Enhancing the Performance of Semi-Supervised Classification Algorithms with Bridging

Jason Chan, Josiah Poon, Irena Koprinska

School of Information Technologies
The University of Sydney, NSW 2006 Australia
{jchan3,josiah,irena}@it.usyd.edu.au

Abstract

Traditional supervised classification algorithms require a large number of labelled examples to perform accurately. Semi-supervised classification algorithms attempt to overcome this major limitation by also using unlabelled examples. Unlabelled examples have also been used to improve nearest neighbour text classification in a method called bridging. In this paper, we propose the use of bridging in a semi-supervised setting. We introduce a new bridging algorithm that can be used as a base classifier in any supervised approach such as co-training or self-learning. We empirically show that classification performance increases by improving the semi-supervised algorithm's ability to correctly assign labels to previously-unlabelled data.

1. Introduction

A variety of supervised classification algorithms have been developed and applied successfully to problems in data mining. In general, these methods require a large amount of labelled data to achieve high performance. That is, supervised algorithms need to be supplied with a large mass of examples, each with the correct class attached to it, to accurately label new instances in the future. These samples have to be manually labelled by a human annotator, which requires an expert who is knowledgeable in the application domain. The process itself is expensive and can be very slow and error-prone.

There have been a number of methods proposed that attempt to either obtain such a labelled set more easily, such as by automatic means, or improve the performance of a supervised algorithm given a small labelled set.

Semi-supervised algorithms attempt to produce a large labelled set automatically given only a small labelled set and many unclassified instances. Algorithms such as co-training (Nigam and Ghani 2000) and expectation maximization (McCullum and Nigam 1998) fall into this category.

Active learning methods attempt to aid the process of building a labelled set by using heuristics to carefully select instances that are most likely to be useful to a classifier, thus reducing the effort needed to build a good quality training set (Muslea, Minton and Knoblock 2002).

Finally, the incorporation of *background knowledge* in a *second-order nearest neighbour* approach has been shown to improve text classification when only a small labelled set is available for training (Zelikovitz and Hirsh 2000). In this technique, external data of a different format to the problem setting is used in a modified version of the k-nearest neighbour algorithm.

In this paper, we propose the use of bridging in a semi-supervised setting. We introduce a new bridging algorithm which can be used as a base classifier in any semi-supervised approach such as co-training or self-learning (Nigam and Ghani 2000). We study the performance of both co-training and self-learning with our new bridging algorithm for text classification. In particular, we consider the problem of classifying short text strings into a set of pre-defined categories, such as the categorization of the titles of physics papers into sub-topics. This is a difficult problem as each individual instance (the title of a physics research paper) contains only a small subset of the large set of possible words. We show that new bridging improves the classification accuracy of both co-training and self-learning.

2. Proposed Algorithm

2.1 Using Background Knowledge as a Bridge

First we discuss the original bridging technique that uses background knowledge as described in (Zelikovitz and Hirsh 2000), which has been shown to improve the performance of text classification over other strong approaches. It involves the use of a small labelled set and a large set of background knowledge consisting of instances that are somehow related to the real instances that are to be classified, but they do not necessarily have to be in the same format as the instances in the problem setting. For example, if we are interested in classifying the titles of physics papers, we can use the abstracts of physics papers (instead of titles) as background knowledge.

A system resembling the nearest neighbour algorithm called WHIRL was used in (Zelikovitz and Hirsh 2000). Given an instance (the target instance) that needs to be classified, the bridging technique works by finding the *k* instances in the set of background knowledge that are most similar to it, where *k* is a user-specified positive integer. Similarly, for each of the instances in the small labelled set, the *k* most similar unlabelled instances are obtained. The labelled instances have class labels assigned to them so that it is possible to assign probabilities of class membership to the target instance to be classified according to the class labels of the labelled data and the similarities between the target instance, the instances in the background knowledge, and the labelled instances.

The idea behind this method is that with traditional methods, poor results will be obtained if the target instance were compared directly to the labelled set, since the labelled set is so small. It may be very hard to find similar instances to the target instance, in which case classification performance will deteriorate. However, if a large dataset of related background knowledge is used as a bridge between the target instance and the labelled set, it can be very useful because there are likely to be many instances that are similar to both the target instance and the labelled set, so a connection can be made between the two.

Clearly, for the background knowledge to be useful as a bridge, it must satisfy certain characteristics. Firstly, the background knowledge must be related to the test and labelled instances themselves. Otherwise, it will be difficult to obtain instances similar to a given target instance as well as instances in the labelled set.

Secondly, the background knowledge should be large. Again this is necessary to ensure that instances similar to the test and labelled set can be obtained.

Thirdly, for best performance, the individual instances in the background knowledge should contain a sizable amount of information. That is, a single instance in the background knowledge should contain more information than an instance in the labelled or test sets. In (Zelikovitz and Hirsh 2000), the problem of classifying the titles of physics papers (which are on average 12 words long) used abstracts of physics papers (on average 140 words long) in the background knowledge.

2.2 Bridging in Semi-Supervised Algorithms

The new approach that is introduced in this paper is to use a technique similar to bridging with background knowledge and to apply it into semi-supervised algorithms rather than using only bridging as a classifier or semi-supervised algorithms on their own. That is, instead of just using unlabelled data as examples that can be assigned a label and hence converted into labelled data, we propose that the unlabelled data also be used to help assign more accurate labels to the instance being assigned a class. This is possible by using the unlabelled data as an intermediary bridge between the target instance and the labelled set.

Existing semi-supervised algorithms convert labelled data to unlabelled data by using at least one supervised classifier. We propose that our *new bridging* classifier (as opposed to the *original bridging* algorithm in section 2.1) be used in replacement of these supervised classifiers.

Note that we use the unlabelled data as the intermediary bridge instead of background knowledge. We need to use unlabelled data instead of background knowledge, because when our new bridging algorithm is used in a semi-supervised algorithm, it will convert the unlabelled data into labelled training data. As discussed in section 2.1, background knowledge is more flexible as it is possible to use other types of data that contain more information than the instances in the problem space.

Using the unlabelled data may possibly deteriorate performance of the new supervised classifier because each individual instance in the unlabelled set now contains less information in comparison to background knowledge. However, we expect that the semi-supervised algorithm will improve in performance because the unlabelled data is assisting in labelling the unlabelled data. We will compare the use of background knowledge to unlabelled instances to see what impact it makes to the performance of the bridging algorithm on its own, that is, using bridging simply as a classifier rather than being combined with a semi-supervised algorithm.

Figures 1 and 2 outline the new bridging algorithm. In our experiment, the unlabelled set **U** contains titles instead of abstracts, since the labelled set **L** consists of titles.

```
Input:
L: small labelled set of instances
U: large unlabelled set of instances
t: target instance that we wish to classify
k: a positive integer parameter specified by the user
C: a set of classes to categorize instances in L

Algorithm:
Let $S(x,c)$ be the class membership score of any instance $x$ for class $c$, $c \in C$
$S(t,c) = 0$
Find the $k$ nearest instances $U_t=\{u_i | i:1..k\}$ in $U$ that are most similar to $t$, using a similarity metric (section 3.1)
Let $m_i$ be the similarity score between $t$ and $u_i$
for each instance $u_i$ in $U_t$
    Find the $k$ nearest instances $L_t=\{l_{ij} | i:1..k, j:1..k\}$ in $L$ that are most similar to $u_i$
Let $m_{ij}$ be the similarity score between $u_i$ and $l_{ij}$
for each class $c$ in $C$
        for each instance $u_i$ in $U_t$
                $S(u_i,c) = 0$
                for each instance $l_{ij}$ similar to $u_i$
                        if $l_{ij}$ belongs to class $c$
                                $S(u_i,c) = S(u_i,c) + m_{ij}$
for each class $c$ in $C$
        for each instance $u_i$ in $U_t$
                $S(t,c) = S(t,c) + m_i . S(u_i,c)$
Normalise $S(t,c)$ over all classes $C$ so that the sum over all classes is 1

Output:
Normalized class membership scores $S(t,c)$ for test instance $t$ and all classes $c$ in $C$
```

Figure 1. Pseudocode for the new bridging algorithm.

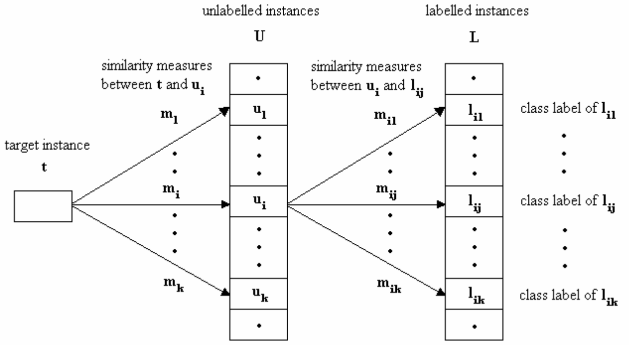

*Figure 2. A diagram illustrating the relationships between the labelled set **L**, the unlabelled set **U**, and the target instance **t** in the new bridging algorithm.*

2.3 Proposed Instance Selection Heuristics

For iterative semi-supervised learning algorithms, an instance must be selected from the unlabelled set to be transferred to the labelled set with an assigned label attached to it. We compared three different instance selection heuristics to be used in self-learning, a simple semi-supervised algorithm with just a classifier labelling unlabelled data and re-training on each iteration.

The *most non-zeroes heuristic* selects the instance that contains the most number of attributes with non-zero value. With this dataset, the instances (titles) do not contain many words, so those titles that contain even fewer words will be labelled with less confidence.

In the *minimum similarity heuristic*, the instance that is selected for labelling is the one that is least similar to the current set of instances in the labelled set, that is, the instance whose sum of similarities with all instances in the labelled set is the least. The idea is to introduce different instances into the labelled set so that performance can be improved in other areas of the learning problem.

The *random heuristic* is our baseline, which selects an instance randomly.

3. Experimental Setup

Our experiments consist of two stages. First, we compare our new bridging algorithm with the original bridging algorithm as described in (Zelikovitz and Hirsh 2000) to see how our changes affect classification performance.

In stage two, we test two semi-supervised algorithms, self-learning and co-training (Nigam and Ghani 2000), using our new bridging algorithm. Their performance is evaluated by comparing supervised classification using the original labelled set with the expanded labelled set that is produced.

3.1 Dataset Preprocessing

We used a dataset consisting of 953 physics technical papers, 493 of which belong in the astro-physics category and 460 in the condensed matter physics class. When implementing the bridging technique introduced in (Zelikovitz and Hirsh 2000), we use as background knowledge 1531 abstracts of physics papers. The entire dataset is the same as used in (Zelikovitz and Hirsh 2000).

There are a total of 1975 unique words present in the titles. After applying a stop list to remove common English words, we used information gain to rank attributes according to how well they distinguish between the different classes. We then apply dimensionality reduction and keep only the 100 most useful attributes. For each of these attributes, a Boolean value represents whether the instance has the corresponding word or not.

For our similarity metric that is needed to compare how similar two strings are, we count the number of attributes for which both strings have a non-zero value. This similarity metric is simply the numerator of the Jaccard coefficient (Tan, Steinbach and Kumar 2006), which is more suitable when dealing with data with attributes whose values are typically asymmetric, such as in our case where the number of zeroes easily outweighs the number of non-zero values. Separate testing (not shown here) showed that this was much more effective than using the regular cosine similarity metric because of the sparse nature of the dataset.

3.2 Supervised Learning Experiment

We compare the previously proposed original bridging algorithm using background knowledge (abstracts of physics papers) with our new bridging algorithm using instances in the unlabelled set in the same format as those in the labelled set. We also compare to JRipper, a conventional supervised algorithm that is an extension of Ripper, an algorithm that is already considered strong in this problem setting (Zelikovitz and Hirsh 2000). 5-fold cross validation is performed so that the test set contains 190 instances, the labelled set consists of 10 instances, and the remaining titles are used as the unlabelled set. Performance is measured on the test set using the standard f1-metric, which is the macro-average of the precision and recall. We try a variety of k values, using all odd numbers from 1 to 29 inclusive.

We attempt as best as possible to replicate the algorithm as introduced in (Zelikovitz and Hirsh 2000) but there are some noticeable differences. Their system uses WHIRL, a system with SQL-like queries to obtain the 30 most similar instances to a given instance, whereas we use a modification of the k-nearest neighbour algorithm for various k values. They do not use feature selection, whereas we have used information gain to keep only the top 100 attributes. We use Boolean values unlike how they have used TF-IDF values. However, the most important characteristics of their algorithm were replicated in our experiment, in particular, the use of unlabelled data and background knowledge as a 'bridge' to relate target instances to the labelled data.

3.3 Semi-Supervised Learning Experiment

In this experiment we evaluate the performance of semi-supervised algorithms using new bridging and compare this to semi-supervised learning without new bridging. We split the dataset into 5 equal-sized parts, with one part forming the test set (hence the test set contains 190 instances) and 10 instances from another part forming the initial labelled set. All other instances are used to form the initial unlabelled set. We repeat this 5 times so that it somewhat resembles 5-fold cross validation.

On each iteration of one of the semi-supervised learning algorithms, an instance is selected and labelled by a new bridging classifier according to what it thinks is the most likely label for that instance, then transferred from the unlabelled set to the labelled set with its newly-assigned label attached to it. This process continues until the labelled set reaches a size of 400; further testing showed that there was no significant improvement for larger set sizes.

The performance of semi-supervised algorithms is evaluated by measuring the quality of its expanded labelled set as follows. For each iteration of a semi-supervised algorithm, there exists a labelled set that contains the originally labelled instances as well as the instances that have been transferred from the unlabelled set by the algorithm. This labelled set is used to train a separate supervised classifier, whose f1 performance measure is determined on a test set. We decided to use the original bridging classifier as this supervised classifier to test the quality of labelled data.

For self-learning, we use various instance-selection heuristics as described in section 2.3. For the results of the *random selection heuristic*, we repeat the experiment 5 times and report the average results. There was generally little deviation between different runs using the *random selection heuristic*.

We also test another semi-supervised algorithm, co-training (Nigam and Ghani 2000), a more conventional and well-known semi-supervised algorithm. Briefly, co-training involves two base classifiers that have access to the same dataset, but the two classifiers have access to a disjoint set of attributes. For example, in email classification, co-training may involve two classifiers, one of which is only able to 'look' at the header information of the email, while the other classifier is only able to see the words in the main body of the email. Hence these classifiers are also known as 'view' classifiers because, in this example, they both 'see' the same emails but have different 'views'. On each iteration of the algorithm, a classifier labels an instance from each class that it considers to be most confident with labelling, and the classifiers take turns on different iterations. The idea behind co-training is that by having two view classifiers, one classifier can confidently label an instance and that newly labelled instance may help the other classifier who may not have been so confident to learn the problem task and vice versa.

In our experiment, we randomly split the features (as done successfully in (Chan, Koprinska and Poon 2004)) into two equal halves before applying the same feature selection methods as discussed in section 3.1 to both halves (resulting in two sets of 100 features). New bridging is used as the supervised learning algorithm by co-training for both base classifiers. Co-training transfers the instance from the unlabelled set that its base classifiers consider to label with most confidence. In the case of the new bridging classifier, confidence is determined by the normalised class membership scores that are output.

These methods are compared to the default method of self-learning and co-training without new bridging. For this baseline, we use JRipper as the base classifier. Finally, we will also compare the semi-supervised algorithms to simply using a large random sample of 400 instances with their correct labels. We re-sampled five times to obtain different sets of 400 instances and took the average supervised classification results using these five labelled sets. This will be referred to as the '*large labelled set*'.

4. Experimental Results

4.1 Supervised Classification Experiment

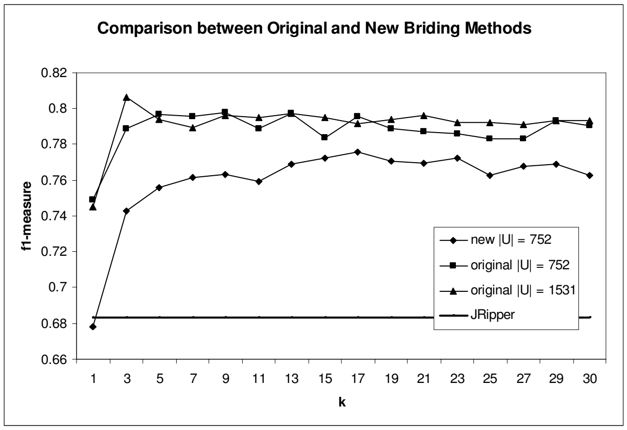

Figure 3: Comparison between original and new bridging algorithms for various k values. The experiment was performed with 5-fold cross validation, resulting in a test set size of 190 instances. The initial labelled set size was 10 instances, 5 of each class.

Figure 3 compares the original bridging algorithm as described in (Zelikovitz and Hirsh 2000) with new bridging as introduced in this paper. The line tagged with 'new |*U*|=*752'* represents the new bridging algorithm, that is, using 752 titles in the unlabelled set. The line tagged *'original* |*U*|=*1531'* is the original bridging algorithm using all the abstracts available from the dataset. Finally, the line marked *'original* |*U*|=*752'* represents the original bridging algorithm using abstracts in the unlabelled set, but restricted to have the same number of instances in the unlabelled set as was used in the new algorithm. The 752 instances were randomly chosen for each of the five different folds. We did this to ensure a fair comparison between the original and new bridging algorithms using the same number of instances in the unlabelled set.

It can be seen that, as expected, the new bridging algorithm as a classifier is inferior to the original bridging algorithm. However, the difference in performance is generally small. Hopefully, this small disadvantage will be outweighed by the improved performance that is associated with generating a larger labelled set by using it with a semi-supervised learning algorithm.

In Figure 3, we also compare the performance of supervised classification using our new bridging algorithm with using JRipper, a rule-based classifier. When the original bridging algorithm was introduced in (Zelikovitz and Hirsh 2000), it was compared to Ripper which was reported to obtain strong results in the tested domain. JRipper is an improved version of Ripper.

We see from the graph that the new bridging algorithm easily outperforms JRipper. The performance of JRipper is independent of different k-values, since it is not a nearest-neighbour approach. Figure 3 clearly shows that the newly-modified supervised bridging algorithm is still a useful classifier in this problem setting.

4.2 Semi-Supervised Learning Experiment

Out of the different instance selection heuristics applied to self-learning, the best performing and most consistent heuristic was easily the *most non-zeroes heuristic*. The *minimum similarity* and *random* heuristics performed very erratic over the different cross validation folds. Further investigation (not shown here) revealed that they were much more sensitive to the initial labelled set than the *most non-zeroes* heuristic. Hence, the results of the minimum similarity and *random* heuristics are not shown in the following graphs.

Also, we omit the results of semi-supervised learning (both self-learning and co-training) without new bridging (that is, using JRipper instead of new bridging as the base classifiers) because their results are, as expected, very poor. In all of these cases, performance deteriorates below that of using the initial labelled set.

As in the Supervised Learning Experiment, we ran the Semi-Supervised Learning Experiment over a wide range of k-values, obtaining consistent results over most k-values.

Figure 4 is an example of a typical graph that we obtained. It illustrates the successful performance of self-learning and co-training combined with new bridging using the *most non-zeroes heuristic* for a k-value of 15, with similar graphs resulting for other k-values (provided they were not too low). As shown, the semi-supervised algorithms using the new bridging approach are able to improve performance very quickly over the baseline (the lower horizontal line in all figures), and at least approach the performance obtained when trained with the labelled set of 400 instances with their correct labels, that is, the performance using the *large labelled set* (the upper horizontal line in all figures). For labelled set sizes of 100 or greater, the performance of self-learning using the *most non-zeroes heuristic* is comparable to co-training.

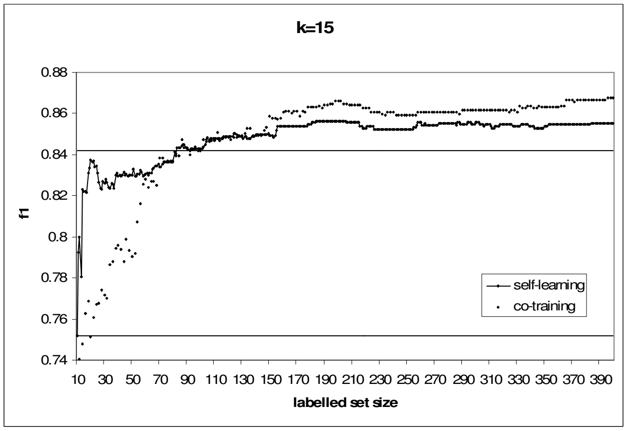

Figure 4: Performance of the new bridging algorithm for a variety of heuristics when k=15. In Figures 5-6, the lower horizontal line is the baseline (performance using initial labelled set), while the upper horizontal line is performance using the large labelled set (a random sample of 400 instances with their correct labels).

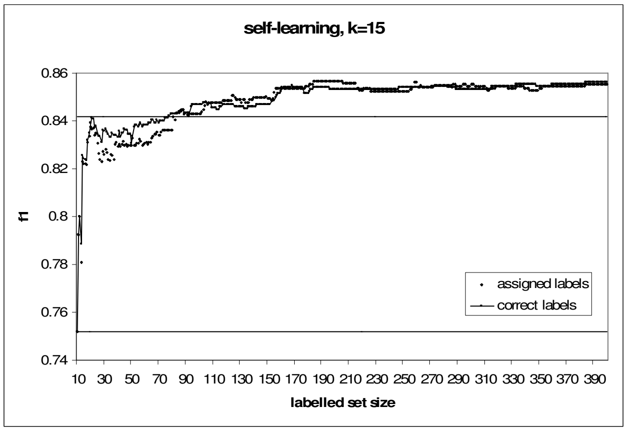

Figure 5: Supervised classification performance using the expanded labelled set. The comparison is between using the labels assigned by the semi-supervised learner against the actual (correct) labels. The semi supervised algorithm is self-learning using the most non-zeroes heuristic when k=15.

Figures 5 and 6 show the performance of a supervised classifier (original bridging) using the expanded labelled sets generated by the semi-supervised learners (self-learning using the *most non-zeroes* heuristic in Figure 5 and co-training in Figure 6). Within each graph, there is a comparison between performance using the labelled set with the labels assigned by the semi-supervised learner against using the actual (correct) labels. In the case of self-learning using the *most non-zeroes* heuristic, there is virtually no difference between the two curves unlike for co-training, especially for the earlier iterations. The *most non-zeroes heuristic* was successful because, by selecting instances with more non-zero values, could label them more accurately during semi-supervised learning.

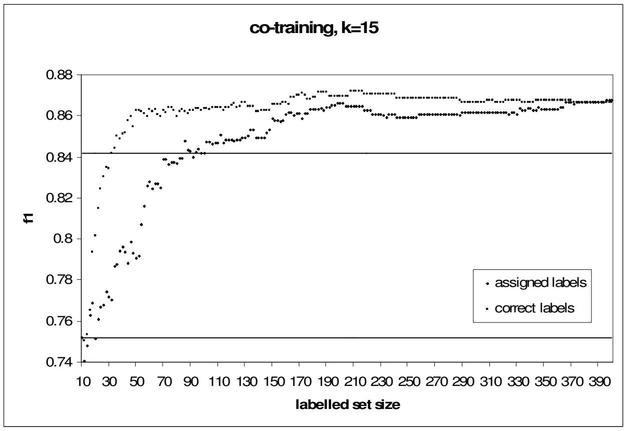

Figure 6: Supervised classification performance using the expanded labelled set. The comparison is between using the labels assigned by the semi-supervised learner (bottom curve) against the actual (correct) labels (top curve). The semi supervised algorithm is co-training when k=15.

Also note that the increase in performance during the early iterations is very steep for self-learning with the *most non-zeroes heuristic* compared with co-training. The performance reaches the expected performance with a random sample of 400 instances with their true labels after only a few iterations. This suggests that the instance selection heuristic selects instances that will be most beneficial to the supervised learner. Hence good performance can be obtained with fewer labelled instances.

5. Conclusions

In this paper, we have introduced a new approach to improve the performance of existing semi-supervised algorithms such as co-training and self-learning. Instead of just converting unlabelled data into labelled data like in existing semi-supervised algorithms, the unlabelled data is also used to provide a link between the labelled data and the instance that is being assigned a label. This results in much better classification performance, and therefore, the production of better quality labelled data. We tested our algorithm on the problem setting of classifying short-text strings (the titles of physics papers into physics categories) and also compared some instance-selection heuristics.

In contrast to a similar approach, bridging with background knowledge, which was used in (Zelikovitz and Hirsh 2000) but in a supervised context, our new bridging algorithm is used by semi-supervised algorithms. Our algorithm has the additional advantage that it is successful with data from only the original problem setting, unlike the original bridging approach, which is dependent on background knowledge from additional sources, complicating the issue of how to select appropriate data.

The semi-supervised experiment showed that both self-learning and co-training using the new bridging algorithm are capable of improving the performance of a supervised classifier trained with a small set of labelled data.

In the case of self-learning with an instance-selection heuristic, good performance was quickly achieved within only a few iterations. This implies that good performance can be obtained with fewer labelled instances because more useful instances are being selected. For our case of short-text strings, which is naturally a very sparse dataset, more can be learnt from instances with more non-zero values. In other domains, other selection heuristics may also speed up improvement in semi-supervised learning.

References

Chan, J., Koprinska, I., and Poon, J. eds. 2004. Co-training on Textual Documents with a Single Natural Feature Set. Proceedings of the 9th Australiasian Document Computing Symposium.

McCullum, A., and Nigam, K. eds. 1998. Employing EM and Pool-Based Active Learning for Text Classification. In proceedings of the Fifteenth International Conference (ICML '98), pp. 359-367.

Muslea, I., Minton, S., and Knoblock, C. A. eds. 2002. Active + Semi-Supervised Learning = Robust Multi-View Learning. Proceedings of ICML.

Muslea, I., Minton, S., and Knoblock, C. A. eds. 2000. Selective Sampling with Redundant Views. Proceedings of the 17th National Conference on Artificial Intelligence (AAAI).

Nigam, K., and Ghani, R. eds. 2000. Analyzing the Effectiveness and Applicability of Co-training. Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM).

Tan, P. N., Steinbach, M., Kumar, V. eds. 2006. Introduction to Data Mining. Addison Wesley.

Zelikovitz, S., and Hirsh, H. eds. 2000. Improving Short-Text Classification using Unlabeled Background Knowledge to Assess Document Similarity. Proceedings of 17th International Conference on Machine Learning.

Zelikovitz, S., and Hirsh, H. eds. 2002. Integrating Background Knowledge into Nearest Neighbor Text Classification. Proceedings of the Sixth European Conference on Case Based Reasoning (ECCBR), 2002.

Zelikovitz, S., and Hirsh, H. eds. 2003. Integrating Background Knowledge into Text Classification. Proceedings of the International Joint Conference of Artificial Intelligence, 2003.