Active Visual Recognition with Expertise Estimation in Crowdsourcing

Chengjiang Long, Gang Hua
Stevens Institute of Technology
Hoboken, NJ 07030
{clong, ghua}@stevens.edu

Ashish Kapoor
Microsoft Research
Redmond, WA 98052
{akapoor}@microsoft.com

Abstract

We present a noise resilient probabilistic model for active learning of a Gaussian process classifier from crowds, i.e.*, a set of noisy labelers. It explicitly models both the overall label noises and the expertise level of each individual labeler in two levels of flip models. Expectation propagation is adopted for efficient approximate Bayesian inference of our probabilistic model for classification, based on which, a generalized EM algorithm is derived to estimate both the global label noise and the expertise of each individual labeler. The probabilistic nature of our model immediately allows the adoption of the prediction entropy and estimated expertise for active selection of data sample to be labeled, and active selection of high quality labelers to label the data, respectively. We apply the proposed model for three visual recognition tasks, i.e, object category recognition, gender recognition, and multi-modal activity recognition, on three datasets with real crowd-sourced labels from Amazon Mechanical Turk. The experiments clearly demonstrated the efficacy of the proposed model.*

1. Introduction

As research on visual recognition evolving more towards an experimental science, partly due to the success of the introduction of machine learning approach to computer vision [21, 13, 14], collecting labeled visual datasets at large scale from crowd-sourcing tools such as Amazon Mechanical Turk has become a common practice [6, 23]. Although it is cheap to obtain large quantity of labels through crowdsourcing, it has been well known that the collected labels could be very noisy. So it is desirable to model the expertise level of the labelers to ensure the quality of the labels [6, 23, 1]. The higher the expertise level a labeler is at, the lower the label noises he/she will produce.

Previous works for modeling the labelers' expertise mainly adopted two approaches. The first approach attempts to evaluate the labelers by adopting a pre-labeled gold standard dataset [1]. When a labeler is constantly generating contradicting labels on data samples from the gold standard dataset, all labels from that labeler may be discarded as he/she is highly likely to be an irresponsible one. The second approach addresses this issue through evaluating the labels by collecting multiple labels for each data sample [6, 23]. Then online or postmortem majority voting, or majority model consistency check is conducted to obtain the more likely ground-truth label of the data sample. The basic assumption is that majority of the labelers are behaving in good faith.

The first approach is able to evaluate the labelers online, which is desirable. But it needs to pre-label a set of data to serve as the gold standard, which may be an obstacle by itself. The second approach focuses on the label noise. It does not explicitly evaluate the labelers, although it may be extended to do so by online tracking how often a labeler is contradicting with the majority. Notwithstanding their demonstrated success, these two approaches are rather *Ad Hoc*. There lacks a principled approach to jointly model the global noise level of the labels and the expertise level of each individual labeler, in the absence of gold standard labels, which is what we want to achieve in this paper.

We present a Bayesian model (Figure 1), which explicitly models the global noise level of the labels and the expertise level of each individual labeler from crowds (*i.e.*, a group of noisy labelers). These two different statistics are modeled hierarchically with two levels of flip models [16]. Expectation propagation [16] is used for approximate Bayesian inference of the posterior of the latent classification function. A generalized Expectation Maximization (GEM) algorithm is conducted to estimate both quantities. The resulting classifier is more resilient to label noises, adapting to the expertise of labelers.

Another improvement that can be made to current crowdsourcing labeling system such as Amazon Mechanical Turk (AMT) is to make it actively guide the labelers for more efficient labeling. The proposed Bayesian model enables not only active selection of data samples to be labeled, but also active selection of quality labelers. These are enabled by the probabilistic nature of our model and the explicit modeling of both global label noise and expertise of each individual labeler, thereby allowing entropy based uncertainty measure to be readily adopted for these purposes.

Several aspects distinguish our work from previous active learning based labeling [23, 2, 11, 15, 9]: first of all, our work deals with active learning with multiple labelers, a topic which has not been sufficiently explored before. Secondly, we do not assume that the labels provided by the la-

1550-5499/13 $31.00 © 2013 IEEE
DOI 10.1109/ICCV.2013.373

belers are absolutely correct. In other words, the labeler may label an example incorrectly. Most previous work on active learning has assumed that the labels provided by the human oracle is noise free. Thirdly, our model allows online evaluation of the quality of the labelers without relying on any additional pre-labeled gold standard data. Hence we can select higher quality labelers and reduce the noise level of the labels we collected.

The main contributions of this paper are: (1) a Bayesian probabilistic formulation to learn a Gaussian process classifiers from multiple noisy labels, which models both the global label noise and expertise of each individual labeler; and (2) an active classifier learning system which determines which users to label which unlabeled examples. We apply our proposed model on datasets with real noisy labels obtained from Amazaon Mechanical Turk on three visual recognition tasks, *i.e.*, object category recognition, gender recognition, and multi-modal activity recognition. The results clearly demonstrated the efficacy of our proposed model.

2. Related work

Related works can be grouped into three categories including noise resilient Gaussian process classifiers [26, 10, 12], approximate Bayesian inference methods [17, 10, 18, 16], and active learning algorithms embracing crowdsourced labels [8, 25, 1, 23].

In the case of Gaussian process classifier, the TAP style mean field approximation [18] is equivalent to the more general Expectation Propagation (EP) for approximate inference, which is firstly proposed by Minka [16]. A noise resilient likelihood model, namely flip noise model, is introduced in [16] to better handle label noises in Gaussian process classifier. More recently, Kim and Ghahramani [12] exploited the flip noise model to explicitly handle outlier labels in Gaussian process classifier. An EM algorithm built on top of EP is proposed to estimate the label noise levels online. None of these methods ever considers the case when a data sample has multiple copies of noisy labels, which is the focus of our proposed approach. Several previous works have explored active learning from noisy crowd-sourced labels [1, 23] in different domains, where the two aforementioned approaches are exploited to handle label noise.

To better mitigate label noises online in the absence of gold standard labels, Donmez *et al*. [8, 7] have explored confidence interval based estimation and sequential Bayesian estimation method to evaluate the label quality of the annotators in both stationary and non-stationary cases. Zhao *et al*. [29] proposed an incremental relabeling mechanism which employed active learning to not only select the unlabeled data to be labeled by the crowds, but also select already labeled data samples to be relabeled until sufficient confidence is built. Raykar *et al*. [20, 19] proposed a probabilistic model, which assumes independence of the annotator judgement given the true label, and alternatively conducts model learning and performance evaluation of the multiple annotators.

Dekel and Shamir [4] adapted the formulation of support vector machines (SVMs) to identify low quality or malicious annotators, which assumes that each annotator is either good or bad. Later, they [5] proposed a method for pruning low-quality lablers by using the model trained from the entire labeled dataset from all labelers as ground truth. Chen *et al*. [3] proposed to identify good annotators by spectral clustering in the worker space. The assumption is that good labelers will behave similarly. Yan *et al.* [27, 28] presented a Bayesian model and adopted a logistic regression function to model the labeler's quality. These works build insights on how to deal with label noises and evaluate labeler quality. They lack explicit joint modeling of both the label noises and the lablers' quality. Our proposed approach models both in a unified way.

For modeling annotators' quality for image labeling from the crowds, the most relevant works to our research is Welinder and Perona [25] and Welinder *et al.* [24]. In both works, a parameter vector of each annotator which represents the annotator's expertise, a feature (parameter) vector associated with each image which encodes each annotator's visual response to the image, and a linear classifier operating on the parameter vector associated with each image, are all inferred from existing labels provided by the annotators through a Bayesian model.

We shall emphasize that in these two pieces of work proposed by Welinder and Perona [25] and Welinder *et al.* [24], no visual features are directly extracted from the set of images and for one image, at least one label is needed to be able to infer the parameter vector associated with it. Hence the classifier induced from their models can not be applied directly to an unlabeled sample, because there is no feature vector to operate on. In this sense, their models provided a more principled way for active data re-labeling. In contrast, our proposed model actively induces a classifier which directly operates on visual features that directly extracted from images, which models the labelers' quality in a principled way to facilitate active selection of annotators for providing better quality labels.

3. Formulation, inference, and learning

Given a set of N data points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^D$, each of which may be labeled by M labelers. We denote s_i to be the latent random variable with a Gaussian process prior. The true label of $\mathbf{x}_i$ is denoted as $y_i \in \{-1, 1\}$, which is hidden. The observed label of $\mathbf{x}_i$ from labeler j is denoted as $t_{ij} \in \{-1, 1\}$, which could be noisy, meaning the t_{ij} may not be consistent with the hidden true label y_i. We denote $\mathbf{t}_i = \{t_{ij}\}_{j=1}^{M}$ as the set of labels from the M labelers for $\mathbf{x}_i$. For notation simplification, we denote $\mathbf{S} = \{s_1, s_2, ..., s_N\}$, $\mathbf{Y} = \{y_1, y_2, ..., y_N\}$ and $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N\}$. Our objective is to build a probabilistic model to robustly infer y_i for each $\mathbf{x}_i$.

3.1. Probabilistic Model

The proposed probabilistic model is illustrated in the graphical model in Figure 1. The conditional joint proba-

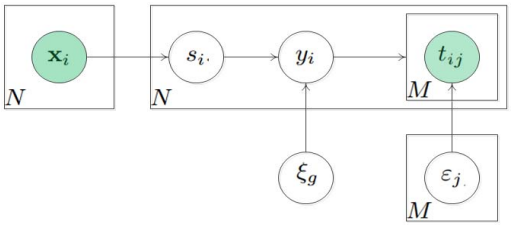


Figure 1: Graphical model of the proposed Gaussian process classifier, with multiple noisy labels from the crowds.

bility of this probabilistic model is defined as

$$p(\mathbf{T},\mathbf{Y},\mathbf{S}|\mathbf{X},\vec{\varepsilon}) \propto \frac{1}{Z}p(\mathbf{S}|\mathbf{X})\prod_i p(y_i|s_i,\xi_g)\prod_{i,j} p(t_{ij}|y_i,\varepsilon_j), \tag{1}$$

where $\vec{\varepsilon}=\{\{\varepsilon_j\}_{j=1}^M,\xi_g\}$, ξ_g is the global label noise measure, and ε_j is the label quality measure for labeler j, and Z is the partition function.

In our model, $p(\mathbf{S}|\mathbf{X})$ is a Gaussian process prior [26] to ensure that similar data samples to have similar prediction scores. Formally, it is defined as

$$p(\mathbf{S}|\mathbf{X}) \sim \mathbf{N}(\mathbf{S}|\mathbf{0},\mathbf{K}), \tag{2}$$

where $\mathbf{K}=[k(\mathbf{x}_i,\mathbf{x}_j]_{i,j=1}^N$ is a kernel matrix defined over the set of all N data samples. Any valid kernel function, which measures the similarity between two data samples, can be used here.

The conditional likelihood probability $p(y_i|s_i)$ is defined as a flip model [16], *i.e.*,

$$p(y_i|s_i,\xi_g) = \xi_g\Theta(y_i s_i) + (1-\xi_g)\Theta(-y_i s_i), \tag{3}$$

where $\Theta(\rho)=1$ if $\rho>0$, and $\Theta(\rho)=0$ otherwise. In other words, the *a posteriori* estimation of y_i takes the sign of the predicted soft label s_i with probability ξ_g. This treatment make the GPC to be resilient to label noise and outliers [12].

The conditional likelihood probability $p(t_{ij}|y_i,\varepsilon_j)$ is also modeled as a flipping noise model, *i.e.*,

$$p(t_{ij}|y_i,\varepsilon_j) = \varepsilon_j\Theta(y_i t_{ij}) + (1-\varepsilon_j)\Theta(-y_i t_{ij}). \tag{4}$$

Intuitively, with probability $1-\varepsilon_j$, t_{ij} will be a flipped version of y_i. Therefore, the larger ε_j is, the higher the probability that t_{ij} will agree with the true label y_i, and vice versa. Hence, ε_j naturally represents the expertise or quality of the labels induced by labeler j. We note here that unlike in [16], we parameterize this model based on label quality, which is one minus the label noise.

3.2. Inference

As a matter of fact, this two-level flip model can be conveniently collapsed by integrating y_i out to obtain the joint probability

$$\begin{aligned} p(\mathbf{t}_i|s_i,\vec{\varepsilon}) \quad &= p(y_i=+1|s_i,\xi_g)\textstyle\prod_j p(t_{ij}|y_i=+1,\epsilon_j) \\ &+p(y_i=-1|s_i,\xi_g)\textstyle\prod_j p(t_{ij}|y_i=-1,\epsilon_j), \end{aligned} \tag{5}$$

We can rewrite the joint probability in Equation 1 as

$$p(\mathbf{T},\mathbf{S}|\mathbf{X},\vec{\varepsilon}) = p(\mathbf{S}|\mathbf{X})\prod_i p(\mathbf{t}_i|s_i,\vec{\varepsilon}), \tag{6}$$

This collapsed joint probability will help us to more conveniently derive the EP inference algorithm.

For the proposed Bayesian framework, we assume that we are given a set of labeled data samples $\mathbf{X}_L=\{\mathbf{x}_1,...,\mathbf{x}_L\}$, and the set of labels are denoted as $\mathbf{T}_L=\{t_{ij}|1\le i\le N, 1\le j\le M\}$. We denote $\mathbf{D}_L=\{\mathbf{X}_L,\mathbf{T}_L\}$, $\mathbf{S}=\{\mathbf{S}_L,s_u\}$, and $\mathbf{X}=\{\mathbf{X}_L,\mathbf{x}_u\}$, where $\mathbf{x}_u$ is an unlabeled data sample. To predict the label y_u of a $\mathbf{x}_u$, we need to solve the following Bayesian inference problem, *i.e.*,

$$\begin{aligned} p(y_u|\mathbf{x}_u,\mathbf{D}_L) &= \int_{\mathbf{S}} p(y_u|s_u)p(\mathbf{S}|\mathbf{D}_L,\mathbf{x}_u)d\mathbf{S} \\ &= \int_{s_u} p(y_u|s_u)\int_{\mathbf{S}_L} p(\mathbf{S}|\mathbf{D}_L,\mathbf{x}_u)d\mathbf{S}_L ds_u \end{aligned} \tag{7}$$

where

$$p(\mathbf{S}|\mathbf{D}_L,\mathbf{x}_u) \propto p(\mathbf{S}|\mathbf{X})\prod_{s_i\in\mathbf{S}_L} p(\mathbf{t}_i|s_i,\vec{\varepsilon}). \tag{8}$$

The integral in Equation 7 is intractable as neither $p(\mathbf{S}|\mathbf{D}_L,\mathbf{x}_u)$ nor $p(y_u|s_u)$ can be integrated in close form. We resort to Expectation Propagation [16] to obtain an approximate integral by approximating $p(\mathbf{S}|\mathbf{D}_L,\mathbf{x}_u)$ to be a Gaussian, *i.e.*,

$$Q(\mathbf{S}) = p(\mathbf{S}|\mathbf{X})\prod_i \tilde{F}_i(s_i) \sim \mathcal{N}(\mathbf{S}|\mathbf{m},\mathbf{\Sigma}),, \tag{9}$$

where $\mathbf{m}=[m_1,m_2,\ldots,m_L]$ and $\mathbf{\Sigma}=[\sigma_{ij}]_{i,j=1}^N$ are the mean vector and covariance matrix of the Gaussian distribution $Q(\mathbf{S})$ each $\tilde{F}_i(s_i)$ is a Gaussian distribution with mean $\tilde{m}_i$, variance v_i, and normalization constant A_i, i.e.,

$$\tilde{F}_i(s_i) = A_i\exp(-\frac{1}{2v_i}(s_i-\tilde{m}_i)^2). \tag{10}$$

which approximates the joint likelihood of the set of all labels obtained for $\mathbf{x}_i$, *i.e.*,

$$\tilde{F}_i(s_i) \approx p(\mathbf{t}_i|s_i,\vec{\varepsilon}). \tag{11}$$

Since the prior $p(\mathbf{S}|\mathbf{X})$ is a Gaussian by definition, hence $Q(\mathbf{S})$ will also be a Gaussian distribution. Note this approximation is in contrast to previous work using EP for inference in GPC in the sense that the approximation is performed over the joint likelihood of a set of labels on a single data. Most previous work only considered the case of a single label for each data. Instead of solving for each $\tilde{F}_i(s_i)$ independently, we use EP [16] to obtain a better overall approximation. The exact steps of the inference algorithm are omitted due to space limit.

EP obtains a Gaussian approximation $Q(\mathbf{S})$ to the posterior distribution $p(\mathbf{S}|\mathbf{D}_L,\mathbf{x}_u)$. Hence the integral over

$\mathbf{S}_L$ in Equation 7 can also be approximated by a Gaussian distribution over s_u, i.e., $\mathcal{N}(s_u|m_u, v_u)$, where m_u and v_u can be obtained in closed form. Denote $\tilde{\mathbf{m}} = [\tilde{m}_1, \tilde{m}_2, \ldots, \tilde{m}_L]^T$ be the concatenation of the mean value of each $\tilde{F}_i(s_i)$, we have

$$m_u = \mathbf{k}_u^T(\mathbf{K}+\mathbf{\Lambda})^{-1}\tilde{\mathbf{m}} \quad (12)$$

$$v_u = k(\mathbf{x}_u, \mathbf{x}_u) - \mathbf{k}_u^T(\mathbf{K}+\mathbf{\Lambda})^{-1}\mathbf{k}_u \quad (13)$$

where $\mathbf{k}_u = [k(\mathbf{x}_u, \mathbf{x}_1), k(\mathbf{x}_u, \mathbf{x}_2), \ldots, k(\mathbf{x}_u, \mathbf{x}_N)]^T$. We immediately have that the whole integral over all $\mathbf{S}$ in Equation 7 can be approximated as

$$p(y_u|\mathbf{x}_u, \mathbf{D}_L) \doteq (2\xi_g - 1)\Phi\left(\frac{y_u\mathbf{k}_u^T(\mathbf{K}+\mathbf{\Lambda})^{-1}\tilde{\mathbf{m}}}{\sqrt{k(\mathbf{x}_u, \mathbf{x}_u) - \mathbf{k}_u^T(\mathbf{K}+\mathbf{\Lambda})^{-1}\mathbf{k}_u}}\right) + 1 - \xi_g \quad (14)$$

where $\Phi(\cdot)$ is the Probit function. We subsequently predict the label y_u of $\mathbf{x}_u$ based on Equation 14.

3.3. Learning $\vec{\varepsilon}$ with Expectation Maximization

To online estimate the quality of both the labels and labelers, we need to online estimate the parameters $\vec{\varepsilon} = \left\{\xi_g, \{\varepsilon_j\}_{j=1}^M\right\}$, which represent the overall label quality and label quality of each labeler. We further develop a generalized Expectation-Maximization algorithm for estimating it. Note $\vec{\lambda}$ is a function of $\vec{\varepsilon}$, so we abuse the notation a bit and use them interchangeably. We start by building the lower bound F of the log likelihood, *i.e.*,

$$\begin{aligned}\log p(\mathbf{T}_L, \mathbf{S}_L|\mathbf{X}_L, \vec{\varepsilon}) &\geq \int_{\mathbf{S}_L} Q(\mathbf{S}_L)\log\frac{p(\mathbf{T}_L, \mathbf{S}_L|\mathbf{X}_L, \vec{\varepsilon})}{Q(\mathbf{S}_L)} \\ &= C + \sum_{i=1}^{L}\int_{s_i} q(s_i)\log p(\mathbf{t_i}|s_i, \vec{\varepsilon})ds_i. \quad (15)\end{aligned}$$

Then the following iterative steps form the EM algorithm

1. **E-Step:** Given the current parameter $\vec{\varepsilon}_p$, conduct the EP inference to obtain an approximate inference of $Q(\mathbf{S}_L) \sim p(\mathbf{S}_L|\mathbf{X}_L, \mathbf{T}_L)$.

2. **M-Step:** Maximize the lower bound of $\log p(\mathbf{T}_L, \mathbf{S}_L|\mathbf{X}_L, \vec{\varepsilon})$ in Equation 15 over $\vec{\varepsilon}$. $\vec{\varepsilon}_p \leftarrow \vec{\varepsilon}$, goto the **E-Step** and iterate until convergence.

For the **M-Step**, closed form solution of ξ_g and ϵ_j is not tractable. Hence we resort to the L-BFGS-B algorithm [30] to find a numerical estimation of them to maximize the lower bound F by gradient ascent, which is guaranteed to obtain a local optimal solution.

4. Bayesian Active Learning

For pool based active learning, we assume that we are given a pool of both labeled and unlabeled data samples $\mathbf{X} = \{\mathbf{X}_L, \mathbf{X}_U\}$, and $\mathbf{T}_L$ is the label set for $\mathbf{X}_L$ from M labelers. The proposed model conveniently allows for both active selection of unlabeled data samples to be labeled, and also active selection of higher quality labelers.

For active sample selection, a criterion that can readily be adopted is the entropy $H(y_u) = -\sum_{y_u\in\{1,-1\}} p(y_u|\mathbf{x}_u, \mathbf{D}_L)\log p(y_u|\mathbf{x}_u, \mathbf{D}_L)$ of the predicted label y_u on unlabeled data $\mathbf{x}_u$. We select the most uncertain unlabeled example to be labeled, i.e.,

$$\mathbf{x}_u^* = \arg\max_{\mathbf{x}_u\in\mathbf{X}_U} H(y_u), \quad (16)$$

where $p(y_u|\mathbf{x}_u, \mathbf{D}_L)$ can be obtained using the EP algorithm introduced in Section 3.2.

Note ϵ_j in our model directly models the labeler j's quality. It can be regarded as the probability that labeler j would label the data correctly. Therefore, the higher ϵ_j is, the better quality the labeler has. In our active learning process, we can naturally select the top $K < M$ labelers with the top K ϵ_j to label a selected data sample, where ϵ_j is estimated by the EP-GEM algorithm presented in Section 3.3. The joint active selection of both labelers and data samples greatly facilitates to obtain higher quality labels.

Another active learning strategy is to only actively select the data sample to be labeled by all M labelers. Our model indeed can benefit from the multiple labels, even though there may be noises. We also compare this strategy with online majority voting in our experiments.

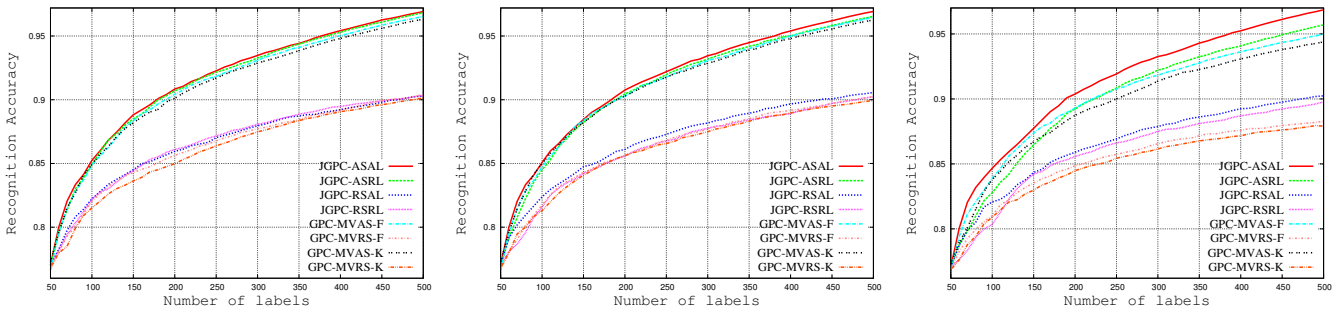
5. Experiments

Our experiments are conducted on three datasets with real crowd-sourced labels. In our experiments, we use RBF kernel unless otherwise specified.

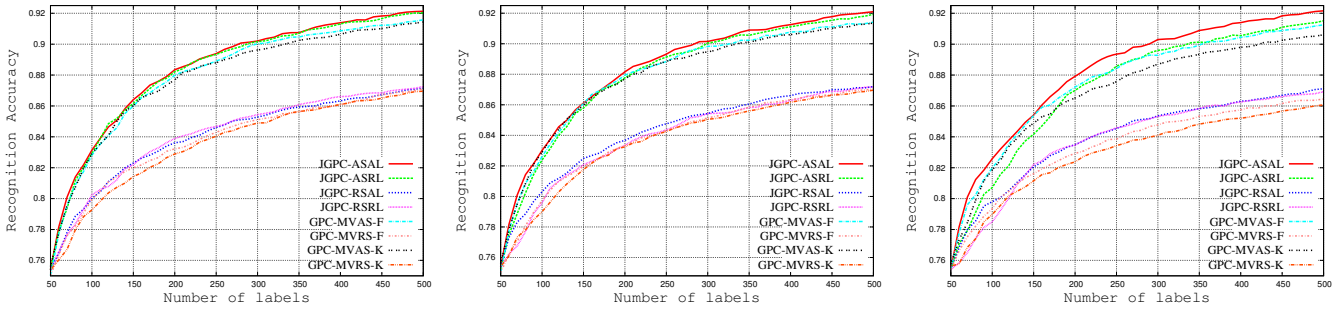
5.1. Datasets

The first dataset is composed of 3 classes of images from the ImageNet grand challenge [6], which includes 2 category of dogs, i.e., "Yorkshire terrier", "English setter" plus the "Meerkat, meerkat" category. These three classes of images are among the top 10 in the ImageNet grand challenge in terms of number of labeled examples, which have 3047, 2426 and 2341 labeled images, respectively. We repush these images back to Amazon Mechanical Turk and obtained 7 copies of labels for each image. The percentages of the labels which are correct (*i.e.*, agreeing with the ground-truth labels from ImageNet) are 97.87%, 96.83% and 99.27%, respectively. The features we used to represent each image is the local coodinate coding (LCC) [14] on densely extracted HoG features with 4096 codewords. The LCC features is pooled in 10 spatial cells, resulting a 40960 dimensional feature.

The second dataset we take experiments on is a subset of the CMU multi-modal action category dataset (CMU-



(a) Active learning pool - 2 bad labelers (c) Active learning pool - 3 bad labelers (e) Active learning pool - 4 bad labelers



(b) Hold-out testing set - 2 bad labelers (d) Hold-out testing set - 3 bad labelers (f) Hold-out testing set - 4 bad labelers

Figure 2: Recognition performance on the "Meerkat, meerkat" class with different number of bad labelers.

MMAC) [22], where crowd-sourced labels have been obtained by Zhao *et al*. [29]. In total there are 2682 labeled video clips, each has 7 copies of labels from Amazon Mechanical Turk. The action labels include: 1. close; 2. crack; 3. open; 4. pour; 5. put; 6. read; 7. spray; 8. stir; 9. switch on; 10. take; 11. twist off; 12. twist on; 13. walk and 14. others. The corresponding number of clips for each action are: 7, 54, 711, 112, 453, 116, 43, 94, 654, 103, 290, 11, 12, and 22, respectively. Since many of the classes have limited labeled clips, and also considering that the raw label accuracy of action 3 and action 5 is less than 50%, which fail all the classifiers we tried. We choose to work on the classification problem of action 9 only, which has sufficient number of labeled clips and its label accuracy is 75.56% . Since the CMU-MMAC dataset incorporates multiple modality, instead of using visual features extracted from video frames, we use the feature extracted from the IMU modality provided by Zhao *et al*. [29]. The feature dimension is 180. We refer the reader to [29] for more details on how the features are extracted.

The last dataset for our experiment is a face gender dataset, where we try to learn a gender classifier from facial features. We collected 5 copies of gender labels for 9441 face images. The face images are all 64×64. We extract a 5408 dimensional features from each face image. This feature extractor is a convolutional neural network trained for gender recognition with a separate small set of labeled gender face images. The feature is the output of the last layer of the convolutional neural network. We will share the features of this data upon publication of our paper.

5.2. Experiments on ImageNet Dataset

Effectiveness of Labeler Selection: The simulation experiment we conducted is on the ImageNet dataset. To demonstrate the effectiveness of our model to avoid low quality labelers, we use the class "Meerkat, meerkat" as an example. We take 1000 images as positive samples from the class "Meerkat, meerkat" and 1000 images from the other two dog classes to serve as the hold-out testing set. The rest of the images in "Meerkat, meerkat" and an equal number of images from the other two classes are put together to form the active learning pool. We simulate the case that there are 2, 3, 4 bad labelers, who would randomly assign a label to the sample, so there is 50% chance that the label from them will be erroneous. For good labelers, we used the noisy labels obtained from Amazon Mechanical Turk. Therefore, We run our proposed active learning algorithm for both active selection of data samples and labelers. 3 labelers with higher estimated label quality ϵ_j are selected to provide the labels for the actively selected samples. A linear kernel function is adopted due to the LCC features we used.

We name our algorithm as JGPC-ASAL, stands for joint learning GPC with active selection of both samples and labelers (we call it joint learning in the sense that the multiple labels of a single example is jointly considered). We compare with a combination of other learning strategies with our model, such as active selection of samples but random selection of labelers, random selection of samples and active selection of labelers, and random selection of both samples and labelers. We call these three algorithms JGPC-ASRL,

JGPC-RSAL and JGPC-RSRL, respectively. For all these online learning algorithms based on JGPC, we select 3 labelers to provide the label using the corresponding criterion for labeler selection.

One algorithm we compare against is an active learning GP classifier with the global flip noise observation model similar to the model in [12]. For this method, at each round, we use the prediction entropy to select the next sample to be labeled and majority voting is performed to obtain a single label from all 7 copies of labels. We name it as majority vote active learning GPC with flip noise model, or in short GPC-MVAS-F. The corresponding algorithm performing random sample selection using majority voted label, is named as GPC-MVRS-F. Another algorithm we compare against is based on the active learning GP classifier proposed by Kapoor *et al.* [2], where a Gaussian observation model is adopted, and a confidence criterion normalized by the variance of the posterior prediction is adopted for active learning. Again, majority voting is performed at each active learning step to obtain a single label from all 7 copies of noisy labels. We name this algorithm as GPC-MVAS-K, and its random sample selection version is named GPC-MVRS-K. Since there are no labeler selection mechanism, we simply gather majority voted labels from all 7 labelers.

Figure 2 presents the recognition accuracy evolving with increasing number of labeled examples for all the competing methods with 2 (Figure 2a and 2b), 3 (Figure 2c and 2d) ,and 4 (Figure 2e and 2f) bad labelers. We have the following observations: (1)overall, in all cases, the recognition accuracy of the proposed JGPC-ASAL is constantly ranked on the top, in both the active learning pool and the hold-out testing set, which is not affected by the number of bad labelers due to the active selection of higher quality labelers. (2). The JGPC-ASRL algorithm achieves more or less the same accuracy than the JGPC-ASAL when there is only 2 bad labelers, and degraded with the increased number of bad labelers. Suggesting that when the labels are less noisy, active selection of samples are more important than active selection of labelers, which intuitively makes sense as the label quality is high. (3). The GPC-MVAS-F algorithm outperformed GPC-MVAS-K, which revalidated the advantage of the flip noise model based observation model over a simple Gaussian observation model in a Gaussian process classifier. (4). In all cases for all algorithms, the active sample selection strategy always outperforms its random sample selection counterpart, which suggest that the proposed active learning criterion is robust against label noises.

Figure 3 visualizes the top three labelers selected at each active learning step when running the proposed JGPC-ASAL algorithm on the "Meerkat, meerkat" class with 4 bad labelers (labeler 4, 5, 6, and 7 are bad labelers). The red, blue, and green color circles represents the top three labelers selected based on the estimated labeler quality measure at each active learning step. As we can clearly observe, at the beginning, the users selected are more or less uniform across the 7 labelers. Then with the progression of the active learning process, the three good labelers (labeler 1,2,and 3) got constantly selected. This demonstrated the efficacy of the proposed model for online modeling of the labelers' quality.

Labelers with different expertise: To better understand the behavior of our algorithm, we run two sets of experiments with simulated label noises from the ground-truth labels on the 3 categories of ImageNet dataset. We use the "Meerkat, meerkat" as the positive class and the other two dog classes as the negative class. In the first experiment, we simulated the case that each labeler produces 10%, 15%, 20%, 25%, 30%, 35% and 40% erroneous labels, respectively. In the second experiment, we increase the label noise level to have each labeler to produce 15%, 20%, 25%, 30%, 35%, 40%, and 45% erroneous labels, respectively. We also impose a naive majority voting consensus based labeler selection scheme to the GPC-MVAS-F and GPC-MVRS-F algorithm. For each labeler, we record online his rate of consistent labels with the corresponding majority voted labels. Intuitively, the larger this rate, the better the labeler's quality. We call the GPC-MVAS-F and GPC-MVRS-F algorithm equipped with this simple active labeler selection scheme as GPC-MVASAL-F and GPC-MVRSAL-F, respectively. To validate its efficacy, we also compare against its corresponding random labeler selection version, namely GPC-MVASRL-F and GPC-MVRSRL-F. Again, at each step of the online learning process, we select 3 labelers to provide the labels.

Figure 4 presents the results of the two experiments. Our observations are: (1) the proposed JGPC-ASAL algorithm achieved slightly better accuracy than the GPC-MVASAL-F algorithm in the first experiment, and much better accuracy in the second experiment, which indicates that our proposed labeler selection criterion is more robust to label noises than the naive majority voting consensus based labeler selection criterion. (2) The naive majority voting consensus based labeler selection criterion is also effective, as it achieved better accuracy than its random labeler selection counterpart.

Experiments with real crowd-sourced labels: We further run experiments with all real crowd-sourced labels on the 3 classes of images. For each classes, we randomly sample 1000 examples from it and another 1000 examples from the other two classes to serve as the hold-out testing set. The rest of the examples in the target class and an equal number of examples from the other two classes are put in the active learning pool. Since the 7 copies of labels we collected from Amazon Mechanical Turk do not really entail labels from bad labelers, we found that active selection of higher quality labelers does not really improve recognition accuracy much. Hence in this experiments we only do active sample selection, and assume all the 7 labelers will all label the selected example. We call this algorithm under our proposed model to be JGPC-AS. We argue that our joint treatment of multiple labels in GPC in general is superior than the majority voting strategy (GPC-MVAS-F and GPC-MVAS-K), as manifested by the results shown in Figure 5. We also compare against two versions of the active learning algorithm proposed by Yan *et al.* [28, 27], namely ML-Bernoulli-AS and ML-Gaussian-AS, respectively.

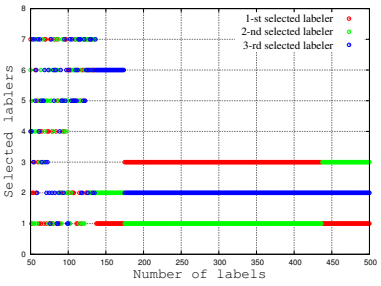
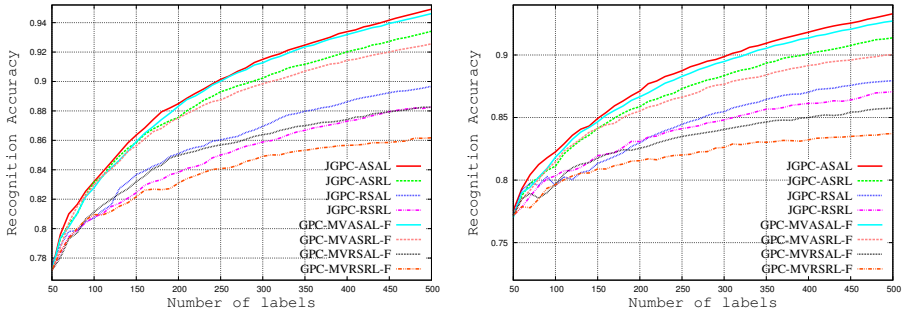
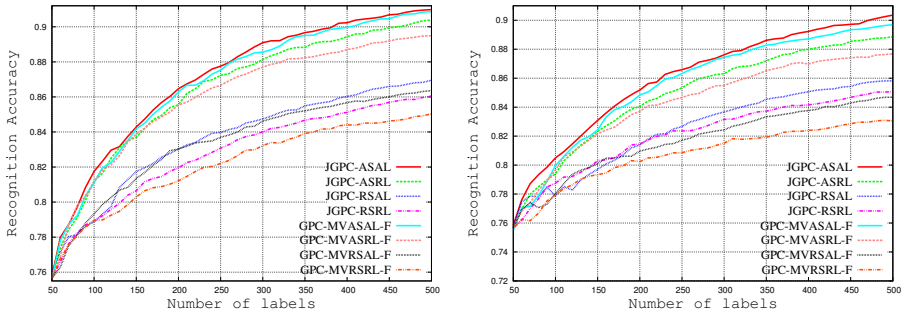


Figure 3: Selected labelers on "Meerkat, meerkat" class with 4 bad labelers.



(a) Active learning pool - 10%-40% for each labeler (c) Active learning pool - 15%-45% for each labeler



(b) Hold-out testing set - 10%-40% for each labeler (d) Hold-out testing set - 15%-45% for each labeler

Figure 4: Experiments with labelers with different level of label noises on the ImageNet dataset: (a)&(b) Each labeler is simulated with 10%, 15%, 20%, 25%, 30%, 35% and 40%, respectively; (c)&(d) Each labeler is simulated with 15%, 20%, 25%, 30%, 35%, 40%, and 45%, respectively.

The figure also presents the results of these competing GPC algorithms with random sample selection strategy, namely JGPC-RS, GPC-MVRS-F, and GPC-MVRS-K, respectively. The curves plotted in Figure 5 are averaged over the three classes over multiple runs with different initial labels to counter the statistic differences. As we can clearly observe, the proposed JGPC-AS is on par with GPC-MVAS-F and outperforms the GPC-MVAS-K algorithm in this dataset. Again, active sample selection always achieves better performance than random sample selection. The ML-Bernoulli-AS and ML-Gaussian-AS performed poorly on this dataset with real crowdsourced labels, which is not surprising as it induce a linear classifier.

5.3. Experiments on CMU-MMAC Dataset

For experiment on CMU-MMAC dataset, we take 250 clips of action 9 and 250 clips of the other actions to form the hold-out testing set. The rest 404 clips of action 9 and the same number of clips from the other actions are used as the active learning pool. As we can clearly observe in Figure 6, our proposed JGPC-AS algorithm consistently outperforms the GPC-MVAS-F and GPC-MVAS-K algorithms. This further validated the efficacy of our model.

5.4. Experiments on Gender Face Dataset

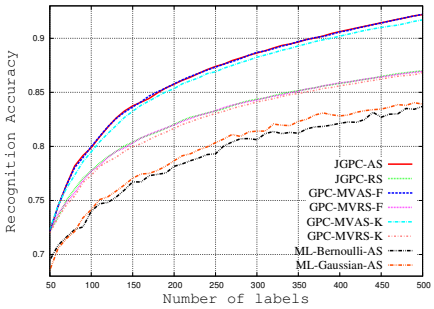
We hold out 2000 face images, for which all 5 copies of labels are in consensus, for testing purpose. The rest of the face images with different percentage of label inconsistency are used as the active learning pool. As shown in Figure 7, the proposed JGPC-AS algorithm again showed superior recognition accuracy when compared with the GPC-MVAS-F and GPC-MVAS-K algorithms, in both the active learning pool and the hold-out testing set. It is also obvious that algorithms performing active learning always achieved better performance when compared with their random learning counterparts.

6. Conclusion

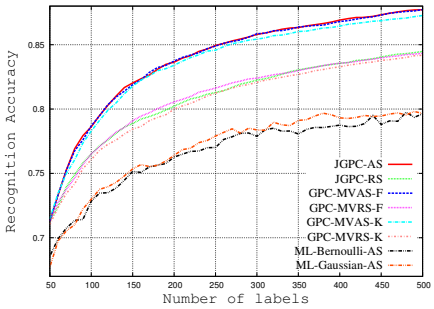
In this paper, we present a hierarchical Bayesian model to learn a Gaussian process classifier from crowd-sourced labels by jointly considering multiple labels instead of taking the majority voting. Our two-level flip model enables us to design principled active learning strategy to not only select data sample, but also select quality labelers. Our experiments on three visual recognition datasets with real-crowdsourced labels clearly demonstrated that the active selection of labelers is beneficial when there are a lot of careless labelers. Our joint treatment of multiple labels for each data sample is also proven to be superior to the online majority voting scheme. The Gaussian process classifier learned from our model consistently outperforms the one learnt using majority voting strategy. Our future work will further explore how to design an active learning machine to jointly select both the user and sample in a single criterion.

Acknowledgement

This work is partly supported by US National Science Foundation Grant IIS 1350763, China National Natural Sci-

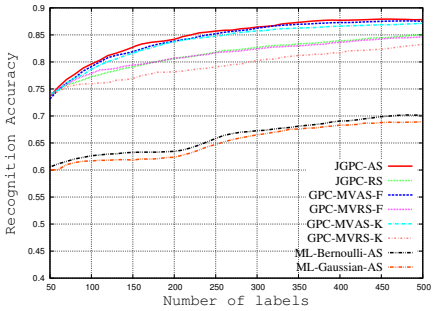


(a) Active learning pool

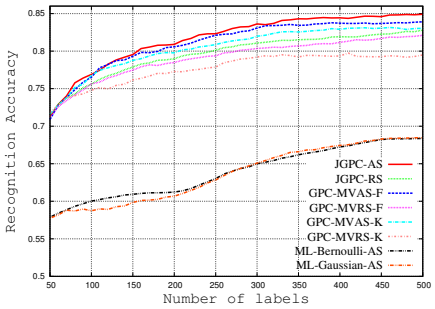


(b) Hold-out testing set

Figure 5: ImageNet dataset.

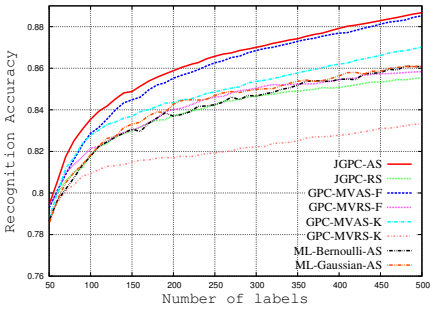


(a) Active learning pool

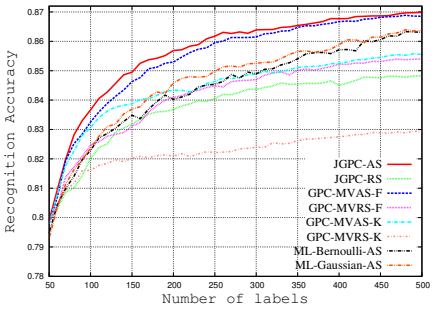


(b) Hold-out testing set

Figure 6: CMU-MMAC dataset.



(a) Active learning pool



(b) Hold-out testing set

Figure 7: Face gender dataset.

ence Foundation Grant 61228303, GH's start-up funds from Stevens Institute of Technology, a Google Research Faculty Award, a gift grant from Microsoft Research, and a gift grant from NEC Labs American.

References

[1] V. Ambati, S. Vogel, and J. Carbonell. Active learning and crowd-sourcing for machine translation. In *LREC*, 2010. 1, 2

[2] R. U. Ashish Kapoor, Kristen Grauman and T. Darrell. Active learning with gaussian processes for object categorization. *ICCV*, 2007. 1, 6

[3] S. Chen, J. Zhang, G. Chen, and C. Zhang. What if the irresponsible teachers are dominating? a method of training on samples and clustering on teachers. In *AAAI*, 2010. 2

[4] O. Dekel and O. Shamir. Good learners for evil teachers. In *ICML*, 2009. 2

[5] O. Dekel and O. Shamir. Vox populi: Collecting high-quality labels from a crowd. In *COLT*, 2009. 2

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248 –255, June 2009. 1, 4

[7] P. Donmez, J. Carbonell, and J. Schneider. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In *SDM*, 2010. 2

[8] P. Donmez, J. G. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *SIGKDD*, 2009. 2

[9] S. Ebert, M. Fritz, and B. Schiele. Ralf: A reinforced active learning formulation for object class recognition. In *CVPR*, 2012. 1

[10] M. Gibbs and D. Mackay. Variational gaussian process classifiers. *T-NN*, 2000. 2

[11] A. Kapoor, G. Hua, A. Akbarzadeh, and S. Baker. Which faces to tag: Adding prior constraints into active learning. *ICCV*, 2009. 1

[12] H.-C. Kim and Z. Ghahramani. Outlier robust gaussian process classification. In *SSPR/SPR*, pages 896–905, 2008. 2, 3, 6

[13] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[14] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale image classification: Fast feature extraction and svm training. In *CVPR*, 2011. 1, 4

[15] C. Loy, T. Hospedales, T. Xiang, and S. Gong. Stream-based joint exploration-exploitation active learning. In *CVPR*, 2012. 1

[16] T. Minka. *A family of algorithms for approximate Bayesian inference*. Ph.d. thesis, MIT, 2001. 1, 2, 3

[17] R. M. Neal. Monte carlo implementation of gaussian process models for bayesian regression and classification. Technical Report CRGTR972, University of Toronto, 1997. 2

[18] M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *NC*, 2000. 2

[19] V. C. Raykar and S. Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *JMLR*, 2012. 2

[20] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *ICML*, 2009. 2

[21] J. Sanchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *CVPR*, 2011. 1

[22] E. H. Spriggs, F. D. L. Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPRW*, 2009. 5

[23] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR*, 2011. 1, 2

[24] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *NIPS*, 2010. 2

[25] P. Welinder and P. Perona. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *CVPRW*, 2010. 2

[26] C. Williams and D. Barber. Bayesian classification with gaussian processes. *T-PAMI*, 20(12):1342 –1351, dec 1998. 2, 3

[27] Y. Yan, R. Rosales, G. Fung, and J. Dy. Active learning from multiple knowledge sources. In *AISTATS*, 2012. 2, 6

[28] Y. Yan, R. Rosales, G. Fung, and J. G. Dy. Active learning from crowds. *ICML*, pages 1161–1168, 2011. 2, 6

[29] L. Zhao, G. Sukthankar, and R. Sukthankar. Incremental relabeling for active learning with noisy crowdsourced annotations. In *SocialCom*, 2011. 2, 5

[30] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 1997. 4