ACTIVE DATA ACQUISITION WITH INCOMPLETE DATA

David Williams, Xuejun Liao, Lawrence Carin

Duke University
Department of Electrical and Computer Engineering
Box 90291, Durham, NC 27708-0291

ABSTRACT

We present a unified framework under which active data acquisition can be performed. This comprehensive framework allows for the acquisition of both labels and features. Moreover, several types of feature acquisition are permitted, including the acquisition of individual or multiple features for individual or multiple data points, which may be either labeled or unlabeled. The algorithm automatically determines the most beneficial type of data to acquire when multiple options exist. The framework also has a principled, intuitive criterion for terminating the active data acquisition process. Experimental results on two real multi-sensor remote-sensing data sets demonstrate the advantages of the proposed approach for sensor deployment tasks.

1. INTRODUCTION

The incomplete-data problem, in which certain features are missing for particular data points, exists in a wide range of fields, including social sciences, computer vision, biological systems, and remote sensing. In many applications involving incomplete data, it is also possible to acquire the missing data at a cost. The fact that acquiring data is usually an expensive, time-consuming task necessitates an intelligent feature acquisition process.

In multi-sensor remote-sensing applications, incomplete data can result when only a subset of physical sensors (*e.g.*, radar, infrared, acoustic) are deployed at certain regions. Data is then acquired by deploying sensors to data points. In this work, we address the sensor deployment problem, answering when, where, and which data-collecting sensors should be deployed to maximize classification performance.

The sensor deployment problem can be viewed as a specific case of an *active feature acquisition* process. In contrast to conventional active learning [1], which selects the most beneficial labels to acquire, this process would intelligently select the most beneficial missing *features* to acquire. The major flaws that plague the scant heuristic active-feature-acquisition approaches in the literature [2–5] stem from two sources: the reliance on complete-data classification algorithms, and the disregard of data acquisition costs. In fact, if no costs are incurred by acquiring additional data, the active acquisition process is moot because all data should be collected. By ignoring the acquisition costs, the methods [2–5] are also forced to adopt *ad hoc* criteria to terminate the active feature acquisition process (*e.g.*, after a fixed, *a priori* number of features have been acquired).

The reliance on complete-data classification algorithms exposes several other shortcomings of methods [2–4] in the active data acquisition literature. First, since only complete data can be handled, not all available data is utilized. Moreover, these methods [2–4] (unrealistically) require complete testing data; the methods cannot handle testing data with missing features. Furthermore, the only capability these methods [2–4] have is to consider acquiring *all* missing features for *single* data points. In general, many different types of data can be acquired: perfect or imperfect labels, as well as individual or multiple features for individual or multiple, labeled or unlabeled data points. In practice, the application at hand will dictate which types of data acquisition are feasible.

In this work, we present a comprehensive active data acquisition framework that allows for the acquisition of both labels and features. The proposed active data acquisition approach does not suffer from any of the numerous limitations that plague the existing methods [2–5] in the literature. Our framework accounts for acquisition costs and has a natural, intuitive termination criterion. Moreover, if different types of data acquisition are feasible for a given application, our framework automatically determines the most beneficial type of data to acquire.

The remainder of this paper is organized as follows. The proposed active data acquisition framework is presented in Section 2. Experimental results are presented in Section 3, followed by concluding remarks in Section 4.

2. ACTIVE DATA ACQUISITION WITH INCOMPLETE DATA

2.1. Notation

Assume we have a set of labeled incomplete data,

$$\mathcal{D}_L = \{(\mathbf{x}_i, y_i, m_i) : \mathbf{x}_i \in \mathbb{R}^d, x_{ia} \quad \text{missing} \quad \forall a \in m_i\}_{i=1}^{N_L}, \tag{1}$$

where $\mathbf{x}_i$ is the i-th data point, labeled as $y_i \in \{-1, 1\}$; the features in $\mathbf{x}_i$ indexed by m_i (*i.e.*, $x_{ia}, a \in m_i$) are missing. Let o_i be the (complementary) set of observed features in $\mathbf{x}_i$. Each $\mathbf{x}_i$ has its own (possibly unique) set of missing features, m_i. Also assume we have a set of unlabeled incomplete data,

$$\mathcal{D}_U = \{(\mathbf{x}_i, m_i) : \mathbf{x}_i \in \mathbb{R}^d, x_{ia} \quad \text{missing} \quad \forall a \in m_i\}_{i=N_L+1}^{N}. \tag{2}$$

Let $\mathbf{w}$ constitute a classifier (*e.g.*, representing weights on the features) that can be learned with incomplete data. Our active data acquisition framework is general in that any incomplete-data classifier that does not impute values for the missing data can be used. In this work, we utilize the incomplete-data logistic regression classifier introduced in [6].

2.2. General Data Acquisition

In addition to the costs involved with acquiring data, there is a different cost due to the misclassification of data, called the *risk*. It is sensible to perform active data acquisition only when the costs of acquiring additional data are outweighed by the benefit accrued. This benefit is simply the reduction in risk resulting from possessing the new data. Thus, the logical objective that should drive active data acquisition is to maximize the expected benefit derived from acquiring additional data. It should be noted that the acquisition and misclassification costs must be in the same units.

Let the (estimated) risk[1] $R(\mathcal{D}_U|\mathcal{D}_L)$ be the risk on the unlabeled data $\mathcal{D}_U$, from using a classifier designed using $\mathcal{D}_L$; it is defined as

$$R(\mathcal{D}_U|\mathcal{D}_L) = \sum_{i=N_L+1}^{N} \min\{C_{[1,-1]}p(y_i = -1|\mathbf{x}_i^{o_i}, \mathbf{w}), C_{[-1,1]}p(y_i = 1|\mathbf{x}_i^{o_i}, \mathbf{w})\} \tag{3}$$

where the weights $\mathbf{w}$ are trained using $\mathcal{D}_L$, and $C_{[a,b]}$ is the cost of misclassifying a data point as belonging to class a instead of the true class b. Let the expected risk on the unlabeled data $\mathcal{D}_U$ after acquiring new data $\mathcal{D}_*$ — which can be features or a label (perfect or imperfect) — be $\mathbb{E}_{\mathcal{D}_*}[R(\mathcal{D}_U|\mathcal{D}_L \cup \mathcal{D}_*)]$. The expected benefit of acquiring data $\mathcal{D}_*$ is then the cost of acquiring the data, $C(\mathcal{D}_*)$, subtracted from the expected decrease in (estimated) risk:

$$\mathbb{E}_{\mathcal{D}_*}[B(\mathcal{D}_*|\mathcal{D})] = -C(\mathcal{D}_*) + \{R(\mathcal{D}_U|\mathcal{D}_L) - \mathbb{E}_{\mathcal{D}_*}[R(\mathcal{D}_U|\mathcal{D}_L \cup \mathcal{D}_*)]\} \tag{4}$$

where $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$. When formulated as in (4), the active data acquisition process has a natural termination criterion: when the expected benefit of all possible acquisitions is no longer positive.

[1]This risk on the unlabeled data is the *estimated* risk because the true labels (which are unavailable) must be known to compute the *true* risk.

2.3. Label Acquisition

If the new data is a label ($\mathcal{D}_* = y_*$), the requisite expectation in (4) can be performed analytically since there are only a finite number — namely two in a binary problem — of possible values the new data can take. Specifically, two classifiers must be built for each unlabeled data point, one for each possible label. The expected benefit is then

$$\begin{aligned}\mathbb{E}_{y_*}[B(y_*|\mathcal{D})] = &-C(y_*) + R(\mathcal{D}_U|\mathcal{D}_L) \\ &- \{p(y_* = +1|\mathbf{x}_*, \mathbf{w})R(\mathcal{D}_U|\mathcal{D}_L \cup \{y_* = +1\}) \\ &+ p(y_* = -1|\mathbf{x}_*, \mathbf{w})R(\mathcal{D}_U|\mathcal{D}_L \cup \{y_* = -1\})\}\end{aligned} \tag{5}$$

where $p(y_* = \pm 1|\mathbf{x}_*, \mathbf{w})$ is obtained from the extant classifier built using $\mathcal{D}_L$.

The label for the data point with the maximum (positive) expected benefit is then acquired. Thereafter, the classifier is re-trained. The process is then repeated until the expected benefit of all possible acquisitions is no longer positive.

2.4. Feature Acquisition

If the new data is a (continuous-valued) feature, computing the new expected risk is intractable. We therefore appeal to an idea motivated by the theory of multiple imputation [7] to compute the expectation. Whereas multiple imputation will impute values for *all* missing data, our method will impute a value for only the single feature[2] under consideration, leaving the other missing data incomplete (and handled by the incomplete-data classifier). This effectively isolates the utility of a single feature. Theoretical work [7] has shown the proximity between an estimate's uncertainty resulting from a small number of imputations and an infinite number of imputations.

In our approach, we impute M samples from the conditional distribution $p(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i})$ (which is estimated as in [6] as a Gaussian mixture model using a Variational Bayesian Expectation-Maximization algorithm); each sample, $\mathcal{D}_*^j$, is a value for the feature under consideration. A classifier is then built using the augmented data set consisting of the original data and one of the imputed values. Each of the imputed values is used in turn so that M unique classifiers are constructed. For each classifier, the (estimated) risk — $R(\mathcal{D}_U|\mathcal{D}_L \cup \mathcal{D}_*^j)$ — can be computed. The requisite expectation from (4) is then approximated as

$$\mathbb{E}_{\mathcal{D}_*}[R(\mathcal{D}_U|\mathcal{D}_L \cup \mathcal{D}_*)] \approx \frac{1}{M}\sum_{j=1}^{M} R(\mathcal{D}_U|\mathcal{D}_L \cup \mathcal{D}_*^j). \tag{6}$$

Thus, the (approximate) expected benefit of acquiring each single missing feature for every data point can be computed.

[2]For concreteness, we explain the method for the case in which the new data is a single feature for a single data point. If the new data are multiple features (rather than a single feature), the same basic framework applies. The only difference is that in this case, values will be imputed for the *group* of features under consideration.

The feature with the maximum (positive) expected benefit is then acquired. Thereafter, the classifier is re-trained. The process is then repeated until the expected benefit of all possible acquisitions is no longer positive.

It should be noted that the above algorithm can be used when considering acquiring features of either labeled or unlabeled data. For a general purely supervised classification algorithm, acquiring an additional feature for an unlabeled data point does not change the classifier, so no classifier retraining must be performed. However, the risk would still change because the unlabeled testing data changes. This observation highlights the fact that active data acquisition can improve performance in two distinct ways: by improving the classifier, or by enhancing the (testing) data to be classified.

3. EXPERIMENTAL RESULTS

We evaluated our proposed active data acquisition algorithm on three different data sets: one synthetic data set and two multi-sensor data sets (UXO and LAND MINE) of real (*i.e.,* measured) data. Prior to beginning data acquisition, we randomly removed features from all data points (both labeled and unlabeled). In the two multi-sensor applications, missing features would result if sensors were deployed to only a subset of data points. To simulate this real problem, for the UXO and LAND MINE data sets, we remove *groups* of features corresponding to the features produced by a sensor, as opposed to randomly removing *individual* features. Data acquisition then corresponds to the deployment of sensors.

The synthetic data set is included to highlight the unified aspect of the proposed framework that allows for the acquisition of both labels and (several different types of) features. For this data set, one action can be the acquisition of a single missing feature for a single labeled or unlabeled data point; the acquisition of *all* missing features for a single labeled or unlabeled data point; the acquisition of a single feature for *all* labeled or unlabeled data points missing it; or the acquisition of a perfect or imperfect label.

The objective of the UXO data set is to discriminate between (*i.e.,* classify) unexploded ordnance (UXO) and clutter. The data set is composed of data from two ground-based sensors: a magnetometer and an electromagnetic induction (EMI) sensor. Because of the nature of the sensors, the only feasible type of data acquisition at each step is the deployment of a single sensor to a single data point (labeled or unlabeled).

The objective of the LAND MINE data set is to discriminate between land mines and clutter. The data set is composed of data collected via four airborne sensors, each mounted on a different aircraft. The four sensors are a ground-penetrating radar (GPR) sensor, an electro-optic infrared (EOIR) sensor, a *Ku*-band synthetic aperature radar (SAR) sensor, and an *X*-band SAR sensor. Because of the airborne nature of the sensors, the only feasible type of data acquisition at each step is the deployment of a single sensor to either *all* training (labeled) data points missing it, or *all* testing (unlabeled) data points missing it.

Table 1 provides details of the three data sets when the data acquisition process commences. An extended version of this paper [8] contains both additional details of the data sets, as well as the data acquisition and misclassification costs used for our experiments. Space limitations prevent inclusion of this information in this paper.

We did not compare our method to the other active data acquisition methods [2–5] in the literature because those heuristic methods do not account for costs, cannot handle incomplete testing data, do not have principled termination criteria, and can only perform one type of data acquisition (all missing features of a single data point). In all of the experiments, we instead compared our proposed active data acquisition method to randomly selecting which data to acquire. Our active data acquisition method terminates automatically when the expected benefit of all possible acquisitions is no longer positive. For the random data acquisition, the same number of actions was taken as in the corresponding active data acquisition case. The random method is also given the advantage that no regard is paid to costs. To compute the expected benefit of potential acquisitions, $M = 1$ imputation was always used (see Section 2). The incomplete-data classifier of [6] is used in all experiments.

The area under a receiver operating characteristic curve (AUC) is given by the Wilcoxon statistic [9]

$$\text{AUC} = (MN)^{-1} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathbf{1}_{x_m > y_n} \tag{7}$$

where $x_1, \ldots, x_M$ are the classifier outputs of data belonging to class 1, $y_1, \ldots, y_N$ are the classifier outputs of data belonging to class -1, and $\mathbf{1}$ is an indicator function. We present the results of the active and random data acquisition algorithms in terms of the AUC.

Experimental results of the data acquisition process are summarized in Table 1. The performance progression of the data acquisition process for each of the three data sets is shown in Figure 1. As can be seen from Table 1 and Figure 1, the proposed active data acquisition process achieves significantly better performance than random data acquisition.

4. CONCLUSION

Our main contribution is the development of a comprehensive unified framework under which active data acquisition can be performed. This framework is the first that allows for the acquisition of both features and labels. When faced with multiple possible types of data acquisition, our approach automatically determines which type of acquisition is the most beneficial. Because every possible type of data acquisition can be handled using this method, the algorithm is suitable for any real application. Furthermore, our proposed method

Table 1. Details of the three data sets when the data acquisition process begins, and experimental results. In all experiments, the performance of the random acquisition cases is the mean value $\pm$ one standard deviation over twenty independent trials.

Data Set	Number of Data Points		Number of Features	Fraction of Features Missing	Final AUC with	
	Labeled	Unlabeled			Active Data Acquisition	Random Data Acquisition
Synthetic	20	180	3	0.43	0.9008	0.7975 ± 0.0522
UXO	25	224	6	0.36	0.8156	0.7561 ± 0.0431
Land Mine	36	677	41	0.48	0.7105	0.5657 ± 0.0984

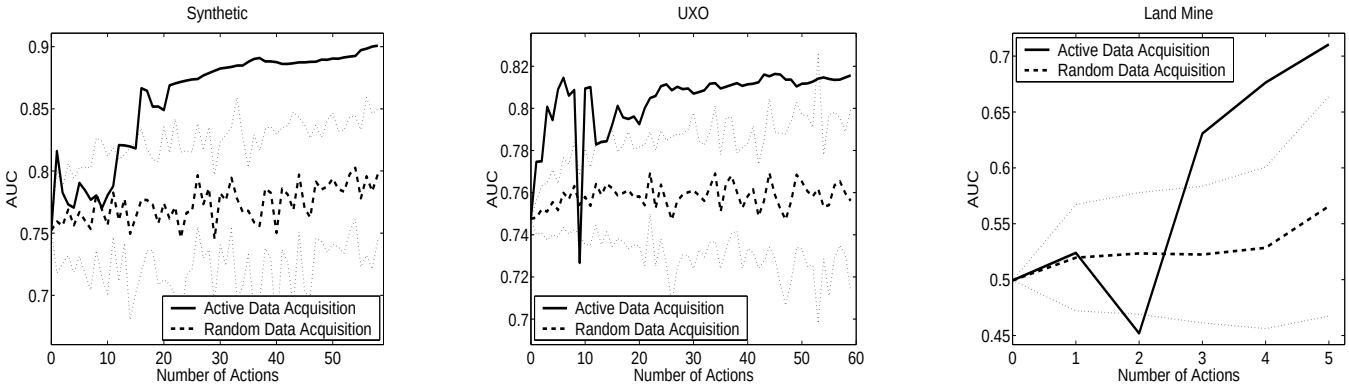

Fig. 1. The progression of the active data acquisition process from Table 1 for the (a) Synthetic, (b) UXO, and (c) Land Mine data sets. Details indicating the type of acquisition that was chosen (automatically) at each step of the active data acquisition cases can be found in [8]. Thin dotted curves in the figures represent one standard deviation about the mean value of the random data acquisition cases.

accounts for data acquisition costs and has a principled termination criterion. Our proposed approach also overcomes several other aforementioned limitations of existing methods. Future work will focus on techniques to ease the computational burden of classifier re-training in the algorithm.

5. REFERENCES

[1] D. Cohn, Z. Ghahramani, and M. Jordan, "Active Learning with Statistical Models," *J. Artificial Intelligence Research*, 4, pp. 129-145, 1996.

[2] P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney, "Active Feature-Value Acquisition for Classifier Induction," *Proc. Int'l Conf. Data Mining (ICDM)*, pp. 483-486, 2004.

[3] Z. Zheng and B. Padmanabhan, "On Active Learning for Data Acquisition," *Proc. Int'l Conf. Data Mining (ICDM)*, pp. 562-570, 2002.

[4] X. Zhu and X. Wu, "Data Acquisition with Active and Impact-Sensitive Instance Selection," *Proc. Int'l Conf. Tools with Artificial Intelligence (ICTAI)*, pp. 721-726, 2004.

[5] B. Krishnapuram, D. Williams, Y. Xue, L. Carin, M. Figueiredo, and A. Hartemink, "Active Learning of Features and Labels," *Proc. Int'l Conf. Machine Learning Workshop on Learning with Multiple Views*, pp. 43-50, 2005.

[6] D. Williams, X. Liao, Y. Xue, and L. Carin, "Incomplete-Data Classification using Logistic Regression," *Proc. Int'l Conf. Machine Learning (ICML)*, pp. 977-984, 2005.

[7] D. Rubin, *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley, 1987.

[8] D. Williams, X. Liao, and L. Carin, "Active Data Acquisition with Incomplete Data," Duke University, Department of Electrical and Computer Engineering, Technical Report, 2005. Available at www.duke.edu/$\sim$dpw5.

[9] J. Hanley and B. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology 143*, pp. 29-36, 1982.