Sparse Multimodal Gaussian Processes

Qiuyang Liu and Shiliang Sun

Department of Computer Science and Technology, East China Normal University,
3663 North Zhongshan Road, Shanghai 200062, China
qiuyangliu2014@gmail.com, slsun@cs.ecnu.edu.cn

Abstract. Gaussian processes (GPs) are effective tools in machine learning. Unfortunately, due to their unfavorable scaling, a more widespread use has probably been impeded. By leveraging sparse approximation methods, sparse Gaussian processes extend the applicability of GPs to a richer data. Multimodal data are common in machine learning applications. However, there are few sparse multimodal approximation methods for GPs applicable to multimodal data. In this paper, we present two kinds of sparse multimodal approaches for multi-view GPs, the maximum informative vector machine (mIVM) and the alternative manifold-preserving (aMP), which are inspired by the information theory and the manifold preserving principle, respectively. The aMP uses an alternative selection strategy for preserving the high space connectivity. In the experiments, we apply the proposed sparse multimodal methods to a recent framework of multi-view GPs, and results have verified the effectiveness of the proposed methods.

Keywords: Classification, Kernel methods, Sparse Gaussian processes, Multimodal learning

1 Introduction

Gaussian processes (GPs) are widely used in machine learning and statistics as a powerful and flexible Bayesian nonparametric tool for probabilistic modeling [1]. However, computational requirements of the GPs grow as the cube of the size of the training set, impeding their widespread use to the scenario of scalable data. In order to address this limitation, researchers have recently proposed some sparse approximations [2–9]. They can be grouped into four classes. The first one uses only a subset of the data and focuses on the strategies of selecting the representative data points to form the subset [4]. The second kind of method concentrates on using a reduced-rank matrix to approximate the covariance matrix [2]. Another kind of method seeks to give a low rank approximation to the covariance matrix based on inducing points [6], while the fourth uses the method of variational inducing points [7, 8]. These methods lead to a significant reduction of the computational complexity, which makes sparse Gaussian processes (SGPs) efficiently applied to a richer class of data [10, 11, 9].

Typically, standard SGPs only pay attention to the scenario where data from a single modality are provided. In practice, multimodal data are common in applications of machine learning. They refer to the kind of data involving associated

descriptive information from multiple domains, which are also called multi-view data. For instance, in speaker recognition, audio and visual data are correlated descriptions as phonemes and lip pose have correlations. In image classification, an image can be described by different features such as texture, shape, and color. As multiple modalities often provide complementary information, better performance is likely to be expected by utilizing multimodal instead of single-modal representations. Therefore, there has been a wealth of interest in multimodal learning recently [12–14]. However, SGPs, as popular and efficient methods in machine learning, are barely applied in multimodal learning. In this paper, our motivation is to study the sparse multimodal methods for GPs applicable to multimodal data.

We propose two kinds of sparse multimodal methods, the maximum informative vector machine (mIVM) and the alternative manifold-preserving (aMP), which are inspired by principles in information theory [4] and manifold learning [15], respectively. In the multimodal setting, the sparse multimodal methods need to consider all modalities together efficiently. On the one hand, the mIVM leverages a Gaussian process (GP) to model data from the same modality. Since data involve multiple related modalities, the mIVM uses multiple GPs, which are potentially correlated with each other as data from different modalities describe the same objective. For every example, it calculates the associated entropy reduction of each modality, and use the maximum entropy reduction among all the modalities as the overall entropy reduction of that data point. At each selection, the data point with the maximum overall entropy reduction is added to the sought sparse set. By using these strategies, the mIVM takes into consideration the entropy reduction of every modality for each data point. Overall, it tries to obtain the maximum of information among all the modalities with the minimum number of examples.

On the other side, for each modality, the aMP constructs a graph using the corresponding data. Vertices in different graphs are corresponding to each other if these vertices represent the same data point. Initially, the candidate set contains all the data points, while the sought sparse set is null. For each data point in the candidate set, the aMP calculates the degree of the corresponding vertex in each graph. To start the selection, it first chooses a modality randomly. Next, it selects a vertex with the maximum degree in the graph corresponding to the chosen modality. At the same time, all the vertices in other graphs corresponding to this chosen vertex are also selected. Then we include the data point associated with the chosen vertex into the sought sparse set and remove it from the candidate set. At the same time, we remove the chosen vertex and all the associated edges from each graph. Another round of selection will start with the alternative chosen modality. Overall, inspired by the manifold-preserving principle, the aMP makes use of data from all modalities by an alternative selection strategy for preserving the high space connectivity. Among the GP related multimodal learning methods [16–19], the two sparse multimodal approaches employ the recent framework of multi-view GPs [19] to evaluate the validity. It

was a straightforward extension of GPs to multimodal learning with convenient implementation.

The contributions of our work are summarized as follows. First, we study the sparse multimodal methods for GPs from two different aspects and propose two kinds of sparse multimodal approaches. On the one hand, we present the mIVM to accommodate the multimodal data from the perspective of information theory. On the other hand, we present the aMP for multimodal sparsity from the manifold-preserving perspective. Secondly, we apply our two sparse multimodal methods to a recent multi-view GP framework. Finally, the proposed sparse multimodal methods can reduce the training time significantly with slight reduction of the accuracy, which extend the multimodal GPs to the scenario of scalable data.

The structure of the remainder of the paper is as follows. In Section 2, we briefly review GPs and propose the mIVM, our first sparse multimodal approach. Section 3 review the manifold-preserving principle, and present the other sparse multimodal method, aMP. A recent framework of multimodal GPs and our novel application are described in Section 4. Experimental results are reported in Section 5. Finally, we conclude this paper in Section 6.

2 Maximum Informative Vector Machine

This section first reviews the GP model, and then introduces the maximum Informative Vector Machine (mIVM), our first sparse multimodal approach. For the sake of clarity, in this section and Section 3, we take the case that data from two modalities are available as an example to illustrate our sparse multimodal methods. Similar algorithms can be mimicked if data concerning more than two modalities are adopted for the sparse multimodal GPs.

2.1 Gaussian Processes

GPs have proven their effectiveness as successful tools for classification and regression. They are frequently applied to describe a distribution over functions, and can be completely specified by its mean function and covariance function [1].

Suppose the training data are $\boldsymbol{X}, \boldsymbol{Y}$ with N points, where $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]^{\mathbf{T}}$, $\boldsymbol{x}_i \in R^M$ is the ith input, $\boldsymbol{Y} = [y_1, y_2, ..., y_N]^{\mathbf{T}}$, and $y_i \in R$ is the ith output. The latent function of the data is denoted as $\boldsymbol{f}$.

Following standard settings for GPs, the prior distribution for $\boldsymbol{f}$ is supposed to be Gaussian with a zero mean and a covariance matrix $\boldsymbol{K}$, $\boldsymbol{f}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$, and the covariance function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ determines the element K_{ij} of $\boldsymbol{K}$. Numerous kernel functions can be applied in GPs. Since the squared exponential kernel is a frequently-used covariance function, we select it as the covariance function in this paper. The Gaussian likelihood for regression is $\boldsymbol{Y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f}, \sigma^2\boldsymbol{I})$, and the marginal likelihood can be written as $\boldsymbol{Y}|\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{K} + \sigma^2\boldsymbol{I})$. The posterior of the latent function is

$$\boldsymbol{f}|\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{\mu} = \boldsymbol{K}(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{Y}$ is the mean of the posterior distribution and $\boldsymbol{\Sigma} = \boldsymbol{K} - \boldsymbol{K}(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{K}$ is the covariance of the posterior distribution.

The prediction of a new point $\boldsymbol{x}^*$ is also Gaussian,

$$f^*|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}^* \sim \mathcal{N}(\bar{f^*}, \mathrm{cov(f^*)}), \tag{2}$$

where $\bar{f^*} = \boldsymbol{k^*}^{\mathbf{T}}[\boldsymbol{K} + \sigma^2\boldsymbol{I}]^{-1}\boldsymbol{Y}$, $\mathrm{cov}(f^*) = k(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{k^*}^{\mathbf{T}}[\boldsymbol{K} + \sigma^2\boldsymbol{I}]^{-1}\boldsymbol{k}^*$, k is the covariance function, and $\boldsymbol{k}^*$ is the vector of covariance function values between $\boldsymbol{x}^*$ and the training data $\boldsymbol{X}$.

Typically, if N is the size of training data, GPs need $\mathcal{O}(N^3)$ time and $\mathcal{O}(N^2)$ memory for training, and at least $\mathcal{O}(N)$ time for prediction on a test point.

2.2 Algorithm

Keeping the GP predictor only on a smaller subset of the data is a simple approximation to the full-sample GP predictor. This kind of approximation method makes sense if the information contained in points of the subset is sufficiently close to the information obtained by the full data set. Clearly, it is pivotal to select the points in the subset, which is called as the sparse set in this paper. Based on the information theory, [4] proposed to select the point with maximum differential entropy score to be included into the sparse set at every selection. In other words, they chose the point with the most information for inclusion. By the most information, it means that for a point the quantity

$$\Delta H_{in_i} = -\frac{1}{2}\log|\boldsymbol{\Sigma}_{in_i}| + \frac{1}{2}\log|\boldsymbol{\Sigma}_{i-1}| \tag{3}$$

is maximized, where $\boldsymbol{\Sigma}_{in_i}$ is the posterior covariance after choosing the n_ith point at the ith selection, and $\boldsymbol{\Sigma}_{i-1}$ is the posterior covariance after the $(i-1)$th choice. This quantity is the reduction in the posterior process entropy associated with selecting the n_ith point at the ith selection [4]. Inspired by these thoughts, we concentrate on the multimodal cases and propose the mIVM for sparse multimodal GPs.

The mIVM use a GP to model data from the same modality, which means that for each modality, the mIVM models data by a GP. In each selection, for each candidate point, it first calculates the entropy reduction associated with every modality. Next, the overall entropy reduction associated with the candidate is determined by the maximum among all the modalities. Then the candidate giving the largest overall reduction in the posterior process entropy is added to the sparse set.

Formally, assume that we have a data set $D = \{(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, y_i)\}_{i=1}^N$ with N examples, where $\boldsymbol{x}_i^1 \in R^{M_1}$ is the ith observation from the first modality, $\boldsymbol{x}_i^2 \in R^{M_2}$ is the ith observation from the second modality, and $y_i \in \{+1, -1\}$ is the corresponding label. Denote $\boldsymbol{X}^1 = [\boldsymbol{x}_1^1, ..., \boldsymbol{x}_N^1]^{\mathbf{T}}$, $\boldsymbol{X}^2 = [\boldsymbol{x}_1^2, ..., \boldsymbol{x}_N^2]^{\mathbf{T}}$, and $\boldsymbol{Y} = [y_1, ..., y_N]^{\mathbf{T}}$. Let T denote the sparse set, I denote the candidate set, and t denote the number of points in the sparse set, namely the size of the sought sparse set.

The mIVM use two GPs to model two modalities of data. Specifically, it uses one GP to modal data from the first modality, i.e. $\{\boldsymbol{X}^1, \boldsymbol{Y}\}$, and uses the other GP to modal data from the second modality, i.e. $\{\boldsymbol{X}^2, \boldsymbol{Y}\}$. That is, the prior distributions for the latent functions $\boldsymbol{f}_1$ on the first modality of data and $\boldsymbol{f}_2$ on the second modality of data are assumed to be Gaussian, i.e. $p(\boldsymbol{f}_1|\boldsymbol{X}^1) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_1)$, and $p(\boldsymbol{f}_2|\boldsymbol{X}^2) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_2)$, where $\boldsymbol{K}_1$ is the covariance matrix about the first modality of data and $\boldsymbol{K}_2$ is the covariance matrix about the second modality of data. As for the likelihood, we use the Gaussian likelihood here. Although the Gaussian noise model is originally developed for regression, it has also been proved effective for classification, and its performance typically is comparable to the more complex probit and logit likelihood models used in classification problems [20]. Therefore, we also use Gaussian noise model for classification tasks in this paper.

Initially, T is a null set and I contains all the N examples. At the ith ($i = 1...t$) selection, the entropy reductions with the n_ith point for the first modality and the second modality are obtained by

$$\Delta H^1_{in_i} = -\frac{1}{2}\log|\boldsymbol{\Sigma}^1_{in_i}| + \frac{1}{2}\log|\boldsymbol{\Sigma}^1_{i-1}|, \tag{4}$$

and

$$\Delta H^2_{in_i} = -\frac{1}{2}\log|\boldsymbol{\Sigma}^2_{in_i}| + \frac{1}{2}\log|\boldsymbol{\Sigma}^2_{i-1}|, \tag{5}$$

respectively, where $\boldsymbol{\Sigma}^1_{in_i}$ is the posterior covariance for the first modality of data after choosing the n_ith point at the ith selection, $\boldsymbol{\Sigma}^1_{i-1}$ is the posterior covariance for the first modality of data after the $(i-1)$th choice, and $\boldsymbol{\Sigma}^2_{in_i}$ and $\boldsymbol{\Sigma}^2_{i-1}$ are defined analogously for the second modality. The overall entropy reduction associated with the n_ith point is given by

$$\Delta H_{in_i} = \max(\Delta H^1_{in_i}, \Delta H^2_{in_i}). \tag{6}$$

Then, at the ith selection, the n^*_ith data point is selected for inclusion at the sparse set T and removed from the candidate set I, where

$$n^*_i = \max_{n_i}(\{\Delta H_{in_i}\}_{n_i \in I}). \tag{7}$$

The selection procedure repeats until t points are added into the sparse set T.

The mIVM explores a sparse representation of multimodal data, which leverages the information from the input data and corresponding output labels. It attempts to obtain the maximum amount of information among all the modalities with the minimum number of data points. The computational complexity of the mIVM is $\mathcal{O}(t^2N)$, where t is the number of data points included in the sparse multimodal representation.

3 Alternative Manifold-Preserving

We first review the principle of manifold-preserving. Then we introduce our second sparse multimodal method, the alternative Manifold-Preserving (aMP).

3.1 Manifold-Preserving

Assume we are given a graph $G(V, E, W)$ corresponding to a manifold with vertex set $V = \{v_i\}_{i=1}^m$, edge set E, and weight matrix W, and the number of vertices reserved in the desired sparse graphs is s. Manifold-preserving seeks a sparse graph G', which is a subgraph of G with s vertices, having a high connectivity with G, that is to say, a candidate that maximizes the quantity

$$\frac{1}{m-s}\sum_{i=s+1}^{m}(\max_{j=1,..,s} W_{ij}), \tag{8}$$

where W_{ij} characterizes the similarity or closeness between the ith vertex and the jth vertex, and a small value denotes a low similarity [15].

The manifold-preserving sparse graph G' focuses on reducing the number of vertices, and the edge weights from the original graph G to sparse graph G' need not change. The high demand for space connectivity inclines to choose vital data points and thus remove outliers and noisy points, which can maintain the manifold structure. The maximum preservation of the manifold structure can be beneficial to machine learning tasks. Inspired by this thought, we propose the aMP in the following section.

3.2 Algorithm

To make this section self-contained, we repeat the data notations. We are given data $D = \{(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, y_i)\}_{i=1}^N$ with N examples, where $\boldsymbol{x}_i^1 \in R^{M_1}$ is the ith observation from the first modality, $\boldsymbol{x}_i^2 \in R^{M_2}$ is the ith observation from the second modality, and $y_i \in \{+1, -1\}$ is the corresponding output. Denote $\boldsymbol{X}^1 = [\boldsymbol{x}_1^1, ..., \boldsymbol{x}_N^1]^{\mathbf{T}}$, $\boldsymbol{X}^2 = [\boldsymbol{x}_1^2, ..., \boldsymbol{x}_N^2]^{\mathbf{T}}$, and $\boldsymbol{Y} = [y_1, ..., y_N]^{\mathbf{T}}$.

We use $\{\boldsymbol{X}^1, \boldsymbol{Y}\}$ to construct the graph $G^1(V^1, E^1, W^1)$, where $V^1 = \{v_i^1\}_{i=1}^N$ is the vertex set of graph G^1. The graph $G^2(V^2, E^2, W^2)$ is constructed by using $\{\boldsymbol{X}^2, \boldsymbol{Y}\}$, where $V^2 = \{v_i^2\}_{i=1}^N$ is the vertex set of graph G^2. Clearly, graph G^1 is associated with the first modality of data, while graph G^2 is associated with the second modality. There are many methods to create the graphs. In this paper, we do not investigate the distinctions of properties of graphs constructed by different methods, but assume that a reasonable graph can be constructed.

Note that the vertex v_i^1 corresponds to the vertex v_i^2 since they are associated with the same example, namely the ith example. In fact, the ith example has observation $(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2)$ and label y_i, and is corresponding to the vertex v_i^1 in graph G^1, and vertex v_i^2 in graph G^2.

The degree $d^1(i)$ associated with vertex v_i^1 is defined to be $d^1(i) = \sum_{i\sim j} W_{ij}^1$, where $i \sim j$ denotes that there is an edge connecting the vertex v_i^1 and vertex v_j^1 (if there is no edge between two vertices, their similarity is regarded as 0). For the degree $d^2(i)$ associated with vertex v_i^2, the definition is similar. Suppose that the number of retained examples is t. Our goal is to seek t examples to form a sparse set $T = \{\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, y_i\}_{i\in T_I}$, where T_I is the index set of the sought

sparse set, from the original N examples. Inspired by the manifold-preserving principle, we present the aMP whose details are described as follows.

The aMP first chooses a modality from the two modalities at random, for example, the second modality. Next, the vertex $v^2_{w_1}$ with the maximum degree in the graph associated with the chosen modality is selected. As we have mentioned above, the vertex $v^2_{w_1}$ is associated with the w_1th data point and vertex $v^1_{w_1}$ in the other graph. Thus, the vertex $v^1_{w_1}$ is also selected. All the edges and weights linked to the vertex $v^2_{w_1}$ from the original graph G_2 are then removed and all the edges and weights associated with the vertex $v^1_{w_1}$ in the other graph G_1 are also removed as they represent the same data point. At the same time, the chosen vertices and edges linking these vertices are added from the original graphs G^1 and G^2 to the corresponding sparse graphs G^1_s and G^2_s (which are null initially), respectively. Add the corresponding example w_1 to the index set T_I. Then a similar selection proceed on the resultant graphs with the first modality as the chosen modality. The alternative selection procedure repeats until t data points are added into the index set T_I. We summarize aMP in Algorithm 1.

Algorithm 1 Alternative Manifold-Preserving

Input: graphs $G^1(V^1, E^1, W^1)$, $G^2(V^2, E^2, W^2)$ with N vertices, t for the size of the sparse set, training data $\{(\boldsymbol{x}^1_i, \boldsymbol{x}^2_i, y_i)\}_{i=1}^N$.
Output: the index set T_I of sparse set, the sparse set T.

1: Initialize: a is randomly set in $\{1, 2\}$; $T_I = \emptyset$.
2: **for** $j = 1, ..., t$ **do**
3: $b = a$, $a = (a \bmod 2) + 1$.
4: compute degree $d^a(i)$ $(i = 1, ..., N - j + 1)$.
5: pick one vertex v^a_w in graph G^a with the maximum degree.
6: remove v^a_w and associated edges from graph G^a, remove v^b_w and associated edges from graph G^b.
7: add w to the index set T_I.
8: **end for**
9: The sparse set is $T = \{\boldsymbol{x}^1_i, \boldsymbol{x}^2_i, y_i\}_{i \in T_I}$.

The aMP focuses on the sparse point selection for multimodal data. Motivated by the manifold preserving, it uses an alternative selection strategy to preserve the high space connectivity. Assume the maximum number of edges linked to a vertex in the original graphs G^1 and G^2 is d_E. The computational complexity of the aMP is

$$\mathcal{O}[d_E(N + (N-1) + .. + (N-t+1))] = \mathcal{O}(d_E N t), \tag{9}$$

Since the aMP is simple and efficient, it is quite straightforward to be applied to scalable data.

4 Application to Multi-view GPs

The framework of multi-view Gaussian processes (MvGPs) has recently been proposed as a straightforward extension of the GPs for multimodal data [19]. The core idea is to impose consistency between the posterior distributions of the functions across modalities.

Taking the data having two modalities as an example, the MvGP first models each modality of data by a GP. Then, it proposes the consistency criterion to regularize the objective function, and optimizes the hyperparameters collaboratively by the two modalities. The objective function of MvGP is

$$\begin{aligned}\min\{&-[a\log p(\boldsymbol{Y}|\boldsymbol{X}^1)+(1-a)\log p(\boldsymbol{Y}|\boldsymbol{X}^2)]\\&+\frac{b}{2}[KL(p(\boldsymbol{f}_1|\boldsymbol{X}^1,\boldsymbol{Y})||p(\boldsymbol{f}_2|\boldsymbol{X}^2,\boldsymbol{Y}))\\&+KL(p(\boldsymbol{f}_2|\boldsymbol{X}^2,\boldsymbol{Y})||p(\boldsymbol{f}_1|\boldsymbol{X}^1,\boldsymbol{Y}))]\},\end{aligned}\tag{10}$$

where $\boldsymbol{X}^1 \in R^{N\times M_1}$ is the data matrix on the first modality, $\boldsymbol{X}^2 \in R^{N\times M_2}$ is the data matrix on the second modality, $\boldsymbol{Y}$ is corresponding label matrix, $\boldsymbol{f}^1$ and $\boldsymbol{f}^2$ are the associated latent functions for the two modalities of data, respectively, and a and b are parameters.

To demonstrate the performances of our proposed sparse multimodal methods, we apply them to the framework of MvGPs and use the combined models to solve the classification problem. For convenience, we denote the mIVM based MvGP as mMvGP, and the aMP based MvGP as aMvGP. The computational complexity of the mMvGP is

$$\mathcal{O}(t^2N+t^3)=\mathcal{O}(t^2N),\tag{11}$$

while the computational complexity of the aMvGP is

$$\mathcal{O}(d_ENt+t^3),\tag{12}$$

where N is the number of original training data points, t is the number of data points included in the sparse multimodal representation (usually, $t << N$), and d_E is the maximum number of edges linked to a vertex in the original graphs from all modalities. The original MvGP needs $\mathcal{O}(N^3)$ time, the same as the GP. From the analysis of computational complexity, it is clear that both the aMvGP and the mMvGP significantly reduce the training time. Thus, applying the mIVM and aMP to multimodal GPs would be quite efficient.

5 Experiment

In this section, experiments are conducted to assess the effectiveness of the two proposed sparse multimodal methods.

Table 1. Statistical information of the data sets.

Data set	size	content dimension	citation dimension	# P class	# N class
Cornell	195	1703	195	83	112
Washington	230	1703	230	107	123
Wisconsin	265	1703	265	122	143
Texas	187	1703	187	103	84
cora	2708	1433	2708	818	1890

5.1 Data

Four Web-Page Data Sets The web-page data sets, as widely used data sets in multimodal learning, consist of two-modalities web pages collected from computer science department websites of four universities: Cornell university, university of Washington, university of Wisconsin, and university of Texas. The two modalities are words occurring in a web page and words appearing in the links pointing to that page. We list the statistical information about the four data sets in Table 1. The web pages are classified into five classes: student, project, course, staff and faculty. In each data set, we set the category with the greatest size to be the positive class (denoted as "P class"), and all the other categories as the negative class (denoted as "N class").

Cora Data Set The cora data set consists of 2708 scientific publications belonging to seven categories, of which the one with the most publications is set to be the positive class, and the rest the negative class. Each publication is represented by words in the content modality, and the numbers of citation links between other publications and itself in the citation modality. The dimensions are 1433 and 2708, respectively.

5.2 Setting

In the experiments, we select two-thirds of data in each data set as the training set, and the rest as the test set. For the four web-page data sets, the sizes of the sparse set are 40%, 60%, and 80% of the corresponding training sets, and we also conduct experiments without sparse approximation. For the cora data set, the sizes of the sparse set are 8%, 10%, and 12% of the size of the training data set. For comparison, we give a random sparse approximation for MvGP, which just randomly selects points to form the sparse set, and denote it as rMvGP. The kernel functions used in mIVM and aMP are the squared exponential kernel functions. After finding the sparse set, aMvGP, mMvGP, and rMvGP employ the similar hyperparameters learning and parameters setting as [19]. We repeat the experiments for all the data sets five times and record the average accuracies and the corresponding standard deviations. The average training times for each model on all the data sets are also reported.

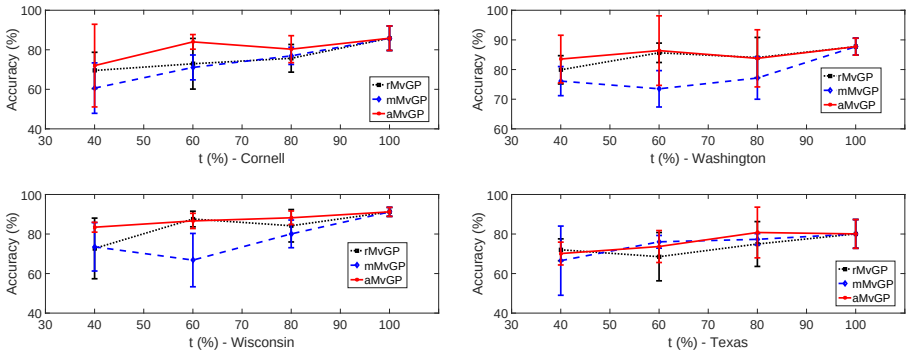

Fig. 1. Classification results for four web-page data sets. The x-axis corresponds to different settings of the size of sparse set, where t represents the percentage of the training set. The figures from top to bottom are results on the Cornell, Washington, Wisconsin, and Texas data set. Error bars represent standard deviations of the accuracies.

5.3 Results on Four Web-Page Data Sets

We first evaluate mMvGP, aMvGP, and rMvGP on four web-page data sets in consideration of making comprehensive comparisons of the accuracies of such sparse methods. The results are shown in Figure 1. Compared with the models without using sparse methods, the models leveraging sparse methods only reduce the classification accuracies slightly. For a range of t values, the classification results on the four data sets show that the aMvGP produce superior classification performance to other sparse models, which verifies the effectiveness of our proposed sparse multimodal method, the aMP. The performances of the mMvGP are not so well as aMvGP.

When the modalities are not necessarily compatible, a variant of the MvGP was given in [19]. We also combine the mIVM, the aMP, and random sparse approximation with this variant and denote the combinations as mMvGP2, aMvGP2, and rMvGP2, respectively. We evaluate mMvGP2, aMvGP2, and rMvGP2 on the four data sets. The corresponding classification results reflect that there is generally no improvement of the performances on accuracy.

The average training times of aMvGP and aMvGP2 on four data sets are presented in Figure 2, which verify the significant reduction of computational complexity. It is shown that the training times increase rapidly with the size of the sparse set, which indicates that the sparse methods effectively reduce the training times. The average training times of other sparse models are comparable to aMvGP and aMvGP2.

Taken the computational complexity and the classification accuracy together, the sparse multimodal models significantly reduce the training times but without obvious loss of accuracies.

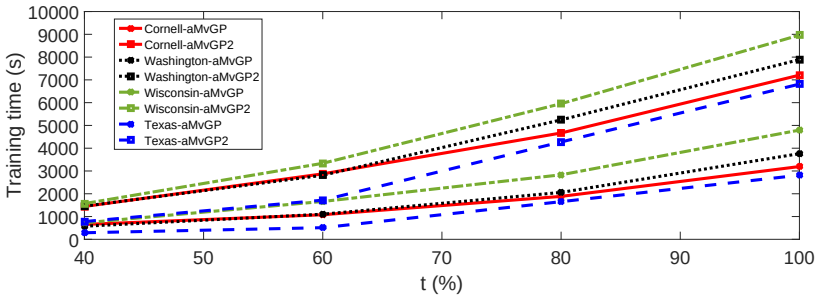

Fig. 2. Average training times on four web-page data sets. The x-axis corresponds to different settings of the size of sparse set, where t represents the percentage of the training set. The y-axis corresponds to the average training times on different settings of the size of sparse set. The upper four lines are average training times of aMvGP2, while the lower four lines are average training times of aMvGP.

Table 2. The accuracies on the cora data set (%).

Model	6%	8%	10%
rMvGP	72.46±4.22	73.32±10.04	78.51±1.61
mMvGP	74.78±3.09	76.42±3.85	75.56±2.89
aMvGP	**76.00**±3.00	**77.11**±2.08	**79.90**±2.22

5.4 Results on the Cora Data Set

On the same experimental setting, it takes about one week for MvGP to train on the cora data set with a normal computer (Intel(R) Core(TM) i7-6700 3.40GHz CPU). The long training time caused by MvGP may come from the cross validation and the grid search for optimizing the hyperparameters. As the computation of MvGP involves matrix inversions, it is unaffordable to be applied to large-scale data sets, such as a dataset with more than ten thousands points. Considering those factors, we choose the cora data set. Since the performances of mMvGP, aMvGP, and rMvGP are generally better than mMvGP2, aMvGP2, and rMvGP2, we only evaluate mMvGP, aMvGP, and rMvGP on the cora data.

The classification results are demonstrated in Table 2. The average training times of mMvGP, aMvGP, and rMvGP are shown in Figure 3. Taking into account of the results in the figure and table, we find that the aMP and mIVM greatly reduce the training times of the multimodal GPs with an acceptable performance on the accuracy.

The accuracy of the MvGP for the cora data set is around 92%. We can see the loss of the accuracy is slight though we only use a tiny proportion of the training set, such as 8%. Specifically, aMvGP can achieve the accuracy of 77.11% with only 8% of the whole training set, which is 83.82% of the accuracy of MvGP trained on the whole training set. It indicates that our sparse methods are scalable on large data sets.

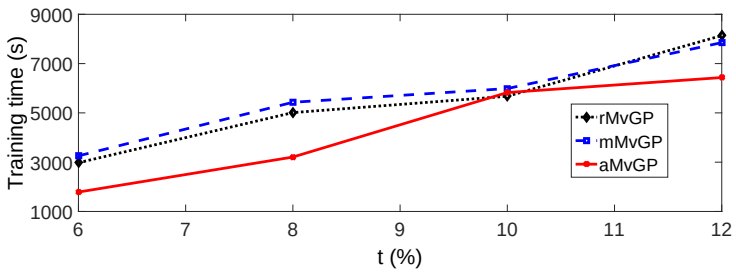

Fig. 3. Average training times on cora data. The x-axis corresponds to different settings of the size of sparse set, where t represents the percentage of the training set.

Combining the results here and classification results on the four web-page data sets, we find that aMvGP achieves the best performance. We have also given the training times of aMvGP, which indicates that when it is unaffordable to train on the original full data set, we can use the aMvGP to approach a good approximation.

6 Conclusion

In this paper, we have proposed the mIVM and aMP as two kinds of sparse multimodal methods for the multimodal GPs. The mIVM is inspired by information theory, seeking the maximum amount of information from all the modalities with the same number of data points. The aMP is more intuitive, which adopts an alternative selection strategy to utilize data from all the modalities for preserving the high space connectivity. We apply the two proposed sparse multimodal methods to multi-view GPs to verify the effectiveness. The classification accuracies on four web-page data sets and the cora data set have shown that aMvGP outperforms other competitive methods. The scalability was also tested on preliminary experiments with tiny proportions of data for training. More experiments will be conducted in the future.

Acknowledgments. The corresponding author Shiliang Sun would like to thank supports from the National Natural Science Foundation of China under Projects 61673179 and 61370175, Shanghai Knowledge Service Platform Project (No. ZF1213), and the Fundamental Research Funds for the Central Universities.

References

1. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
2. Williams, C.K., Seeger, M.: Using the Nyström method to speed up kernel machines. Advances in Neural Information Processing Systems **13** (2000) 661–667
3. Csató, L., Opper, M.: Sparse on-line Gaussian processes. Neural Computation **14** (2002) 641–668

4. Lawrence, N., Seeger, M., Herbrich, R.: Fast sparse Gaussian process methods: The informative vector machine. Advances in Neural Information Processing Systems **15** (2003) 625–632
5. Quiñonero-Candela, J., Rasmussen, C.E.: A unifying view of sparse approximate Gaussian process regression. Journal of Machine Learning Research **6** (2005) 1939–1959
6. Snelson, E., Ghahramani, Z.: Sparse Gaussian processes using pseudo-inputs. Advances in Neural Information Processing Systems **18** (2006) 1257–1264
7. Titsias, M.K.: Variational learning of inducing variables in sparse Gaussian processes. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. (2009) 567–574
8. Hensman, J., Fusi, N., Lawrence, N.D.: Gaussian processes for big data. In: Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence. (2013) 282–290
9. Cheng, C.A., Boots, B.: Incremental variational sparse Gaussian process regression. Advances in Neural Information Processing Systems **29** (2016) 4410–4418
10. Gal, Y., van der Wilk, M., Rasmussen, C.E.: Distributed variational inference in sparse Gaussian process regression and latent variable models. Advances in Neural Information Processing Systems **27** (2014) 3257–3265
11. Deisenroth, M.P., Ng, J.W.: Distributed Gaussian processes. In: Proceedings of the 32nd International Conference on Machine Learning. (2015) 1481–1490
12. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning. (2011) 689–696
13. Sun, S.: A survey of multi-view machine learning. Neural Computing and Applications **23** (2013) 2031–2038
14. Rao, D., De Deuge, M., Nourani-Vatani, N., Williams, S.B., Pizarro, O.: Multimodal learning and inference from visual and remotely sensed data. The International Journal of Robotics Research **36** (2016) 24–43
15. Sun, S., Hussain, Z., Shawe-Taylor, J.: Manifold-preserving graph reduction for sparse semi-supervised learning. Neurocomputing **124** (2014) 13–21
16. Shon, A.P., Grochow, K., Hertzmann, A., Rao, R.P.N.: Learning shared latent structure for image synthesis and robotic imitation. Advances in Neural Information Processing Systems **19** (2005) 1233–1240
17. Yu, S., Krishnapuram, B., Rosales, R., Rao, R.B.: Bayesian co-training. Journal of Machine Learning Research **12** (2011) 2649–2680
18. Xu, C., Tao, D., Li, Y., Xu, C.: Large-margin multi-view Gaussian process for image classification. In: Proceedings of the 5th International Conference on Internet Multimedia Computing and Service. (2013) 7–12
19. Liu, Q., Sun, S.: Multi-view regularized Gaussian processes. In: Proceedings of the 21st Pacific-Asia Conference on Knowledge Discovery and Data Mining. (2017) 655–667
20. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. International Journal of Computer Vision **88** (2010) 169–188