Intelligent Information Acquisition for Improved Clustering

Duy Vu
University of Texas at Austin
duyvu@cs.utexas.edu

Prem Melville
IBM T.J. Watson Research Center
pmelvil@us.ibm.com

Mikhail Bilenko
Microsoft Research
mbilenko@microsoft.com

Maytal Saar-Tsechansky
University of Texas at Austin
maytal@mail.utexas.edu

1. Introduction and motivation

In many data mining and machine learning tasks, datasets include instances that have missing feature values that 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. For example, clustering algorithms are commonly used to identify users with similar preferences, so as to produce personalized product recommendations. With instances corresponding to individual consumers and features describing consumers' ratings of a given product/service, individual features of particular instances may be missing as customers may have not provided feedback on all the items they purchased. Furthermore, because consumers are often reluctant to provide feedback, acquiring feedback on unrated items may entail costly incentives, such as free or discounted products or services. However, obtaining different feature values may have varying effect on accuracy of subsequently obtained clustering of consumers. Thus, choosing which ratings to acquire via incentives that will benefit the clustering task most cost-effectively is an important decision --- as acquiring feedback for all missing ratings is prohibitively expensive.

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 focused either on acquiring pairwise distances ([3],[4]) or cluster labels for complete instances [1]. Prior work addressed the AFA task for supervised learning, where missing feature values are acquired in a cost-effective manner for training classification models [6]. However, this approach exploits supervised information to estimate the expected improvement in model accuracy for prospective acquisitions. The primary challenge addressed in this paper lies in a priori estimation of the value of a potential acquisition in the absence of any supervision (i.e., it is not known to which cluster each instance actually belongs).

We employ an expected utility acquisition framework and present an instantiation of our overall framework for K-means, where the value of prospective acquisitions is derived from their expected impact on the clustering configuration (see [8] for an instantiation of our framework for hierarchical agglomerative clustering algorithm). Empirical results demonstrate that the proposed utility function effectively identifies acquisitions that improve clustering quality per unit cost significantly better than acquisitions selected uniformly at random. In addition, we show that our policy performs well for different feature cost structures.

2. Task definition and algorithm

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. We focus our attention on clustering in domains where instances include missing feature values that can be acquired at a cost. A

dataset consisting of *m* *n*-dimensional instances is represented 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 quality per unit cost.

The overall framework for the generalized AFA problem is presented in Algorithm 1. Information is acquired iteratively, where at each step all possible queries are ranked 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. While this framework is intuitive, the crux of the problem lies in devising effective measures for the utility of acquisitions. In subsequent sections, we address challenges related to performing this task accurately and efficiently.

Algorithm 1: Active Feature-value Acquisition for Clustering

Given: X – initial (incomplete) instance-feature matrix, L – clustering algorithm, b – size of query batch, C – cost matrix for all instance-feature pairs. Output: M = 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 = {qij : xij is missing}.
3. Repeat until stopping criterion is met
4. Generate a clustering M = L(X)
5. $\forall q_{ij} \in Q$ compute utility score
6. Select a subset S of b queries with the highest scores
7. $\forall q_{ij} \in S$: Acquire values for $x_{ij} : X = X \wedge x_{ij}$. TotalCost = TotalCost + C_{ij}
8. Remove *S* from *Q*
9. Return M = L(X)

At every step of the AFA algorithm, the feature value which in expectation will result in the highest clustering improvement per unit cost is acquired. Fundamental to our approach is a utility function $U\left(x_{ij} = x, C_{ij}\right)$ which quantifies the benefit from a specific value *x* for feature x_{ij} acquired via the corresponding query q_{ij} at cost C_{ij}. Then, expected utility for query q_{ij}, $EU(q_{ij})$, is defined as the expectation of the utility over the marginal distribution for the feature x_{ij}: $EU(q_{ij}) = \int_x U(x_{ij} = x, C_{ij}) \cdot P(x_{ij} = x)$. Since the true marginal distribution of each missing feature value is unknown, an empirical estimate of $P(x_{ij} = x)$ can be obtained using probabilistic classifiers. For example, in the case of discrete (categorical) data, for each feature *j*, a näive 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, the expectation can be easily computed by piecewise summation over the possible values. For continuous attributes, computation of expected utility can be performed either using computational methods such as Monte Carlo estimation, or via discretizing them and using probabilistic classifiers as described above.

2.1 Capturing the utility from a prospective acquisition

Devising a utility function *U* to capture the benefits of possible acquisition outcomes is the critical component of the AFA framework. Acquisitions aim to improve clustering *quality*. Clustering quality measures proposed in prior work can be loosely divided into external measures, such as pairwise F-measure [7], which are derived from a category distribution unseen at clustering time, and internal measures, e.g., ratio between average inter-cluster and intra-cluster distances, which use only data that is available to the clustering algorithm. Since external measures cannot be assessed at the time of clustering, an acquisition policy must capture the value of acquisitions using merely the dataset at hand.

Most clustering algorithms optimize a specific objective function, which allows defining utility as improvement in this objective per unit cost. For example, the objective of the popular K-Means algorithm [5] is to minimize the sum of squared distances between every instance x_i and the centroid of the instance's cluster, $\mu_{y_i} : J_{KM}(X) = \sum_{x_i \in X} \left(x_i - \mu_{y_i}\right)^2$, 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. Thus, the objective-based utility from acquisition outcome $x_{ij} = x$ can be defined as the cost-normalized reduction in the value of the objective function: $U_{KM}^{Obj}(x_{ij}, C_{ij}) = \frac{J_{KM}(X \wedge x_{ij} = x) - J_{KM}(X)}{C_{ij}}$, where the objective function value after the acquisition $J_{KM}(X \wedge x_{ij} = x)$ is estimated following the relocation of cluster centroids caused by the acquisition.

While an objective-based utility function provides a well-motivated acquisition strategy, it may select feature values that improve cluster centroid locations without significantly changing cluster *assignments* which often underlie external measures of clustering outcome. The effect of such wasteful acquisitions can be significant, rendering an objective-based utility a suboptimal strategy for improving external evaluation measures.

Because internal objective functions may not relate well to external measures we propose an alternative utility measure which approximates the *qualitative* impact on clustering *configuration* caused by the acquisition. We define this utility as the number of instances for which cluster *membership changes* as the result of an acquisition, given a certain value of the acquired feature. 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, the perturbation-based utility of acquiring value x for feature x_{ij} is defined as follows: $U^{Pert}(x_{ij} = x, C_{ij}) = \frac{\sum_{i=1}^{M} y_i^{(X \wedge x_{ij} = x)} \neq y_i^{(X)}}{C_{ij}}$. For K-Means, the cluster assignments after the acquisition, $Y_i^{(X \wedge x_{ij} = x)} = \left\{ y_i^{(X \wedge x_{ij} = x)} \right\}_{i=1}^{M}$ can be obtained by re-estimating the cluster centroid to which instance x_i is currently assigned, assuming the value *x* for feature x_{ij}. Then, performing a single assignment step for all points would provide the new set of cluster assignments $Y^{(X \wedge x_{ij} = x)}$. As we show below, this utility measure identifies highly informative acquisitions. Henceforth, we refer to this perturbation-based utility as Expected Utility (EU); we refer to the use of the objective-based utility as Expected-Utility-Objective (EU-Objective).

2.2 Efficiency considerations: Instance-based sampling

A significant challenge lies in the fact that exhaustively evaluating all potential acquisitions is computationally infeasible for datasets of even moderate size. We propose to 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 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. The value of α can be set depending on the amount of time the user is willing to spend on this process. One approach is to draw this sample uniformly at random to make the computation feasible. However, it may be possible to improve performance by applying Expected Utility estimation to a particularly *informative* sample of queries. In particular, because the goal of clustering is to define *boundaries* between potential classes, instances near these boundaries have the most impact on cluster formation. Consequently, missing features of these instances give us the most decisive information to adjust the clustering boundaries. Formally, if μ_{y_i} and μ' are respectively the closest and second closest centroids for instance x_i in the current clustering, we define the margin $\delta(x_{ij})$ of instance x_i as the difference between their distances from x_i, according to the distance metric D being used for clustering: $\delta(x_{ij})$ = D(x_i,μ') - D(x_i,μ_{y_i}). Given incomplete information about the position of instances in the feature space, smaller margins for instances correspond to lower confidence in their current cluster assignment. For these instances, obtaining a better estimate of their position in the feature space is more likely to improve our ability to assign them to the correct cluster than for instances with large margins. Following this rationale, we rank all *instances* in ascending order of their margins based on the current cluster assignments. Then, a set of *αb* queries from the top-ranked instances are selected for evaluation; where *b* is the desired batch size and *α* is the exploration parameter. This candidate set of queries is then subjected to the same expected utility sampling described in the previous section. We refer to this approach as Instance-Based Sampling Expected Utility (IBS-EU).

3. Experimental evaluation

We evaluated our proposed approach on four datasets from the UCI repository [2]: *iris*, *wine*, *letters-ijl*, and *protein*, which have been previously used in a number of clustering studies. Features with continuous values in these datasets were discretized into 10 bins of equal width. Since feature acquisition costs are not available for these datasets, in our first set of experiments, we assume that acquisition costs are uniform for all feature values followed by experiments for other cost distributions. Discrete feature values enable the use of a piecewise summation for the expectation calculation and is computationally preferable. However, in principle, continuous values can also be used. We compare the proposed acquisition policies with a strategy that selects queries uniformly at random, and using the K-means clustering algorithm. The sampling parameter *α* of our methods is set to 10. We report results obtained from 100 runs for each active acquisition policy. In each run, a small fraction of features is randomly selected for initialization for each instance in the dataset[1], and we evaluate clustering performance after each acquisition step. Lastly, because the datasets we consider have underlying class labels, we employ an external metric, pairwise F-measure, to evaluate clustering quality. We have found that empirically there are no qualitative differences in our results for different external measures. Given a clustering and underlying class labels, pairwise precision and recall are defined 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, respectively. 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 the average percentage increase in pairwise F-measure of A over B over all acquisition phases. We refer to this metric as the average % F-measure increase.

4. Results

Table 1 presents summary results for EU, EU-Objective, IBS-EU, and IBS-Random, which acquires feature values drawn uniformly at random from informative instances selected by IBS. Let us

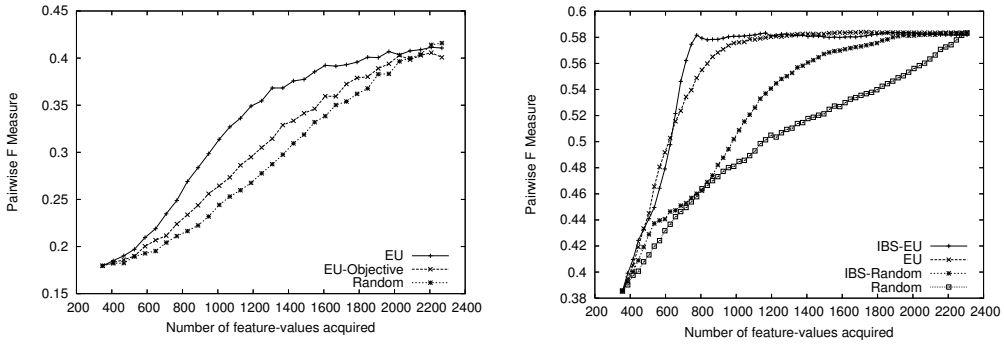
[1]We randomly selected 1 out of 4 features for each instance in the *iris* dataset, 2 out of 4 features for *wine*, and 3 out of 16 and 20 features for the *letter-ijl* and *protein* data sets, respectively.

first examine the relative performance of the EU policy which identifies acquisitions that are likely to impact the cluster assignments and EU-Objective which targets acquisitions which are expected to improve the clustering algorithm's internal objective function. Figure 1(a) presents clustering performance as a function of acquisition costs for the *protein* data set, obtained with EU, EU-Objective, and random sampling.

For all data sets, EU leads to better clustering than random query sampling. The improvements in performance range from a 10% to 32% increase in F-measure on the top 20% of acquisition phases. One can also observe the cost benefits of using EU to obtain a desired level of performance. For example, on the *iris* data set, Expected Utility 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. In contrast to Expected Utility, using the objective-based utility function in EU-Objective is rather ineffective in improving pairwise F-measure. This is because 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.

Data Set	% F-measure Increase over Random			
	EU-Objective	EU	IBS-EU	IBS- Random
iris	-0.81	6.19	7.96	3.14
wine	-8.42	10.92	11.41	4.58
letters	0.47	6.22	5.55	0.16
protein	4.78	14.19	14.93	2.90

Table 1: Performance of different acquisition policies for clustering



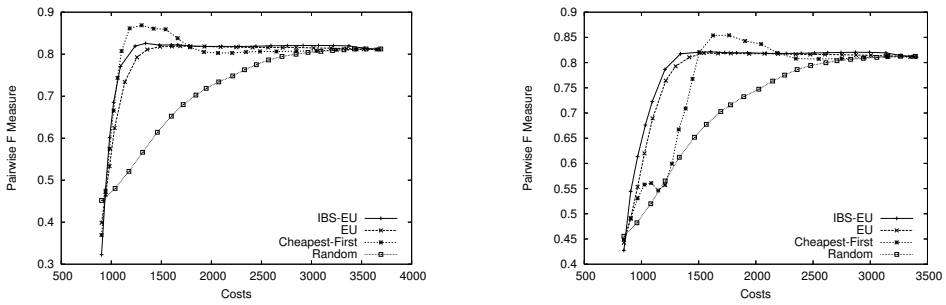
(a) (b)

Figure 2: Learning curves for alternative acquisition policies

Now, let us examine the benefit to EU from evaluating a subset of acquisitions from particularly informative instances as captured by our Instance-Based sampling approach. Table 1 presents summary performance for EU-IBS and IBS-Random, and for the *iris* data set, Figure 1(b) show clustering quality after each acquisition phase obtained by EU, IBS-EU, and IBS-Random. On 3 of the 4 datasets, IBS-EU produces the highest average increase in pairwise F-measure compared to random sampling. On these datasets, IBS-Random also performs substantially better than random. These results demonstrate that our margin measure effectively identifies particularly informative instances for acquisition. Consequently, IBS-EU focuses the evaluation of Expected Utility to a more promising set of queries, leading to better models on average. However, the improvements of IBS-EU over EU are not very large.

Lastly, we evaluated the policies when applied to the *iris* dataset under different cost distributions. We assigned each feature a cost drawn uniformly at random from a range between 1 and 100. For this evaluation we include a cost-sensitive benchmark policy, Cheapest-first, which selects acquisitions in order of increasing cost. The results for all randomly assigned cost distributions show that

IBS-EU and Expected Utility consistently results in better clustering than random acquisition for a given cost. Figure 3 presents F-measure versus acquisition costs for two representative cost distributions. As shown, in settings where features have varying information value with non-negligible costs, EU's ability to capture the value of different feature values per unit cost is more critical. In such cases, acquiring an uninformative feature value for a substantial cost results in a significant loss and, as shown, EU and IBS-EU are more likely to avoid such losses. In contrast, the performance of Cheapest-first is inconsistent. It performs well when its underlying assumption holds and the cheapest features are also informative. In such cases, EU does not perform as well, since it imperfectly estimates the expected improvement from each acquisition. When many inexpensive features are also uninformative Cheapest-first can perform poorly, as shown by the early acquisition stages of Figure 3. EU, 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.



(a) Inexpensive features are also informative (b) Some expensive feature are informative

Figure 3: Performance under different feature-value cost structures

5. 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. Experiments show that the Expected Utility approach consistently leads to better clustering than random sampling for a given acquisitions cost.

6. References

[1] S. Basu, A. Banerjee, and R. J. Mooney. Active semi-supervision for pairwise constrained clustering. In Proceedings of the 2004 SIAM International Conference on Data Mining (SDM-04), Apr. 2004.

[2] C. L. Blake and C. J. Merz. UCI repository of machine learning databases. http://www.ics.uci.edu/˜mlearn/MLRepository.html, 1998.

[3] J. M. Buhmann and T. Zller. Active learning for hierarchical pairwise data clustering. In ICPR, pages 2186–2189, 2000.

[4] T. Hofmann and J. M. Buhmann. Active data clustering. In Advances in Neural Information Processing Systems 10, 1998.

[5] J. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, pages 281–297, 1967.

[6] P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. An expected utility approach to active feature-value acquisition. In Proceedings of the International Conference on Data Mining, pages 745–748, Houston, TX, November 2005.

[7] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In Proceedings of the KDD-2000 Workshop on Text Mining, 2000.

[8] D. Vu, M. Bilenko, P, Melville, M. Saar-Tsechansky. “Active information acquisition for improved clustering”, Working Paper, McCombs School of Business, May 2007.