Statistica Sinica **24** (2014), 1547-1570
doi:http://dx.doi.org/10.5705/ss.2012.331

OPTIMAL RANKING IN MULTI-LABEL CLASSIFICATION USING LOCAL PRECISION RATES

Ci-Ren Jiang, Chun-Chi Liu, Xianghong J. Zhou and Haiyan Huang

Academia Sinica, National Chung Hsing University
University of Southern California and University of California at Berkeley

Abstract: Multi-label classification is increasingly common in modern applications such as medical diagnosis and document categorization. One important issue in multi-label classification is the existence of statistical difference of classifier scores among different classes. When not accounted for properly, such differences can lead to poor classification decisions on some classes. We address this issue by developing a strategy based on a new concept, Local Precision Rate (LPR), under the assumption that classifiers learned for each class are given and corresponding classifier scores for a set of training objects and a set of objects to be classified are available. Under certain conditions, we show that transforming the classifier scores into LPRs and making classification decisions by comparing LPR values for all objects against all classes can theoretically guarantee the maximum of precision at any recall rate. We also show that LPR is mathematically equivalent to 1-ℓFDR, where ℓFDR stands for local false discovery rate. This equivalence and the Bayesian interpretation of ℓFDR provide an alternative justification for the theoretical optimal property of LPR. We propose a new estimation method for 1-ℓFDR (or LPR) based on the formulation of LPR, since the original formulation of 1-ℓFDR has limitations for estimation when data are noisy. Numerical studies are conducted based on both simulation and real data to demonstrate the superior performance of LPR over existing methods.

Key words and phrases: False discovery rate, local false discovery rate, local precision rate, multilabel classification, optimization, smoothing.

1. Introduction

The traditional problem of single-label classification, which concerns assigning each input object in a sample to exactly one class, has generated a broad literature in both statistics and computer science (Hastie, Tibshirani, and Friedman (2008)). The related but more complicated problem of multi-label classification assigns an object to one or multiple classes, among which dependencies and hierarchies may exist. Multi-label classification is increasingly common in modern applications. Typical examples include medical diagnosis (Karalic and Pirnat (1991); Maimon and Rokach (2010)), gene/protein function prediction (Alves, Delgado, and Freitas (2008); Cerri, da Silva, and de Carvalho (2009),

Schietgat et al. (2010); Vens et al. (2008)) and document (or text) categorization (Rousu et al. (2006), Geurts, Wehenkel, and d'AlchéBuc (2006)). More specifically, a patient may suffer from diabetes and prostate cancer at the same time; some *Drosophila pumilio* genes are known to play multiple functional roles in germline development, gonadogenesis, oogenesis and embryogenesis (Parisi and Lin (1999)); a newspaper article talking about layoffs during the recent recession can be classified as both financial news and social news. A tutorial on multi-label classification can be found in de Carvalho and Freitas (2009).

In multi-label classification, many methods are binary relevance approaches, in which the classifiers of each class are learned independently. These classifiers often have different statistical properties due to varying quality and quantity of training data for different classes. When not accounted for properly, such differences can lead to poor classification decisions on some classes. To address this issue, a second step is often followed. A simple, intuitive idea is to cogitate different thresholds of classifier scores for each class, rather than using a common threshold for all classes as in conventional score-based methods. Existing class-dependent thresholding approaches include rank-based and proportion-based methods (Yang (2001)). Rank-based approaches (Joachims (1998)) rank the objects by their classifier scores and assign a fixed number (say k) of top-ranking objects to each class. Proportion-based approaches (Lewis (1992)) only differ in that they assign k_j top-ranking objects to each class, where j is the class index and k_j is proportional to the prior probability of a random object to be in class j. Recent developments are largely focused on finding class-dependent thresholds through a global optimal criterion. A greedy algorithm ("cyclic optimization") was proposed in Fan and Lin (2007) to find thresholds that can maximize a given measure. Its key idea was to iteratively select an optimal threshold for one class while keeping the thresholds for the other classes unchanged until the measure cannot be improved. Pillai, Fumera, and Roli (2013) proposed a new strategy with a low computational cost to search for optimal thresholds that theoretically guarantee a global maximum F-measure (Dembczyński et al. (2011), also see (A.2)). There are also efforts to approach the problem from a bayesian point of view. In Quevedo, Luaces, and Bahamonde (2012), a family of thresholding strategies, called probabilistic thresholds, was introduced, and the posterior probabilities for an object to be in each of the classes were inferred and used for classification decisions.

In this paper, we introduce a strategy, based on the local precision rate (LPR), to determine the thresholds of classifier scores for all classes. We show that transforming the classifier scores into LPRs and making classification decisions by comparing LPR values for all objects against all classes can theoretically guarantee the maximum of precision at any recall rate. Here an optimal precision-recall curve (with decisions on all objects against all classes pooled together) can

be generated by applying LPR. LPR is derived by optimizing the pooled precision rate, see (2.1), at any recall but turns out to be mathematically equivalent to (1 - ℓFDR), where ℓFDR stands for local false discovery rate (Efron et al. (2001); Efron and Tibshirani (2002); Efron (2005, 2010)). This equivalence and the Bayesian interpretation of ℓFDR provide an alternative justification for the theoretical optimal property of LPR. We propose a new estimation method for 1-ℓFDR (or LPR) based on the formulation of LPR, since the original formulation of 1-ℓFDR has limitations for estimation when the training data are noisy and complex. The advantages of our estimation method (derived through LPR) over the traditional estimation methods for 1-ℓFDR is further demonstrated by numerical studies in Section 3. Our study is based on the assumption that classifiers learned for each class are given and corresponding classifier scores for a set of training objects and a set of objects to be classified are available, like other efforts discussed previously. That is, we aim to alleviate the differences among classes without giving up existing classifiers, especially those carefully learned ones. Our work is thus complementary to the existing work on constructing high-performance classifiers.

The rest of this article is organized as follows. In Section 2, the new method based on LPR is introduced, and its asymptotic properties are discussed. In Section 3, two simulation studies are reported on that demonstrate the performance of the LPR approach. Section 4 presents a comparison between our approach and a few others, based on their applications to the NCBI Gene Expression Omnibus (GEO) disease diagnosis database. In Section 5, a comparison between the method and the optimal thresholding method in Pillai, Fumera, and Roli (2013) is conducted by applying both methods to three MLC benchmark datasets. Section 6 discusses and summarizes the main results.

2. Method

2.1. Problem formulation and notations

Assume that K classifiers have been (independently) learned for K classes, and the statistical distributions of the classifier scores are known. Denote the classifier scores for M to-be-classified objects as $s_{k,x}$'s with $x = 1, \ldots, M$ and $k = 1, \ldots, K$. Our goal is to find an "optimal" strategy to assign the M objects to the K classes, based on the $s_{k,x}$'s, with each object assigned to zero or more classes. A formal definition of "optimal" strategy is given through a statistical model.

We need some basic notations. Let $S_{k,x}$ denote a random classifier score (with cdf F_k) for object x against class k. Let λ_{k,u_k} be a cutoff score such that any objects with higher classifier scores is assigned to class k; u_k in λ_{k,u_k} indicates the chance that the random object x is not assigned to class k or, equivalently,

$u_k = P(S_{k,x} \le \lambda_{k,u_k}) = F_k(\lambda_{k,u_k})$. Let $Q_{k,x}$ denote the true (unknown) binary class label: $Q_{k,x} = 1$ if object x truly belongs to class k and $Q_{k,x} = 0$ otherwise.

For a single class k, it is natural to evaluate its classifier's performance by the precision function $G_k(u_k) = P(Q_{k,x} = 1 | S_{k,x} > \lambda_{k,u_k})$, the conditional probability for a random object to be truly in class k given that its classifier score is above the threshold. In multi-label classification, the classifier's performance for all individual classes need to be jointly considered. With the classification decisions over all classes pooled together, a pooled precision rate (*ppr*) can be defined correspondingly. With $u_1, \ldots, u_K$, we define

$$ppr = \frac{\sum_{k=1}^{K}(1-u_k)G_k(u_k)}{\sum_{k=1}^{K}(1-u_k)}, \tag{2.1}$$

where the denominator is the expected number of classes (out of the K classes) that object x is assigned to, and among these assignments, the numerator is the expected number of correct ones. The pooled recall rate (*prr*) can be similarly defined as $prr = \sum_{k=1}^{K}(1-u_k)G_k(u_k)/\sum_{k=1}^{K} Q_{k,x}$. Here $\sum_{k=1}^{K} Q_{k,x}$ is constant (though unknown). Thus, when $\sum_{k=1}^{K}(1-u_k)$ is specified, maximizing $\sum_{k=1}^{K}(1-u_k)G_k(u_k)$ maximizes both *ppr* and *prr*. This leads to our "optimal" decision strategy with the objective function

$$\max_{\substack{u_1,\ldots,u_K,\\ \sum_{k=1}^{K}(1-u_k)=c}} \sum_{k=1}^{K}(1-u_k)G_k(u_k),$$

for any $c \in [0, K]$. This optimization is equivalent to

$$\min_{\substack{u_1,\ldots,u_K,\\ \sum_{k=1}^{K}(1-u_k)G_k(u_k)=c'}} \sum_{k=1}^{K}(1-u_k),$$

to maximize precision given any recall rate. We note that *ppr* can also be considered as the approximate population version of the micro-averaging precision rate (Pillai, Fumera, and Roli (2013)) because the micro-averaging precision can be expressed as

$$\frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K}(TP_k + FP_k)} = \frac{\sum_{k=1}^{K} TP_k/M}{\sum_{k=1}^{K}(TP_k/M + FP_k/M)} \approx ppr,$$

where M is the number of objects. The micro-averaging recall (Pillai, Fumera, and Roli (2013)) can be approximated similarly.

2.2. Optimization based on local precision

We aim to solve the optimization of $\sum_{k=1}^{K}(1-u_k)G_k(u_k)$ given $\sum_{k=1}^{K}(1-u_k)$ fixed. We assume that the precision functions $G_k(u_k)$'s are sufficiently smooth.

We discuss this smoothness assumption in Appendix C. Roughly speaking, the extreme value of (2.1) given $\sum_{k=1}^{K}(1-u_k) = c$ can be obtained when $\frac{d}{du_k}\sum_{k=1}^{K}(1-u_k)G_k(u_k) = 0$, for $k = 1, \ldots, K-1$. This leads to the concept of Local Precision Rate (LPR). Intuitively, the absolute change rate $|\triangle(1-u_k)G_k(u_k)|$ indicates how more likely a random object x is a correct assignment to class k as u_k increases. We define the LPR function for class k as

$$LPR_k(u_k) = -\frac{d}{du_k}\{(1-u_k)G_k(u_k)\}. \tag{2.2}$$

Here are some useful properties of $G_k(u_k)$ and $LPR_k(u_k)$. For a random object x, $(1-u_k)G_k(u_k)$ is a monotonically decreasing function, $LPR_k(u_k)$ is non-negative and, for well-learned classifiers, $LPR_k(u_k)$ is monotonically increasing.

Theorem 1. *Suppose $K = 2$, and that $LPR_1(u)$ and $LPR_2(u)$ are monotonically increasing with u for $0 \le u \le 1$. Take $LPR_1(1) \ge LPR_2(1)$. Then, given $u_1 + u_2 = c$, the ppr at (2.1) is maximized at $u_1 = \min\{u; u \in [0,1], c-u \in [0,1],$ and $LPR_1(u) \ge LPR_2(c-u)\}$.*

Thus for $K = 2$, LPR can be used to compare the classification scores across classes to guarantee the optimum of precision and recall rates given a fixed number of class assignment decisions. This optimization also leads to an optimal precision-recall curve. The proof is deferred to Appendix B. We can easily generalize Theorem 1 to more than two classes. Basically, $(K-1)$ partial derivatives similar to the one in (A.3) can be used to show that the optimal result happens if the LPRs are used to order the candidates across all classes.

2.3. LPR, ℓFDR and the MAP rule

This section relates LPR to the ℓFDR statistic and the maximum a posteriori probability (MAP) rule. For this purpose, we model the distribution of classification scores as a mixture of two distributions. Let $F_k = \pi_{0,k}F_{0,k} + \pi_{1,k}F_{1,k}$, where $F_{1,k}$ is the cdf of classification score when the object is in class k and $F_{0,k}$ is the cdf of classification score when the object is not in class k. $\pi_{1,k}$ is the probability that a randomly selected object is in class k, and $\pi_{0,k} = 1 - \pi_{1,k}$. Then

$$G_k(u) = \frac{\pi_{1,k}\{1 - F_{1,k}(F_k^{-1}(u))\}}{1-u},$$

and

$$LPR_k(u) = \pi_{1,k}\frac{f_{1,k}(F_k^{-1}(u))}{f_k(F_k^{-1}(u))}, \tag{2.3}$$

where f_k, $f_{0,k}$, and $f_{1,k}$ are the derivatives of F_k, $F_{0,k}$, and $F_{1,k}$, respectively. We see that (2.3) is identical to $(1 - \ell\text{FDR})$, also called local true discovery rate

(ℓtdr) in Efron (2010). In Cai and Sun (2009), ℓFDR was used to achieve a similar statistical optimization for the multiple testing of grouped hypotheses.

For a given class k, the classification problem can also be viewed as a simple hypothesis testing problem with two possible values (0 and 1) for the parameter. In a Bayesian framework, when the MAP rule is applied to estimate this binary parameter, it leads to the same statistic as (2.3). Specifically, an object is classified as positive to class k if (2.3)$\geq 1/2$, by the MAP rule.

2.4. Estimation of LPR

Two strategies can be applied to estimate LPRs. One is through (2.2), which can be rewritten as

$$LPR_k(u) = G_k(u) - (1-u)G_k'(u), \tag{2.4}$$

with $G_k'(u)$ the derivative of $G_k(u)$. In (2.4), $G_k(u)$ and $G_k'(u)$ can be estimated simultaneously by applying a local polynomial (quadratic) smoother to a set of precisions and cdf's (the u_i) for any given u. In practice, the true precisions and cdf's are unknown, and thus are replaced by empirical estimates. For any given u we have

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \sum_{i=1}^{M} \mathbb{K}\Big(\frac{u-u_i}{h}\Big)\Big[v_{k,i} - \big\{b_0 + b_1(u-u_i) + b_2(u-u_i)^2\big\}\Big]^2, \tag{2.5}$$

where M is the sample size, $\mathbb{K}(\cdot)$ is a kernel function, h is the smoothing parameter (the bandwidth, selected via leave-one-out cross-validation in our simulation studies and real data analysis), and $u_i \left(= \sum_{j=1}^{M} \mathbb{I}(s_{k,j} \leq s_{k,i})/M\right)$ is the empirical cdf. The bandwidth h is a function of sample size M (see C2 in Appendix for details). Finally, $v_{k,i}$ is the empirical precision using $s_{k,i}$, the score of the i^{th} sample, as a cutoff. As a result, $\hat{G}_k(u) = \hat{b}_0$ and $\hat{G}_k'(u) = \hat{b}_1$ and $LPR_k(u)$ can be estimated by plugging in the two pointwise estimates at (2.4). Asymptotic results on $\hat{G}_k(u)$ and $\hat{G}_k'(u)$ have been well studied (e.g. Bhattacharya and Müller (1993); Fan and Gijbels (1996, 2000)). By combining their asymptotic properties and that $\hat{G}_k'(u)$ has the slower convergence rate $O_p(M^{-2/7})$, the asymptotic properties of $\widehat{LPR}_k(u)$ can be obtained.

Corollary 1. *Under* C1−C5 *in Appendix* C*, we have*

$$\sqrt{Mh^3}\left(\widehat{LPR}_k(u) - LPR_k(u)\right) \xrightarrow{\mathcal{D}} N(-(1-u)\xi, (1-u)^2\delta^2),$$

where

$$\xi = \frac{d}{6}\frac{\int K(t)t^4dt}{\|K\|_2^2}G_k^{(3)}(u), \text{ and } \delta^2 = \frac{\operatorname{var}(V_k|u)}{f(u)}\int K^2(t)t^2dt.$$

The convergence rate of $\widehat{LPR}_k(u)$ is obtained by plugging the optimal bandwidth into $\sqrt{Mh^3}$ and the optimum balances the orders of bias square and variance. The convergence rate is $O_p(M^{-2/7})$.

The other strategy is through (2.3), mathematically identical to ℓtdr, and can be rewritten as a function of $x(= F_k^{-1}(u))$:

$$\ell tdr_k(x) = 1 - \pi_{0,k}\frac{f_{0,k}(x)}{f_k(x)}. \tag{2.6}$$

Equation (2.6) can be estimated by any density estimator in R, Matlab or other software. We do not estimate $f_{1,k}$ directly, since $\ell tdr_k(x) = \pi_{1,k}f_{1,k}(x)/f_k(x)$, but $\pi_{1,k}$ is usually small in practice (e.g., $< .15$ for 95% of the classes in the GEO disease diagnosis dataset), and $\ell tdr_k(x)$ can be expressed as (2.6). We obtain a better estimate of $\ell tdr_k(x)$ via the estimate of $f_{0,k}$. To show the asymptotic property of $\widehat{\ell tdr}_k(x)$ under the conditions in Corollary 2, we first observe that

$$\widehat{\ell tdr}_k(x) = 1 - \frac{1}{1 + \{\hat{\pi}_1\hat{f}^*_{1,k}(x)\}/\{\hat{\pi}_0\hat{f}_{0,k}(x)\}} + O_p\left((Mh)^{-1/2} + h^2\right), \tag{2.7}$$

where $\hat{f}_{0,k}(x)$ and $\hat{f}^*_{1,k}(x)$ are the kernel density estimators with bandwidths h_0 and h, respectively, with h_0 and h of the same order. Here h_0 is the optimal bandwidth for $\hat{f}_{0,k}(x)$ while h used in $\hat{f}^*_{1,k}(x)$ is the optimal bandwidth for $\hat{f}_k(x)$ due to the estimator of (2.6). Since $(M_0h_0)^{1/2}\hat{f}_{0,k}(x)$ and $(M_1h)^{1/2}\hat{f}^*_{1,k}(x)$ in (2.7) are independent and asymptotically normally distributed, we get the following.

Corollary 2. *Under* C1, C2 (*with* $(\nu, \kappa) = (0, 2)$) *and* C4 *in Appendix* C,

$$\widehat{\ell tdr}_k(x) = \ell tdr_k(x) + O_p\left((Mh)^{-1/2} + h^2\right).$$

The asymptotic result in Corollary 2 is obtained under the bandwidth assumption C2 with $(\nu, \kappa) = (0, 2)$, and thus the convergent rate of $\widehat{\ell tdr}_k(x)$ is the same as that of a one dimensional kernel density estimator, $O_p(M^{-2/5})$.

The two corollaries show that the convergence rate of $\widehat{\ell tdr}_k(\cdot)$ is faster than that of $\widehat{LPR}_k(\cdot)$ and theoretically, $\widehat{\ell tdr}_k(\cdot)$ should perform better. However, its performance is sensitive to data layout as well as to the data's distribution complexity. Specifically, if the samples are observed densely in one or two short intervals and very sparsely elsewhere, the estimated $\widehat{\ell tdr}_k(\cdot)$ would show considerable variability and be unreliable. In contrast, $G_k(u)$ in (2.4) is usually smooth and can be estimated more reliably since the u values are always densely observed and evenly spaced. In addition, $\widehat{\ell tdr}_k(\cdot)$ involves estimating the ratio between two probability densities, $f_{0,k}(x)$ and $f_k(x)$, and this might not lead to satisfactory results. These arguments suggest that estimating LPR through (2.4)

is likely a better choice in many problems. In the simulation section, we further demonstrate these points.

2.5. Algorithm for LPR

The details of estimating LPR are summarized in Algorithm 1. With the $\{\widehat{LPR}_k(u_i) : 1 \le i \le M, 1 \le k \le K\}$ and a threshold, the corresponding precision and recall rates (see (A.1)) can be obtained and thus the F-measure (see (A.2)) of a given β.

Algorithm 1 LPR for a given class k

Require: The training set of raw classifier scores $\{s_{k,j}; j = 1, \ldots, M\}$;

Ensure: The estimated LPR $\{\widehat{LPR}_k(u_i) : i = 1, \ldots, M\}$;

1: **Calculate the empirical cdf u_i's and precision $v_{k,i}$'s:**

- $u_i = \sum_{j=1}^{M} \mathbb{I}(s_{k,j} \le s_{k,i})/M$;
- $v_{k,i} = TP_k/(TP_k + FP_k)$, when $s_{k,i}$ is the threshold;

2: **Calculate estimated LPRs:**

- $(\hat{G}_k(u_i), \hat{G}'_k(u_i)) = (\hat{b}_0, \hat{b}_1)$ by (2.5);
- $\widehat{LPR}_k(u_i) = \hat{G}_k(u_i) + (1 - u_i)\hat{G}'_k(u_i)$.

3. Simulation Studies

Given classifier scores for a set of to-be-classified objects, we propose to transform these scores into class-specific LPRs, and then to make classification decisions by comparing the LPR values across all classes for all objects. To evaluate our approach, we compared the two estimators $\widehat{LPR}_k(\cdot)$ from (2.4) and $\widehat{\ell tdr}_k(\cdot)$ from (2.6) to the commonly used methods based on raw classifier scores, p-values, and FDRs, where p-values and FDRs are transformed from raw classifier scores. We also compared the best F-measure (at $\beta = 1$) obtained by our LPR approach with the one from the optimal thresholding approach in Pillai, Fumera, and Roli (2013). We denote the method in Pillai, Fumera, and Roli (2013) as OT.

To estimate ℓtdr, we used the "`ksdensity`" function in Matlab, with the Epanechnikov kernel and the default bandwidth. We do not use the R package for estimating ℓFDR ("fdrtool" Strimmer (2008a,b)) here, because this package was built for hypothesis testing problems and does not utilize any training data. Therefore, to have a fair comparison we decided not to include the results of "`fdrtool`" in the following numerical studies (we found that its performance

Table 1. Parameters of five independent classes.

Class	1	2	3	4	5
X_0	$B(0.5, 3)$	$B(1, 1)$	$B(0.5, 4)$	$B(0.5, 5)$	$B(0.5, 0.5)$
X_1	$B(10, 10)$	$B(4, 0.9)$	$B(10, 10)$	$B(4, 0.9)$	$B(16, 0.1)$

was much worse than all other methods in Table 2 and Table 4, which methods used training data in their estimation.)

3.1. Simulation I

Our first test case used simulated objects that could belong to five independent classes. For simplicity, we assumed that the true LPRs were monotonically increasing. The score of each sample object with respect to class k was randomly generated from the mixed beta distribution,

$$S_k = \frac{X_0}{1.25}\mathbb{I}(\text{cluster}_k = 0) + X_1\mathbb{I}(\text{cluster}_k = 1),$$

where $\mathbb{I}(\cdot)$ is an indicator function and X_0 and X_1 are beta distributions with the parameters given in Table 1. We used beta distributions because they can be right-skewed, left-skewed or centered depending on the chosen parameters.

The range of the generated scores is $[0, 1]$. To make classification easier, we reduced the range of scores generated from the null distribution to $[0, 0.8]$ by dividing the null score by 1.25. In practice, the proportion of alternative (e.g., disease) cases is often much smaller than that of null cases. Therefore, we chose four small values for the sample ratios, $\pi_{1,k} = 0.05$, 0.10, 0.15, and 0.20. We used sample sizes 100, 200, and 500. The number of runs was 100 for each simulation. For each run, we generated two independent samples of the stated size: the training set was used to estimate necessary statistics (such as LPR, $F_{0,k}$, and FDR), while the test set was used to compare the performance of the different approaches. The results are summarized in Tables 2 and 3.

For the five methods, we see that the areas under the precision-recall curves depend strongly on $\pi_{1,k}$, and results are better when $\pi_{1,k}$ is larger as expected. We also see that the performance of the methods generally improves with sample size, except for the raw-classifier-score method. From Table 2, we see that the method using only raw classifier scores performs the worst, while LPR (2.4) does the best. This is not surprising as the statistic p-value is a measure under the null hypothesis and only controls for false positives, and the FDR measures the false discovery rate but ignores the false + true discoveries rate. We note that ℓtdr (2.6) performed worse than p-values in this study though, in theory, it

Table 2. Areas under the overall Precision-Recall curves (Simulation I). The values given here are the average of 100 runs with standard errors in the brackets.

M	$\pi_{1,k}$	Original	LPR	ℓtdr	p-value	FDR
100	0.05	0.573 (0.085)	0.771 (0.087)	0.651 (0.098)	0.688 (0.104)	0.659 (0.104)
	0.1	0.667 (0.053)	0.875 (0.048)	0.793 (0.074)	0.812 (0.064)	0.785 (0.063)
	0.15	0.724 (0.042)	0.909 (0.040)	0.845 (0.060)	0.857 (0.047)	0.837 (0.051)
	0.2	0.766 (0.031)	0.930 (0.027)	0.882 (0.042)	0.894 (0.032)	0.882 (0.035)
200	0.05	0.592 (0.065)	0.817 (0.052)	0.718 (0.070)	0.767 (0.063)	0.734 (0.070)
	0.1	0.666 (0.036)	0.889 (0.031)	0.826 (0.042)	0.849 (0.039)	0.827 (0.042)
	0.15	0.722 (0.030)	0.919 (0.019)	0.877 (0.035)	0.890 (0.028)	0.871 (0.034)
	0.2	0.769 (0.025)	0.938 (0.014)	0.902 (0.026)	0.915 (0.022)	0.902 (0.024)
500	0.05	0.584 (0.047)	0.835 (0.032)	0.769 (0.040)	0.800 (0.038)	0.769 (0.041)
	0.1	0.660 (0.026)	0.900 (0.016)	0.864 (0.025)	0.872 (0.021)	0.849 (0.026)
	0.15	0.721 (0.018)	0.926 (0.012)	0.900 (0.017)	0.905 (0.016)	0.889 (0.019)
	0.2	0.767 (0.015)	0.942 (0.008)	0.922 (0.013)	0.924 (0.009)	0.911 (0.011)

should perform at least as well as LPR (2.4). The order of performance in this simulation setting was

$$LPR \succ \text{p-value} \succ \ell tdr \approx \text{FDR} \succ \text{Raw Scores}\ .$$

When estimating ℓtdr (2.6), one can find both $\hat{f}_{0,k}(\lambda)$ and $\hat{f}_k(\lambda)$ zero. We used some heuristic rules to choose the ℓtdr value for such cases. That worked well for us.

There are efforts to determine class-specific thresholds of classifier scores by optimizing a given global criterion, such as $F_\beta = (1+\beta^2)/\{(1/precision) + (\beta^2/recall)\}$. We compared the best F-measure (with $\beta = 1$) obtained by LPR with that from the OT method based on this simulation data. It was shown in Pillai, Fumera, and Roli (2013) that the maximum F-measure with a given β can be guaranteed by the OT method. The comparison results are presented in Table 3. We see that the two methods perform similarly; the OT method looks slightly better than LPR here but the difference is not significant. The main advantage of LPR over OT is that LPR can achieve the optimum F-measure for different β's all at once, while OT achieves the optimum of a single F-measure each time.

3.2. Simulation II

In practice, there can be sub-populations within classes. Vascular diseases and heart diseases are both cardiovascular diseases, for example, while central nervous system (CNS) infections, brain diseases, and movement disorder are all in the class of CNS disorders. Accordingly, it is reasonable to model the alternatives with mixed distributions. Consider three clusters (0, 1, and 2) that are all

Table 3. F-measure (Simulation I). The values given here are the average of 100 runs with standard errors in the brackets.

M	$\pi_{1,k}$	Raw Scores	LPR	OT
100	0.05	0.615 (0.097)	0.745 (0.079)	0.753 (0.073)
	0.1	0.622 (0.063)	0.830 (0.040)	0.832 (0.042)
	0.15	0.639 (0.042)	0.862 (0.029)	0.865 (0.026)
	0.2	0.673 (0.028)	0.873 (0.024)	0.878 (0.025)
200	0.05	0.610 (0.076)	0.760 (0.047)	0.765 (0.047)
	0.1	0.623 (0.044)	0.834 (0.029)	0.836 (0.030)
	0.15	0.630 (0.030)	0.862 (0.023)	0.867 (0.022)
	0.2	0.666 (0.022)	0.879 (0.019)	0.887 (0.019)
500	0.05	0.622 (0.041)	0.781 (0.032)	0.785 (0.031)
	0.1	0.624 (0.029)	0.836 (0.022)	0.845 (0.021)
	0.15	0.621 (0.021)	0.863 (0.014)	0.872 (0.013)
	0.2	0.665 (0.014)	0.882 (0.012)	0.890 (0.011)

contained within one class k. Cluster 0 corresponds to null cases, and cluster 1 and 2 correspond to two sub-populations of alternative (e.g., disease) cases. For the classifier score associated with class k, a model mimics this situation:

$$S_k = \frac{X_0}{1.25}\mathbb{I}(\text{cluster}_k = 0) + X_1\mathbb{I}(\text{cluster}_k = 1) + X_2\mathbb{I}(\text{cluster}_k = 2),$$

where $\mathbb{I}(\cdot)$ is an indicator function, X_0 and X_1 are beta distributions with parameters given in Table 1, and $X_2 = 0.5$ if $k = 1, 3, 5$, and 1 otherwise.

This model allows us to evaluate the robustness of our approach, as the distributions of cluster 2 are discrete. This property may result in LPRs not monotonically increasing. In this simulation data, the LPR of the fifth class ($k = 5$) is not monotonically increasing because the scores of the first and the third clusters are both centered at 0.5. Since the total proportion of disease cases $\pi_{1,k} + \pi_{2,k}$ should be small, we chose values of $\pi_{1,k}$ and $\pi_{2,k}$ such that $\pi_{1,k} + \pi_{2,k}$ was 0.05, 0.10, 0.15, and 0.20 and $\pi_{2,k}/(\pi_{1,k} + \pi_{2,k})$ was 0.1, 0.3, and 0.5 for $k = 1, \ldots, 5$. As with Simulation I, we repeated this simulation for sample sizes 100, 200, and 500 and produced 100 runs for each simulation. Two independent samples of the specified size were generated for each run: the training set and the test set. The results are summarized in Tables 4 and 5.

Similar to the results of Simulation I, the method that uses raw classifier scores performed worst in all cases. The LPR (2.4) had the best overall performance, and was least affected by adding the cluster 2 (X_2) to the alternative samples. ℓtdr performed similarly to p-values in general, but performed significantly better than FDRs. Also interestingly, p-values and FDRs performed worse when $\pi_{2,k}/(\pi_{1,k} + \pi_{2,k})$ increased. This does not happen to the LPR and ℓtdr,

suggesting that our approach is more robust against distribution noise. The order of performance in this simulation setting was

$$LPR \succ \ell tdr \approx \text{p-value} \succ \text{FDR} \succ \text{Raw Scores}\ .$$

We evaluated the performance of LPR in terms of F-measure. In Table 5, we see that LPR slightly outperforms OT. More specifically, when the sample size is larger and when the data distribution is more complex, LPR tends to perform better than OT. This is consistent with the discussions in Pillai, Fumera, and Roli (2013) on OT's potential over-fitting problem.

3.3. LPR v.s. ℓtdr

We find it better to estimate local precision rates through LPR (2.4) than through ℓtdr (2.6). One possible explanation to LPR's good performance is that the precision function $G_k(u)$ in (2.4) is usually smooth and can be estimated reliably as the u values are always densely observed and evenly spaced. This is true no matter how different and complicated the null and alternative distributions are. In contrast, the performance of $\widehat{\ell tdr}_k(\cdot)$ is very sensitive to data layout as well as the data's distribution complexity. Specifically, if the training samples are observed densely in one or two short intervals and very sparsely elsewhere, the estimated $\widehat{\ell tdr}_k(\cdot)$ can be unreliable. When $\hat{f}_0$ and $\hat{f}$ are estimated using different bandwidths, they have different levels of bias and variance, and the estimated ℓtdr may not be always between 0 and 1. Thus, even though $\hat{G}'(u)$ has a slower convergence rate in some situations LPR (2.4) can still lead to better and more reliable results.

In brief, when the two pdf's (f and f_0) can be well estimated from the data, the performance of ℓtdr should be at least comparable to LPR, while the LPR likely performs better when the density functions are skewed or multimodal.

4. Application to Disease Diagnosis

We use the NCBI GEO datasets in Huang, Liu, and Zhou (2010) to validate our approach. The details of data preprocessing strategies and methods can be found there. Briefly, 100 GEO datasets, consisting of about 9,000 microarrays related to a total of 110 human disease concepts, were collected. Each of the disease concepts is associated with ≥ 3 and ≤ 30 GEO datasets.

We compared the LPR (2.4) to the p-value and the FDR methods, as well as to the "first stage" and "two stage" approaches described in Huang, Liu, and Zhou (2010) (the methods in Huang, Liu, and Zhou (2010) are based on a common cutoff on classifier scores for all classes; see Supplementary Material for more details). Leave-one-out cross-validation was applied to evaluate the performance of the approaches. The results are shown in Figure 1 and Table 6.

Table 4. Area under the overall Precision-Recall curves (Simulation II). The values given here are the average of 100 runs with standard errors in the brackets.

M	$\pi_{1,k}$	$\pi_{2,k}$	Raw Scores	LPR	ℓtdr	p-value	FDR
100	0.045	0.005	0.572(0.101)	0.747(0.107)	0.632(0.120)	0.675(0.122)	0.631(0.124)
	0.035	0.015	0.555(0.078)	0.769(0.108)	0.657(0.115)	0.679(0.110)	0.650(0.110)
	0.025	0.025	0.536(0.091)	0.762(0.119)	0.655(0.130)	0.627(0.111)	0.609(0.115)
	0.090	0.010	0.648(0.056)	0.860(0.047)	0.779(0.073)	0.795(0.062)	0.765(0.071)
	0.070	0.030	0.638(0.062)	0.871(0.044)	0.773(0.065)	0.781(0.069)	0.766(0.072)
	0.050	0.050	0.619(0.060)	0.885(0.047)	0.776(0.068)	0.757(0.073)	0.746(0.071)
	0.135	0.015	0.711(0.040)	0.900(0.033)	0.832(0.045)	0.838(0.044)	0.823(0.049)
	0.105	0.045	0.699(0.040)	0.915(0.029)	0.826(0.055)	0.829(0.047)	0.816(0.050)
	0.075	0.075	0.681(0.043)	0.921(0.030)	0.829(0.050)	0.812(0.049)	0.801(0.050)
	0.180	0.020	0.758(0.030)	0.928(0.027)	0.876(0.040)	0.886(0.034)	0.872(0.038)
	0.140	0.060	0.752(0.033)	0.933(0.023)	0.861(0.042)	0.856(0.040)	0.846(0.044)
	0.100	0.100	0.732(0.035)	0.938(0.023)	0.863(0.043)	0.839(0.043)	0.832(0.045)
200	0.045	0.005	0.575(0.068)	0.814(0.058)	0.720(0.072)	0.753(0.066)	0.723(0.071)
	0.035	0.015	0.559(0.061)	0.825(0.058)	0.717(0.074)	0.712(0.076)	0.684(0.077)
	0.025	0.025	0.558(0.063)	0.851(0.053)	0.733(0.076)	0.708(0.076)	0.689(0.077)
	0.090	0.010	0.643(0.041)	0.889(0.033)	0.825(0.045)	0.833(0.044)	0.814(0.046)
	0.070	0.030	0.635(0.042)	0.895(0.032)	0.811(0.045)	0.804(0.043)	0.786(0.048)
	0.050	0.050	0.629(0.040)	0.914(0.029)	0.816(0.047)	0.795(0.045)	0.785(0.049)
	0.135	0.015	0.717(0.036)	0.923(0.021)	0.868(0.034)	0.877(0.033)	0.858(0.037)
	0.105	0.045	0.705(0.030)	0.934(0.018)	0.863(0.034)	0.857(0.033)	0.845(0.036)
	0.075	0.075	0.683(0.028)	0.935(0.024)	0.862(0.033)	0.833(0.033)	0.824(0.037)
	0.180	0.020	0.766(0.024)	0.942(0.016)	0.898(0.027)	0.904(0.023)	0.893(0.025)
	0.140	0.060	0.748(0.026)	0.945(0.015)	0.888(0.022)	0.884(0.025)	0.873(0.027)
	0.100	0.100	0.733(0.025)	0.948(0.016)	0.892(0.027)	0.866(0.027)	0.861(0.029)
500	0.045	0.005	0.573(0.041)	0.847(0.031)	0.768(0.043)	0.789(0.035)	0.754(0.036)
	0.035	0.015	0.559(0.045)	0.873(0.028)	0.774(0.042)	0.767(0.044)	0.743(0.048)
	0.025	0.025	0.546(0.042)	0.883(0.026)	0.794(0.038)	0.740(0.045)	0.723(0.044)
	0.090	0.010	0.654(0.031)	0.908(0.018)	0.854(0.028)	0.862(0.024)	0.842(0.025)
	0.070	0.030	0.639(0.023)	0.920(0.013)	0.853(0.022)	0.841(0.021)	0.824(0.022)
	0.050	0.050	0.622(0.027)	0.929(0.015)	0.869(0.025)	0.814(0.027)	0.805(0.031)
	0.135	0.015	0.718(0.020)	0.933(0.011)	0.891(0.017)	0.897(0.017)	0.882(0.018)
	0.105	0.045	0.702(0.020)	0.942(0.010)	0.890(0.018)	0.878(0.016)	0.866(0.018)
	0.075	0.075	0.688(0.020)	0.944(0.011)	0.904(0.015)	0.856(0.017)	0.851(0.019)
	0.180	0.020	0.764(0.016)	0.949(0.009)	0.917(0.013)	0.919(0.012)	0.907(0.013)
	0.140	0.060	0.748(0.012)	0.952(0.008)	0.911(0.013)	0.898(0.012)	0.890(0.013)
	0.100	0.100	0.737(0.016)	0.955(0.008)	0.926(0.012)	0.882(0.014)	0.877(0.015)

Figure 1 indicates that LPR significantly outperforms the "first stage" and "two stage" methods from Huang, Liu, and Zhou (2010). LPR also outperforms p-value at almost all recall rates, and significantly outperforms FDR at recall rates < 0.42. While at recall rates between 0.42 and 0.61, FDR performs slightly

Table 5. F-measures (Simulation II). The values given here are the average of 100 runs with standard errors in the brackets.

M	$\pi_{1,k}$	$\pi_{2,k}$	Raw Scores	LPR	OT
100	0.045	0.005	0.616(0.100)	0.747 (0.073)	0.752 (0.073)
	0.035	0.015	0.610(0.089)	0.757 (0.070)	0.765 (0.072)
	0.025	0.025	0.586(0.083)	0.777 (0.082)	0.767 (0.080)
	0.090	0.010	0.614(0.062)	0.822 (0.044)	0.826 (0.045)
	0.070	0.030	0.614(0.062)	0.845 (0.042)	0.831 (0.042)
	0.050	0.050	0.600(0.055)	0.863 (0.043)	0.830 (0.042)
	0.135	0.015	0.630(0.036)	0.855 (0.033)	0.856 (0.033)
	0.105	0.045	0.638(0.041)	0.873 (0.032)	0.850 (0.035)
	0.075	0.075	0.643(0.032)	0.884 (0.037)	0.848 (0.034)
	0.180	0.020	0.678(0.023)	0.876 (0.026)	0.872 (0.027)
	0.140	0.060	0.683(0.031)	0.891 (0.024)	0.866 (0.026)
	0.100	0.100	0.702(0.026)	0.900 (0.029)	0.865 (0.030)
200	0.045	0.005	0.611(0.068)	0.780 (0.049)	0.779 (0.050)
	0.035	0.015	0.598(0.058)	0.786 (0.049)	0.773 (0.050)
	0.025	0.025	0.596(0.069)	0.825 (0.053)	0.798 (0.047)
	0.090	0.010	0.613(0.041)	0.833 (0.028)	0.831 (0.030)
	0.070	0.030	0.606(0.048)	0.862 (0.028)	0.835 (0.031)
	0.050	0.050	0.600(0.047)	0.875 (0.027)	0.839 (0.029)
	0.135	0.015	0.627(0.031)	0.868 (0.022)	0.865 (0.022)
	0.105	0.045	0.629(0.028)	0.885 (0.023)	0.858 (0.022)
	0.075	0.075	0.636(0.023)	0.895 (0.024)	0.851 (0.025)
	0.180	0.020	0.672(0.022)	0.887 (0.018)	0.883 (0.018)
	0.140	0.060	0.685(0.021)	0.898 (0.018)	0.873 (0.021)
	0.100	0.100	0.700(0.023)	0.904 (0.020)	0.866 (0.021)
500	0.045	0.005	0.612(0.044)	0.790 (0.030)	0.791 (0.029)
	0.035	0.015	0.606(0.046)	0.827 (0.027)	0.803 (0.028)
	0.025	0.025	0.603(0.041)	0.849 (0.029)	0.804 (0.028)
	0.090	0.010	0.617(0.032)	0.845 (0.017)	0.840 (0.017)
	0.070	0.030	0.608(0.030)	0.864 (0.018)	0.836 (0.017)
	0.050	0.050	0.595(0.027)	0.885 (0.019)	0.840 (0.018)
	0.135	0.015	0.618(0.019)	0.877 (0.016)	0.869 (0.015)
	0.105	0.045	0.623(0.016)	0.892 (0.012)	0.865 (0.013)
	0.075	0.075	0.634(0.016)	0.901 (0.014)	0.858 (0.016)
	0.180	0.020	0.669(0.014)	0.894 (0.012)	0.887 (0.012)
	0.140	0.060	0.681(0.013)	0.901 (0.011)	0.877 (0.012)
	0.100	0.100	0.702(0.013)	0.908 (0.013)	0.870 (0.012)

better than LPR. However, there is one critical issue with the p-value and FDR methods: they fail to obtain results with high precision. We also applied ℓtdr to the same data. It performed only slightly better than the "two stage" approach, and generated much worse results than the ones from the LPR, p-value, and FDR approaches. The precision rates at various recall rates for all five approaches can

Table 6. Precision rates of different approaches at given recall rates for the NCBI GEO datasets.

	first stage	two stage	LPR	p-value	FDR
0.2	0.25	0.82	0.91	N.A.	N.A.
0.3	0.13	0.58	0.89	0.65	0.74
0.4	0.11	0.27	0.87	0.54	0.68
0.5	0.09	0.20	0.32	0.33	0.57
0.6	0.08	0.18	0.19	0.21	0.21

Table 7. Precision, recall and F-measure of the NCBI GEO datasets.

Method	F-measure	Precision	Recall
LPR	0.531	0.727	0.418
OT	0.545	0.702	0.445

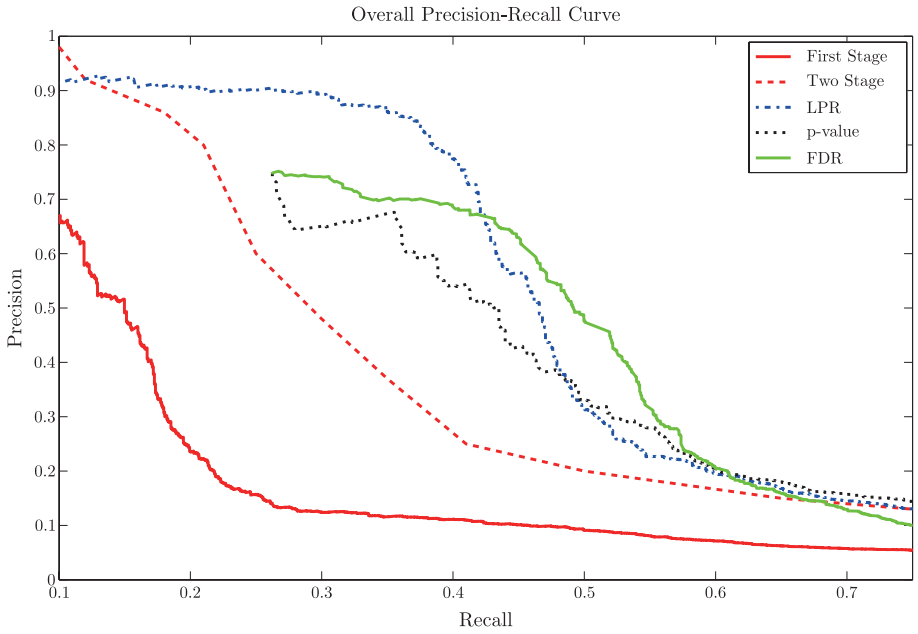

Figure 1. The precision-recall curves of the data.

be found in Table 6.

As in the simulation study, we further compared LPR with the OT method (Pillai, Fumera, and Roli, 2013) in terms of F-measure. The results are summarized in Table 7. We see that the performance of the OT method is comparable or slightly better in terms of F-measure (with $\beta = 1$) than the LPR approach, though the results from LPR have a better precision.

5. Benchmark Datasets

To evaluate the performance of LPR in terms of F-measure, we applied it to three benchmark datasets: Reuters RCV1v2 (text categorization, Lewis et al. (2004)), Scene (image annotation, Boutell et al. (2004)) and Yeast (gene annotation, Elisseeff and Weston (2002)). These datasets were obtained from the LIBSVM Chang and Lin (2011) website (`http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.) Scene and Yeast datasets were originally subdivided into a training and testing set, while RCV1v2 consists of five pairs of training and testing sets. For Scene and Yeast, we pooled the training and testing sets and randomly divided the data into 10 groups. A ten-fold cross-validation was then applied to evaluate the performance of LPR and OT in terms of F-measure. For RCV1v2, we evaluated the performance of LPR and OT based on the given paired training and testing datasets, then switched the role of training and testing and repeated the analysis. Three β's (1/2,1,2) were considered and the numerical results are summarized in Table 8.

Overall, LPR and OT preformed quite similarly on these datasets. For the results related to RCV1v2, LPR looked to have significantly better precision than OT, and LPR had a better F-measure at a smaller β (i.e., at $\beta = 1/2$).

6. Discussion and Conclusions

In this paper, we introduced LPR and demonstrated that transforming classifier scores into LPR values and then making classification decisions accordingly can achieve a globally optimal precision rate at any given recall. Large-margin classifiers such as support vector machines (Cristianini and Shawe-Taylor (2000)) and distance-weighted discrimination (Marron, Todd, and Ahn (2007)) often produce sparse solutions, and their decision functions may not provide estimated class-assignment probabilities (continuous classifier scores) of a query sample. Some techniques (Platt (1999); Bartlett and Tewari (2007); Wang, Shen, and Liu (2008)) exist to reproduce the class-assignment probabilities and then LPR can be applied.

The main advantage of LPR over OT is that LPR can achieve the optimum of F-measure for different β's all at once, while OT only achieves the optimum of a single F-measure each time.

The LPR and ℓtdr are not without flaws. The LPR uses the entire sample and works well if the precision functions G_k's are smooth and densely observed, while ℓtdr uses only a portion of the sample to estimate $f_{0,k}$. Here the variation is larger and the performance depends strongly on the distribution of observed objects. Since smoothing techniques are applied in both estimators, a boundary effect exists, but we see significant improvements after applying LPR in both simulation studies and data analysis.

Table 8. Precisions, recalls and F-measures of the three benchmark datasets.

Dataset	β		OT	LPR
Yeast	1/2	Precision	0.678(0.046)	0.646(0.051)
		Recall	0.440(0.056)	0.514(0.038)
		F-measure	0.610(0.042)	0.613(0.038)
	1	Precision	0.569(0.020)	0.547(0.037)
		Recall	0.683(0.014)	0.716(0.064)
		F-measure	0.620(0.014)	0.617(0.013)
	2	Precision	0.497(0.027)	0.398(0.035)
		Recall	0.829(0.015)	0.937(0.026)
		F-measure	0.731(0.017)	0.736(0.016)
Scene	1/2	Precision	0.799(0.041)	0.806(0.078)
		Recall	0.522(0.063)	0.541(0.087)
		F-measure	0.720(0.032)	0.728(0.043)
	1	Precision	0.654(0.044)	0.661(0.049)
		Recall	0.761(0.063)	0.766(0.049)
		F-measure	0.701(0.035)	0.708(0.030)
	2	Precision	0.545(0.024)	0.543(0.064)
		Recall	0.878(0.018)	0.878(0.052)
		F-measure	0.782(0.016)	0.777(0.017)
RCV1v2	1/2	Precision	0.818(0.028)	0.904(0.008)
		Recall	0.707(0.016)	0.619(0.021)
		F-measure	0.793(0.022)	0.828(0.006)
	1	Precision	0.752(0.046)	0.836(0.020)
		Recall	0.765(0.009)	0.695(0.016)
		F-measure	0.757(0.023)	0.759(0.011)
	2	Precision	0.669(0.051)	0.787(0.037)
		Recall	0.802(0.008)	0.719(0.011)
		F-measure	0.771(0.011)	0.731(0.007)

It might be too strong to assume that the G_k's are smooth, but it is reasonable to assume that they are piecewise smooth, with at most a few discontinuous points. Some nonparametric procedures to address this change-point issue are in McDonald and Owen (1986), Hall and Titterington (1992), and Lee (2002). In practice, one can examine G_k by scatter plots and justify whether assumptions regarding its shape (smooth or piecewise smooth) are satisfied. For simplicity, we estimated G_k by applying a one-dimensional smoother, since the smoothness assumption does not seem to be violated in our data.

We have not considered the complicated hierarchy of classes. Ensuring consistency in a hierarchy is an equally important but separate issue in multi-label classification problems, one beyond the scope of this paper. In future work, we will develop an approach to estimating LPR that incorporates useful information from the LPRs generated by neighboring classifiers in the hierarchy.

Acknowledgements

The authors would like to express gratitude to the referee, an associate editor and the Editor for their constructive comments and suggestions which led to many improvements. Also, the authors wish to thank SAMSI for providing a great research environment that allowed Ci-Ren Jiang to continue participating in this project, and Wayne Lee (a graduate student at UC Berkeley) for his insightful comments and proofreading. Ci-Ren Jiang's research is support in part by NIH Grant EY019094. Xianghong J. Zhou's research is supported in part by NHLBI MAPGen U01HL108634. Haiyan Huang's research is supported in part by NIH Grant EY019094.

Appendix

A. Precision, Recall and F-measure

The precision and recall rates are two sensible summary measures of a confusion matrix that reports the number of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). They have been commonly used to evaluate the performance of a classification approach due to the imbalanced structure of the data. The (micro-averaging) precision and recall rates are

$$Precision = \frac{\sum_k TP_k}{\sum_k (TP_k + FP_k)}, \text{ and } Recall = \frac{\sum_k TP_k}{\sum_k (TP_k + FN_k)}, \tag{A.1}$$

where the subscript k indicates the k-th class. The generalized F-measure is a weighted harmonic mean of precision and recall, and the weight β defines a trade-off between precision and recall. Specifically,

$$F_\beta = \frac{1 + \beta^2}{1/Precision + \beta^2/Recall}. \tag{A.2}$$

Optimizing this measure is a challenging problem, since no closed-form maximizer exists (Dembczyński et al. (2011)). Approaches from different directions have been proposed, such as Pillai, Fumera, and Roli (2013).

B. Proof of Theorem 1

Proof. Given $u_1 + u_2 = c$, the *ppr* in (2.1) can be re-expressed as

$$ppr = \frac{(1 - u_1)G_1(u_1) + \{1 - (c - u_1)\}G_2(c - u_1)}{2 - c}.$$

To determine the condition for *ppr* to be maximized, we check its first derivative with respect to u_1 (we have assumed that the precision functions are sufficiently smooth) and get

$$\frac{dppr}{du_1} = \frac{-LPR_1(u_1) + LPR_2(c - u_1)}{2 - c}. \tag{A.3}$$

- Case I: Suppose $u^* = \min\{u; u \in [0,1], c-u \in [0,1], \text{ and } LPR_1(u) = LPR_2(c-u)\}$. This implies $u^* = \min\{u; u \in [0,1], c-u \in [0,1], \text{ and } LPR_1(u) \geq LPR_2(c-u)\}$ since $LPR_1(1) \geq LPR_2(1)$, and both $LPR_1(\cdot)$ and $LPR_2(\cdot)$ are monotonically increasing. Now we show that *ppr* can be maximized at $u_1 = u^*$. We note that $dppr/du_1$ in (A.3) is monotonically decreasing with u_1. Therefore, $dppr/du_1 = 0$ when $u_1 = u^*$, $dppr/du_1 \geq 0$ when $u_1 < u^*$, and $dppr/du_1 \leq 0$ when $u_1 > u^*$. Therefore, *ppr* can be maximized at $u_1 = u^*$.
- Case II: Suppose that no $u_1 \in [0,1]$ and $c - u_1 \in [0,1]$ satisfy $LPR_1(u_1) = LPR_2(c-u_1)$ or $dppr/du_1 = 0$. Since $dppr/du_1$ is monotonically increasing, we have $dppr/du_1 \leq 0$ always or $dppr/du_1 \geq 0$ always. With $LPR_1(1) \geq LPR_2(1)$, $dppr/du_1$ is always ≤ 0. Then *ppr* is monotonically decreasing with u_1, and should be maximized at $u_1 = \min\{u; u \in [0,1] \text{ and } c-u \in [0,1]\}$. Also, since $dppr/du_1 \leq 0$ always, we have $LPR_1(u) \geq LPR_2(c-u)$ for $u \in [0,1]$ and $c-u \in [0,1]$. Therefore,

  $$u_1 = \min\{u; u \in [0,1] \text{ and } c-u \in [0,1]\}$$

  is equivalent to

  $$u_1 = \min\{u; u \in [0,1], c-u \in [0,1], \text{ and } LPR_1(u) \geq LPR_2(c-u)\}.$$

The above arguments and results tell us that the *ppr* can be maximized for any given call rate by selecting the candidates with the top LPRs from both classes.

C. Local Polynomial Regression

We suppress the subscript k for convenience here.

C.1. Assumption

Let K be a kernel function satisfying the following.

C1

C1.1 K is compactly supported, $\|K\|_2^2 = \int K^2(t)dt < \infty$.

C1.2 K is of order (ν, κ),

$$\int u^\ell K(u) = \begin{cases} 0, & 0 \leq \ell < \kappa, \ell \neq \nu, \\ (-1)^\nu \nu!, & \ell = \nu, \\ \neq 0, & \ell = \kappa. \end{cases}$$

Let $h = h(M)$ be a sequence of bandwidths satisfying the following.

C2 $h \to 0$, $Mh^{\nu+1} \to \infty$ and $Mh^{2\kappa+1} \to d^2 < \infty$.

Suppose that $(u_1, v_1), \ldots, (u_M, v_M)$ are i.i.d. observations of (U, V) with $v = G(u) + \varepsilon$, where ε is an independent random error. Assume that the joint

density, $g(u, v)$, of (U, V) exists. Let $f(u)$ be the marginal density of U. Let $N(u)$ be a neighborhood of u satisfying the following.

C3 The first three derivatives of $G(x)$ exist and are continuous for $x \in N(u)$.

C4 The first two derivatives of $f(x)$ exist and are continuous, and $f(x) > 0$ for $x \in N(u)$.

C5

C5.1 The joint density $g(x, v)$ is continuous on $N(u) \times \mathbb{R}$.

C5.2 The derivatives of $g(x, v)$ exist and are continuous on $N(u) \times \mathbb{R}$.

The smoothness assumptions are on the probability density functions (pdf's) and the precision functions, which are functions of cumulative density functions (cdf's). The smoothness assumptions on pdf's and cdf's are common in the literature of nonparametric statistics. In practice, the empirical precision functions are step functions of cutoff values and the step size depends on the sample size and randomness. It does not appear to be unrealistic to assume that the latent true precision functions are smooth.

C.2. Proof of Corollary 1

We have $LPR = G(u) - (1-u)G'(u)$, where $G(u)$ and $G'(u)$ are estimated via applying a local quadratic regression. The asymptotic distributions of $\hat{G}(u)$ and $\hat{G}'(u)$ can be obtained by applying Theorem 1 in Fan and Gijbels (2000). Kernel functions with different orders can be applied to estimate $G(u)$ and $G'(u)$ to obtain smaller biases, but the asymptotic properties of $\widehat{LPR}$ are dominated by those of $\hat{G}'(u)$ since its convergence rate is $O_p(M^{-2/7})$ while the convergence rate of $\hat{G}(u)$ is $O_p(M^{-2/5})$. Therefore, we can focus on the asymptotic properties of $\hat{G}'(u)$. Under Assumptions C1–C5, let $(\nu, \kappa) = (1, 3)$ in C2. By Theorem 1 in Fan and Gijbels (2000), we get

$$\sqrt{Mh^3}\left(\hat{G}'(u) - G'(u)\right) \xrightarrow{\mathcal{D}} N(\xi, \delta^2),$$

where

$$\xi = \frac{d}{6}\frac{\int K(t)t^4 dt}{\|K\|_2^2} G^{(3)}(u),$$

$$\delta^2 = \frac{\mathrm{var}(V|u)}{f(u)} \int K^2(t)t^2 dt.$$

Thus, the proof is complete.

C.3. Proof of Corollary 2

Suppose the first M_0 observations are from null class while the rest are from the alternative class. The two kernel density estimators are

$$\hat{f}_0(x) = \frac{1}{M_0 h_0} \sum_{i=1}^{M_0} K(\frac{x - X_i}{h_0}),$$

$$\hat{f}(x) = \frac{1}{Mh} \sum_{i=1}^{M} K(\frac{x - X_i}{h}).$$

Both estimators are asymptotically normally distributed under C1, C2 (with $(\nu, \kappa) = (0, 2)$), and C4. Specifically,

$$\sqrt{M_0 h_0} \left(\hat{f}_0(x) - f_0(x) \right) \xrightarrow{\mathcal{D}} N \left((\pi_0 \rho^5)^{1/2} \beta_0, \sigma_0^2 \right),$$

$$\sqrt{Mh} \left(\hat{f}(x) - f(x) \right) \xrightarrow{\mathcal{D}} N \left(\beta, \sigma^2 \right),$$

where $\beta_0 = (d/2)\sigma_K^2 f_0''(x)$, $\beta = (d/2)\sigma_K^2 f''(x)$, $\sigma_0^2 = \|K\|_2^2 f_0(x)$, $\sigma^2 = \|K\|_2^2 f(x)$, $\sigma_K^2 = \int K(t) t^2 dt$, $(h_0/h) \to \rho$, and $(Mh^5)^{1/2} \to d$.

We observe that

$$\begin{aligned}
\hat{f}(x) &= \frac{1}{Mh} \sum_{i=1}^{M} K(\frac{x - X_i}{h}) \\
&= \frac{M_0}{M} \frac{1}{M_0 h} \sum_{i=1}^{M_0} K(\frac{x - X_i}{h}) + \frac{M_1}{M} \frac{1}{M_1 h} \sum_{i=M_0+1}^{M} K(\frac{x - X_i}{h}) \\
&\equiv \hat{\pi}_0 \hat{f}_0^*(x) + \hat{\pi}_1 \hat{f}_1^*(x),
\end{aligned}$$

and therefore

$$\begin{aligned}
\widehat{\ell pr}(x) &= 1 - \frac{1}{\{\hat{\pi}_0 \hat{f}_0^*(x) + \hat{\pi}_1 \hat{f}_1^*(x)\} / \{\hat{\pi}_0 \hat{f}_0(x)\}} \\
&= 1 - \frac{1}{1 + \{\hat{\pi}_1 \hat{f}_1^*(x)\} / \{\hat{\pi}_0 \hat{f}_0(x)\}} + O_p(\frac{1}{\sqrt{Mh}} + h^2),
\end{aligned}$$

because

$$\frac{\hat{f}_0^*(x)}{\hat{f}_0(x)} = 1 + O_p(\frac{1}{\sqrt{Mh}} + h^2).$$

Since $(M_0 h_0)^{1/2} \hat{f}_0(x)$ and $(M_1 h)^{1/2} \hat{f}_1^*(x)$ are independent and asymptotically normally distributed, we have

$$\begin{aligned}
\widehat{\ell pr}(x) =& 1 - \frac{1}{1 + \{\pi_1 f_1^*(x) + O_p((M_1 h)^{-1/2} + h^2)\} / \{\pi_0 f_0(x) + O_p((M_0 h_0)^{-1/2} + h_0^2)\}} \\
&+ O_p \left((Mh)^{-1/2} + h^2 \right) \\
=& \ell pr(x) + O_p \left((Mh)^{-1/2} + h^2 \right).
\end{aligned}$$

Thus, the proof is complete.

References

Alves, R. T., Delgado, M. R. and Freitas, A. A. (2008). Multi-label hierarchical classification of protein functions with artifical immune systems. In *Advances in Bioinformatics and Computational Biology*, Volume 5167, 1-12. Springer, Berlin.

Bartlett, P. and Tewari, A. (2007). Sparseness vs estimating conditional probabilities: Some asymptotic results. *J. Machine Learning Research* **8**, 775-790.

Bhattacharya, P. K. and Müller, H. G. (1993). Asymptotics for nonparametric regression. *Sankhyā* **55**, 420-441.

Boutell, M. R., Luo, J., Shen, X. and Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition* **37**, 1757-1771.

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* **104**, 1467-1481.

Cerri, R., da Silva, R. R. O. and de Carvalho, A. C. (2009). Comparing methods for multilabel classification of proteins using machine learning techniques. In *Advances in Bioinformatics and Computational Biology* **5676**, 109-120. Springer-Verlag, Berlin Heidelberg.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2**, 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press.

de Carvalho, A. and Freitas, A. (2009). *A Tutorial on Multi-label Classification Techniques*, Volume 5 of *Studies in Computational Intelligence* **205**, 177-195. Springer-Verlag, Berlin Heidelberg.

Dembczyński, K., Waegeman, W., Cheng, W. and Hüllermeier, E. (2011). An exact algorithm for f-measure maximization. In *Advances in Neural Information Processing Systems*, **24**, 1404-1412.

Efron, B. (2005). Local false discovery rates. Available at `http://www-stat.stanford.edu/~ckirby/brad/papers/2005LocalFDR.pdf`.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70-86.

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96**, 1151-1160.

Elisseeff, A. and Weston, J. (2002). A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems* (Edited by T. G. Dietterich, S. Becker and Z. Ghahramani), **14**.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

Fan, J. and Gijbels, I. (2000). Local Polynomial Fitting. In *Smoothing and Regression: Approaches, Computation, and Application* (Edited by M. G. Schimek), 229-276. Wiley.

Fan, R.-E. and Lin, C.-J. (2007). A study on threshold selection for multi-label classification. Technical report, Department of Computer Science, National Taiwan University.

Geurts, P., Wehenkel, L. and d'AlchéBuc, F. (2006). Kernelizing the output of tree-based methods. In *Proceedings of the* 23*rd International Conference on Machine Learning*, 345-352.

Hall, P. and Titterington, D. M. (1992). Edge-preserving and peak-preserving smoothing. *Technometrics* **34**, 429-440.

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd edition. Springer.

Huang, H., Liu, J. C.-C. and Zhou, X. J. (2010). Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc. Natl. Acad. Sci. USA* **107**, 6823-6828.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*, 137-142.

Karalic, A. and Pirnat, V. (1991). Significance level based multiple tree classification. *Informatica* **5**, 54–58.

Lee, T. C. M. (2002). Automatic smoothing for discontinuous regression functions. *Statist. Sinica* **12**, 823–842.

Lewis, D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In 15*th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'92), 37-50.

Lewis, D. D., Yang, Y., Rose, T. G. and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research* **5**, 361–397.

Maimon, O. and Rokach, L. (Eds.) (2010). *Data Mining and Knowledge Discovery Handbook*. 2nd edition. Springer.

Marron, J. S., Todd, M. J. and Ahn, J. (2007). Distance-weighted discrimination. *J. Amer. Statist. Assoc.* **102**, 1267-1271.

McDonald, J. A. and Owen, A. B. (1986). Smoothing with split linear fits. *Technometrics* **28**, 195-208.

Parisi, M. and Lin, H. (1999). The drosophila pumilio gene encodes two functional protein isoforms that play multiple roles in germline development, gonadogenesis, oogenesis and embryogenesis. *Genetics* **153**, 235-250.

Pillai, I., Fumera, G. and Roli, F. (2013). Threshold optimisation for multi-label classifiers. *Pattern Recognition* **46**, 2055-2065.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers* (Edited by A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans), MIT Press.

Quevedo, J. R., Luaces, O. and Bahamonde, A. (2012). Multilabel classifiers with a probabilistic thresholding strategy. *Pattern Recognition* **45**, 876-883.

Rousu, J., Saunders, C., Szedmak, S. and Shawe-Taylor, J. (2006). Kernel-based learning of hierarchical multilabel classification models. *J. Mach. Learn. Res.* **7**, 1601-1626.

Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D. and Dzeroski, S. (2010). Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics* **11**, doi:10.1186/1471–2105–11–2.

Strimmer, K. (2008a). fdrtool: a versatile r package for estimating local and tail area-based false discovery rates. *Bioinformatics Applications Note* **24**, 1461-1462.

Strimmer, K. (2008b). A unified approach to false discovery rate estimation. *BMC Bioinformatics* **9**, 303.

Vens, C., Struyf, J., Schietgat, L., Deroski, S. and Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning* **73**, 185-214.

Wang, J., Shen, X. and Liu, Y. (2008). Probability estimation for large-margin classifiers. *Biometrika* **96**, 149-167.

Yang, Y. (2001). A study of thresholding strategies for text categorization. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 137-145.

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan.

E-mail: cirenjiang@stat.sinica.edu.tw

Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung 40227, Taiwan.

E-mail: jimliu@nchu.edu.tw

Department of Biological Sciences, University of Southern California, Los Angeles, California 90033, U.S.A.

E-mail: xjzhou@usc.edu

Department of Statistics, University of California, Berkeley, California 94720, U.S.A.

E-mail: hhuang@stat.berkeley.edu

(Received November 2012; accepted October 2013)