The Interaction Between Supervised Learning and Crowdsourcing

Anthony Brew, Derek Greene, Pádraig Cunningham
School of Computer Science & Informatics
University College Dublin
{anthony.brew,derek.greene,padraig.cunningham}@ucd.ie

Abstract

In this paper we report insights on combining supervised learning methods and crowdsourcing to annotate the sentiment of a large number of economic news articles. The application entailed using annotations from a group of non-expert annotators on a small subset of articles to train a classifier that would annotate a large corpus of articles. This presents an *active learning* problem where the challenge is to make the best use of the annotators' efforts. We discuss the trade-off between determining consensus annotations and maximizing coverage on the training data. We also demonstrate that classifier uncertainty (a popular criterion for example selection in active learning) and disagreement between annotators are not the same thing. This finding provides an important insight into the interplay between supervised learning and crowdsourcing.

1 Introduction

Social media monitoring entails tracking huge volumes of data in order to track what is being said about products, people, and items of public policy. In such scenarios it will be impossible to annotate all content by hand. This challenge can be addressed by using machine learning techniques to train a classifier on a small subset of annotated data, and then use that classifier to annotate the remainder of the collection [4]. In the past the annotation might have been performed by expert annotators. Recently, it has become more common to acquire annotations via crowdsourcing, either from a pool of volunteer annotators, or through a micro-task market such as Amazon's Mechanical Turk.

The combination of crowdsourcing and machine learning for large-scale annotation has broad applicability in social web analytics, and has recently attracted a significant level of research interest [1, 5]. Raykar et *al.* addressed the problem of training a supervised learning system in the absence of ground truth data, when all that is available is noisy label information from non-expert annotators. They estimate the sensitivity and specificity of each of the annotators, and also annotate unlabeled examples. It is important to distinguish the work of Raykar et *al.* from the earlier work of Dawid and Skene [2], which addressed the problem of establishing a ground truth for a set of noisy labels, rather than the further problem of annotating unlabeled examples.

While Raykar et *al.* have provided an elegant solution to the problem of how to train a classifier from data with multiple noisy labels, they do not address the problem of managing the annotation process itself. We show that in some circumstances it is *not* useful to acquire multiple annotations for each label, instead it is better to use the annotation effort to maximize coverage in the dataset. This observation is particularly relevant in scenarios where the *annotation budget* will be strictly limited, such as in micro-task market settings. In Section 2 we provide some evidence that this decision concerning the trade-off between consensus and coverage will depend on the level of agreement between annotators. This point has already been raised by Sheng et *al.* [6]. However, they perform their analysis on datasets that do not contain multiple labels. Instead they synthesize multiple labels

using a noise model where the correct label is assigned with probability p. A key contribution in this paper is the provision of a real-world dataset[1] with multiple noisy labels provided by crowdsourcing. It is interesting to note that the noise in our data does not fit well with the model used by Sheng et *al.* as the noise is heteroskedastic, i.e. it is not constant across examples – there are many examples on which the annotators are in strong agreement and others on which they disagree completely.

The final message in this paper, relating to the management of the annotation budget, concerns the difference between annotator uncertainty and classifier uncertainty. We know from research on active learning that training examples on which the classifier is uncertain can be informative for the classifier [3]. In Section 3 we show that classifier uncertainty is quite different to annotator uncertainty, because consensus examples (from an annotator perspective) are more beneficial for training than examples on which there is significant disagreement.

2 Consensus versus Coverage

The application with which we are concerned [1] entails two related classification tasks: the identification of topically-relevant articles from a broad corpus, and the classification of relevant articles into positive and negative subsets. The distributions of the annotations for both tasks are shown in Figure 1. We observe that both annotation tasks are skewed – the relevance task towards the relevant side, and the sentiment task towards the negative side. For both tasks there are a number of consensus examples. For instance, there are almost 270 examples on which all annotations are negative and almost 150 examples on which all annotations are positive. It is important to note that the level of consensus varies considerably across articles.

Next we look at the impact of consensus on sentiment classifier accuracy and the related question of how to spend the annotation budget most effectively. Figure 2(a) shows learning curves as articles are added to a classifier – these curves correspond to three different annotation policies. The 'First to 3' policy is essentially a 'Best of 5' strategy, except that fourth or fifth annotations are not sought once a majority of three is attained. As expected, articles with more than one annotation yield a higher level of accuracy. It is perhaps surprising that 'First to 2' is as good as 'First to 3'. We believe this is due to the comparatively high level of agreement among annotators (see Figure 1(b)).

The situation is different when we look at learning curves plotted against annotation budget, as shown in Figure 2(b). In this situation the 'First to 1' policy is most effective because it achieves better coverage of the training data. This is surprising as it shows that it is not worthwhile establishing consensus annotations, rather it is better to spread the annotation effort across many examples. It seems that this is the case because of the reasonably high level of consensus in the annotations. Indeed, if we artificially add noise by flipping a portion of the annotations, the situation changes and 'First to 3' becomes the best way to spend the annotation budget.

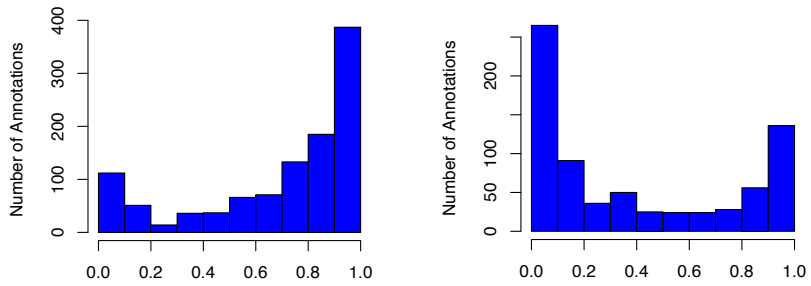
3 User Consensus and Classifier Uncertainty

Early research on active learning suggested that examples close to the decision surface will be most useful for training the classifier [3]. It is to be expected that annotators will disagree on marginal cases, while classifiers will be uncertain about cases that are close to the decision boundary. This train of thought suggests that annotator disagreement and classifier uncertainty are correlated.

In previous work we reported that we can achieve a greater increase in classification performance by adding training data starting with examples that have attained the highest consensus among users, rather than by starting with contentious examples for which user agreement is low [1]. The logic here is that the classifier derives most benefit from clear-cut examples on which annotators agree.

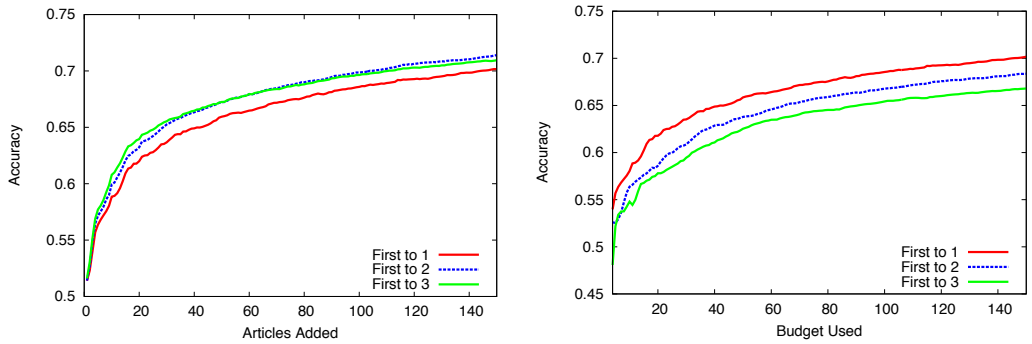
In order to reconcile the benefit of consensus examples reported in [1] with the perceived usefulness of examples on which classifiers are uncertain, we examined the correlation between classifier uncertainty and annotator disagreement. To do this we collected examples with more than 10 annotations and repeatedly randomly split these annotations into two groups. We calculated the consensus annotation agreement and the consensus score correlations between both sides of the splits. These are shown as accuracy and Spearman's rank correlation scores in Table 1. There is 95% agreement

[1] `http://mlg.ucd.ie/sentiment`



(a) Consensus on Relevance (b) Consensus on Positive Sentiment

Figure 1: The levels of consensus on the relevance annotation and sentiment annotation tasks.



(a) Increase in accuracy as articles are added. (b) Increase in accuracy as annotations are used up.

Figure 2: The impact of multiple annotations on classifier accuracy using alternative annotation policies. For instance, the policy 'First to 3' indicates that three annotations are in agreement.

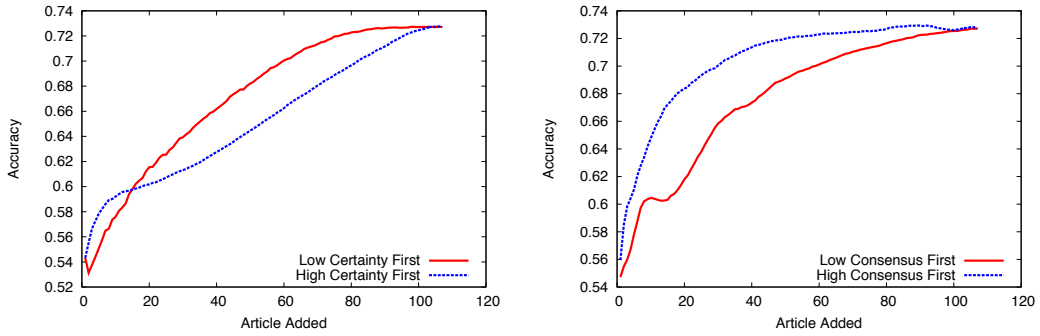
across the splitson the majority annotation. The correlation between the consensus ranking is 0.74 for the relevance classification task and 0.82 for sentiment classification.

	Relevance		**Sentiment**	
Labeler	Accuracy	Correlation	Accuracy	Correlation
User	*0.95*	*0.74*	*0.95*	*0.82*
Naïve Bayes	0.85	0.60	0.73	0.32
SVM	0.84	0.57	0.73	0.32

Table 1: Spearman's rank correlation between classifier uncertainty and user consensus, for the relevance and sentiment classification tasks. The "User" row reports inter-annotator group agreement.

The *User* scores in Table 1 serve as a baseline for our analysis on classifier uncertainty. We evaluated support vector machine (SVM) and naïve Bayes classifiers in a cross-validation framework, which allows us to calculate classifier uncertainty scores for all training examples. We then examined the rank correlation of these classifier uncertainty scores against annotator disagreement (one minus annotator consensus). The correlations are moderate for the relevancy classification task, and surprisingly low (0.32) for the sentiment classification problem. We also observed that the accuracy figures show that the sentiment classification is harder than the relevance classification. Poor correlation on the sentiment data between classifier uncertainty and annotator disagreement can be explained in part by the fact that the two classes are not well-separated in the bag-of-words representation used by the classifier.

To explore the apparent discrepancy between the two notions of uncertainty in more detail, we framed the sentiment classification task in an active learning framework. For the choice of example selection policy, we consider both classifier uncertainty and user consensus – examining low-to-high



(a) Classifier Certainty (b) User Consensus

Figure 3: Active learning curves for the sentiment classification task, using different example selection strategies. Examples are added to the learner using an ordering based on (a) classifier uncertainty, (b) consensus between user annotations.

and high-to-low in both cases. The learning curves for the different strategies are shown in Figure 3. It is apparent from Figure 3(a) that, with the exception of the earliest stages of the process, examples with low classifier certainty are most beneficial to the learning process. In contrast, in Figure 3(b) we see that the learner benefits most from the addition of examples about which users are strongly in agreement. The marked difference in the two sets of learning curves supports the view that classifier certainty and annotator consensus are quite different things.

4 Conclusions

We have examined the idea of using crowdsourcing to annotate a subset of a large unlabeled collection and then use supervised machine learning to label the remainder of the pool. Recently Raykar et *al.* [5] have presented a comprehensive strategy for the management of the learning process in the presence of more than one (noisy) annotation per object. In the work presented here we addressed the remaining challenge of managing the annotation process so as to use the available annotation budget as effectively as possible. We have reported two important findings. Firstly, it will not always be beneficial to gather more than one annotation per example – rather this decision will depend on the level of agreement between annotators. Secondly, although annotator disagreement and classifier uncertainty may be easily conflated with one another, we have shown that in practice they represent two very different concepts. In summary, classifier uncertainty can be useful to guide annotation, while annotator disagreement is an indicator of poor training data.

References

[1] A. Brew, D. Greene, and P. Cunningham. Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In *Proc. 19th European Conference on Artificial Intelligence (ECAI'10)*, pages 1–11, 2010.

[2] A.P. Dawid and A.M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

[3] D.D. Lewis and W.A. Gale. A sequential algorithm for training text classifiers. In *Proc. 17th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

[4] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proc. ACL-02 conference on Empirical methods in natural language processing*, volume 10, pages 79–86, 2002.

[5] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning From Crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[6] V.S. Sheng, F. Provost, and P.G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proc. 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 614–622, 2008.