An Efficient Method to Estimate Labelled Sample Size for Transductive LDA(QDA/MDA) Based on Bayes Risk

Han Liu[1], Xiaobin Yuan[2], Qianying Tang[3], and Rafal Kustra[2]

[1] Department of Computer Science, University of Toronto
M5S 3G4 Toronto, Canada
hanliu@cs.toronto.edu
[2] Department of Statistics, University of Toronto, M5S 3G3 Toronto, Canada
{yuanx,r.kustra}@utoronto.ca
[3] Department of Electronics and Computer Engineering, University of Toronto
tangq@eecg.utoronto.ca

Abstract. As semi-supervised classification drawing more attention, many practical semi-supervised learning methods have been proposed. However,one important issue was ignored by current literature–how to estimate the exact size of labelled samples given many unlabelled samples. Such an estimation method is important because of the rareness and expensiveness of labelled examples and is also crucial in exploring the relative value of labelled and unlabelled samples given a specific model. Based on the assumption of a latent gaussian-distribution to the domain, we described a method to estimate the number of labels required in a dataset for semi-supervised linear discriminant classifiers (Transductive LDA) to reach an desired accuracy. Our technique extends naturally to handle two difficult problems: learning from gaussian distributions with different covariances, and learning for multiple classes. This method is evaluated on two datasets, one toy dataset and one real-world wine dataset. The result of this research can be used in areas such text mining, information retrieval or bioinformatics.

1 Introduction

Machine learning falls into two broad categories: supervised learning and unsupervised learning, primarily distinguished by the use of labelled and unlabelled data. Semi-supervised learning has received considerable attention in the literature due to its potential in reducing the need for expensive labelled data [1]. A general strategy is to assume that the distribution of unlabelled data is linked to their labels. In fact, this is a necessary condition for semi-supervised learning to work. Existing approaches make different assumptions within this common framework. Generative mixture model method[2] assumes that data comes from some identifiable mixture distribution, with unlabelled data, the mixture components can be identified independently. Transductive Support Vector Machines[3] take the metrics space to be a high-dimensional feature space defined by the

J.-F. Boulicaut et al. (Eds.): ECML 2004, LNAI 3201, pp. 274–285, 2004.

kernel, and by maximizing the margin based on the unlabelled data, effectively assume that the model is maximally smooth with respect to the density of unlabelled data in feature space. Co-training[4] assumes that data attributes can be partitioned into groups that individually sufficient for learning but conditionally independent given the class, and working by feeding classifications made by one learner as examples for the other and vice versa. Graph methods[5] assume a graph structure underlying the data and the graph structure coincide with classification goal. The nodes of the graph are data from the dataset, and edges reflect the proximity of examples. The intuition is that close examples tend to have similar labels, and labels can propagate along dense unlabelled data regions.

While many practical semi-supervised classification algorithms have been proposed, an important issue is ignored: Given many unlabelled samples, what is the minimum labelled samples we need while achieving a desired classification performance? Given labels to more training samples lowered the classification errors, but increased the cost when obtaining those labels. Thus, a method to estimation the minimum labelled sample size becomes a necessity. Moreover, a detailed analysis of labelled sample size under specific model assumption can improve our understanding of the relative values of labelled and unlabelled samples.

In this paper, our labelled sample size estimation method was derived by computing the bays error for a binary semi-supervised linear classifier(i.e. transductive LDA), and estimating the appropriate number of labels necessary for the a certain classification accuracy . We chose transductive LDA in our setting since as a generalization of regular LDA and based on the assumption of a latent gaussian-distribution to the domain,it has a relatively large bias but little variance, and avoid overfitting effectively when sample size is small. Besides a theoretical underpinning, we also developed a computationally tractable implementation based on simple parameter vector space transformation for our estimation method. Detailed discussion and analysis are presented to show that the technique could extend naturally to quadratic classifiers and to multi-class classifiers. The next section provided background information on transductive LDA as discussed in the literature of semi-supervised learning. Section 3 detailed a mathematical derivation of the labelled sample size estimation method for transductive LDA classifiers. Section 4 discussed the extension of our estimation method to the case of quadratic discriminant analysis and multi-class classification. Experimental results on both toy data and real world dataset were shown in Section 5. Summarization and future work were discussed in section 6. the last part was the acknowledge.

2 Transductive Linear Discriminant Analysis

2.1 Formalization of Labelled Sample Size Estimation Problem

We begin with a domain of objects $X = \{x_1, ..., x_n\}$. Each object x_i is associated with a vector of observable features(also denoted x_i). We are given labels $Y_l = \{y_1, ..., y_l\}$ for the first l objects($l << n$), and our goal is to infer the labels

$Y_u = \{y_{l+1}, ..., y_n\}$ of the $n-l$ unlabelled data. We refer to $X_l = \{x_1, ...x_l\}$ as the "labelled sample", and the complement $X_u = X - X_l$ as the "unlabelled sample". For now we assume a binary classification problem, with $Y = Y_l \cup Y_u \in \{1, 0\}^{|X|}$, The labels y_i are independent random variables satisfying $Pr\{y_i = 1\} = \pi_1$, $Pr\{y_i = 0\} = 1 - \pi_1$; generalization to the multi-class is straightforward. We use $R(l, u)$ to denote the classification error with l labelled data and u unlabelled data while R^* denotes bayes risk.

$$R^* = \int min\{\pi_1 f_1(x), (1 - \pi_1) f_2(x)\} dx$$

The labelled sample size estimation problem is formulated as given an acceptable additional probability of error $\triangle_{err}$ of any Bayesian solution to a classification problem with a smooth prior, $0 < \triangle_{err} < 1$, from current X and Y_l, how many more x_k should be given labels, so that the difference between real classification error and bayes error $R(l, u) - R^*$ will be less then $\triangle_{err}$.

We also assume that the underlying distributions of the samples are mixture of gaussian with identical covariance matrix, i.e., each class conditional distribution has the form $p_{y_i}(\cdot) \sim N(\mu_i, \varSigma)$, where $p_{y_i}(\cdot)$ denotes the distribution of samples for class i. Given a large number of unlabelled samples, this assumption is reasonable because of central limited theorem as a theoretical foundation. The assumption of identical covariance is crucial for LDA to work. Even though such an assumption is very strong, LDA is shown to work well in many applications in real-word datasets. In addition, we can extend LDA to Quadratic Discriminant Analysis(QDA) naturally to relax this assumption.

2.2 Transductive LDA

Transductive LDA is a semi-supervised version of common LDA. Expectation-Maximization(EM) approach can be applied to this learning problem, since the labels of unlabelled data can be treated as missing values. Let the entire training dataset D be a union of labelled dataset L and unlabelled dataset U, and assume each sample is independent, the joint probability density of the hybrid dataset can be written as:

$$p(x|\theta) = \prod_{x_i \in X_u} \sum_{k=1}^{K} p(y_i = k|\theta) p(x_i|y_i = k; \theta) \cdot \prod_{x_i \in X_l} p(y_i = k) p(x_i|y_i = k; \theta)$$

where $k = 1$ or $k = 2$, representing the categories. The first part of this equation is for the unlabelled dataset, and the second part is for the labelled data.

The parameters $\theta = (\pi_1, \mu_k, \varSigma)^T$ can be estimated by maximizing *a posteriori* probability $p(\theta|D)$. Equivalently,this can be done by maximizing the log likelihood $\log p(D|\theta)$ when the prior probability is uniform. The likelihood is given as:

$$L(\pi_1, \mu_k, \varSigma) = (2\pi)^{-\frac{np}{2}} |\varSigma|^{-\frac{n}{2}} \cdot \prod_{x_i \in Y_u} (\pi_1 e_{1i} + \pi_2 e_{2i})$$
$$\prod_{x_i \in Y_l} \{(\pi_1 \cdot e_{1i})^{y_i} \cdot [(1 - \pi_1) e_{2i}]^{1 - y_i}\}$$

where $e_{ki} = \exp\{-\frac{1}{2}(x_i - \mu_k)^T \Sigma^{-1}(x_i - \mu_k)\}$. Thus the log likelihood is $l(\theta; D) = \log L(\pi_1, \mu_k, \Sigma)$. Assume that $l_{uk} = \log p(x^u, y = k|\mu_k, \Sigma)$ and $l_u = \log p(x^u) = \log \sum_k p(x^u, y = k|\mu_k, \Sigma)$, where x^u denotes all the $x \in X_u$. When using EM algorithm to estimate the probability parameters π_1, μ_k, Σ by an iterative hill climbing procedure, the E-step and M-step are designed as follows, for the E-step:

$$r_{uk} = p(y = k|x^u) = \frac{p(x^u, y = k)}{p(x^u)} = \exp\{l_{uk} - l_u\}$$

For the M-step: $\pi_1 = \frac{\sum_u r_{uk} + l_1}{u+l}$, $\pi_2 = 1 - \pi_1$, $\mu_k = \frac{\sum_u r_{uk} x^u + \sum x_k^l}{\sum_u r_{uk} + l_k}$, and $\mu_k' = \frac{\sum_u r_{uk} x^u}{\sum_u r_{uk}}$. Σ is computed as:

$$\Sigma = \frac{\sum_k (\sum_u r_{uk}(x^u - \mu_k')(x^u - \mu_k')^T + \sum_l (x_k^l - \bar{x}_k^l)(x_k^l - \bar{x}_k^l)^T)}{\sum_u \sum_k r_{uk} + l}$$

where l_k denotes the number of labelled data in class k, x_k^l denotes the number of x^l in class k, $\bar{x}_k^l$ is the mean of all the x^l in class k. When the size of the labelled dataset is small, EM basically performs an unsupervised learning, except that the labelled data are used to identify the components. Detailed analysis for this issue could be found in [2]. After the EM progress, all the parameters needed for linear discriminant analysis are tuned, and discriminant functions for conducting classification can be obtained based on this parameters(as described in the next section).

3 Labelled Sample Size Estimation Technique

3.1 Bayes Risk for LDA Rule Classifier

In this section we present the equation for calculating bayes risk R^* of a LDA classifier $\hat{G}$. For all feature vectors X and a class membership G, we let $L(G, \hat{G}(X))$ be the loss function of a misclassification, and furthermore assume it only has 0-1 values, meaning all misclassification are charged a single unit, as in the case of many discriminant analysis. Next, we model each class conditional density as multivariate Gaussian, i.e. $X|G = g_k \sim N(\mu_k, \Sigma)$, where g_k is the class label, and the discriminant functions are therefore given by $\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$ and $G(x) = arg \max_k \delta_k(x)$. For two classes, the LDA rule classifies to class 2 if $\delta_2(x) > \delta_1(x)$ and class 1 otherwise. With $\pi_2 = 1 - \pi_1$, the Bayes risk R^* is given by

$$R^* = \pi_1 P_1(X^T \Sigma^{-1}(\mu_2 - \mu_1) > \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 - \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \log \frac{\pi_1}{\pi_2})$$
$$+\pi_2 P_2(X^T \Sigma^{-1}(\mu_2 - \mu_1) < \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 - \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \log \frac{\pi_1}{\pi_2})$$

In class g_1, $X \sim N(\mu_1, \Sigma)$, and $Z_1 = X^T \Sigma^{-1}(\mu_2 - \mu_1)$ is transformation of X and is a univariate Gaussian random variable with mean $\mu_2 - \mu_1)^T \Sigma^{-1} \mu_1$, and variance $\sigma^2 = (\mu_2 - \mu_1)^T \Sigma^{-1}(\mu_2 - \mu_1))$. Z can be transformed to a standard Gaussian random variable $\frac{Z-(\mu_2-\mu_1)^T\Sigma^{-1}\mu_1}{\sigma} \sim N(0,1)$. In class g_2 , Z has similar distribution except for its mean and $\frac{Z-(\mu_2-\mu_1)^T\Sigma^{-1}\mu_2}{\sigma} \sim N(0,1)$. Thus, the Bayes risk can be calculated as

$$\pi_1 P_1(\frac{Z-(\mu_2-\mu_1)^T\Sigma^{-1}\mu_1}{\sigma} > a_1) + \pi_2 P_2(\frac{Z-(\mu_2-\mu_1)^T\Sigma^{-1}\mu_2}{\sigma} > a_2)$$

where

$$a_1 = \frac{\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\frac{\pi_1}{\pi_2} - (\mu_2-\mu_1)^T\Sigma^{-1}\mu_1}{\sigma}$$

$$a_2 = \frac{\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\frac{\pi_1}{\pi_2} - (\mu_2-\mu_1)^T\Sigma^{-1}\mu_2}{\sigma}$$

Let Φ denote the cumulative distribution function for a standard Gaussian model. R^* can be written as $\pi_1(1 - \Phi(\frac{\frac{1}{2}\sigma^2+\log\frac{\pi_1}{\pi_2}}{\sigma})) + \pi_2\Phi(\frac{-\frac{1}{2}\sigma^2+\log\frac{\pi_1}{\pi_2}}{\sigma})$.

3.2 Labelled Sample Size Estimation Method

The estimation of an appropriate size of the labelled samples is determined by the required reduction in $R(l,u) - R^*$, which is affected by the current size of unlabelled data,dimensionality of the sample space and the separability of the two classes. We first derive a way to calculate $R(l,u) - R^*$. $R(l,u)$ is a function of θ, where $\theta = (\pi_1, \mu_1, \mu_2, \Sigma^{-1})^T$. We let θ^* denotes the true value of θ and $\hat{\theta}$ denotes the estimated value, and using Taylor series expansion of $R(\hat{\theta})$ up to second term, we obtain

$$R(\hat{\theta}) = R(\theta^*) + \frac{\partial R(\theta)^T}{\partial\theta}|_{\theta=\theta^*}(\hat{\theta}-\theta^*) + \frac{1}{2}tr\{\frac{\partial^2 R(\theta)^T}{\partial\theta^2}|_{\theta=\theta^*}(\hat{\theta}-\theta^*)(\hat{\theta}-\theta^*)^T\}$$

where $tr(A)$ denotes the trace of a matrix A. The term $\frac{\partial R(\theta)}{\partial\theta}|_{\theta=\theta^*}$ is zero since θ^* is an extreme point of $R(\theta)$. Assuming the bias of $\hat{\theta}$ is negligible, i.e. $(E\{\hat{\theta}\} = \theta^*)$, $R(l,u) - R^*$ can be approximated as $\frac{1}{2}tr\{\frac{\partial^2 R(\theta)}{\partial\theta^2}|_{\theta=\theta^*}cov(\hat{\theta})\}$. By asymptotic theory, as the sample size approaches infinity, $\hat{\theta} \sim N(\theta^*, J^{-1}(\theta))$, where $J(\theta) = -\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}$, with $l(\theta)$ representing the log likelihood of θ, is the observed fisher information matrix of θ, and an approximation of the covariance matrix $cov(\hat{\theta})$. $J(\theta)$ is calculated by the summation of two parts: $J_l(\theta)$ from the labelled data, and $J_u(\theta)$ from the unlabelled data. Let n be the total sample size; $\overline{J_l}(\theta)$ and $\overline{J_u}(\theta)$ be the observed information of a single observation for labelled and unlabelled data respectively. If given an required reduction in classification error $\triangle_{err}$, we can find the labelled sample size l needed from

$$tr\{\frac{\partial^2 R(\theta)}{\partial\theta^2}|_{\theta=\hat{\theta}}(l\overline{J_l}(\hat{\theta}) + (n-l)\overline{J_u}(\hat{\theta}))^{-1}\} < 2 \cdot \triangle_{err} \qquad (*)$$

Since $\overline{J_l}(\theta) = J_l(\theta)/l'$ and $\overline{J_u}(\theta) = J_u(\theta)/(n-l')$, where l' is current labelled sample size, and since $n - l \approx n - l'$ ($n >> l'$ and $n >> l$), formula (*) can be simplified as follows,

$$tr\{\frac{\partial^2 R(\theta)}{\partial\theta^2}|_{\theta=\hat{\theta}}(l\overline{J_1}(\hat{\theta}) + J_2(\hat{\theta}))^{-1}\} < 2 \cdot \triangle_{err} \qquad (**)$$

3.3 Computational Consideration

According to formula (*), quantities $\frac{\partial^2 R(\theta)}{\partial\theta^2}|_{\theta=\hat{\theta}}$, $\overline{J_l(\hat{\theta})}$ and $\overline{J_u(\hat{\theta})}$ need to be computed first when estimating the desired labelled sample size l. However, when the dimensionality p of the feature space is very large, the computation is very intensive. To illustrate, if $R(\theta)$ is a continuous function,the $\frac{\partial^2 R(\theta)}{\partial\theta^2}|_{\theta=\hat{\theta}}$ is a $(p^2+5p+2)/2 \times (p^2+5p+4)/2$ dimensional matrix, not mentioning the product of two such high-scale matrix! Finding a computationally tractable method to compute formula (**) is therefore of great practical benefit. In this section, we develop a method to reduce the computational load of $\frac{\partial^2 R(\theta)}{\partial\theta^2}|_{\theta=\hat{\theta}}$ to a 2×2 matrix calculation based on simple vector space transformation, while at the same time reduce the matrix production calculation of two $(p^2+5p+2)/2 \times (p^2+5p+4)/2$ dimensional matrix production to two 2×2 matrix production. According to the proof in [7], in the case of LDA, $R(l,u)$ only depends on π_1 and $\sigma^2(\sigma^2 = (\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1))$ instead of the full parameter set. Let $\varphi = (\sigma^2, \pi_1)$, and let φ^* denotes the true value of φ and $\hat{\varphi}$ denotes the estimation, by Taylor series expansion of $R(\hat{\varphi})$ up to second order, we obtain

$$R(\hat{\varphi}) = R(\varphi^*) + \frac{\partial R(\varphi)^T}{\partial\varphi}|_{\varphi=\varphi^*}(\hat{\varphi}-\varphi^*) + \frac{1}{2}tr\{\frac{\partial^2 R(\varphi)^T}{\partial\varphi^2}|_{\varphi=\varphi^*}(\hat{\varphi}-\varphi^*)(\hat{\varphi}-\varphi^*)^T\}$$

Again, the term$\frac{\partial R(\varphi)}{\partial\varphi}|_{\varphi=\varphi^*}$ is zero since φ^* is an extreme point of $R(\varphi)$. If the bias of $\hat{\varphi}$ is negligible,i.e., $(E\{\hat{\varphi}\} = \varphi^*)$, $R(l,u) - R^*$ can be approximated as,

$$R(l,u) - R^* = \frac{1}{2}tr\{\frac{\partial^2 R(\varphi)}{\partial\varphi^2}|_{\varphi=\varphi^*}cov(\hat{\varphi})\}$$

The approximated covariance matrix of $\hat{\varphi}$ can be obtained from the inverse of observed fisher information matrix $J(\tilde{\theta}) = -\frac{\partial^2 l(\tilde{\theta})}{\partial\tilde{\theta}\tilde{\theta}^T}|_{\tilde{\theta}=\tilde{\theta}^*}$, where $l(\tilde{\theta})$ is the log likelihood of $\tilde{\theta}$ based on the ladled and unlabel samples. $\tilde{\theta}$ is reparameterized from the old parameter $\theta = (\pi_1, \mu_{12}, ...\mu_{1p}, \mu_{22}, ..., \mu_{2p}, a_{ij})^T$, where a_{ij} is the elements of the matrix Σ^{-1}, $i = 1, ..., p$,$j = 1, ..., p$.

We map the elements in the original parameter space of θ to the new parameter space $\tilde{\theta}$ by letting $\tilde{\theta} = (\pi_1, \sigma^2, \mu_{12}, ...\mu_{1p}, \mu_{22}, ..., \mu_{2p}, a_{ij}|_{i,j\neq 1})^T$, i.e., removing the term a_{11} from θ and adding the term σ^2. The vector space of φ is a subspace of vector space $\tilde{\theta}$. Since θ and $\tilde{\theta}$ have the same dimensionality, the mapping is guaranteed to be one to one transformation, with a_{11} expressed by the elements of $\tilde{\theta}$ as

$$a_{11} = \frac{\sigma^2 - \sum_{i,j\neq 1} a_{ij}(\mu_{1i}-\mu_{2i})(\mu_{1j}-\mu_{2j})}{(\mu_{11}-\mu_{21})^2}$$

Again a_{ij} is the element of Σ^{-1}. The new log likelihood $l(\tilde{\theta})$ can be easily obtained from the original $l(\theta)$ and can be differentiated with respect to $\tilde{\theta}$. The information $J(\tilde{\theta})$ is also the summation of two parts: $J_l(\tilde{\theta})$ from the labelled data and $J_u(\tilde{\theta})$ from the unlabelled data. We let n be the total sample size, $\overline{J_l}(\tilde{\theta})$ and $\overline{J_u}(\tilde{\theta})$ be the observed information of a single observation for labelled and unlabelled data respectively. Similar to the derivation above, $J(\tilde{\theta}) \approx l\overline{J_l}(\tilde{\theta}) + J_u(\tilde{\theta})$ and thus $J^{-1}(\tilde{\theta}) \approx (l\overline{J_l}(\tilde{\theta}) + J_u(\tilde{\theta}))^{-1}$. Let $I(l)$ denotes the matrix made of the first two columns and first two rows, under our differentiation order of $\tilde{\theta}$, we can prove that $I(l) \approx cov(\hat{\varphi})$. The detailed proof is omitted here. After the vector space transformation, given $\triangle_{err}$, formula (**) is equivalent to

$$tr\{\frac{\partial^2 R(\varphi)}{\partial \varphi^2}|_{\varphi=\hat{\varphi}}(I(l)\} < 2 \cdot \triangle_{err} \qquad (***)$$

which is computationally tractable.

From the mathematical derivation above, we can see that the labelled sample size l is determined by several factors. Because the log likelihood is the likelihood for both labelled and unlabelled data, the final l is affected by the number of unlabelled data and also the dimensionality of the sample space. Furthermore, σ^2 and μ_k determines whether the two classes are easily classifiable, and it is an important factor in our estimation equation.

4 Relax the Strong Assumptions to Transductive QDA and Transductive MDA

4.1 From Transductive LDA to Transductive QDA

In this section, we discuss how to relax the strong assumption of identical covariance in gaussian mixtures by extending Transductive LDA to Transductive QDA. The modification of EM algorithm is trivial, requiring changes only in the M-step, i.e, $\Sigma_k^{new} = \frac{(\sum_u r_{uk}(x^u - \mu_k')(x^u - \mu_k')^T + \sum_l (x_k^l - \bar{x_k^l})(x_k^l - \bar{x_k^l})^T)}{\sum_u r_{uk} + l_k}$. We also need modifications to the estimation method when relaxing the identical covariance assumption. For quadratic discriminant analysis, each class conditional density is modelled as $X|G = g_k \sim N(\mu_k, \Sigma_k)$, and the discriminant functions are given as $\delta_k(x) = x^T \Sigma_k^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma_k^{-1} \mu_k + \log \pi_k$ and $G(x) = arg \max_k \delta_k(x)$. In the case of two classes, the corresponding Bayes risk R^* is $\pi_1 P_1(X^T(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1) > \frac{1}{2}\mu_2^T \Sigma_2^{-1}\mu_2 - \frac{1}{2}\mu_1^T \Sigma_1^{-1}\mu_1 + \log\frac{\pi_1}{\pi_2}) + \pi_2 P_2(X^T(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1) < \frac{1}{2}(\mu_2^T \Sigma_2^{-1}\mu_2 - \mu_1^T \Sigma_1^{-1}\mu_1) + \log\frac{\pi_1}{\pi_2})$. In class g_1, $X \sim N(\mu_1, \Sigma_1)$, $Z = X^T(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1)$ is a univariate gaussian random variable, which is a transformation of X, with the distribution, $Z \sim N((\mu_2^T \Sigma_2^{-1} - \mu_1^T \Sigma_1^{-1})\mu_1, \sigma_1)$ where $\sigma_1 = (\mu_2^T \Sigma_2^{-1} - \mu_1^T \Sigma_1^{-1})\Sigma_1(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1)$. By defining $mean_1 = (\mu_2^T \Sigma_2^{-1} - \mu_1^T \Sigma_1^{-1})\mu_1$ and $mean_2 = (\mu_2^T \Sigma_2^{-1} - \mu_1^T \Sigma_1^{-1})\mu_2$, $\sigma_2 = (\mu_2^T \Sigma_2^{-1} - \mu_1^T \Sigma_1^{-1})\Sigma_2(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1)$ We have $\frac{Z - mean_1}{\sigma_1} \sim N(0,1)$. In class g_2 , Z has similar distribution such that $\frac{Z - mean_2}{\sigma_2} \sim N(0,1)$. Thus, the Bayes risk for QDA can be calculated as

$$R^* = \pi_1 P_1(\frac{Z - mean_1}{\sigma_1} > a_1) + \pi_2 P_2(\frac{Z - mean_2}{\sigma_2} > a_2)$$

where $a_1 = \frac{\frac{1}{2}\mu_2^T \Sigma_2^{-1}\mu_2 - \frac{1}{2}\mu_1^T \Sigma_1^{-1}\mu_1 + \log\frac{\pi_1}{\pi_2} - mean_k}{\sigma_k}$, $k = 1, 2$. Let Φ denote the cumulative distribution function for a standard Gaussian model. R^* can then be calculated by $\pi_1(1 - \Phi(a_1)) + \pi_2\Phi(a_2)$. The above method is a theoretical analysis, in fact, after tuning out the covariance matrix Σ_1 and Σ_2, we can simply use a very naive method to merge these two different matrices into one single matrix $\Sigma_{common} = (\pi_1^2 \cdot \Sigma_1 + \pi_2^2 \cdot \Sigma_2)/(\pi_1^2 + \pi_2^2)$, which is still a semi-positive definite and can be used instead to apply the estimation technique for Semi-supervised LDA directly. Another important point to note is that without the assumption of identical covariance matrix, $R(l, u)$ does not depend on (σ^2, π_1) only. Consequently, our computational tractable approach dose not hold. Yet, one can still use formula (**) to estimate the appropriate size of labelled samples. For a domain with a very flexible distribution but relatively small dimensionality, applying Semi-Supervised QDA would be more suitable.

4.2 From Two-Class Classification to Multi-class Classification

Some limitations exist when applying transductive LDA to multi-class classification problems, especially when the size of the labelled set is small. In such situation, the EM algorithm may fail if the distribution structure of the data set is unknown. A natural solution in dealing with multi-class classification is to map the original data samples into a new data space such that they are well clustered in the new space, in which case the distributions of the dataset can be captured by simple gaussian mixtures and LDA can be applied in the new space.

Transductive Multiple Discriminant Analysis(MDA)[8] used this idea and defined a linear transformation W of the original p_1-dimension data space to a new p_2-dimension space such that the ratio of between-class scatter S_b to within-class scatter S_w is maximized in the new space, mathematically, $W = \arg max_W \frac{|W^T S_b W|}{|W^T S_w W|}$. suppose x is an p-dimensional random vector drawn from C classes in the original data space. The kth class has a probability P_k and a mean vector μ_k. thus $S_w = \sum_{k=1}^{C} P_i E[(x - \mu_k)(x - \mu_k)^T | c_i]$ and $S_b = \sum_{k=1}^{C} P_i(\mu_k - \sum_{i=1}^{C} P_i\mu_i)(\mu_k - \sum_{i=1}^{C} P_i\mu_i)^T$. The main advantage of transductive MDA is that the data are clustered to some extent in the projected space, which simplifies the selection of the structure of Gaussian mixture models. The EM algorithm for semi-supervised MDA can be found in [9]. Because semi-supervised MDA is just another perspective of semi-supervised LDA, our labelled sample size estimation method also applies well in this setting. Assuming there are k classes altogether, the bayes risk is calculated as follows,

$$R^* = \pi_1 P_1(\delta_1 < \max(\delta_2, ..., \delta_k)) + \pi_2 P_2(\delta_1 < \max(\delta_1, \delta_3, ..., \delta_k)) + \cdots\cdots +$$

$$\pi_i P_i(\delta_1 < \max(\delta_1, ..., \delta_i, \delta_{i+1}, ..., \delta_k)) + \pi_k P_k(\delta_1 < \max(\delta_2, ..., \delta_{k-1}))$$

which can be computed quite efficiently.

5 Experiments and Results

5.1 Toy Data of Mixture of Gaussian with Six Components

To illustrate the performance of our estimation method, first we show an example of no obvious practical significance. Consider Gaussian observation (X, Y) taken from six classes $g_1, g_2, ..., g_6$. We know that X and Y are Gaussian variables, and we know exactly the means of $(X, Y$ is (μ_{ix}, μ_{iy}) and variance-covariance matrices is $|Sigma_i$ given that the class $G = g_i$. We need to estimate the mixing parameter $p_i = p(G = g_i)$. The data is sampled from a distribution with mixing parameter α_i. The total number of our data is 900, with dimensionality equals to two, and are divided into 6 classes: 100 data for class g_1, 100 for class g_2,150 for class g_3,150 for class g_4,200 for class g_5 and 200 for class g_6. The means and covariance matrices are shown as follows: $\mu_1 = (-\frac{3}{2}, 1/2)^T$, $\mu_2 = (2, 2)^T$, $\mu_3 = (-\frac{1}{2}, \frac{5}{2})^T$, $\mu_4 = (\frac{1}{2}, 0)$, $\mu_5 = (\frac{1}{3}, -2)^T$ and $\mu_6 = (\frac{7}{2}, -\frac{1}{2})^T$. For their covariance matrices:

$$\Sigma_1 : \begin{pmatrix} 3 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} \Sigma_2 : \begin{pmatrix} 3 & \frac{1}{2} \\ \frac{1}{5} & 1 \end{pmatrix} \Sigma_3 : \begin{pmatrix} 3 & 1 \\ \frac{1}{5} & \frac{1}{2} \end{pmatrix} \Sigma_4 : \begin{pmatrix} \frac{5}{2} & \frac{1}{3} \\ 2 & \frac{3}{2} \end{pmatrix} \Sigma_5 : \begin{pmatrix} 3 & 1 \\ \frac{1}{5} & 1 \end{pmatrix} \Sigma_6 : \begin{pmatrix} \frac{1}{3} & \frac{1}{10} \\ \frac{1}{2} & \frac{5}{2} \end{pmatrix}$$

and their mixing parameters $\pi_1 = \frac{1}{9}$,$\pi_2 = \frac{1}{9}$, $\pi_3 = \frac{1}{6}$, $\pi_4 = \frac{1}{6}$,$\pi_5 = \frac{2}{9}$, $\pi_6 = \frac{1}{9}$. Based on these information, 900 data were randomly generated. For our experiment, the initial number of the labelled data is 4 for each class. Applying semi-supervised QDA on the data, we obtained a classification result shown in the first plot of figure 1. Given the desired $\triangle_{err} = 0.05$, by our algorithm, the estimated labelled data number is 90, thus at least 15 labels are needed for each class.

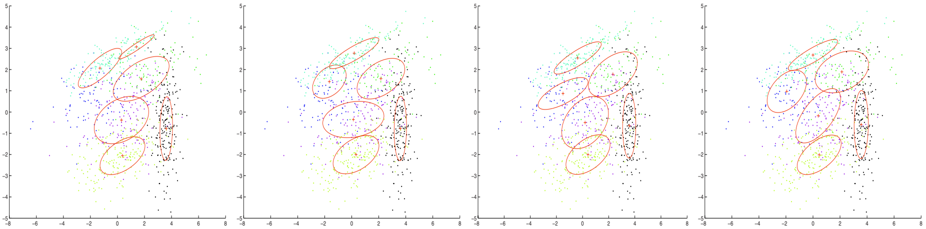

Fig. 1. Illustration of the fitting with 4 labels, 7 labels, 15 labels and 22 labels for each class respectively, in this first 2 cases, the number of labelled data is not large enough, thus can not give out enough information for fitting, while in the latter 2 cases,the number of labelled data is enough,the fitting is good, but given more data can not improve the fitting significantly

From the first plot in figure 1, it is easy to see that the fitting is not quite good. the shape of the gaussian and the position is quite different from the original figure. plot 2 was generated from 7 labels per class, the fitting is still not good enough. While plot 3 represents the fitting condition with 15 labels for each class, the fitting result is satisfiable; plot 4 is generated with 22 labels for each classes, we can see that the fitting performance does not improve much from the 15 labels case. The classification error for each class is shown in table 1 below: Running for every label size 10 times. The box plot for these 4 conditions are shown in figure 2, From the box plot above, we can see that with the increase number of the labelled data, the overall error rate was reduced significantly at first, but slightly after it exceeds a threshold. Normally, we use 5% as this threshold.

Table 1. Classification error for each classes of toy data

label number	4/class	7/class	15/class	22/class
error for g_1	0.70515	0.47788	0.54221	0.55176
error for g_2	0.77217	0.73579	0.65222	0.55411
error for g_3	0.18903	0.10350	0.14149	0.04770
error for g_4	0.53220	0.60979	0.60518	0.57308
error for g_5	0.17077	0.10841	0.10666	0.10176
error for g_6	0.16513	0.13630	0.10500	0.12819
overall error	0.35856	0.31040	0.29419	0.28778

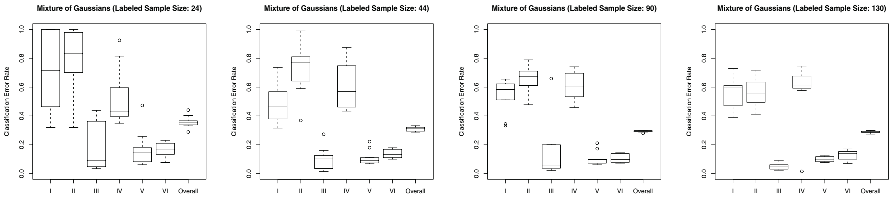

Fig. 2. Illustration of the box plot for 4 labels, 7 labels 15 labels and 22 labels per class respectively

5.2 Real World Dataset: Wine Recognition Data

In order to test the idea of our estimation method, we applied it to the problem of wine recognition. These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines[11]. There are 59 data in the class g_1, 71 data in the class g_2 and 48 data in the class g_3, with 13 predictors, i.e., the dimensionality is 13. After randomly choosing 16 labelled data for every class, and requiring $\triangle_{err} = 0.05$, our estimated number of the labels needed is $l = 60$, meaning at least 20 labels are need for each class. the classification errors for each class were shown in table 2 above.

From which, we can see the data is easily to be fitted by QDA$-$ thus it satisfies the gaussian assumptions well. The computed bayes risk for this data set is about 0.06948-0.05=0.01948. From the original data set, compute pairwise about the bayes risk R^*, it's 0.01966. the result is very close. The box plot for the classification of each class is shown below:

Table 2. Classification error for each classes of wine data

label number	16/class	20/class	22/class	24/class
error for g_1	0.08093	0.01281	0.00000	0.01143
error for g_2	0.13134	0.04256	0.04887	0.01190
error for g_3	0.37941	0.20619	0.08066	0.04138
overall error	0.13646	0.06948	0.04191	0.01980

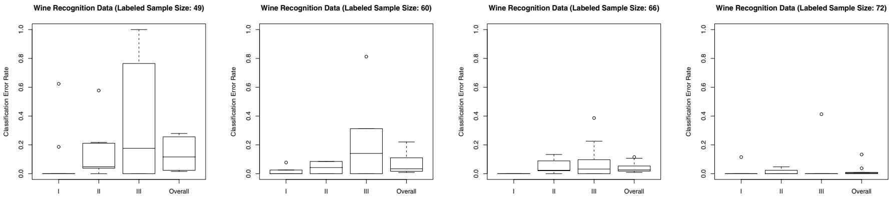

Fig. 3. Illustration of the box plot for 16 labels, 20 labels, 22 labels and 24 labels per class respectively

6 Conclusion and Future Extension

We have examined a labelled sample size estimation problem under a specific model, i.e., semi-supervised LDA. Given an additional probability of error $\triangle_{err}$ of any Bayesian solution to the classification problem with respect to a smooth prior, $\triangle_{err} = R(l,u) - R^*$, under the gaussian-distribution domain assumption, we presented a practical labelled sample size estimation method and a computationally tractable approach. Possible extensions and future work are discussed below. This research result could be applied in different semi-supervised learning domain with probability model type 1 [10].

Linear discriminant analysis and logistic regression are two main representatives of these two classes. Our labelled sample size estimation method applies on semi-supervised LDA well, but not on logistic regression. We believe that similar research on logistic regression model would be very meaningful. Based on the detailed analysis for these two types of models, a common labelled sample size estimation framework maybe built and based on this architecture, interesting research topic can be found.

Acknowledgment

The authors would like to thank Dr Sam Roweis's comment and Kannan Achahan's suggestions. This work is also supported by scholarship from University of Toronto.

References

1. M. Seeger *Learning with labeled and unlabeled data*, (Technical Report), Institute for Adaptive and Neural Computation, University of Edinburgh, Edinburgh, United Kingdom, pp. 609-616, 2001
2. Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell *Text classification from labeled and unlabeled documents using EM*, Machine Learning, 39(2/3): pp.103-134, 2000
3. Kristin Bennett and A.Demiriz *Semi-supervised support vector machines*,Advances in Neural Information Processing Systems (NIPS) [NIPS99] pp1-7. 1999.
4. Avrim Blum and Tom Mitchell *Combing labeled and unlabeled data with co-training*,Proc. Of the 1998 Conference on Computational Learning Theory, pp.1-10, 1998
5. A.Blum and S.Chawla *Learning from labeled and unlabeled data using graph mincut*,In proc. 17th Intl Conf. on Machine Learning (ICML) ,pp.1181-1188, 2001
6. Mardia,K., Kent, J. and Bibby,J. *Multivariate Analysis*, Academic Press. 1979
7. T.O'Neil. *Normal discrimination with unclassified observations*, Journal of American Statistical Association, Volume 73, no. 364, pp 821-826, Dec. 1978
8. R.Duda and P.Hart. *Pattern Classification and Scene Analysis*, New York: Wiley, 1973
9. Ying Wu, Qi Tian, Thomas S. Huang. *Discriminant-EM Algorithm with Application to Image Retrieval*, Technical Report,UIUC,USA 1999
10. T.Zhang and F.Oles. *A probability Analysis on the value of unlabeled data for classification problem.*, ICML pp.1191-1198 2000
11. Forina, M.et al, PARVUS. *An Extensdible Package for Data Exploration, Classification and Correlation.*,Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, Italy.