Learning Classifiers When The Training Data Is Not IID

Murat Dundar, Balaji Krishnapuram, Jinbo Bi, R. Bharat Rao
Computer Aided Diagnosis & Therapy Group, Siemens Medical Solutions,
51 Valley Stream Parkway, Malvern, PA-19355, USA
{murat.dundar, balaji.krishnapuram, jinbo.bi, bharat.rao}@siemens.com

Abstract

Most methods for classifier design assume that the training samples are drawn independently and identically from an unknown data generating distribution, although this assumption is violated in several real life problems. Relaxing this i.i.d. assumption, we consider algorithms from the statistics literature for the more realistic situation where batches or sub-groups of training samples may have internal correlations, although the samples from different batches may be considered to be uncorrelated. Next, we propose simpler (more efficient) variants that scale well to large datasets; theoretical results from the literature are provided to support their validity. Experimental results from real-life *computer aided diagnosis* (CAD) problems indicate that relaxing the i.i.d. assumption leads to statistically significant improvements in the accuracy of the learned classifier. Surprisingly, the simpler algorithm proposed here is experimentally found to be even more accurate than the original version.

1 Introduction

Most classifier-learning algorithms assume that the training data is independently and identically distributed. For example, *support vector machine* (SVM), back-propagation for Neural Networks, and many other common algorithms implicitly make this assumption as part of their derivation. Nevertheless, this assumption is commonly violated in many real-life problems where sub-groups of samples exhibit a high degree of correlation amongst both features and labels.

In this paper we: (a) experimentally demonstrate that accounting for the correlations in real-world training data leads to statistically significant improvements in accuracy; (b) propose simpler algorithms that are computationally faster than previous statistical methods and (c) provide links to theoretical analysis to establish the validity of our algorithm.

1.1 Motivating example: CAD

Although overlooked because of the dominance of algorithms that learn from i.i.d. data, sample correlations are ubiquitous in the real world. The machine learning community frequently ignores the non-i.i.d. nature of data, simply because we do not appreciate the benefits of modeling these correlations, and the ease with which this can be accomplished algorithmically. For motivation, consider *computer aided diagnosis* (CAD) applications where the goal is to detect structures of interest to physicians in medical images: *e.g.* to identify potentially malignant tumors in CT scans, X-ray images, etc. In an almost universal paradigm for CAD algorithms, this problem is addressed by a 3 stage system: identification of potentially unhealthy candidate *regions of interest* (ROI) from a medical image, computation of descriptive features for each candidate, and classification of each candidate (*e.g.* normal or diseased) based on its features. Often many candidate ROI point to the same underlying anatomical structure at slightly different spatial locations.

Under this paradigm, correlations clearly exist among both the features and the labels of candidates that refer to the same underlying structure, image, patient, imaging system, doctor/nurse, hospital etc. Clearly, the candidate ROI acquired from a set of patient images cannot be assumed to be IID. Multiple levels of hierarchical correlations are commonly observed in most real world datasets (see Figure 1).

1.2 Relationship to Previous Work

Largely ignored in the machine learning and data mining literature, the statistics and epidemiology communities have developed a rich literature to account for the effect of correlated samples. Perhaps the most well known and relevant models for our purposes are the random effects model (REM) [3], and the generalized linear mixed effects models (GLMM) [7]. These models have been mainly studied from the point of view of *explanatory data analysis* (*e.g.* what is the effect of smoking on the risk of lung cancer?) not from the point of view of *predictive modeling*, *i.e.* accurate classification of unseen test samples. Further, these algorithms tend to be computationally impractical for large scale datasets that are commonly encountered in commercial data mining applications. In this paper, we propose a simple modification of existing algorithms that is computationally cheap, yet our experiments indicate that our approach is as effective as GLMMs in terms of improving classification accuracy.

2 Intuition: Impact of Sample Correlations

Simplified thought experiment Consider the estimation of the odds of heads for a biased coin, based on a set of ob-

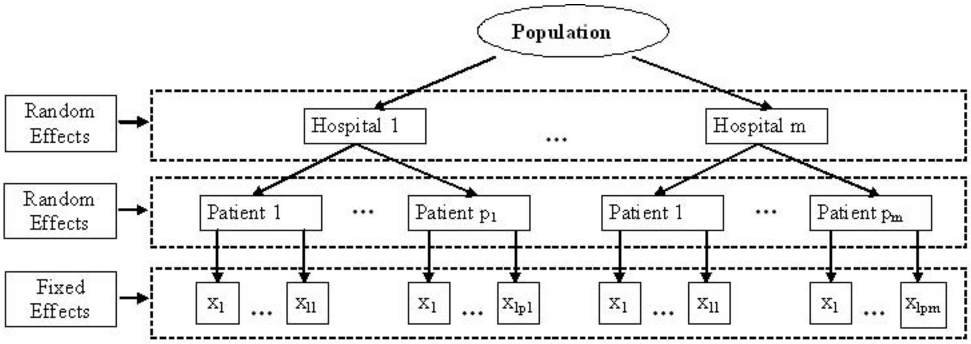

Figure 1: A Mixed effects model showing random and fixed effects

servations. If every observation is an independent flip of a coin, then this corresponds to an i.i.d. assumption for the data; based on this assumption one can easily build estimators for our binomial data. This would work fine so long as the observations reported to the statistician are really true to the underlying coin flips, and provided the underlying data generating mechanism was truly binomial (and not from some other distribution that generates "outliers" as per the binomial model).

However, suppose that the experimenter performing the experiment reports something else to the statistician. After certain coin flips, the experimenter reports the result of one coin flip observation as if it occurred many times (but only if he observed "heads" on those occasions). For other observations he reports the occurrences exactly once. Then the simple binomial estimator (designed using the i.i.d. assumption) would not be appropriate. On the other hand, if every occurrence of an i.i.d. sample is repeated the same number of times, then we are essentially immune to this effect. Although this example may seem simplistic, one should bear in mind that logistic regression, Gaussian Processes and most other classifiers essentially rely on the same binomial distribution to derive the likelihood that they maximize in the training step.

Implications for Classifier Learning The implicit assumption in the machine learning community seems to be that even if IID assumptions are violated, the algorithms would work well in practice. When would this *not* be the case?

The first intuition is that outliers (*e.g.* mis-labeled samples), do not systematically bias the estimation of the classifier during training, provided they are truly i.i.d. and drawn from a fairly symmetric distribution. These outliers introduce a larger variance in the estimation process, but this can be largely overcome simply by increasing the sample sizes of the training set. On the other hand, if outliers (and samples) are systematically correlated, they do introduce a systemic bias in the estimation process, and their effect remains even if we have a large amount of training data.

For explaining the second intuition, let us first consider a practical situation occurring in CAD problems. Due to the way the candidate generation (CG) algorithms for identifying ROI are designed, some diseased structures (*e.g.* wall attached nodules in a lung) may be identified by many candidates that are spatially close in the image; i.e., all these candidates will refer to the same underlying physiological region (*e.g.* the same lung nodule). Moreover, some other types of structures (*e.g.* non-wall attached nodules) may be associated with only one or two candidates, again due to the fundamental properties and biases of the candidate generation (CG) algorithm. The occurrence frequency & other statistical properties of the underlying structural causes of the disease are systematically altered if we naively treated all data as being produced from i.i.d. data sources. As a result, a systematic bias is introduced into the statistical classifier estimation algorithm that learns to diagnose diseases, and this bias remains even when we have large amounts of training data.

When does the violation of i.i.d. assumptions *not* matter? Clearly, if the correlations between samples is very weak, we can effectively ignore them, treating the data as i.i.d. . Further, even if the correlations are not weak, if we do not much care for outlier immunity, and if each sub-type or subpopulation occurs with similar or almost identical frequency, then we should be able to ignore these effects.

3 Random & Mixed Effects Models

Consider the problem shown in Figure 1. We are given a training dataset $\mathcal{D} = \{(x_{ijk}, y_{ijk})\}$, where $x_{ijk} \in \Re^d$ is the feature vector for the i^{th} candidate in the j^{th} patient of the k^{th} hospital and $y_{ijk} \in \{-1, 1\}$ are class labels. Let us also assume without loss of generality that the indexing variable follow $i = \{1, \dots, \ell_j\}, \; j = \{1, \dots, p_k\}, \; k = \{1, \dots, K\}$. In *generalized linear models* (GLM) [7] such as logistic or probit regression, one uses a nonlinear link function—*e.g.* the logistic sigmoid function $\sigma(r) = 1/(1 + \exp(-r)))$—to express the posterior probability of class membership as:

$$P(y_{ijk} = 1 | x_{ijk}, \alpha) = \sigma(\alpha' x_{ijk} + \alpha_0). \tag{1}$$

However, this mathematically expresses conditional independence between the classification of samples and ignores the correlations between samples from the same patient (j, k) or from the same hospital k. This limitation is overcome in a *generalized linear mixed effects model* (GLMM) [7] by postulating the existence of a pair of random variables that explain the patient specific effect $\delta_{j,k}$ and a hospital specific effect

δ_k. During training, one not only estimates the fixed effect parameters of the classifier α, but also the distribution of the random effects $p(\delta_k, \delta_{j,k}|\mathcal{D})$. The class prediction becomes:

$$P(y_{ijk} = 1|x_{ijk}, \alpha) = \int \sigma(\alpha' x_{ijk} + \alpha_0 + \delta_k + \delta_{j,k}) p(\delta_k, \delta_{j,k}) d\delta_k d\delta_{j,k}. \quad (2)$$

In Bayesian terms, classification of a new test sample x with a previously trained GLM involves marginalization over the posterior of the classifier's fixed effect parameters:

$$\mathbb{E}[P(y = 1|x)] = \int \sigma(\alpha' x + \alpha_0) p(\alpha, \alpha_0|\mathcal{D}) d\alpha d\alpha_0. \quad (3)$$

The extension to find the classification decision on a test sample in a GLMM is quite straight-forward:

$$\mathbb{E}[P(y_{ijk} = 1|x_{ijk})] = \int \sigma(\alpha' x_{ijk} + \alpha_0 + \delta_k + \delta_{j,k}) p(\alpha, \alpha_0, \delta_k, \delta_{j,k}|\mathcal{D}) \; d\alpha d\alpha_0 d\delta_k d\delta_{j,k}. \quad (4)$$

Notice that the classification predictions of sets of samples from the same patient (j, k) or from the same hospital k are no longer independent. During the testing phase, if one wants to classify samples from a new patient or a new hospital not seen during training (so the appropriate random effects are not available), one may still use the GLM approach, but rely on a marginalized version of the posterior distributions learnt for the GLMM:

$$\mathbb{E}[P(y = 1|x)] = \int \sigma(\alpha' x + \alpha_0) p(\alpha, \alpha_0, \delta_k, \delta_{j,k}|\mathcal{D}) \; d\alpha d\alpha_0 d\delta_k d\delta_{j,k}. \quad (5)$$

A note of explanation may be useful to explain the above equation. The integral in (5) is over δ_k and $\delta_{j,k}$, *for all j and k*, *i.e.* we are marginalizing over all the random effects in order to obtain $p(\alpha, \alpha_0|\mathcal{D})$ and then relying on (2). Clearly, training a GLMM amounts to the estimation of the posterior distribution, $p(\alpha, \alpha_0, \delta_k, \delta_{j,k}|\mathcal{D})$.

3.1 Avoiding Bayesian integrals

Since the exact posterior can not be determined in analytic form, the calculation of the above integral for classifying every test sample can prove computationally difficult. In practice, we can use *Markov chain monte carlo* (MCMC) methods but they tend to be too slow for data mining applications. We can also use approximate strategies for computing Bayesian posteriors such as the Laplace approximation, variational methods or *expectation propagation* (EP). However, as we see using the following lemma (adapted from a different context [5]), if this posterior is approximated by a symmetric distribution(*e.g.* Gaussian) and we are only interested in the strict classification decision rather than the posterior class membership probability, then a remarkable result holds: The GLMM classification involving a Bayesian integration can be replaced by a point classifier.

First we define some convenient notation (which will use bold font to distinguish it from the rest of the paper). Let us denote the vector combining α, α_0 and all the random effect variables as

$$\boldsymbol{w} = [\alpha', \alpha_0, \delta_1, \delta_2, \ldots, \delta_K, \delta_{11}, \ldots, \delta_{p_K,K}]'.$$

Let us define the augmented vector combining the feature vector x, the unit scalar 1 and a vector $\mathbf{0}$ of zeros—whose length corresponds to the number of random effect variables—as $\boldsymbol{x} = [x', 1, \; \mathbf{0}']'$. Next, observe that if one only requires a hard classification (as in the SVM literature) the sign$(\bullet)$ function is to be used for the link. Finally, note that the classification decision in (5) can be expressed as: $\mathbb{E}[P(y = 1|x)] = \int \text{sign}(\boldsymbol{w}'\boldsymbol{x}) p(\boldsymbol{w}|\mathcal{D}) d\boldsymbol{w}$.

Lemma 1 *For any sample $\boldsymbol{x}$, the hard classification decision of a Point Classifier $f_{PC}(\boldsymbol{x}, \widehat{\boldsymbol{w}}) = \text{sign}(\widehat{\boldsymbol{w}}'\boldsymbol{x})$ is identical to that of a Bayesian Voting Classification*

$$f_{BVC}(\boldsymbol{x}, q) = \text{sign}\left(\int \text{sign}(\boldsymbol{w}'\boldsymbol{x}) q(\boldsymbol{w}) d\boldsymbol{w}\right) \quad (6)$$

if $q(\boldsymbol{w}) = q(\boldsymbol{w}|\widehat{\boldsymbol{w}})$ is a symmetric distribution with respect to $\widehat{\boldsymbol{w}}$, that is, if $q(\boldsymbol{w}|\widehat{\boldsymbol{w}}) = q(\widetilde{\boldsymbol{w}}|\widehat{\boldsymbol{w}})$, where $\widetilde{\boldsymbol{w}} \equiv 2\,\widehat{\boldsymbol{w}} - \boldsymbol{w}$ is the symmetric reflection of $\boldsymbol{w}$ about $\widehat{\boldsymbol{w}}$.

Proof: For every $\boldsymbol{w}$ in the domain of the integral, $\widetilde{\boldsymbol{w}}$ is also in the domain of the integral. Since $\widetilde{\boldsymbol{w}} = 2\,\widehat{\boldsymbol{w}} - \boldsymbol{w}$, we see that $\boldsymbol{w}'\boldsymbol{x} + \widetilde{\boldsymbol{w}}'\boldsymbol{x} = 2\widehat{\boldsymbol{w}}'\boldsymbol{x}$. Three cases have to be considered:

Case 1: When $\text{sign}(\boldsymbol{w}'\boldsymbol{x}) = -\text{sign}(\widetilde{\boldsymbol{w}}'\boldsymbol{x})$, or $\boldsymbol{w}'\boldsymbol{x} = \widetilde{\boldsymbol{w}}'\boldsymbol{x} = 0$, the total contribution from $\boldsymbol{w}$ and $\widetilde{\boldsymbol{w}}$ to the integral is 0, since

$$\text{sign}(\boldsymbol{w}'\boldsymbol{x}) q(\boldsymbol{w}|\widehat{\boldsymbol{w}}) + \text{sign}(\widetilde{\boldsymbol{w}}'\boldsymbol{x}) q(\widetilde{\boldsymbol{w}}|\widehat{\boldsymbol{w}}) = 0.$$

Case 2: When $\text{sign}(\boldsymbol{w}'\boldsymbol{x}) = \text{sign}(\widetilde{\boldsymbol{w}}'\boldsymbol{x})$, since $\boldsymbol{w}'\boldsymbol{x} + \widetilde{\boldsymbol{w}}'\boldsymbol{x} = 2\widehat{\boldsymbol{w}}'\boldsymbol{x}$, it follows that $\text{sign}(\widehat{\boldsymbol{w}}'\boldsymbol{x}) = \text{sign}(\boldsymbol{w}'\boldsymbol{x}) = \text{sign}(\widetilde{\boldsymbol{w}}'\boldsymbol{x})$. Thus,

$$\text{sign}\left[\text{sign}(\boldsymbol{w}'\boldsymbol{x}) q(\boldsymbol{w}|\widehat{\boldsymbol{w}}) + \text{sign}(\widetilde{\boldsymbol{w}}'\boldsymbol{x}) q(\widetilde{\boldsymbol{w}}|\widehat{\boldsymbol{w}})\right] = \text{sign}(\widehat{\boldsymbol{w}}'\boldsymbol{x}).$$

Case 3: When $\boldsymbol{w}'\boldsymbol{x} = 0, \widetilde{\boldsymbol{w}}'\boldsymbol{x} \neq 0$ or $\boldsymbol{w}'\boldsymbol{x} \neq 0, \widetilde{\boldsymbol{w}}'\boldsymbol{x} = 0$, we have again from $\boldsymbol{w}'\boldsymbol{x} + \widetilde{\boldsymbol{w}}'\boldsymbol{x} = 2\widehat{\boldsymbol{w}}'\boldsymbol{x}$ that

$$\text{sign}\left[\text{sign}(\boldsymbol{w}'\boldsymbol{x}) q(\boldsymbol{w}|\widehat{\boldsymbol{w}}) + \text{sign}(\widetilde{\boldsymbol{w}}'\boldsymbol{x}) q(\widetilde{\boldsymbol{w}}|\widehat{\boldsymbol{w}})\right] = \text{sign}(\widehat{\boldsymbol{w}}'\boldsymbol{x}).$$

Unless $\widehat{\boldsymbol{w}} = 0$ (in which case the classifier in undefined), case 2 or 3 will occur at least some times. Hence proved. □

Assumptions & Limitations of this approach: This Lemma suggests that one may simply use point classifiers and avoid Bayesian integration for linear hyper-plane classifiers. However, this is only true if our objective is only to obtain the classification: if we used some other link function to obtain a soft-classification then the above lemma can not be extended for our purposes exactly (it may still approximate the result).

Secondly, the lemma only holds for symmetric posterior distributions $p(\alpha, \alpha_0, \delta_k, \delta_{j,k}|\mathcal{D})$. However, in practice, most approximate Bayesian methods would also use a Gaussian approximation in any case, so it is not a huge limitation. Finally, although any symmetric distribution may be approximated by a point classifier at its mean, the estimation of the mean is still required. However, for computing a Laplace approximation to the posterior, one simply chooses to approximate the mean by the mode of the posterior distribution: the mode can be obtained simply by maximizing the posterior during *maximum a posteriori* (MAP) estimation. This is computationally expedient & works reasonably well in practice.

4 Proposed Algorithm

4.1 Augmenting Feature Vectors (AFV)

In the previous section we showed that full Bayesian integration is not required and a suitably chosen point classifier would achieve an identical classification decision. In this section, we make practical recommendations about how to obtain these equivalent point classifiers with very little effort, using already existing algorithms in an original way.

We propose the following strategy for building linear classifiers in a way that uses the known (or postulated) correlations between the samples. First, for all the training samples, construct augmented feature vectors with additional features corresponding to the indicator functions for the patient & hospital identity. In other words, to the original feature vector of a sample x, append a vector that is composed of zero in all locations except the location of the patient-id, and another corresponding to the location of the hospital-id, and augment the feature vector with the auxiliary features by, $\bar{x} = [x' \ \tilde{x}']'$.

Next use this augmented feature vector to train a classifier using any standard training algorithm such as Fisher's Discriminant or SVMs. This classifier will classify samples based not only on their features, but also based on the ID of the hospital and the patient. Viewed differently, the classifier in this augmented feature space not only attempts to explain how the original features predict the class membership, but it also simultaneously assigns a patient and hospital specific "random-effect" explanation to de-correlate the training data.

During the test phase, for new patients/hospitals, the random effects may simply be ignored (implicitly this is what we used in the lemma in the previous section). In this paper we implement the proposed approach for Fisher's Discriminant.

4.2 Fisher's Discriminant

In this section we adopt the convex programming formulation of Fisher's Discriminant(FD) presented in [8],

$$\begin{array}{lrcl} \min\limits_{\alpha,\,\alpha_0,\xi} & \mathcal{L}(\xi|c_\xi) & + & \mathcal{L}(\alpha|c_\alpha) \\ \text{s.t.} & \xi_i + y_i & = & \alpha^{'} x_i + \alpha_0 \\ & e'\xi^C & = & 0,\, C \in \{\pm\} \end{array} \qquad (7)$$

where $\mathcal{L}(z) \equiv -\log p(z)$ be the negative log likelihood associated with the probability density p. The constraint $\xi_i + y_i = \alpha^{'} x_i + \alpha_0 \; \forall i, i = (1, 2, \ldots, \ell)$ where ℓ is the total number of labeled samples, pulls the output for each sample to its class label while the constraints $e'\xi^C = 0, \; C \in \{\pm\}$ ensure that the average output for each class is the label, i.e. without loss of generality the between class scattering is fixed to be two. The first term in the objective function minimizes the within class scattering whereas the second term penalizes models α that are *a priori* unlikely.

Setting $\mathcal{L}(\xi|c_\xi) = \|\xi\|^2$, $\mathcal{L}(\alpha|c_\alpha) = \|\alpha\|^2$, i.e. assuming a zero mean Gaussian density model with unit variance for $p(\xi|c_\xi)$ and $p(\alpha|c_\xi)$ and using the augmented feature vectors we obtain the following optimization problem. *Fisher's Discriminant with Augmented Feature Vectors (FD-AFV):*

$$\begin{array}{lc} \min\limits_{\alpha,\,\alpha_0,\xi,\delta,\gamma} & \bar{c}_{\xi+} \left\|\xi^+\right\|^2 + \bar{c}_{\xi-} \left\|\xi^-\right\|^2 + \bar{c}_\alpha \left\|\alpha\right\|^2 + \bar{c}_\delta \left\|\delta\right\|^2 \\ \text{s.t.} & \xi_{ijk} + y_{ijk} = \alpha' x_{ijk} + \delta' \tilde{x}_{ijk} + \alpha_0 \\ & e'\xi^C = 0,\; C \in \{\pm\} \end{array}$$

where for the first set of constraints, k runs from 1 to K, j from 1 to p_k for each k, i from 1 to ℓ_{jk} for each pair of (j, k), ξ^C is a vector of ξ_{ijk} corresponding to class C and δ is the vector of model parameters for the auxiliary features.

4.3 Parameter Estimation

Depending on the sensitivity of the candidate generation mechanism, a representative training dataset might have on the order of few thousand candidates of which only few are positive making the training data very large and unbalanced between classes. To account for the unbalanced nature of the data we used different tuning parameters, $\bar{c}_{\xi^+}$, $\bar{c}_{\xi^-}$ for the positive and negative classes. To estimate these and $\bar{c}_\alpha$, $\bar{c}_\delta$, we first coarsely tune each parameter independently and determine a range of values for that parameter. Then for each parameter we consider a discrete set of three values. We use 10-fold cross validation on the training set as a performance measure to find the optimum set of parameters.

5 Experimental Studies

For the experiments in this section, we compare three techniques: naive Fisher's Discriminant (FD), FD-AFV, and GLMM (implemented using approximate *expectation propagation* inference). We compare FD-AFV and GLMM against FD to see if these algorithms yield statistically significant improvements in the accuracy of the classifier. We also study if the computationally much less expensive FD-AFV is comparable to GLMM in terms of classifier sensitivity.

For most CAD problems, it is important to keep the number of false positives per volume at a reasonable level, since each candidate that is marked as positive by the classifier will then be visually inspected by a physician. For the projects described below, we focus our attention on the region of the Receiver Operating Characteristics (ROC) curve with fewer than 5 false positives per volume. This roughly corresponds to 90% specificity for both datasets.

5.1 Experiment 1: Colon Cancer

Problem Description: Colorectal cancer is the third most common cancer in both men and women. It is estimated that in 2004, nearly 147,000 cases of colon and rectal cancer will be diagnosed in the US, and more than 56,730 people would die from colon cancer [4]. While there is wide consensus that screening patients is effective in decreasing advanced disease, only 44% of the eligible population undergoes any colorectal cancer screening. There are many factors for this, key being: patient comfort, bowel preparation and cost.

Non-invasive virtual colonoscopy derived from computer tomographic (CT) images of the colon holds great promise as a screening method for colorectal cancer, particularly if CAD tools are developed to facilitate the efficiency of radiologists' efforts in detecting lesions. In over 90% of the cases colon cancer progressed rapidly is from local (polyp adenomas) to advanced stages (colorectal cancer), which has very poor survival rates. However, identifying (and removing) lesions (polyp) when still in a local stage of the disease, has very high survival rates [1], hence early diagnosis is critical.

This is a challenging learning problem that requires the use of a random effects model due to two reasons. First, the sizes of polyps (positive examples) vary from 1 mm to all the way up to 60 mm: as the size of a polyp gets larger, the number of candidates identifying it increases (these candidates are highly correlated). Second, the data is collected from 152 patients across seven different sites. Factors such as patient anatomy/preparation, physician practice and scanner type vary across different patients and hospitals—the data from the same patient/hospital is correlated.

Dataset: The database of high-resolution CT images used in this study were obtained from NYU Medical Center, Cleveland Clinic Foundation, and five EU sites in Vienna, Belgium, Notre Dame, Muenster and Rome. The 275 patients were randomly partitioned into training (n=152 patients, with 126 polyps among 15596 candidates) and test (n=123 patients, with 104 polyps among 12984 candidates) groups. The test group was sequestered and only used to evaluate the performance of the final system. A combined total of 48 features are extracted for each candidate.

Experimental Results: The ROC curves obtained for the three techniques on the test data are shown in Figure 2. To better visualize the differences among the curves, the enlarged views corresponding to regions of clinical significance are plotted. We performed pair-wise analysis of the three curves to see if the difference between each pair for the area under the ROC-curve is statistically significant (p values computed using the technique described in [2]). Statistical analysis indicates that FD-AFV is more accurate than FD with a p-value of 0.01, GLMM will be more accurate than FD with a p-value of 0.07 and FD-AFV will be more accurate than GLMM with a p-value of 0.18. The run times for each algorithm are shown in Table 1.

5.2 Experiment 2: Pulmonary Embolism

Data Sources and Domain Description

Pulmonary embolism (PE), a potentially life-threatening condition, is a result of underlying venous thromboembolic disease. An early and accurate diagnosis is the key to survival. Computed tomography angiography (CTA) has merged as an accurate diagnostic tool for PE. However, there are hundreds of CT slices in each CTA study. Manual reading is laborious, time consuming and complicated by various PE look-alikes (false positives) including respiratory motion artifact, flow-related artifact, streak artifact, partial volume artifact, stair step artifact, lymph nodes, vascular bifurcation among many others [6], [10]. Several Computer-Aided Detection(CAD) systems are developed to assist radiologists in this process by helping them detect and characterize emboli in an accurate, efficient and reproducible way [9], [11].

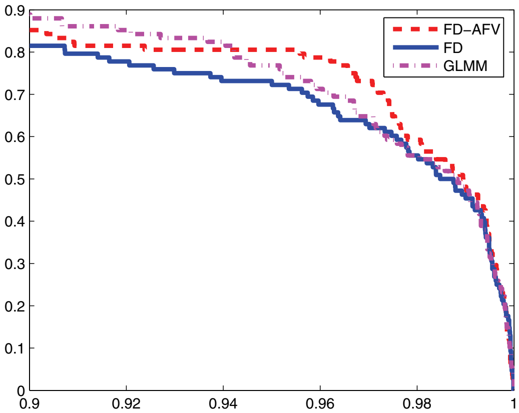

Figure 2: Comparing FD, FD-AFV on the Colon Testing data, $p^{FD}_{FD-AFV} = 0.01, p^{FD}_{GLMM} = 0.07, p^{GLMM}_{FD-AFV} = 0.18$.

Table 1: Comparison of training times in seconds for FD, FD-AFV and GLMM for PE and Colon training data

Algorithm	Time (Colon)	Time (PE)
FD	5	26
FD-AFV	7	28
GLMM	518	329

An embolus forms with complex shape characteristics in the lung making the automated detection very challenging. The candidate generation (CG) algorithm searches for intensity minima. Since each embolus is usually broken into several smaller units, CG picks up several points of interest in the close neighborhood of an embolus from which features are extracted. Thus multiple candidates are generated while characterizing an embolus by the CG algorithm.

In addition to usual patient and hospital level random effects, here we observe a more concrete example of random effects in CAD; the samples within the close neighborhood of an embolus in a patient are more strongly correlated than those far from the embolus. This constitutes a special case of patient-level random effects. All the samples pointing to the same PE are labeled as positive in the ground truth and are assigned the same PE id. We have collected 68 cases with 208 PEs marked by expert chest radiologists at two different institutions. They are randomly divided into two sets: training (45 cases with 142 clots generating a total of 3017 candidates) and testing (23 cases with 66 clots generating a total of 1391 candidates). The test group was sequestered and only used to evaluate the performance of the final system. A combined total of 115 features are extracted for each candidate.

Experimental Design and Results:

The ROC curves obtained for the three techniques on the test data are shown in Figure 3. The statistical analysis indicate

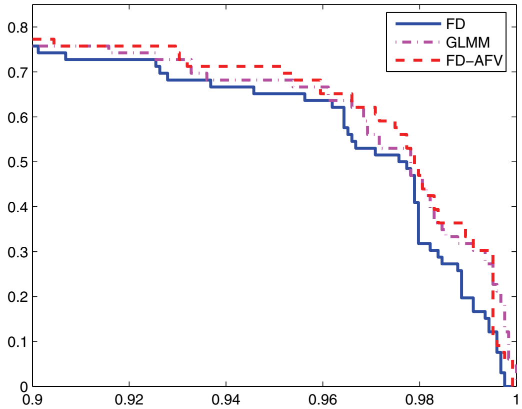

Figure 3: Comparing FD, FD-AFV on the PE Testing data, $p^{FD}_{FD-AFV} = 0.08, p^{FD}_{GLMM} = 0.25, p^{GLMM}_{FD-AFV} = 0.38$.

that FD-AFV will be more sensitive than FD with a p-value of 0.08, GLMM will be more sensitive than FD with a p-value of 0.25 and FD-AFV will be more sensitive than GLMM with a p-value of 0.38. Even though statistical significance can not be proved here (a p-value less than 0.05 is required), a p-value of 0.08 clearly favors FD-AFV over FD. Note also that the ROC curve corresponding to FD-AFV clearly dominates that of FD in the region of clinical interest. When the ROC curves for FD-AFV and GLMM are compared we see that the difference is not significant. However, using the proposed augmented feature vector formulation we were able account for the random effects with much less of computational effort than is required for GLMM.

6 Conclusion

Summary: The basic message of this paper is: there is a standard i.i.d. assumption that is implicit in the derivation of most classifier-training algortihms. By allowing the classifier to explain the data based on both the random effects (common to many samples) and the fixed effects specific to a each sample, the learning algorithm can achieve better performance. If training data is limited, fully Bayesian integration via MCMC algorithms can help improve accuracy slightly, but significiant gains can be reaped without even that effort. All that a practitioner needs to do is to create a categorical indicator variable for each random effect.

Contributions: Generalized Linear Mixed effects models have been extensively studied in the statistics and epidemiology communities. The main contribution of this paper is to highlight the problem and solutions to the Machine learning & Data Mining community: in our experiments, both GLMMs and the proposed approach improve the classification accuracy as compared to ignoring the inter-sample correlations. Our secondary contribution is to propose a very simple and extremely efficient solution that can scale well to very large data mining problems, unlike the current statistics approach to GLMMs that are computationally infeasible for the large datasets. Despite the computational speedup, the classification accuracy of the proposed method was no worse than that of GLMMs in our real-life experiments.

Applicability: Although we describes our approach in a medical setting in order to make it more concrete, the algorithm is completely general and is capable of handling any hierarchical correlations structure among the training samples. Though our experiments focussed on Fisher's Discriminants in this study, the proposed model can be incorporated into any linear discriminant function based classifier.

References

[1] L. Bogoni, P. Cathier, M. Dundar, A. Jerebko, S. Lakare, J. Liang, S. Periaswamy, M. Baker, and M. Macari. Cad for colonography: A tool to address a growing need. *British Journal of Radiology*, 78:57–62, 2005.

[2] J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843, 1983.

[3] H. Ishwaran. Inference for the random effects in bayesian generalized linear mixed models. In *ASA Proceedings of the Bayesian Statistical Science Section*, pages 1–10, 2000.

[4] D. Jemal, R. Tiwari, T. Murray, A. Ghafoor, A. Saumuels, E. Ward, E. Feuer, and M. Thun. Cancer statistics, 2004.

[5] B. Krishnapuram, L. C. M. Figueiredo, and A. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27:957968, 2005.

[6] Y. Masutani, H. MacMahon, and K. Doi. Computerized detection of pulmonary embolism in spiral ct angiography based on volumetric image analysis. *IEEE Transactions on Medical Imaging*, 21:1517–1523, 2002.

[7] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, 1989.

[8] S. Mika, G. Rätsch, and K.-R. Müller. A mathematical programming approach to the kernel fisher algorithm. In *NIPS*, pages 591–597, 2000.

[9] M. Quist, H. Bouma, C. V. Kuijk, O. V. Delden, and F. Gerritsen. Computer aided detection of pulmonary embolism on multi-detector ct, 2004.

[10] C. Wittram, M. Maher, A. Yoo, M. Kalra, O. Jo-Anne, M. Shepard, and T. McLoud. Ct angiography of pulmonary embolism: Diagnostic criteria and causes of misdiagnosis, 2004.

[11] C. Zhou, L. M. Hadjiiski, B. Sahiner, H.-P. Chan, S. Patel, P. Cascade, E. A. Kazerooni, and J. Wei. Computerized detection of pulmonary embolism in 3D computed tomographic (CT) images: vessel tracking and segmentation techniques. In *Medical Imaging 2003: Image Processing. Edited by Sonka, Milan; Fitzpatrick, J. Michael. Proceedings of the SPIE, Volume 5032, pp. 1613-1620 (2003).*, pages 1613–1620, May 2003.