

Cost-Effective Clustering through Active Feature-value Acquisition

Abstract

Many datasets include feature values that are missing but may be acquired at a cost. In this paper, we consider the clustering task for such datasets, and address the problem of acquiring missing feature values that improve clustering quality in a cost-effective manner. Since acquiring all missing information may be unnecessarily expensive, we propose a framework for iteratively selecting feature values that result in highest improvements in clustering quality per unit cost. Our framework can be adapted to different clustering algorithms, and we illustrate it in the context of two popular methods, K-Means and hierarchical agglomerative clustering. Experimental results on several datasets demonstrate clustering accuracy improvements provided by the proposed framework over random acquisition. Additional experiments demonstrate the performance of the framework for different cost structures, and explore several alternative formulations of the acquisition strategy.

Introduction

In many data mining and machine learning tasks, datasets include instances that have missing feature values which are initially unknown, but can be acquired at a cost. However, both the acquisition cost and the usefulness with respect to the learning task may vary dramatically for different feature values. While this observation has inspired a number of approaches for active and cost-sensitive learning, most work in these areas has focused on classification settings. Yet, the problem of obtaining most useful missing data cost-effectively is equally important in unsupervised settings, such as clustering, since the amount by which acquired information may improve performance varies significantly across instances and features.

In business domains, for example, clustering algorithms are commonly used to discover potential groups of customers for which companies can employ different direct marketing or pricing policies. However, some individual features values of particular instances (customers) may be missing, which results in the degradation of clustering performance. This missing information can often be acquired by contacting customers or by purchasing it from third party

vendors. Since individual feature values may provide varying levels of information about the underlying clustering of customers, choosing which feature values to acquire for each customer is an important decision; as buying information for all customers with missing features may be prohibitively expensive, while randomly querying a subset of these customers is often not optimal.

In this paper, we address the problem of *active feature-value acquisition* (AFA) for clustering: given a clustering of incomplete data, the task is to select feature values which, when acquired, are likely to provide the highest improvement in clustering quality with respect to acquisition cost. To the best of our knowledge, this general problem has not been considered previously, as prior research on active learning for clustering focused either on acquiring pairwise distances (Hofmann & Buhmann 1998; Buhmann & Zoller 2000) or cluster labels for complete instances (Basu, Banerjee, & Mooney 2004). Both of these scenarios can be viewed as instantiations of the general AFA task, with missing feature values corresponding to either pairwise distances or latent cluster labels.

The primary challenge of the overall AFA problem in the clustering setting lies in the difficulty of estimating the value of a potential acquisition in the absence of any supervision. In previous work, the AFA task was studied only for *supervised* learning, where missing feature values are acquired in a cost-effective manner for training classification models (Melville *et al.* 2005). Since this approach was using supervised data to estimate the expected improvement in model accuracy for the possible acquisitions, it cannot be applied to clustering where no supervision is available. This paper describes several avenues for formulating acquisition strategies based on expected utility in clustering settings.

We consider several utility measures for estimating the expected contribution of an acquisition to clustering quality, and demonstrate that the optimal choice of a utility measure is affected by the particular clustering algorithm and the desired performance metric. Specifically, we present an instantiation of our overall framework for K-means for two classes of utility measures: one based on the objective function for the clustering algorithm at hand, and another based on the stability of the current clustering configuration. Empirical results on several datasets demonstrate that such utility functions are effective at identifying missing feature val-

ues which, when acquired, improve clustering quality per unit cost significantly faster than random acquisitions. In additional experiments, we show that our framework performs consistently for different costs distributions and study how to adapt the framework to the hierarchical agglomerative clustering algorithm.

Task definition and algorithm

AFA framework for clustering

The clustering task is traditionally defined as the problem of partitioning a set of instances into disjoint subsets, or clusters, where each cluster contains similar instances. As discussed in the previous section, we focus our attention on clustering in domains where instances include missing feature values. Given a dataset consisting of m n -dimensional instances, we represent it by an m -by- n data matrix X , where x_{ij} corresponds to the value of the j -th feature of the i -th instance. Initially, the data matrix X is incomplete, i.e., its elements corresponding to missing values are undefined. For each missing feature value x_{ij} , there is a corresponding cost C_{ij} at which it can be acquired. Let q_{ij} refer to the query for the value of x_{ij} . Then, the general task of active feature-value acquisition is the problem of selecting the instance-feature query that will result in the highest increase in clustering performance per unit cost.

The overall framework for the generalized AFA problem is presented in Algorithm 1. Because AFA is defined as an iterative task, at each step the clustering algorithm partitions the current (incomplete) dataset and ranks all possible queries based on their expected contribution to clustering quality normalized by cost. The highest-ranking query is then selected, and the feature value corresponding to this query is acquired. The dataset is appropriately updated, and this process is repeated until some stopping criterion is met, e.g., desirable clustering quality has been achieved. To reduce computational costs, multiple queries can be selected at each iteration, resulting in batch acquisitions, as often done in active learning settings for classification tasks.

While the overall framework of Algorithm 1 is intuitive, the crux of the problem lies in ranking queries by their expected contribution to clustering quality. In subsequent sections, we address challenges related to performing this task accurately and efficiently.

Estimating expected utility

At every step of the AFA algorithm, the next best feature to acquire is the one that will result in the highest improvement in clustering quality per unit cost. Since the true values of missing features are unknown prior to acquisition, it is necessary to estimate the potential impact of every acquisition for all possible outcomes. Hence, the optimal policy is to ask for feature values which, once incorporated into the data, will result in the highest increase in clustering quality in *expectation*. Therefore, our approach is based on defining a *utility function* $\mathcal{U}(x_{ij} = x, C_{ij})$ which quantifies the anticipated benefit arising from obtaining a specific value x for feature x_{ij} via the corresponding query q_{ij} at cost C_{ij} . Then, expected utility for query q_{ij} , $EU(q_{ij})$, is defined as

Algorithm 1 Active Feature-value Acquisition for Clustering

Given:

- X – initial (incomplete) instance-feature matrix
- \mathcal{L} – clustering algorithm
- b – size of query batch
- C – cost matrix for all instance-feature pairs

Output:

$M = \mathcal{L}(X)$ – final clustering of the dataset incorporating acquired values

1. Initialize $TotalCost$ to initial cost of X
 2. Initialize set of possible queries $Q = \{q_{ij} : x_{ij} \text{ is missing}\}$.
 3. Repeat until stopping criterion is met
 4. Generate a clustering, $M = \mathcal{L}(X)$
 5. $\forall q_{ij} \in Q$ compute $score(M, q_{ij}, \mathcal{L}, X)$
 6. Select a subset S of b queries with the highest $score$
 7. $\forall q_{ij} \in S$,
 8. Acquire values for x_{ij} : $X = X \wedge x_{ij}$
 9. $TotalCost = TotalCost + C_{ij}$
 10. Remove S from Q
 11. Return $M = \mathcal{L}(X)$
-

the expectation of the utility function over the marginal distribution for the feature x_{ij} :

$$EU(q_{ij}) = \int_x \mathcal{U}(x_{ij} = x, C_{ij})P(x_{ij} = x) \quad (1)$$

While ranking queries using the expected utility defined above is the optimal acquisition strategy, the true marginal distribution of each missing feature value is unknown. Instead, an empirical estimate of $P(x_{ij} = x)$ in Eq.(1) can be obtained using probabilistic classifiers. In the case of discrete (categorical) data, for each feature j , a naïve Bayes classifier M_j can be trained to estimate the feature’s probability distribution based on the values of other features of a given instance. Then, when evaluating the query q_{ij} , the classifier M_j is applied to the corresponding instance x_i to estimate the distribution of possible values for the missing feature $\hat{P}_j(x_{ij} = x|x_i)$, conditioned on all known feature values for the instance. Then, the expectation in Eq.(1) can be easily computed by piecewise summation over the possible values. For continuous attributes, computation of expected utility can be performed either using statistical methods such as Markov Chain Monte Carlo, or via discretizing them and using probabilistic classifiers as described above.

Computing the utility function

Selecting an appropriate utility function \mathcal{U} to estimate the benefits of possible acquisition outcomes in Eq.(1) is the critical component of the AFA framework. Expected utility formulations have been previously employed in active learning for classification settings (Roy & McCallum 2001; Melville *et al.* 2005), where utility functions can be naturally derived from expected error reduction (or generalization accuracy), since classification error is a well-defined

evaluation measure for supervised tasks. Since clustering is an unsupervised problem for which there is no undisputed evaluation measure or intrinsic generalization objective, several alternatives exist for defining the utility function.

A number of clustering quality measures have been proposed and employed in prior work. They can be loosely divided into two large groups: external measures that evaluate clustering accuracy with respect to some category distribution unseen at clustering time, and internal measures that use only data that is available to the clustering algorithm. Examples of external measures are Rand Index (Rand 1971), cluster purity and pairwise F-measure (Steinbach, Karypis, & Kumar 2000), while a large number of internal measures exists, e.g., ratio between average inter-cluster and intra-cluster distances, clustering compactness, and average partition density (Halkidi, Batistakis, & Vazirgiannis 2001). Because in unsupervised settings external labels are unavailable at clustering time, we are limited to defining utility functions based on criteria internal to the dataset at hand.

Most clustering algorithms optimize a specific objective function, which allows defining utility as *reduction in objective function* per unit cost. For example, the objective function minimized by the very popular K-Means algorithm (MacQueen 1967) is the sum of squared distances between every instance x_i and the centroid of the instance’s cluster, μ_{y_i} :

$$J_{KM}(X) = \sum_{x_i \in X} |x_i - \mu_{y_i}|^2 \quad (2)$$

where y_i is the index of the cluster to which instance x_i is assigned, $y_i \in \{h\}_{h=1}^k$, and missing feature values are omitted from the squared distance computation. Then, the objective-based utility function for K-Means can be defined as the cost-normalized reduction in the value of the objective function due to acquisition outcome $x_{ij} = x$:

$$\mathcal{U}_{KM}^{(obj)}(x_{ij} = x, C_{ij}) = \frac{J_{KM}(X \wedge x_{ij} = x) - J_{KM}(X)}{C_{ij}} \quad (3)$$

where the objective function value after the acquisition $J_{KM}(X \wedge x_{ij} = x)$ can be estimated as the change in objective function due to the relocation of cluster centroids caused by the acquisition, without incorporating the actual acquired value into computation of Eq.(2), which would bias the computation.

While an objective-based utility function provides a well-motivated acquisition strategy, it may select feature values that reduce the objective function by improving cluster centroid locations without significantly changing actual cluster assignments. This effect can be significant, and we will see in the next section that objective-based utility is frequently a suboptimal strategy for improving performance of K-Means with respect to external evaluation measures.

Besides using the objective function to guide the acquisition process, one may also choose a utility function that approximates the qualitative difference in clustering configuration caused by the acquisition. Such a utility function would measure the benefit of each acquisition by perturbation in the actual cluster assignments caused by it, thus approximating

how robust the clustering configuration is with respect to a missing value. Such a utility can be defined as the number of points for which cluster membership changes as the result of an acquisition. Formally, given the current data matrix X , let $y_i^{(X)}$ be the cluster assignment of the point x_i before the acquisition, and $y_i^{(X \wedge x_{ij} = x)}$ be the cluster assignment of x_i after the acquisition. Then, perturbation-based utility of acquiring value x for feature x_{ij} is defined as follows:

$$\mathcal{U}^{(pert)}(x_{ij} = x, C_{ij}) = \frac{\sum_{i=1}^m \mathbb{1}(y_i^{(X \wedge x_{ij} = x)} \neq y_i^{(X)})}{C_{ij}} \quad (4)$$

Estimating actual cluster assignments after the acquisition, $Y^{(X \wedge x_{ij} = x)} = \{y_i^{(X \wedge x_{ij} = x)}\}_{i=1}^m$, is a problem that is specific to a particular clustering algorithm at hand. For K-Means, this computation can be performed without re-clustering for every possible acquisition outcome by re-estimating the centroid of the cluster to which instance x_i is currently assigned, assuming the value x for x_{ij} . Then, performing a single assignment step for all points would provide the new set of cluster assignments $Y^{(F \wedge x_{ij} = x)}$. Henceforth, we refer to this perturbation-based utility as *Expected Utility-Perturbation (EU-Perturbation)*; and we refer to the use of the objective-based utility as *Expected-Utility-Objective (EU-Objective)*.

A major challenge in implementing both utility functions is the computational complexity of exhaustively evaluating all potential acquisitions. We make this selection tractable by evaluating only a sub-sample of the available queries. We specify an exploration parameter α which controls the complexity of the search. To select a batch of b queries, first a random sub-sample of αb queries is selected from the available pool, and then the expected utility of each query in this sub-sample is evaluated.

Experimental evaluation

Methodology

We evaluated our proposed approach on four datasets from the UCI repository: *iris*, *wine*, *letters-ijl*, and *protein*, which have been previously used in a number of studies for comparing the performance of clustering algorithms. Features with continuous values in these datasets were discretized using equal-width binning. Since feature acquisition costs are not naturally available for these datasets, we consider several cost structures. In our first set of experiments, we assume that acquisition costs are uniform for all feature values. Later, we describe experiments for other cost distributions.

To evaluate the usefulness of the AFA framework, we compare its instantiations with different utility functions with random feature acquisition strategy that selects queries uniformly at random. The sampling parameter α of our methods is set to 10 which means a subset of 100 queries is sampled and evaluated to acquire 10 queries with highest scores. All results are obtained over 100 runs for each active acquisition strategy. In each run, a fraction of features is randomly selected for initialization for each instance in the dataset (ranging from 15-25% of the full dataset). The active feature-value acquisition process is run until the number

of missing features is smaller than the acquisition step size, and clustering performance is evaluated after each acquisition step.

Because the datasets we consider have underlying class labels, they can be used for evaluating clustering quality. As discussed earlier, commonly used external clustering evaluation measures include pairwise F-measure, Rand index, and normalized mutual information. We have found that empirically there are no qualitative differences between these measures, and report results for pairwise F-measure which is defined analogously to non-pairwise F-measure commonly used in information retrieval. Given a clustering and underlying class labels, pairwise precision and recall are estimated as the proportion of same-cluster instances that have the same class, and the proportion of same-class instances that have been placed in the same cluster. Then, F-measure is the harmonic mean of precision and recall: $F1 = (2 \times Precision \times Recall) / (Precision + Recall)$.

The performance comparison for any two acquisition schemes A and B can be summarized by computing the average percentage increase in pairwise F-measure of A over B over all acquisition phases on the learning curve. We refer to this metric as the *average % F-measure increase*. Also, the impact of an effective acquisition policy is most important early on the acquisition curve, where a large number of feature values is missing, since useful acquisitions then have a significant impact on clustering quality. To capture this, we also report the average percentage error reduction over the 20% of points on the learning curve where the largest improvements are observed (Saar-Tsechansky & Provost 2004). We refer to this as the *top-20% average % F-measure increase*.

Results

The results for each dataset are summarized in Table 1, where the two EU variants are compared to random query sampling. Figure 1 presents the results obtained using different utility functions for active feature-value acquisition for K-means on *iris*. The plot shows clustering performance as a function of acquisition costs, obtained with *EU-Perturbation*, *EU-Objective* and random sampling. The plots for other datasets have similar trends.

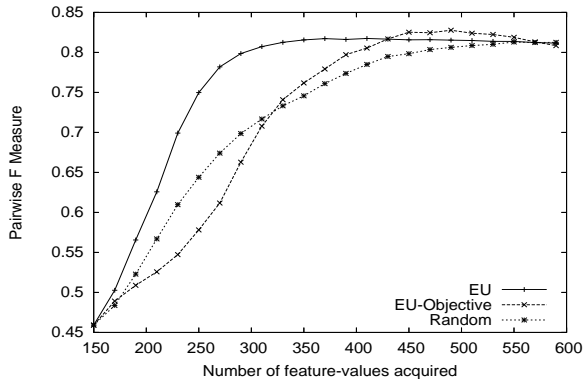


Figure 1: Comparing different utilities for KMeans on *iris*.

The results show that for all data sets *EU-Perturbation* leads to better clustering than random query sampling. The improvements in performance range from a 10% to 29% increase in F-measure on the top 20% of acquisition phases. These results demonstrate that measuring acquisition utility using the expected perturbation of the clustering configuration is very effective for AFA with K-Means, as it identifies missing features which, when acquired, rapidly improve clustering quality. As a result, *EU-Perturbation* selects queries that on average are more informative for the clustering algorithm than queries selected uniformly at random.

To judge the effectiveness of using *EU-Perturbation*, one can also observe the cost benefits of using EU to obtain a desired level of performance as compared to using random acquisitions. For example, on the *iris* data set, *EU-Perturbation* achieves a pairwise F-measure of 0.8 with less than 300 feature values, while random sampling requires twice as many acquisitions to achieve the same result. When data-acquisition costs are non-trivial, this translates to substantial savings in the cost of effective clustering.

Table 1: Performance of *EU-Perturbation* and *EU-Objective* with respect to Random

| Dataset | % F-measure Increase | | Top-20% % F-measure Increase | |
|-------------|----------------------|----------------|------------------------------|----------------|
| | <i>EU-Obj</i> | <i>EU-Pert</i> | <i>EU-Obj</i> | <i>EU-Pert</i> |
| iris | -0.81 | 6.19 | 2.89 | 14.81 |
| wine | -8.42 | 10.92 | -1.73 | 19.25 |
| letters-ijl | 0.47 | 6.22 | 2.58 | 9.63 |
| protein | 4.78 | 14.19 | 10.03 | 28.87 |

In contrast to *EU-Perturbation*, using the objective-based utility function in *EU-Objective* is rather ineffective in improving pairwise F-measure. *EU-Objective* can sometimes do better than random sampling, but in general it has no advantage or performs worse. By using the objective-based utility measure, *EU-Objective* selects feature values that are most likely to reduce the K-means objective function. However, as discussed before, feature values that reduce a clustering objective function may not be the best for improving performance on external metrics. In particular, the K-means objective is focused on producing tighter clusters, and the acquisition strategy based on it may select feature values that reduce this objective without changing any cluster assignments, resulting in no improvement with respect to external evaluation measures.

The results in this section show the effectiveness of the Expected Utility approach to active feature-value acquisition, and also highlight the need for selecting utility measures that correlate well with the desired performance metric. Since the datasets we use have underlying class labels, we find external clustering measures to be more appropriate than objective-value evaluation. As such, for the rest of this paper we will use pairwise F-measure as our performance metric.

Alternative cost distributions

As actual feature-acquisition costs are not available for our datasets, we assumed uniform feature costs for our first set of experiments. However, the goal of our overall approach is to leverage the trade-off between cost and expected improvement in clustering quality. To this end, we tested the performance of our acquisition strategy when applied to the *iris* dataset under different cost distributions. In these experiments, each feature was assigned a cost selected uniformly at random between 1 and 100. We then evaluated performance for 5 such random assignments of feature costs.

Since random feature-acquisition does not take costs into account, we introduced a second baseline method that does. This approach, *Cheapest-first*, selects feature values for acquisition in order of increasing cost. Figure 2 presents plots of F-measure versus acquisition costs for two representative cost distributions. The gains in performance between our approach and random acquisition are greater than those observed with uniform costs. This is because *EU-Perturbation*'s ability to distinguish between uninformative and informative feature-values per unit cost is more critical when there are features of varying information value with non-negligible costs. In such cases, acquiring an uninformative feature-value for a substantial cost results in a significant loss and, as shown, *EU-Perturbation* is more likely to avoid such losses.

In contrast, the performance of *Cheapest-first* has a much higher variance under different cost assignments. When the cost distribution assigns low costs to highly informative features, *Cheapest-first* performs quite well, since its underlying assumption holds — that the cheapest features are also informative. This phenomenon can be seen in the rapid increase in performance of *Cheapest-first* in the early stages of acquisition in Figure 2(a). In such cases, *EU-Perturbation* does not perform as well, since the expected improvement from each acquisition that it computes is an approximation. On the other hand, when many inexpensive features are also uninformative (which is a more realistic scenario), *Cheapest-first* can perform quite poorly, as is demonstrated by the early acquisition stages of Figure 2(b). *EU-Perturbation* however, estimates the trade-off between cost and expected improvement in clustering quality, and although the estimation is imperfect, it consistently selects better queries than random acquisitions for all cost structures.

Hierarchical agglomerative clustering

In this section, we demonstrate the generality of the *EU-Perturbation* approach by using it with an alternative clustering algorithm, hierarchical agglomerative clustering (HAC) (Jain, Murty, & Flynn 1999). HAC is a bottom-up algorithm which initializes clusters with individual instances, and proceeds to iteratively merge clusters that are closest according to a chosen cluster distance measure. Most popular cluster distance measures are single-link, complete-link and group-average distance, which define distance between two clusters as the minimal, maximal, and average distance between all pairs of points that belong to the two clusters. In

this section, we consider group-average HAC, which can be viewed as a trade-off between single-link and complete-link variants of HAC.

The internal objective of group-average HAC is bounded by the maximum variance of the highest-variance cluster, which follows from the probabilistic model-based interpretation provided by Kamvar et al. (2002). Hence, the objective-based utility function for group-average HAC can be approximated by the the difference between total cluster variances for clusterings obtained before and after the acquisition:

$$U_{HAC-GA}^{(obj)}(x_{ij}=x, C_{ij}) = \frac{\sum_{h=1 \dots k} \frac{1}{X_h} \sum_{x_i \in X_h} (x_i - \mu_h)^2}{C_{ij}} \quad (5)$$

where X_h denotes the h -th cluster: $X_h = \{x_i : y_i = h\}$.

As for K-means, we compare this objective-based utility with a perturbation-based utility. Unlike K-Means, HAC is a bottom-up algorithm where point-level perturbation is a sub-optimal utility function; therefore, another complex measure of perturbation that takes the entire cluster hierarchy into account is needed. One such measure of perturbation is the change in the maximum radius R_i of the K clusters before and after a feature-value acquisition:

$$U^{(pert)}(x_{ij}=x, C_{ij}) = \frac{|\max_{i=1}^K R_i^{(X \wedge x_{ij}=x)} - \max_{i=1}^K R_i^{(X)}|}{C_{ij}} \quad (6)$$

Figure 3 demonstrates that both utility measures lead to effective feature acquisition compared to random query selection; and that *EU-Perturbation* is more cost-effective when the evaluation criterion is pairwise F-measure.

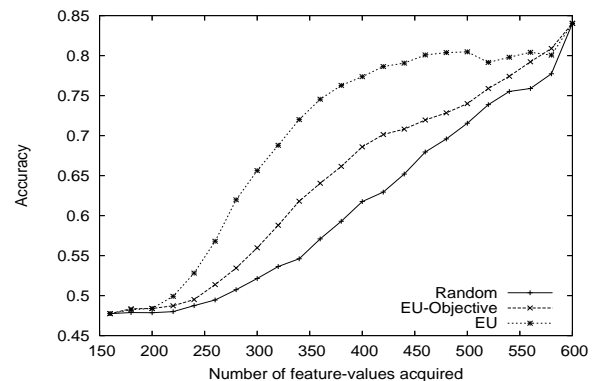


Figure 3: Comparison of acquisition strategies for HAC on *iris*

Related work

The active feature-value acquisition task was studied previously by Melville et al. (2005) and Lizotte et al. (2003) for purely supervised learning settings. The problem of feature-value acquisition that we study is also distinct from traditional active learning (Cohn, Atlas, & Ladner 1994) where class labels rather than feature values are missing. Hofmann et al. (1998) and Buhmann et al. (2000) considered

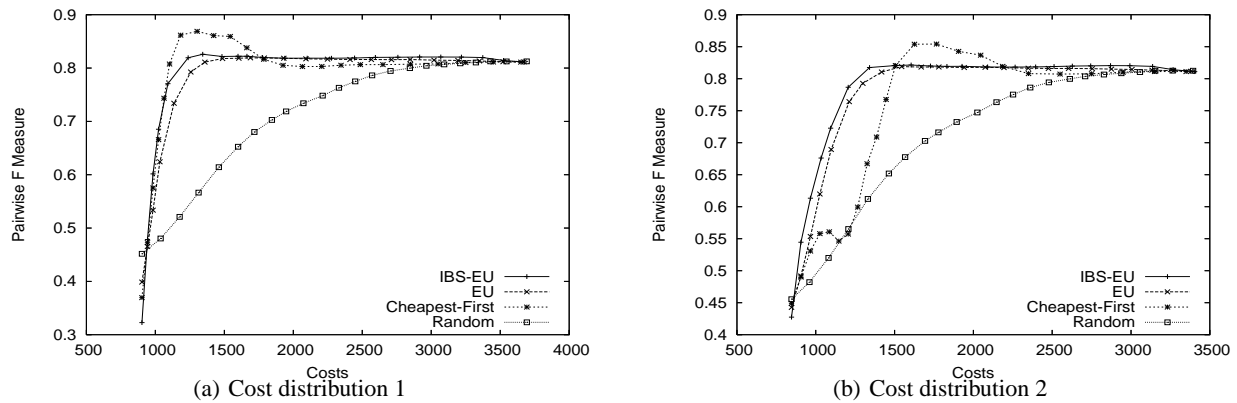


Figure 2: Comparing different algorithms on *iris* under different cost distributions

active learning for clustering in a different setting from ours, where partial information on the pairwise similarities between instances is available. Basu et al. (2004) and Klein et al. (2002) study the task of active learning in the setting of semi-supervised clustering. In addition to unlabeled data, they assumed supervision is available in the form of pairwise *must-link* and *cannot-link* constraints. In contrast to our setting, these studies do not consider instance-feature level queries, and rely on availability of supervised data.

Conclusions

In this paper, we proposed an expected utility approach to active feature-value acquisition for clustering, where informative feature values are obtained based on the estimated expected improvement in clustering quality per unit cost. This approach can be adapted for different clustering algorithms, which we demonstrated by presenting its instantiations for K-means and hierarchical agglomerative clustering. Experiments with uniform feature costs show that the *EU-Perturbation* approach consistently leads to better clustering than random sampling for the same number of feature-value acquisitions. Additional experiments on different cost distributions demonstrate that the performance of the proposed framework is more consistent than that of a simple cost-sensitive method which acquires feature values in order of increasing cost.

References

- Basu, S.; Banerjee, A.; and Mooney, R. J. 2004. Active semi-supervision for pairwise constrained clustering. In *Proc. of the 2004 SIAM Intl. Conf. on Data Mining (SDM-04)*.
- Buhmann, J. M., and Zoller, T. 2000. Active learning for hierarchical pairwise data clustering. In *ICPR*, 2186–2189.
- Cohn, D.; Atlas, L.; and Ladner, R. 1994. Improving generalization with active learning. *Machine Learning* 15(2):201–221.
- Halkidi, M.; Batistakis, Y.; and Vazirgiannis, M. 2001. On clustering validation techniques. *Journal of Intelligent Information Systems* 17(2-3):107–145.
- Hofmann, T., and Buhmann, J. M. 1998. Active data clustering. In *Advances in Neural Info. Processing Systems 10*.
- Jain, A. K.; Murty, M. N.; and Flynn, P. J. 1999. Data clustering: A review. *ACM Computing Surveys* 31(3).
- Kamvar, S. D.; Klein, D.; and Manning, C. D. 2002. Interpreting and extending classical agglomerative clustering algorithms using a model-based approach. In *Proc. of 19th Intl. Conf. on Machine Learning (ICML-2002)*, 283–290.
- Klein, D.; Kamvar, S. D.; and Manning, C. 2002. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proc. of 19th Intl. Conf. on Machine Learning (ICML-2002)*, 307–314.
- Lizotte, D.; Madani, O.; and Greiner, R. 2003. Budgeted learning of naive-Bayes classifiers. In *Proc. of 19th Conf. on Uncertainty in Artificial Intelligence (UAI-2003)*.
- MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symp. on Math. Statistics and Probability*, 281–297.
- Melville, P.; Saar-Tsechansky, M.; Provost, F.; and Mooney, R. 2005. An expected utility approach to active feature-value acquisition. In *Proc. of the Intl. Conf. on Data Mining*, 745–748.
- Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66:622–626.
- Roy, N., and McCallum, A. 2001. Toward optimal active learning through sampling estimation of error reduction. In *Proc. of 18th Intl. Conf. on Machine Learning (ICML-2001)*, 441–448. Morgan Kaufmann, San Francisco, CA.
- Saar-Tsechansky, M., and Provost, F. 2004. Active sampling for class probability estimation and ranking. *Machine Learning* 54:153–178.
- Steinbach, M.; Karypis, G.; and Kumar, V. 2000. A comparison of document clustering techniques. In *Proceedings of the KDD-2000 Workshop on Text Mining*.