2010-01-01

Off to a Good Start: Using Clustering to Select the Initial Training Set in Active Learning

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

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

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

Recommended Citation

Hu, R., Mac Namee, B. & Delany, S.J. (2010) Off to a good start: Using clustering to select the initial training set in active learning. *Twenty-Third International Florida Artificial Intelligence Research Society Conference*, Florida, 19 -21 May. doi:10.21427/D7Q89W

2010-01-01

Off to a Good Start: Using Clustering to Select the Initial Training Set in Active Learning

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

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

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

Off to a Good Start: Using Clustering to Select the Initial Training Set in Active Learning*

Rong Hu
School of Computing
Dublin Institute of Technology
Kevin Street, Dublin 8
rong.hu@dit.ie

Brian Mac Namee
School of Computing
Dublin Institute of Technology
Kevin Street, Dublin 8
brian.macnamee@dit.ie

Sarah Jane Delany
Digital Media Centre
Dublin Institute of Technology
Aungier Street, Dublin 2
sarahjane.delany@dit.ie

Abstract

Active learning (AL) is used in textual classification to alleviate the cost of labelling documents for training. An important issue in AL is the selection of a representative sample of documents to label for the initial training set that seeds the process, and clustering techniques have been successfully used in this regard. However, the clustering techniques used are non-deterministic which causes inconsistent behaviour in the AL process. In this paper we first illustrate the problems associated with using non-deterministic clustering for initial training set selection in AL. We then examine the performance of three deterministic clustering techniques for this task and show that performance comparable to the non-deterministic approaches can be achieved without variations in behaviour.

Introduction

The competence of a supervised machine learning system relies significantly on the quality of the training data used. Building a training set requires a large number of historical labelled examples. Gathering such labelled collections can be laborious, time consuming, often expensive, and prone to human error; which can make it a barrier to the creation of classification systems. Fortunately, this is not an insurmountable problem. Creating labelled datasets can be addressed using *active learning* (AL) (Cohn, Atlas, and Ladner 1994), a semi-supervised machine learning technique that can be used to build accurate classifiers from collections of unlabelled data with minimal effort in labelling. To achieve this AL identifies for labelling those examples that are deemed to be most informative to the training process.

There are three significant issues of concern in AL. First, a technique is required to choose a small initial training set to seed the AL process. Second, a *selection strategy* is required to select the examples that will be labelled throughout the AL process. These should be the examples for which labels will prove most informative as the training process progresses. Third, criteria must be established to determine when the AL process should stop. Most existing research focuses on the second problem. The question of how best to populate the initial training set has received little consideration in the AL community. In fact, most approaches ignore the problem and randomly choose examples.

In a review of 206 AL papers from conferences including *NIPS*, *ICCV*, *CVPR*, *ICML*, *UAI* and *ECML*; journals including *Machine Learning*, *Pattern Recognition*, and *Data Mining and Knowledge Discovery*; and technical reports, over 94% of researchers use a randomly selected initial training set or failed to specify their initial training set selection method. Fewer than 6% used a targeted approach to populating their initial training set. This ignores an opportunity to improve the effectiveness of the AL process.

To populate the initial AL training set in a more targeted way, clustering techniques can be used. According to Nguyen and Smeulders (2004) the most representative examples in a collection are likely to be those at the centres of clusters and these should be used as initial training examples to seed the AL process. In the AL literature there are some examples that take this approach, typically using k-Means (Zhu, Wang, and Tsou 2008) or k-Medoids (Nguyen and Smeulders 2004) clustering. However, both the k-Means and k-Medoids algorithms are non-deterministic and "*can often lead to highly inconsistent results over many trials*" (Greene 2006). This causes inconsistent performance when running the same AL system on the same dataset several times, and so comparison results can be unreliable. This problem is exacerbated in text classification as the datasets are of extremely high-dimensionality which leads to considerable variability in the clustering results.

In this paper we illustrate these problems with AL, showing how non-deterministic clustering methods can result in inconsistent behaviour in the AL process, and we propose the use of deterministic clustering techniques to populate the initial training set. We compare various deterministic clustering techniques and the aforementioned non-deterministic ones, and show that deterministic clustering algorithms are as good as non-deterministic clustering algorithms at selecting initial training examples for the AL process. More importantly, we show that the use of deterministic approaches stabilises the AL process.

The remainder of this paper is organised as follows: firstly we discuss AL and related work which uses clustering. We then briefly review the basic properties of the clustering algorithms used in the remainder of the paper before present-

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

ing and discussing experimental results. Finally, we draw conclusions and outline future work.

Related Work

Active learning attempts to overcome the difficulty and expense in obtaining labelled datasets. It builds labelled datasets by selecting only the *most informative* examples from a larger unlabelled collection for labelling by an *oracle*, typically a human expert. Active learning has been used for labelling large collections of different types of data, including textual datasets (Tong and Koller 2001), image datasets (Cebron and Berthold 2006) and video datasets (Yan, Yang, and Hauptmann 2003).

The most common AL scenario is *pool-based* AL (Lewis and Gale 1994) which assumes that the learner has access to a large pool of unlabelled examples from the beginning of the process, and this is the scenario considered in this work. A conceptual view of pool-based active learning has been modelled as a quintuple, $<\mathcal{C},\mathcal{L},\mathcal{S},\mathcal{P},\mathcal{O}>$ (Ma et al. 2006). The learning starts with a classifier, $\mathcal{C}$, trained on a small labelled training dataset, $\mathcal{L}$. The selection strategy, $\mathcal{S}$, is used to select the most informative examples from the unlabelled pool, $\mathcal{P}$, and request their true label from the oracle, $\mathcal{O}$. Following this, a new classifier, $\mathcal{C}$, is built using all of the examples labelled so far, and the process repeats as long as the oracle will continue to provide labels, or some other stopping criteria is reached — for example, the current classifier has achieved a particular goal. The most popular selection strategy for picking these most informative examples is *uncertainty sampling* (Lewis and Gale 1994) in which examples are selected based on the certainty with which the classifier can classify them.

The AL process begins with a small set of initially labelled examples. While this initial training set can be populated at random, it offers an opportunity to prime the AL process through informed population. Clustering techniques have been used for this task. The most common approaches are based on the k-Means (Kang, Ryu, and Kwon 2004) and k-Medoids (Nguyen and Smeulders 2004) algorithms, although there has been work using fuzzy c-means (Cebron and Berthold 2006). Clustering has also been used in AL for selection strategy design. Tang, Luo, and Roukos (2002) use a k-Means clustering algorithm to calculate the density of each example to quantify its '*representativeness*' which is combined with '*usefulness*' to select the examples to present for labeling. Xu et al. (2003) also proposed a representative sampling approach which explores the clustering structure of uncertain documents and used it to select the examples for labelling. However, in both of these situations the initial examples used to seed the active learner were randomly selected, with clustering only used in the selection strategy.

On the other hand Cebron and Berthold (2006) use an extended version of fuzzy c-means clustering with noise detection to cluster the data space initially and, after refining the clustering with *learning vector quantisation* (LVQ), to choose examples at cluster boundaries for labelling. In extensions to their work, instead of using preclustering of the data they propose a new self-controlled exploration/exploitation strategy which combines measures of example representativeness and uncertainty to select the examples to be labelled (Cebron and Berthold 2008).

Clustering

Clustering is an unsupervised learning method which groups together data examples that are similar to each other into a *cluster*. Clustering techniques have proven to be useful in understanding the structure of data, and a variety of document clustering algorithms have been proposed in the literature (Greene 2006). Deterministic clustering algorithms are those that produce stable clusters which are defined as clusters that "*can be confirmed and reproduced to a high degree*" (Mucha 2006).

The remainder of this section will describe the clustering techniques used in our experiments. The non-deterministic algorithms described are *k-Means*, *KMeans+ME*, and *k-Medoids*, all of which have been used in AL systems before. In the descriptions of these algorithms we will highlight the sources of their non-determinism. The deterministic algorithms described are *furthest-first-traversal* (FFT), *agglomerative hierarchical clustering* (AHC), and *affinity propagation clustering* (APC). To the best of our knowledge, these have not been used in AL systems for initial training set selection before.

k-Means Clustering

The k-Means algorithm (Duda and Hart 1973) groups a collection of examples into k clusters so as to minimise the sum of squared distances to the cluster centres. It can be implemented as a simple procedure that initially selects k random *centroids*, assigns each example to the cluster whose centroid is closest, and then calculates a new centroid for each cluster. Examples are reassigned to clusters and new centroids are re-calculated repeatedly until there is no change in clusters.

When k-Means clustering is used in AL the examples closest to the cluster centroids are selected as the members of the initial training set. The non-determinism in k-Means is introduced by the fact that the starting centroids are randomly selected. Different starting centroids can result in vastly different clusterings of the data, and this is exacerbated when the number of clusters k is large or when the data is high-dimensional. Although there have been efforts at making k-Means clustering deterministic (Likas, Vlassis, and Verbeek 2001; Su and Dy 2004), there is no agreed best technique for doing this and so the problem remains.

K-Means clustering has been used by Zhu, Wang, and Tsou (2008) to generate initial training sets for AL. In a variation on the k-Means approach for initial training set selection, artificial examples built from the virtual centroids, named *model examples*, are also added to the initial training set (Kang, Ryu, and Kwon 2004). This approach is named *KMeans+ME* and leads to an initial training set twice the size of that created when using just k-Means. However, KMeans+ME suffers from the same non-determinism as k-Means clustering.

k-Medoids Clustering

The k-Medoids algorithm (Kaufman and Rousseeuw 1990) is similar to k-Means except that it uses actual examples, *medoids*, as the centre of each cluster instead of artificially generated examples (centroids).

After the k-Medoids algorithm converges the k medoids are used as the initial AL training set. The random selection of the initial k medoids is again the source of non-determinism.

Furthest-First-Traversal (FFT)

The Furthest-First-Traversal clustering technique selects the most diverse examples in a dataset as cluster centres. The algorithm begins by selecting the example closest to the centre of the dataset and then iteratively chooses the example that is located furthest away from the current *centres* as the next centre. Often, *ties* can occur (where more than one example is equi-distant from the current centres) and in these situations the example in the densest area of the dataset is preferred. The density of example e is measured by the number of examples within a region (specified by a threshold d, typically set to the mean of the pair-wise distances) that have e as its centre. A standardised approach to handling ties ensures that the FFT algorithm remains deterministic.

In the same way as in the previous approaches the cluster centres found by the FFT algorithm are used as the initial training set for the AL process. The FFT algorithm has been used before in AL (Baram, El-Yaniv, and Luz 2004), but as part of a novel selection strategy rather than to prime the initial training set.

Agglomerative Hierarchical Clustering (AHC)

Agglomerative hierarchical clustering (Voorhees 1986) is a bottom-up clustering method which constructs a tree of clusters. Each example is initially assigned to its own individual cluster and the procedure repeatedly combines the two closest clusters until there is only one left. Each step creates a level in a dendrogram tree structure.

AHC can be used to select k clusters by pruning the tree so as to retain k leaf nodes in the hierarchy. The examples closest to the centres of these clusters are then selected and labelled to be included in the initial AL training set.

A variety of agglomerative algorithms have been proposed using different strategies to calculate distance between two clusters. Greene (2006) found *Min-Max linkage* (Ding and He 2002) to work well on text data and so this approach is used in our experiments. The AHC algorithm is entirely deterministic and has been used before in AL by Dasgupta and Hsu (2008), again as part of a selection strategy.

Affinity Propagation Clustering (APC)

Affinity propagation clustering (Frey and Dueck 2007) operates by simultaneously considering all examples in a collection as potential cluster centres, or *exemplars*, and exchanging messages indicative of the suitability of an example as an exemplar between them until a good set of exemplars emerges. APC has been shown to be a deterministic technique that can obtain very good clusterings (Frey and Dueck 2007). The APC algorithm has a parameter *preference*, p, which is used to control the number of clusters obtained. Broadly, a higher value of p results in more clusters and a lower value of p in less. To find a specific number of clusters in a dataset the value of p must be tuned experimentally, which is a disadvantage of the technique. The exemplars found are used as the initial training set.

Evaluation

This section reports on the experiments performed to validate the use of deterministic clustering algorithms in selecting initial training sets for AL. There are two objectives to the evaluations described here. The first is to show that k-Means and k-Medoids are non-deterministic and the impact of this on the AL process. The second is to compare the three deterministic clustering algorithms: FFT, AHC and APC; and confirm their superiority in selecting initial training examples.

Datasets

Four datasets were generated from the Reuters-21578 (Lafferty and Lebanon 2005), RCV1 (Lewis et al. 2004) and 20 Newsgroups[1] collections for this work. Similar to (Kang, Ryu, and Kwon 2004), 250 documents each were selected from two topics (*earn* and *acq*) of the Reuters-21578 collection, to form the *Reuters* dataset. From the 20 Newsgroups collection two datasets, *Comp* (consisting of 250 articles each from *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware*) and *WinXwin* (consisting of 250 articles each from *comp.os.ms-windows.misc* and *comp.windows.x*), were generated. 500 documents from the RCV1 collection make up the *RCV1* dataset which includes 250 documents from each of the *internal market* (g151) and *external relations* (g158) topics.

Each dataset was pre-processed to remove stop-words and stemmed using Porter stemming. Normalised *tf-idf* weighted word frequency vectors were used to represent documents. 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 general difficulty of each classification problem).

Table 1: Benchmark Datasets.

Dataset	Task	Features	Accuracy
Reuters	acq vs. earn	3692	0.8956
Comp	ibm.pc vs. mac	7044	0.8556
WinXwin	win. vs. win.x	8557	0.9114
RCV1	g151 vs. g158	6135	0.9536

Evaluation Measures and Experimental Method

In all of the experiments that follow the base classifier used in the AL process was a k-nearest neighbour classifier using

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

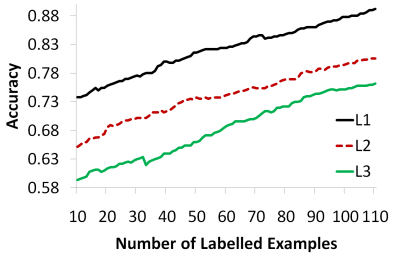


Figure 1: The learning curves produced by three runs of the AL process when the initial training set is populated using KMeans+ME clustering on the Comp dataset

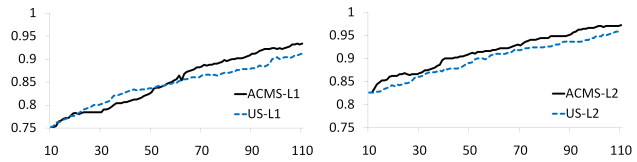
distance-weighed voting, with $k = 5$ and cosine similarity. For the AL process the initial training set size was set to 10 and *uncertainty sampling* (US) was used as the selection strategy (Hu, Mac Namee, and Delany 2008). The stopping criterion used by the AL process was that it continues until 100 manual labels are supplied. However, as the actual labels in all of the datasets used in these experiments are known the AL process is simulated, i.e. there are no real human experts involved.

Each time a new example is labelled by the oracle the labelling accuracy is calculated as $Accuracy = c/n$, where n is the number of examples in the entire collection (including the examples in the initial training set) and c is the number of correctly classified examples. Both manually and automatically labelled examples are included in this calculation to measure labelling accuracy over the entire collection, and to ensure that the measure remains stable as the process continues. For each experiment a learning curve is plotted with the number of labels given by the oracle on the *x-axis* and labeling accuracy on the *y-axis*. In each experiment the initial training set contained 10 examples and the process ran until the oracle had provided 100 labels. Hence, in each learning curve the number of labels given begins at 10 and runs until 110. Figure 1 shows three such learning curves.

Illustrating the Impact of Non-Determinism

The first set of experiments sought to confirm that the k-Means, KMeans+ME and k-Medoids clustering techniques were indeed non-deterministic, and to illustrate the impact of this. The AL process was run repeatedly using each of these algorithms to select the initial training examples. Because the initial cluster centres are selected randomly each time, the initial training set for the same dataset is different on subsequent runs. This results in differing performance for the AL process on each run.

This is illustrated in Figure 1 which shows the performance of the AL process on the Comp dataset when three different KMeans+ME clusterings are used to select the initial training set. This graph shows first that the KMeans+ME technique can produce very different clusterings even when applied to the very same dataset, and second that different initial training sets can have a significant impact on the outcome of the AL process.



(a) Initial training set I (b) Initial training set II

Figure 2: Comparison of the ACMS and US selection strategies on the WinXwin dataset

This lack of determinism is especially damaging when comparing the performance of different selection strategies within the AL process. Figure 2 illustrates this problem. Here, two selection strategies, *uncertainty sampling* (US) and *aggregated confidence measures sampling* (ACMS) (details of which can be found in (Hu, Delany, and Mac Namee 2009)), are compared on the WinXwin dataset using KMeans+ME clustering to select the initial training set. The results for two different runs of the experiment are shown in Figure 2 from which it is clear that due to the slightly different initial training sets selected each time it is very difficult to decide which selection strategy, if either, is performing better.

Comparison of Different Clustering Techniques

The second group of experiments compared the performance of all of the clustering techniques involved in the study — both deterministic and non-deterministic. As a base-line these techniques are also compared against a randomly selected initial training set. For each of the algorithms under consideration the AL process was run to completion using the selected clustering technique to populate the initial training set. For those algorithms containing a random component (i.e. random selection, k-Means, k-Medoids and KMeans+ME) the process was repeated 15 times and average results are presented. For clarity the results of this comparison are split into two groups.

Firstly, Figure 3 shows the results for random initial training set selection, and initial training set selection using k-Means, KMeans+ME, and k-Medoids clustering on two of the datasets. Based on these results it is clear that when any of the clustering techniques are used to select the initial training set the learning curves tend to dominate that achieved when the initial training examples are selected randomly, and that amongst the clustering techniques the learning curve from the initial training set selected using KMeans+ME tends to dominate the others. The learning curve due to the use of KMeans+ME is also flatter than the others indicating that the extra *model examples* smooth the learning process. This same pattern is also seen in the other two datasets used in this study and confirms results presented in (Kang, Ryu, and Kwon 2004).

Figure 4 shows how the use of FFT, AHC and APC clustering in the AL process compare to each other, and to the use of the KMeans+ME algorithm (the best non-deterministic approach). For KMeans+ME standard devi-

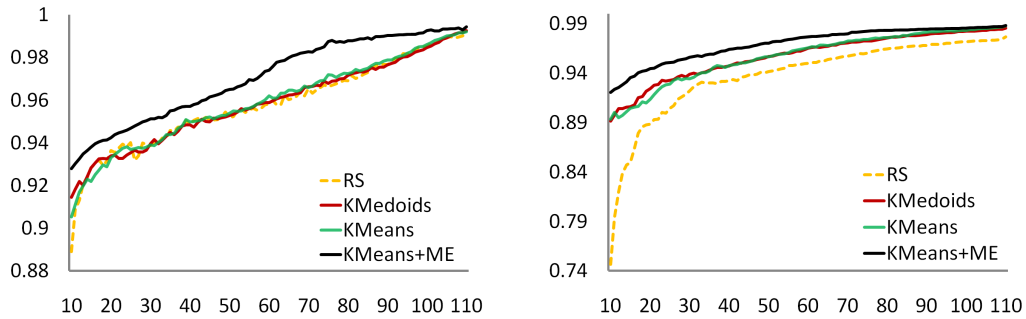


Figure 3: The learning curves produced by the AL process when the initial training set is chosen using random selection (RS), k-Means, KMeans+ME, and k-Medoids on the (a) Reuters and (b) RCV1 datasets

ation error bars are also shown to indicate the variation in the different runs of the process used to calculate this average. The first observation from these results is that the FFT algorithm is not well suited to this task. This is not unexpected since, by choosing examples that are furthest away from each other, this algorithm is particularly susceptible to noise and outliers.

The second observation is that the AHC and APC algorithms perform comparably to KMeans+ME, and on the Comp dataset both clearly dominate the non-deterministic technique. Given that AHC and APC are deterministic, this comparative performance makes them a better solution for selecting the initial training set for the AL process. Overall, the results for AHC are slightly better than for APC suggesting it should be given a slight preference.

Conclusions & Future Work

It has been previously established (and confirmed in our work) that using clustering to populate the initial training set can improve the performance of AL systems. However, the clustering techniques commonly used for this task (k-Means, KMeans+ME and k-Medoids) are non-deterministic which is problematic in its own right, and causes inconsistent performance when different AL selection strategies are compared.

After demonstrating the problems caused by using non-deterministic clustering approaches, this paper examined the use of three deterministic techniques for populating the initial training set in the AL process. Our experiments, on a variety of textual datasets, show that comparable labelling accuracy to that achieved using the best of the non-deterministic approaches, can be achieved using the deterministic clustering algorithms AHC and APC. Furthermore, our experiments indicate a slight preference for AHC. This comparable performance, and the determinism of AHC and APC, clearly indicate that they are the correct solution for selecting the initial training data in the AL process.

The first way we intend to expand this work is to further examine deterministic versions of the k-Means algorithm. KMeans+ME performs very well when compared to AHC and APC but suffers from the fact that it is non-deterministic. Although there is no agreed best technique for doing so, it is possible to modify the KMeans+ME algorithm to perform deterministically. The performance of deterministic versions of KMeans+ME will be compared against AHC and APC.

Secondly, we intend to introduce clustering techniques to some AL selection strategies, such as ACMS (Hu, Delany, and Mac Namee 2009)). Similar to the work of (Xu et al. 2003) and (Shen and Zhai 2003) mixing clustering information into the selection strategy allows it to search not only for the *most useful* examples for labelling by the oracle, but also for the *most representative* ones.

Acknowledgments.

Thanks to Derek Greene, Jun Sun, Jaeho Kang and Delbert Dueck for very helpful discussions.

References

Baram, Y.; El-Yaniv, R.; and Luz, K. 2004. Online choice of active learning algorithms. *J. Mach. Learn. Res.* 5.

Cebron, N., and Berthold, M. R. 2006. Adaptive active classification of cell assay images. *LNCS, PKDD 2006*.

Cebron, N., and Berthold, M. 2008. Active learning for object classification: from exploration to exploitation. *Data Mining and Knowledge Discovery*.

Cohn, D.; Atlas, L.; and Ladner, R. 1994. Improving generalization with active learning. *Mach. Learn.* 15.

Dasgupta, S., and Hsu, D. 2008. Hierarchical sampling for active learning. In *Proc. of ICML'08*.

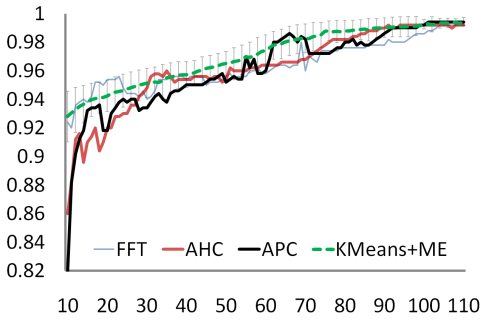
Ding, C., and He, X. 2002. Cluster merging and splitting in hierarchical clustering algorithms. In *Proc. of ICDM'02*.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc.

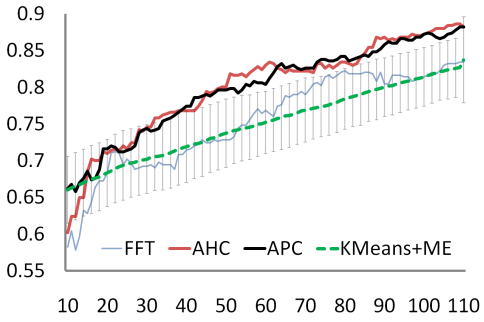
Frey, B. J., and Dueck, D. 2007. Clustering by passing messages between data points. *Science* 315.

Greene, D. 2006. *A State-of-the-Art Toolkit for Document Clustering*. Ph.D. Dissertation, Trinity College Dublin.

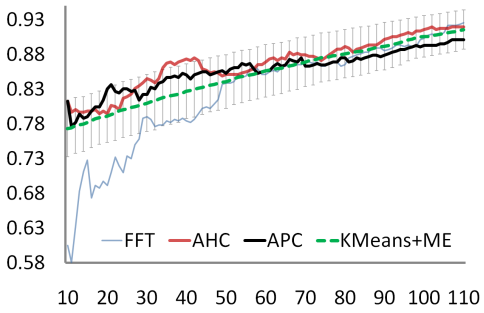
Hu, R.; Delany, S. J.; and Mac Namee, B. 2009. Sampling with confidence: Using k-nn confidence measures in active learning. In *Proc. of the UKDS Workshop at ICCBR'09*.



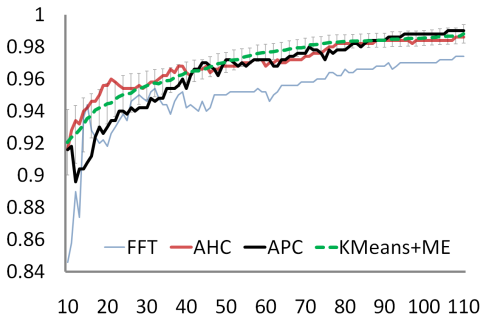
(a)



(b)



(c)



(d)

Figure 4: The learning curves produced by the AL process when the initial training set is chosen using furthest-first-traversal (FFT), agglomerative hierarchical clustering (AHC), affinity propagation clustering (APC), random selection (RS), and KMeans+ME on the (a) Reuters, (b) Comp, (c) WinXwin, and (d) RCV1 datasets

Hu, R.; Mac Namee, B.; and Delany, S. J. 2008. Sweetening the dataset: Using active learning to label unlabelled datasets. In *Proc. of the AICS'08*.

Kang, J.; Ryu, K. R.; and Kwon, H.-C. 2004. Using cluster-based sampling to select initial training set for active learning in text classification. In *Advances in Knowledge Discovery and Data Mining*, volume 3056.

Kaufman, L., and Rousseeuw, P. J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Lafferty, J., and Lebanon, G. 2005. Diffusion kernels on statistical manifolds. *J. Mach. Learn. Res.* 6.

Lewis, D. D., and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *Proc. of SIGIR-94*.

Lewis, D. D.; Yang, Y.; Rose, T. G.; and Li, F. 2004. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* 5.

Likas, A.; Vlassis, N.; and Verbeek, J. J. 2001. The global k-means clustering algorithm. *Pattern Recognition* 36.

Ma, A.; Patel, N.; Li, M.; and Sethi, I. 2006. Confidence based active learning for whole object image segmentation. *LNCS, MRCS 2006*.

Mucha, H. 2006. Finding meaningful and stable clusters using local cluster analysis. In *Data Science and Classification*.

Nguyen, H. T., and Smeulders, A. 2004. Active learning using pre-clustering. In *Proc. of ICML'04*.

Shen, X., and Zhai, C. 2003. Active feedback - uiuc trec-2003 hard experiments. In *Proc. of TREC'03*.

Su, T., and Dy, J. 2004. A deterministic method for initializing K-Means clustering. In *Proc. of ICTAI'04*.

Tang, M.; Luo, X.; and Roukos, S. 2002. Active learning for statistical natural language parsing. In *Proc. of ACL'02*.

Tong, S., and Koller, D. 2001. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* 2.

Voorhees, E. M. 1986. *The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval*. Ph.D. Dissertation, Cornell University.

Xu, Z.; Yu, K.; Tresp, V.; Xu, X.; and Wang, J. 2003. Representative sampling for text classification using support vector machines. In *Advances in Information Retrieval*.

Yan, R.; Yang, J.; and Hauptmann, A. 2003. Automatically labeling video data using multi-class active learning. In *Proc. of ICCV'03*.

Zhu, J.; Wang, H.; and Tsou, B. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proc. of COLING'08*.