabelled datasets. In: Proc. of AICS '08. pp. 53–62 (2008)

11. Hu, R., Mac Namee, B., Delany, S.J.: Off to a good start: Using clustering to select the initial training set in active learning. In: Proc. of FLAIRS '10. p. to appear. (2010)
12. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proc. of SIGIR '94. pp. 3–12 (1994)
13. Li, Y., Guo, L.: An active learning based TCM-KNN algorithm for supervised network intrusion detection. Computers and Security 26, 459–467 (2007)
14. Lindenbaum, M., Markovitch, S., Rusakov, D.: Selective sampling for nearest neighbor classifiers. Machine Learning 54(2), 125–152 (Feb 2004)
15. McCallum, A., Nigam, K.: Employing EM and pool-based active learning for text classification. In: Proc. of ICML '98. pp. 350–358 (1998)
16. Mustafaraj, E., Hoof, M., Freisleben, B.: Learning semantic annotations for textual cases. In: Workshop Proc. of 6th ICCBR. pp. 99–109 (2005)
17. Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: Proc. of ICML '04. pp. 623–630 (2004)
18. Ontañón, S., Plaza, E.: Collaborative case retention strategies for CBR agents. In: Proc. of ICCBR '03. pp. 392–406 (2003)
19. Osugi, T., Kun, D., Scott, S.: Balancing exploration and exploitation: A new algorithm for active machine learning. In: Proc. of ICDM '05. pp. 330–337 (2005)
20. Roy, N., McCallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: Proc. of ICML '01. pp. 441–448 (2001)
21. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proc. of EMNLP '08. pp. 1069–1078 (2008)
22. Shen, D., Zhang, J., Su, J., Zhou, G., Tan, C.L.: Multi-criteria-based active learning for named entity recognition. In: Proc. of ACL '04. p. 589 (2004)
23. Shen, X., Zhai, C.: Active feedback in ad hoc information retrieval. In: Proc. of SIGIR '05. pp. 59–66. ACM (2005)
24. Tang, M., Luo, X., Roukos, S.: Active learning for statistical natural language parsing. In: Proc. of ACL '02. pp. 120–127 (2002)
25. Tomanek, K., Wermter, J., Hahn, U.: An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In: Proc. of EMNLP '07. pp. 486–495 (2007)
26. Wiratunga, N., Craw, S., Massie, S.: Index driven selective sampling for CBR. In: Proc. of ICCBR '03. pp. 637–651 (2003)
27. Xu, Z., Yu, K., Tresp, V., Xu, X., Wang, J.: Representative sampling for text classification using support vector machines. In: Proc. of ECIR '03. pp. 393–407 (2003)
28. Xu, Z., Akella, R.: Active relevance feedback for difficult queries. In: Proc. of CIKM '08. pp. 459–468 (2008)
29. Xu, Z., Akella, R., Zhang, Y.: Incorporating diversity and density in active learning for relevance feedback. In: Advances in Information Retrieval, pp. 246–257 (2007)
30. Zhang, Q., Hu, R., Namee, B.M., Delany, S.J.: Back to the future: Knowledge light case base cookery. In: Workshop Proc. of 9th ECCBR. pp. 239–248 (2008)
31. Zhu, J., Wang, H., Tsou, B.: Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In: Proc. of COLING '08. pp. 1137–1144 (2008)
32. Zhu, J., Wang, H., Tsou, B.K.: A density-based re-ranking technique for active learning for data annotations. In: Proc. of ICCPOL '09. pp. 1–10 (2009)