Expert Systems With Applications 45 (2016) 438–449

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Active learning for text classification with reusability☆

Rong Hu[a], Brian Mac Namee[b], Sarah Jane Delany[a,*]

[a] Applied Intelligence Research Centre, Dublin Institute of Technology, Ireland
[b] School of Computer Science, University College Dublin, Ireland

ARTICLE INFO

Keywords:
Active learning
Machine learning
Reusability problem
Text classification

ABSTRACT

Where active learning with uncertainty sampling is used to generate training sets for classification applications, it is sensible to use the same type of classifier to select the most informative training examples as the type of classifier that will be used in the final classification application. There are scenarios, however, where this might not be possible, for example due to computational complexity. Such scenarios give rise to the reusability problem—are the training examples deemed most informative by one classifier type necessarily as informative for a different classifier types? This paper describes a novel exploration of the reusability problem in text classification scenarios. We measure the impact of using different classifier types in the active learning process and in the classification applications that use the results of active learning. We perform experiments on four different text classification problems, using the three classifier types most commonly used for text classification. We find that the reusability problem is a significant issue in text classification; that, if possible, the same classifier type should be used both in the application and during the active learning process; and that, if the ultimate classifier type is unknown, support vector machines should be used in active learning to maximise reusability.

1. Introduction

Automated *text classification* (or *text categorisation*) (Yang, 1999) is the task of automatically assigning predefined categories to textual documents based on their contents. *Spam filtering* (Drucker, Wu, & Vapnik, 1999) applications that sift through a user's incoming emails and identify those that are unsolicited, unwanted or inappropriate – those that are considered *spam* to the user – are a typical example. Another example is sentiment analysis (Pang, Lee, & Vaithyanathan, 2002) which aims to assist in the evaluation of documents – such as product reviews – by determining their overall sentiment (positive or negative). The relatively recent explosion of textual data from sources such as social network feeds and micro-blogging posts, on top of the already voluminous older sources such as SMS messages, online news articles, and blogs has made text classification an especially important problem within the machine learning community.

Text classification systems typically employ supervised learning approaches (Joachims, 1999; Yang & Liu, 1999) and, so, are reliant on the quality of the labelled historic datasets used to train them. Without a good dataset it is difficult to build an accurate classifier. Unfortunately, generating quality datasets usually requires manual labelling which is a time-consuming and, because experts are usually involved, expensive task. This can be a real barrier to the creation of classification systems but, fortunately, is not an insurmountable problem. *Active learning* (AL) (see Settles, 2009, for a review) is an iterative, semi-supervised learning process that can be used to build high-performance classifiers or labelled datasets 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 train a classifier or infer the labels for the remainder of the unlabelled data. Previous work (Lewis & Catlett, 1994; Tong & Koller, 2001; Yu, Zhu, Xu, & Gong, 2008) has shown that active learning can reduce the number of labelled examples needed to build an accurate text classifier by as much as 90% and, so, makes feasible the prospect of building text classification systems that would otherwise require prohibitively expensive amounts of manual data labelling.

The key consideration in active learning is the design of *selection strategies* that select the most informative examples that will be presented to the oracle for labelling. *Uncertainty sampling* (Cohn, Atlas, & Ladner, 1994; Cohn, Ghahramani, & Jordan, 1996; Lewis & Gale, 1994; Tong & Koller, 2001) is the most commonly used selection strategy. When uncertainty sampling is used in active learning, each time new examples are labelled by the oracle a classifier is trained using these and all of the other examples labelled so far. This classifier is used to classify the remaining unlabelled examples and the certainties associated with these classifications is recorded. The examples with the

☆ This material is based upon works supported by Science Foundation Ireland under Grant no. 07/RFP/CMSF718.

* Corresponding author. Tel.: +353 1 402 2840; fax: +353 1 402 4985.
E-mail addresses: amy.hur@gmail.com (R. Hu), brian.macnamee@ucd.ie (B. Mac Namee), sarahjane.delany@dit.ie (S.J. Delany).

http://dx.doi.org/10.1016/j.eswa.2015.10.003
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

lowest certainties associated with their classification are then presented for labelling and the process repeats until the maximum number of labels offered by the oracle is reached or some other stopping criteria has been met.

In many instances the classification algorithm used in the uncertainty sampling process is the same as the classification algorithm that will ultimately be used in the text categorisation system being constructed. Sometimes, however, the classification algorithm required for the final text categorisation system is not suitable for use in uncertainty sampling, or vice versa, and so the classification algorithms used will be different. There are a number of reasons that this scenario can arise including (i) that a classification algorithm might be too computationally expensive for use in an active learning selection strategy; (ii) that a text classification application might have particular classification algorithm requirements such as a capacity for explanation; or (iii) that the final form of the text classification system is not known at the time a labelled training set is created using active learning. This scenario gives rise to the *reusability problem* (Baldridge & Osborne, 2004; Tomanek, Wermter, & Hahn, 2007): "*is a set of labelled examples that is deemed most informative using one classification algorithm necessarily informative for another classification algorithm?*"

While the reusability problem has been studied before – Tomanek and Olsson (2009) go so far as to suggest that the reusability problem is a barrier to the widespread adoption of active learning – there is no detailed, formal analysis of the problem in the context of text classification in the literature. This paper presents such an analysis. Using the classification algorithms most commonly used in text classification, we consider the suitability of different pairs of classification algorithms used to select examples during active learning and then to perform classification in a resulting text classification application. We consider the following questions:

Q1: Does the reusability problem exist?
Q2: Does a homogeneous system in which the same classification algorithm is used for both uncertainty sampling in active learning and the final text classification application always perform best?
Q3: Are there text classification algorithms that are particularly well suited to active learning selection regardless of the algorithms that will be used in final text classification applications?
Q4: Are there text classification algorithms that are particularly well suited to text classification applications built using data generated using active learning?

Based on the analyses of the questions listed above recommendations are made for the use of active learning in text classification. Ancillary issues such as computational efficiency are also considered.

The structure of this paper is as follows. Section 2 first presents a comprehensive review of active learning and the reusability problem. Section 3 describes the methodology used in our experiments. Section 4 presents the evaluation performed to address the questions outlined above. Finally, Section 5 presents a set of recommendations based on this evaluation and discusses the directions in which the work will be taken in the future.

2. Related work

Active learning first garnered serious research attention in the 1980s (Angluin, 1988) and since then has remained a vibrant research area. Active learning is widely used in situations where there are vast amounts of unlabelled data available (e.g. classification of astrophysical data (Schneider, 2009), image classification (Tong & Chang, 2001), natural language processing (Baldridge & Osborne, 2004) and text classification (McCallum & Nigam, 1998)) or where labelled training examples are expensive or time consuming to obtain (e.g. bioinformatics (Cebron & Berthold, 2006) or medical applications (Warmuth et al., 2003)). Active learning has also recently been used in the field of chemometrics (spectroscopic-based data analysis) where it lead to a significant time and cost saving in an online production platform (Cernuda et al., 2014). Although there are other approaches (active learning with membership queries (Angluin, 1988), and stream-based active learning (Chu, Zinkevich, Li, Thomas, & Tseng, 2011; Freund, Seung, Shamir, & Tishby, 1997; Lughofer, 2012b)), *pool-based* active learning is by far the most common approach to active learning (particularly when active learning is applied to text classification problems (Lewis & Gale, 1994; McCallum & Nigam, 1998; Tong & Koller, 2001)) and is the approach considered in this work.

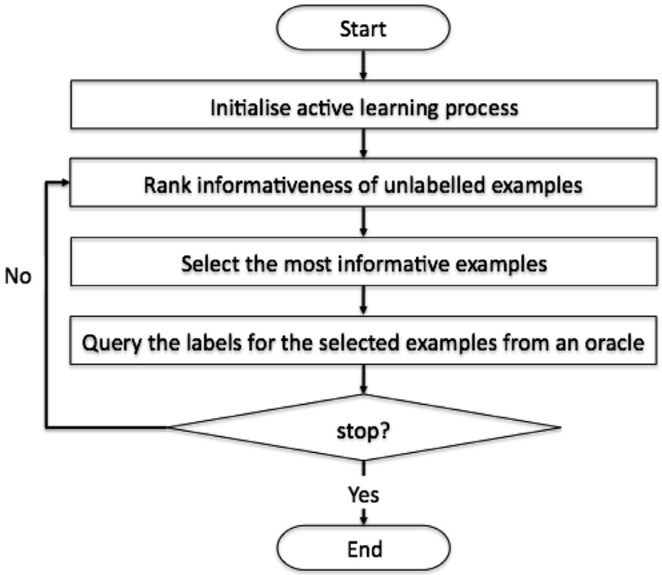
Start
Initialise active learning process
Rank informativeness of unlabelled examples
Select the most informative examples
Query the labels for the selected examples from an oracle
stop?
No
Yes
End

Fig. 1. A flow-chart of the pool-based active learning process.

Pool-based active learning assumes that the learner has access to a large pool of unlabelled examples from the beginning of the process. The goal is to build either an effective classifier or a fully labelled dataset (which will be most likely used at some point to train a classifier) by labelling only a small subset of the examples in the pool. This is achieved by selecting those examples from the unlabelled pool that are deemed to be *most informative* for labelling by an oracle (typically a human expert). Fig. 1 shows a flow-chart of the active learning process. After an initialisation step, the informativeness of each example in the pool is ranked and those deemed most informative are selected for labelling by the oracle. The informativeness ranking of each unlabelled example is then updated and the process is repeated until some stopping criterion has been met.

The key elements of the pool-based active learning process can be more formally modelled as a quintuple: $<\mathcal{S}, \mathcal{O}, \mathcal{L}, \mathcal{U}, \mathcal{SC}>$ (Baram, El-Yaniv, & Luz, 2004). A small set of seeded examples, $\mathcal{L}$, that are labelled by an oracle, $\mathcal{O}$, is used to initialise a selection strategy, $\mathcal{S}$. The selection strategy first involves assigning each member of the unlabelled pool, $\mathcal{U}$, a value indicating how informative a label for that example would be to the active learning process. The examples for which labels are deemed most informative are then selected for presentation to the oracle, $\mathcal{O}$, for labelling. The labelled examples are removed from the pool, $\mathcal{U}$, and added to the set of labelled examples, $\mathcal{L}$, and the informativeness values associated with each unlabelled example in $\mathcal{U}$ are updated. The process repeats as long as the oracle will continue to provide labels, or until some other stopping criterion, $\mathcal{SC}$, is reached. The final labelled set is then typically used to a build a classifier. This classifier itself can be the output of the active learning process or, alternatively, it can be used to label the remainder of