Active Learning for Interactive Visualization

Tomoharu Iwata
University of Cambridge

Neil Houlsby
University of Cambridge

Zoubin Ghahramani
University of Cambridge

Abstract

Many automatic visualization methods have been proposed. However, a visualization that is automatically generated might be different to how a user wants to arrange the objects in visualization space. By allowing users to relocate objects in the embedding space of the visualization, they can adjust the visualization to their preference. We propose an active learning framework for interactive visualization which selects objects for the user to relocate so that they can obtain their desired visualization by re-locating as few as possible. The framework is based on an information theoretic criterion, which favors objects that reduce the uncertainty of the visualization. We present a concrete application of the proposed framework to the Laplacian eigenmap visualization method. We demonstrate experimentally that the proposed framework yields the desired visualization with fewer user interactions than existing methods.

1 Introduction

With the emergence of large and high dimensional data sets, the task of data visualization has become increasingly important in both machine learning and data mining. Visualization is helpful for analyzing and exploring large-scale complex data; it allows one to combine human abilities, such as visual perception, creativity and general knowledge, with the abilities of machines, that is large memories and fast calculation, to the task of understanding data (Keim et al., 2002). One application of visualization is in information retrieval, where users can search objects intuitively in the visualization space (Venna et al., 2010).

A large number of visualization methods have been proposed, such as multi-dimensional scaling (Torgerson, 1958), Isomap (Tenenbaum et al., 2000), locally linear embedding (Roweis and Saul, 2000), stochastic neighbor embedding (Hinton and Roweis, 2002), and Laplacian eigenmap (Belkin and Niyogi, 2003). These algorithms map objects from a high dimensional observation space to a low dimensional 'visualization space'. They find an embedding such that objects in the visualization space preserve their pairwise distances from the high-dimensional observation space. Therefore similar objects are automatically located closer together in the visualization space. However, a visualization that is generated automatically in such a manner may differ from the user's desired visualization who may want to locate objects with a particular meaning to particular areas of visualization space. For example, when visualizing images the user may desire clusters of images of animals to be located in one region of visualization space, inanimate objects in another, and sceneries in another. Alternatively if the objects exhibit a natural ordering, such as digits or letters, then the user may wish to preserve this ordering in visualization space.

To address this problem, interactive visualization systems have been proposed (Wills, 1999; Johansson and Johansson, 2009; Paulovich et al., 2011; Endert et al., 2011). Here, we consider interactive systems in which users can re-locate objects to obtain their desired visualization. When there are a large number of objects it is difficult for users to select which objects to relocate; if many of the moves are redundant, that is, they provide no new information about the user's desired visualization, then even after many queries the visualization may not reflect the intended result.

The goal of this paper is to select objects to re-locate so that the user can obtain their desired visualization by moving as few as possible. For this purpose we propose an active learning framework for visualization. Active learning (Cohn et al., 1996) is a machine learning framework for selecting objects that improve performance with minimum possible labelings. Active learning methods are useful when the cost for obtaining la-

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31. Copyright 2013 by the authors.

beled data is high. Most active learning algorithms were proposed in supervised learning settings (McCallum and Nigam, 1998; Tong and Koller, 2002). We develop an information theoretic active learning criterion that selects objects to re-locate so as to reduce the uncertainty of the visualization the most.

We present our proposed active visualization framework with the widely used Laplacian eigenmap method for nonlinear dimensionality reduction and visualization (Belkin and Niyogi, 2003). Here, we can analytically calculate the objective function for selecting objects, permitting the proposed algorithm to be used in a fast, online, interactive system. Note, however, that we can use many other visualization methods within our framework.

The paper is organized as follows: In Section 2 we propose an active learning framework for visualization based on an information theoretic criterion. In Section 3 we present an implementation of the proposed framework with the Laplacian eigenmap visualization method. In Section 4 we outline related work. In Section 5 we demonstrate the effectiveness of the proposed framework by comparing to existing methods. Finally, we present concluding remarks and a discussion of future work in Section 6.

2 Active Visualization

The task of visualization is, given a set of observations $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$, to find an embedding $\mathbf{Y} = \{\mathbf{y}_n\}_{n=1}^N$ that reveals structure in the data when viewed by the user. Here, $\mathbf{x}_n \in \mathcal{R}^D$ is the feature vector of object n in the observation space, and $\mathbf{y}_n \in \mathcal{R}^K$ is the location of object n in the visualization space. Normally the observation dimensionality is much higher than the visualization dimensionality $D \gg K$. The visualization dimensionality is typically $K = 2$ or $K = 3$; although this constraint arises from our technical ability to view objects in higher dimensional space, the proposed framework is mathematically and computationally applicable with any visualization dimensionality.

In an active learning setting, the algorithm sequentially selects objects for the user to re-locate in $\mathcal{R}^K$, from a given set of N objects. The ground truth locations for the selected objects are obtained from an oracle, i.e. the user, who places the objects within an interactive visualization environment. Given the desired location of the selected object obtained from user feedback, the system changes the visualization of all of the objects, incorporating the new information.

Let $\mathbf{Y}_\mathsf{s}$ be the data that has been labeled by the user, or the set of locations of the selected objects that are associated with the ground truth locations in visualization space, and $\mathbf{Y}_\mathsf{u} = \mathbf{Y} \setminus \mathbf{Y}_\mathsf{s}$ be the unlabeled data, or the set of locations of the unselected objects. The information theoretic approach to active learning selects objects that reduce the uncertainty about the parameters, measured by Shannon's entropy (Cover et al., 1991; Lindley, 1956). In the context of visualization, the variables of interest are the locations of the unlabeled data in visualization space, we may think of these as the 'parameters' about which we want optimally using active learning. Therefore, the objective is to select an object i from the pool of unlabeled data that maximizes the decrease in the entropy of our distribution over the locations of the remaining unlabeled data as follows:

$$\arg\max_i H[p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{Y}_\mathsf{s})] - \mathbb{E}_{p(\mathbf{y}_i|\mathbf{Y}_\mathsf{s})} H[p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{y}_i, \mathbf{Y}_\mathsf{s})], \tag{1}$$

where $\mathbf{Y}_{\mathsf{u}\setminus i}$ is unlabeled data excluding object i, $H[p(\cdot)]$ represents differential entropy of the probability distribution p, and $\mathbb{E}_{p(\cdot)}$ represents expectation under distribution p. For notational simplicity we omit the set of observations $\mathbf{X}$ from the conditioning of all of the probability distributions, e.g. $p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{Y}_\mathsf{s})$ should read $p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{Y}_\mathsf{s}, \mathbf{X})$. The first term,

$$H[p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{Y}_\mathsf{s})] = -\int p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{Y}_\mathsf{s}) \log p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{Y}_\mathsf{s}) d\mathbf{Y}_{\mathsf{u}\setminus i}, \tag{2}$$

is the entropy of the distribution over the unlabeled data given the labeled data; that is, it represents the system's uncertainty in the location of the unlabeled data in the visualization space. The second term,

$$\begin{aligned} &\mathbb{E}_{p(\mathbf{y}_i|\mathbf{Y}_\mathsf{s})} H[p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{y}_i, \mathbf{Y}_\mathsf{s})] \\ &= -\int p(\mathbf{y}_i|\mathbf{Y}_\mathsf{s}) \int p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{y}_i, \mathbf{Y}_\mathsf{s}) \\ &\quad \times \log p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{y}_i, \mathbf{Y}_\mathsf{s}) d\mathbf{Y}_{\mathsf{u}\setminus i} d\mathbf{y}_i, \end{aligned} \tag{3}$$

is the entropy of the distribution over the unlabeled objects after obtaining the true location of object i, where we take its expectation over the location of object to be queried, $\mathbf{y}_i$, because we do not know yet its true location. Further discussion for the exact form of (1) is given in Section 4.

We can gain useful intuition about (1) by rearranging the objective function as follows:

$$\arg\max_i H[p(\mathbf{y}_i|\mathbf{Y}_\mathsf{s})] - \mathbb{E}_{p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{Y}_\mathsf{s})} H[p(\mathbf{y}_i|\mathbf{Y}_{\mathsf{u}\setminus i}, \mathbf{Y}_\mathsf{s})], \tag{4}$$

where we use an insight that the objective in (1) is equivalent to the mutual information between the unlabeled data and the location of the selected object, $\mathbf{I}(\mathbf{Y}_{\mathsf{u}\setminus i}, \mathbf{y}_i)$, given the labeled data. The first term in

(4) favors objects about which we have high uncertainty; this term alone corresponds to a classic objective known as 'uncertainty sampling' or 'maximum entropy sampling' (Sebastiani and Wynn, 2000). The second term has a separate role; it penalizes objects that have high entropy if all of the remaining objects $\mathbf{Y}_{\mathsf{u}\backslash i}$ were observed. This means that if the unobserved objects were seen, we would be confident in the location of object i, alternatively put, the term favors objects that are highly correlated with the remaining unobserved objects. In summary (4) seeks objects about whose location in visualization space we are uncertain, but also correlate with the remaining unlabeled objects, and hence their label provides information about the other unlabeled points' locations also.

If we know that we will query the user with a number of objects $\mathbf{J}$, then it is optimal to maximize our querying strategy over the entire set $\mathbf{J}$. When we select multiple objects to place the objective function becomes

$$\arg\max_{\mathbf{J}} H[p(\mathbf{Y}_{\mathsf{u}\backslash\mathbf{J}}|\mathbf{Y}_{\mathsf{s}})] - \mathbb{E}_{p(\mathbf{Y}_{\mathbf{J}}|\mathbf{Y}_{\mathsf{s}})} H[p(\mathbf{Y}_{\mathsf{u}\backslash\mathbf{J}}|\mathbf{Y}_{\mathbf{J}}, \mathbf{Y}_{\mathsf{s}})]. \quad (5)$$

However, for our active learning criterion, and sequential decision making tasks in general, this problem is NP-hard. As is common in active learning we take a myopic, or greedy approach, performing optimization of (4) assuming that each query is the last. However, the mutual information function is submodular, and the myopic strategy is known to perform near-optimally for submodular functions (Guestrin et al., 2005; Dasgupta, 2005; Golovin and Krause, 2010). Intuitively, this means that it satisfies the property of 'diminishing returns'; that is the gain in information when adding new labeled data point to a smaller pool of observations $\mathbf{Y}_{\mathsf{s}}^{\text{small}}$ is greater than, or equal to, the gain in information when adding the data point to a larger pool $\mathbf{Y}_{\mathsf{s}}^{\text{large}}$.

3 Laplacian eigenmap based active visualization

We present the procedures of our active learning framework for use with the Laplacian eigenmap visualization method (Belkin and Niyogi, 2003). The Laplacian eigenmap is widely used for dimensionality reduction, and it benefits from having a criterion for visualization which can be globally optimized. In this setting we can analytically calculate the objective function for selecting objects to re-locate with our active learning framework.

3.1 Laplacian eigenmap

We outline first the original Laplacian eigenmap algorithm. The Laplacian eigenmap is a nonlinear dimensionality reduction method that has locality-preserving properties based on spectral techniques.

Firstly, a k-nearest neighbor graph is constructed by using observations $\mathbf{X}$ based on the Euclidean distance. One may also use a ϵ-neighborhood graph instead of k-nearest neighbor graph.

Secondly, we set the weight between objects i and j so that $w_{ij} = 1$ if they are connected, $w_{ij} = 0$ otherwise.

Finally, embedding locations $\mathbf{Y}$ that minimize the following function are obtained by solving a generalized eigenvalue problem,

$$\arg\min_{\mathbf{Y}} \mathsf{tr}(\mathbf{Y}^\top \mathbf{L} \mathbf{Y}), \qquad \mathsf{s.t.} \mathbf{Y}\mathbf{D}\mathbf{Y}^\top = \mathbf{I}, \quad (6)$$

where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sum_j w_{ji}$, $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix of $\mathbf{W}$, and $\mathbf{W}$ is an $N \times N$ matrix whose element is w_{ij}.

3.2 Probabilistic interpretation

In order to employ our active learning framework we need a probabilistic interpretation of the Laplacian eigenmap from which we can calculate the relevant entropies and expectations. We make the Laplacian positive definite by adding a small diagonal matrix $\mathbf{\Lambda} = \mathbf{L} + \alpha \mathbf{I}$. When the noise level α is small, we can approximate the minimization of the objective function for the Laplacian eigenmap, $\mathsf{tr}(\mathbf{Y}^\top \mathbf{L} \mathbf{Y})$, by maximizing the likelihood of the following Gaussian distribution:

$$p(\mathbf{Y}) = \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^{-1}), \quad (7)$$

where $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Lambda}^{-1})$ represents a Gaussian with mean $\boldsymbol{\mu}$ and precision, or inverse covariance, $\mathbf{\Lambda}$. The relation between the graph Laplacian and Gaussian Markov random fields is further discussed in (Zhu et al., 2003b).

3.3 Active visualization

We present the procedures of our active learning framework based on objective function (4) with the probabilistic interpretation of the Laplacian eigenmap. Without loss of generality we sort the location vector $\mathbf{Y}$ into labeled then unlabeled data. We then partition the precision matrix $\mathbf{\Lambda}$ into four parts corresponding to the labeled and unlabeled data as follows:

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_{\mathsf{ss}} & \mathbf{\Lambda}_{\mathsf{us}} \\ \mathbf{\Lambda}_{\mathsf{su}} & \mathbf{\Lambda}_{\mathsf{uu}} \end{pmatrix}. \quad (8)$$

By using the fact that

$$p(\mathbf{y}_i|\mathbf{Y}_\mathsf{s}) = \mathcal{N}(-(\mathbf{\Lambda}_\mathsf{uu}^{-1}\mathbf{\Lambda}_\mathsf{us}\mathbf{Y}_\mathsf{s})_{ii}, (\mathbf{\Lambda}_\mathsf{uu}^{-1})_{ii}), \tag{9}$$

and that the entropy of a Gaussian with dimensionality K is

$$H[\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Lambda}^{-1})] = -\frac{1}{2}\log|\mathbf{\Lambda}| + \frac{K}{2}(\log(2\pi)+1), \tag{10}$$

the first term of (4) is obtained by

$$H[p(\mathbf{y}_i|\mathbf{Y}_\mathsf{s})] = \log|(\mathbf{\Lambda}_\mathsf{uu}^{-1})_{ii}| + \frac{K}{2}(\log(2\pi)+1). \tag{11}$$

Similarly, the second term is obtained by

$$\begin{aligned}&\mathbb{E}_{p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{Y}_\mathsf{s})}H[p(\mathbf{y}_i|\mathbf{Y}_{\mathsf{u}\setminus i}, \mathbf{Y}_\mathsf{s})]\\ &\quad = -\log|\mathbf{\Lambda}_{ii}| + \frac{K}{2}(\log(2\pi)+1).\end{aligned} \tag{12}$$

Therefore, (4) based on the Laplacian eigenmap becomes:

$$\arg\max_i \log|(\mathbf{\Lambda}_\mathsf{uu}^{-1})_{ii}| + \log|\mathbf{\Lambda}_{ii}|. \tag{13}$$

Since we can calculate analytically the entropy, the conditional distribution, and the marginal distribution of a Gaussian, we can calculate the objective function for active learning with the Laplacian eigenmap analytically. We note that locally linear embedding (LLE) (Roweis and Saul, 2000) can also been interpreted by a Gaussian model (Verbeek and Vlassis, 2006). Therefore, we may use exactly the same expressions as for the Laplacian eigenmap if we were to use LLE for visualization. When we use visualization methods that are not modeled by a Gaussian, we can use similar procedures by exploiting the Laplace approximation.

After the selected object is re-located by the user we need to re-calculate the visualization given the new labeled datapoint, and show it to the user before selecting the next point to label. For labeled objects, we use the ground truth location given by the user. For the locations for unlabeled objects, we estimate their locations as follows:

$$\hat{\mathbf{Y}}_\mathsf{u} = -\mathbf{\Lambda}_\mathsf{uu}^{-1}\mathbf{\Lambda}_\mathsf{us}\mathbf{Y}_\mathsf{s}, \tag{14}$$

because the distribution of unlabeled data conditioned on the labeled data is given by

$$p(\mathbf{Y}_\mathsf{u}|\mathbf{Y}_\mathsf{s}) = \mathcal{N}(\hat{\mathbf{Y}}_\mathsf{u}, \mathbf{\Lambda}_\mathsf{uu}^{-1}), \tag{15}$$

from the probabilistic interpretation of the Laplacian eigenmap. The estimated locations $\hat{\mathbf{Y}}_\mathsf{u}$ can be seen as a semi-supervised Laplacian eigenmap visualization result, where we have label information for some objects.

3.4 Learning Hyperparameters

We can estimate hyperparameters, such as the number of neighbors and the noise level, α, by maximizing the likelihood $p(\mathbf{Y}_\mathsf{s}) = \mathcal{N}(\mathbf{0}, (\mathbf{\Lambda}^{-1})_\mathsf{ss})$ given the labeled data (Verbeek and Vlassis, 2006).

When the hyperparameters are fixed, inspection of (13) reveals that the optimal object i does not depend on the location of the supervised data in visualization space, just which ones have been selected. This is not a general property of our active framework (1). A consequence of this is that we can pre-compute the optimal (myopic) set of objects to be presented to the user, before the user has re-located any objects. However, when we update the hyperparameters the mapping changes, the precision matrices $\mathbf{\Lambda}$ become implicit functions of the supervised data, and so we must wait for the user to move each object before computing the optimal new object to present.

4 Related Work

Let us first consider our objective in its reformulated form (4). Suppose we were to consider only the first term, the objective would become

$$\arg\max_i H[p(\mathbf{y}_i|\mathbf{Y}_\mathsf{s})], \tag{16}$$

that is we would select the object whose predictive distribution has highest entropy, or uncertainty. This corresponds to one of the most ubiquitous strategies in active learning, uncertainty sampling (Lewis and Gale, 1994; Sebastiani and Wynn, 2000; Settles, 2009), which selects the object for which one is least certain how to label. When the uncertainty measure used is Shannon's entropy, this corresponds exactly to (16). This strategy is used in (Verbeek and Vlassis, 2006) in the context of the locally linear embedding for semi-supervised regression. In the context of visualization this strategy considers only the uncertainty in object to be selected. However, our strategy (1) considers the uncertainty of all of the unlabeled objects; the second term in (4) favors objects that assist in determining the location of other unlabeled objects. We demonstrate experimentally the advantage of our framework over the uncertainty sampling in Section 5.

Now let us consider the initial formulation of our objective (1). It may seem sensible to minimize the absolute value of the entropy of the unseen data, that is to consider only the second term in (1),

$$\arg\min_i \mathbb{E}_{p(\mathbf{y}_i|\mathbf{Y}_\mathsf{s})}H[p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{y}_i, \mathbf{Y}_\mathsf{s})], \tag{17}$$

rather than the expected *decrease* in predictive entropy. However, this criterion turns out to be equivalent to uncertainty sampling (16) because

$$\mathbb{E}_{p(\mathbf{y}_i|\mathbf{Y}_\mathsf{s})}H[p(\mathbf{Y}_{\mathsf{u}\setminus i}|\mathbf{y}_i,\mathbf{Y}_\mathsf{s})] \\ = H[p(\mathbf{Y}_\mathsf{u}|\mathbf{Y}_\mathsf{s})] - H[p(\mathbf{y}_i|\mathbf{Y}_\mathsf{s})], \quad (18)$$

and the first term of the left hand side $H[p(\mathbf{Y}_\mathsf{u}|\mathbf{Y}_\mathsf{s})]$ does not depend on i.

Many information theoretic algorithms for active learning were proposed in the context of supervised learning, where the objective function is equal to the change in entropy of model parameters after receiving the label (Lindley, 1956; MacKay, 1992; Guestrin et al., 2005; Houlsby et al., 2011). The criterion in supervised learning is given by

$$\arg\max_i H[p(\boldsymbol{\theta}|\mathbf{D})] - \mathbb{E}_{p(\mathbf{t}_i|\mathbf{x}_i,\mathbf{D})}H[p(\boldsymbol{\theta}|\mathbf{t}_i,\mathbf{x}_i,\mathbf{D})], \quad (19)$$

where $\boldsymbol{\theta}$ is a set of parameters, $\mathbf{D}$ is a training data set, $\mathbf{x}_i$ is an input variable to be labeled, and $\mathbf{t}_i$ is its target variable. If we were to interpret the unknown locations in visualization space $\mathbf{Y}_{\mathsf{u}\setminus i}$ as our 'parameters of interest' $\boldsymbol{\theta}$, the point to be labeled $\mathbf{y}_i$ as the target variable $\mathbf{t}_i$, and the labeled points $\mathbf{Y}_\mathsf{s}$ as the training data $\mathbf{D}$, then this classical information theoretic approach for supervised learning (19) becomes equivalent to our objective function (1).

Finally, an alternative approach to active learning is to use decision theory in which one selects objects that reduce the expected loss at test-time (Roy and McCallum, 2001; Zhu et al., 2003a), that is, in a Bayesian framework to minimize the 'Bayes posterior risk'. In the context of active visualization, the decision task at hand is to select the location of the unlabeled objects. If we were to choose the log-loss on the probability of placing $\mathbf{Y}_u$ at a particular location as our loss function, then the optimal Bayesian decision at test-time (visualization-time) is to place the objects at the MAP estimate of their locations, as we do in our framework (14). This corresponds to a Bayes risk equal to the expected entropy over unlabeled data. If one seeks to maximize the decrease in Bayes risk then we arrive again at our objective (1). It is interesting that in our context of active visualization (1) has both an information-, and decision-theoretic interpretation, where in general these approaches result in different algorithms.

5 Experiments

5.1 Setting

We evaluated our active learning framework on one synthetic data set, and five real data sets: Wine, Iris, Vowel, Glass and Mnist, which are obtained from LIBSVM multi-class data sets (Chang and Lin, 2011). The synthetic data set, Synth, was generated as follows: 1) for a ground truth visualization we located objects in a two-dimensional grid and added small Gaussian noise as shown in Figure 2 (a), and 2) we generated observation feature vectors with a Gaussian process latent variable model (Lawrence, 2004), using the ground truth as the latent variables. For the five real data sets we generated the ground truth visualization by using class information. In all of the real data sets, each object has a class label; we located objects around a circle, ordered according to their class, and added Gaussian noise. The set up is depicted in Figure 2 (b) and (c), the color of each node represents its class. We summarize the statistics of data sets used for evaluation in Table 1.

We compared our active visualization framework with uncertainty sampling for active visualization, as described in (Verbeek and Vlassis, 2006), and a random sampling baseline method. We used the Laplacian eigenmap for the visualization method.

5.2 Results

The performance metric used was the average mean squared error between the estimated and true locations. To obtain statistailcally meaningful results, an average was taken over 1000 experimental runs with each data set, each using a different ground truth visualization. The noise parameter was set to $\alpha = 10^{-3}$. We selected the optimal number of neighbors k from the set $\{2, \cdots, 20\}$ using maximum likelihood as described in Section 3.4; k was update after every batch of five labeled objects was obtained.

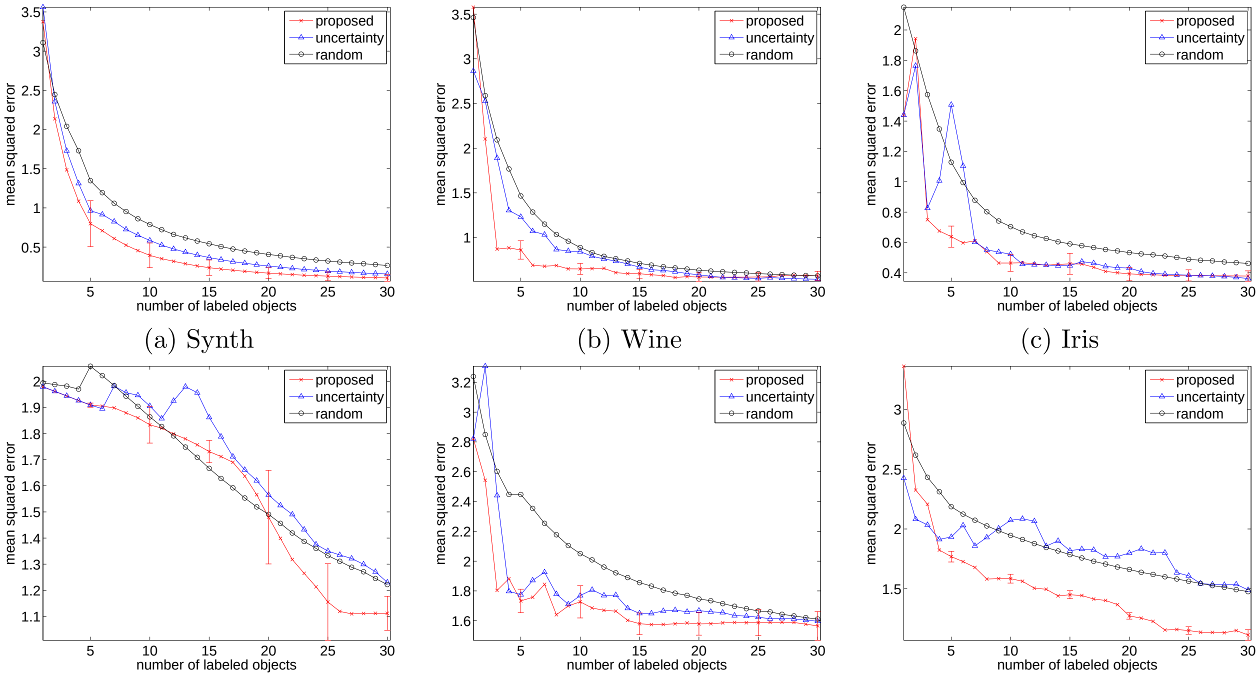
Figure 1 shows the results. For all of the methods, as the number of labeled data points increases, the error decreases. However, in most cases, our method decreases the error faster than uncertainty and random sampling. This indicates the importance of considering the relationship between the point to be labeled and the remaining unlabeled points that is represented by the second term in (4), which is not considered in uncertainty sampling.

Table 2 shows the statistical significance of the results when the number of neighbors k is updated based on the maximum likelihood estimation (a), and it is fixed at $k = 3$ (b). In both of the cases, the proposed method achieved the lowest error for all data sets. And except for Iris and Vowel data sets when $k = 3$, the proposed method was significantly better than uncertainty and random sampling.

Figure 3 shows the visualization attained using the Laplacian eigenmap in an unsupervised setting with Synth, Wine and Mnist data sets. We used three neighbors for constructing the neighbor graph. The

Table 1: The statistics of data sets used for evaluation.

	Synth	Wine	Iris	Vowel	Glass	Mnist
number of objects N	400	178	150	528	214	1000
observed dimensionality D	100	13	4	10	9	784
number of classes C	-	3	3	11	7	10



(a) Synth (b) Wine (c) Iris

(d) Vowel (e) Glass (e) Mnist

Figure 1: Average mean squared error between the estimated locations and the true locations for different numbers of labeled objects achieved by the proposed method, uncertainty sampling, and random sampling. We also show error bars depicting the standard deviation of the proposed method, but omit them from the other methods for visual clarity (see Table 2 for statical significance).

Table 2: Average mean squared error given ten labeled objects when (a) the number of neighbors k is updated based on the maximum likelihood estimation, and (b) it is fixed at $k = 3$. Values in bold typeface are statistically better (at the 5% level) from those in normal typeface as indicated by a paired t-test.

(a) number of neighbors is updated

	Synth	Wine	Iris	Vowel	Glass	Mnist
Proposed method	**0.395**	**0.649**	**0.464**	**1.833**	**1.727**	**1.585**
Uncertainty sampling	0.585	0.842	0.523	1.906	1.769	2.073
Random sampling	0.788	0.888	0.704	1.864	2.050	1.946

(b) number of neighbors is fixed at $k = 3$

	Synth	Wine	Iris	Vowel	Glass	Mnist
Proposed method	**0.382**	**0.648**	**0.455**	**1.820**	**1.770**	**1.594**
Uncertainty sampling	0.597	0.834	**0.456**	**1.820**	2.266	2.073
Random sampling	0.892	0.898	0.698	1.876	2.044	1.951

goal is to obtain a visualization that is similar to the ground truth (Figure 2) by labeling as few objects as possible. Without any labeled objects, the locations differ greatly from the ground truth as shown Figure 3.

Figure 4 shows the visualization when 20 objects are labeled by random sampling (top), uncertainty sampling (middle) and our active learning framework (bottom). The random sampling method sometimes selects

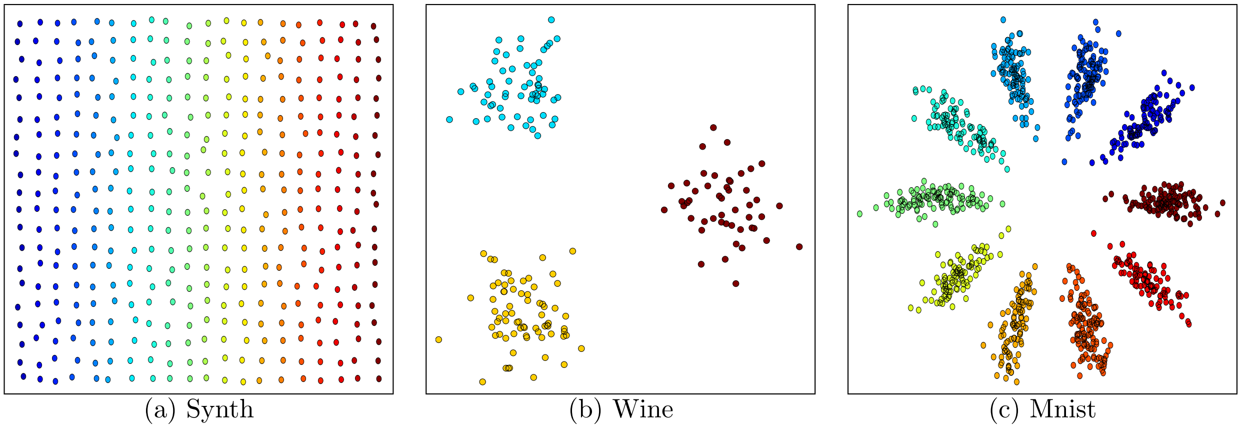


Figure 2: Ground truth, or user's desired visualization. In the Synth data set (a), the color similarity of each node related to the closeness in the ground truth visualization. In the Wine (b) and Mnist (c) data sets, the color of each node represents the class information.

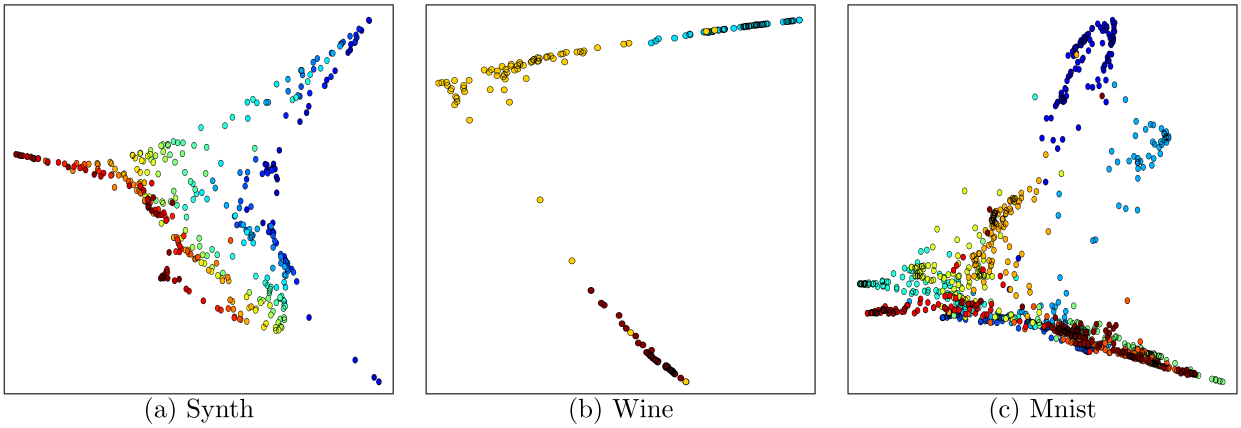


Figure 3: Visualization results by the unsupervised Laplacian eigenmap in Synth (a), Wine (b) and Mnist (c) data sets. The color of each node is the same as those in Figure 2.

objects located close together in visualization space, or similar objects, which is not effective because the locations can be inferred easily by using the locations of those similar objects. Uncertainty sampling tends to select objects that are located at the edges of set of objects, as shown in Figure 4 (a) middle. This is because the entropy of objects that are located as far from other objects is high (Ramakrishnan et al., 2005; Guestrin et al., 2005). On the other hand, our method selects a diverse set of objects by maximizing the decrease of the uncertainty for unlabeled data, and we can obtain visualizations that are more similar to the ground truth with fewer labels than random and uncertainty sampling.

6 Conclusion

We have proposed an active learning framework for data visualization based on an information theoretic criterion where the object that reduces the uncertainty of the unlabeled data is selected. We have confirmed experimentally that our framework can obtain the user's desired visualization with fewer labeled objects than existing active visualization methods.

Although our results have been encouraging, our framework can be further improved upon in a number of ways. Firstly, we plan to use other visualization methods with our framework, such as the Gaussian process latent variable model (Lawrence, 2004) and stochastic neighbor embedding (Hinton and Roweis, 2002). Secondly, we would like to extend our framework to incorporate other types of supervised information. In the current framework, a user re-locates objects to indicate its desired location. However, the user might want to provide information about the desired visualization by selecting two objects that should be located close together, or far apart.

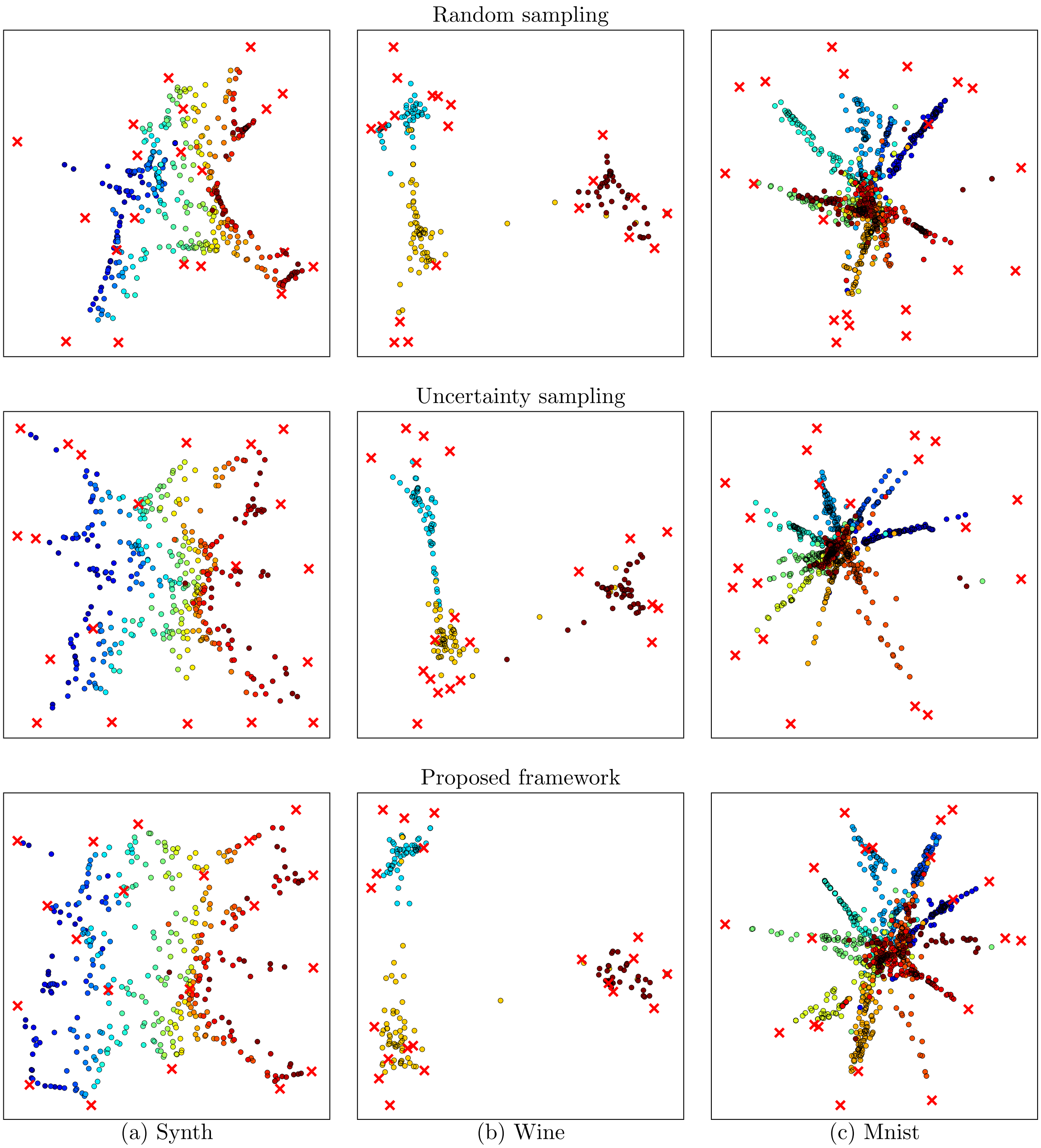


Figure 4: Visualization results with 20 labeled objects selected by random sampling (top), uncertainty sampling (middle) and the proposed method (bottom) in Synth (a), Wine (b) and Mnist (c) data sets. The '×' shows the location selected.

References

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, 2011.

D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

T. Cover, J. Thomas, J. Proakis, M. Salehi, and R. Morelos-Zaragoza. *Elements of Information Theory.* edition: John Wiley & Sons Inc, 1991.

S. Dasgupta. Analysis of a greedy active learning strategy. *Advances in neural information processing systems*, 17:337–344, 2005.

A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North. Observation-level interaction with statistical models for visual analytics. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 121–130. IEEE, 2011.

D. Golovin and A. Krause. Adaptive submodularity: A new approach to active learning and stochastic optimization. In *Proceedings of International Conference on Learning Theory (COLT)*, 2010.

C. Guestrin, A. Krause, and A. Singh. Near-optimal sensor placements in gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 265–272, 2005.

G. Hinton and S. Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:833–840, 2002.

N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):993–1000, 2009.

D. Keim et al. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

N. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in Neural Information Processing Systems*, 16:329–336, 2004.

D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

D. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

D. MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

A. McCallum and K. Nigam. Employing em in pool-based active learning for text classification. In *Proceedings of ICML-98, 15th International Conference on Machine Learning*, pages 350–358, 1998.

F. V. Paulovich, D. Eler, J. Poco, C. P. Botha, R. Minghim, and L. Nonato. Piece wise laplacian-based projection for interactive data exploration and organization. *Computer Graphics Forum*, 30:1091–1100, 2011.

N. Ramakrishnan, C. Bailey-Kellogg, S. Tadepalli, and V. Pandey. Gaussian processes for active data mining of spatial aggregates. In *Proceedings of the Fifth SIAM International Conference on Data Mining*, volume 119, page 427. Society for Industrial Mathematics, 2005.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000.

N. Roy and A. McCallum. Toward optimal active learning through monte carlo estimation of error reduction. In *International Conference on Machine Learning*, pages 441–448, 2001.

P. Sebastiani and H. Wynn. Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000.

B. Settles. Active learning literature survey. Technical report, University of Wisconsin, Madison, 2009.

J. Tenenbaum, V. De Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

W. Torgerson. *Theory and methods of scaling.* Wiley, 1958.

J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *The Journal of Machine Learning Research*, 11:451–490, 2010.

J. Verbeek and N. Vlassis. Gaussian fields for semi-supervised regression and correspondence learning. *Pattern Recognition*, 39(10):1864–1875, 2006.

G. Wills. Nicheworksinteractive visualization of very large graphs. *Journal of Computational and Graphical Statistics*, 8(2):190–212, 1999.

X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 58–65, 2003a.

X. Zhu, J. Lafferty, and Z. Ghahramani. Semi-supervised learning: From Gaussian fields to Gaussian processes. Technical report, Carnegie Mellon University, 2003b.