ICU Mortality Prediction: A Classification Algorithm for Imbalanced Datasets

Sakyajit Bhattacharya, Vaibhav Rajan, Harsh Shrivastava
Xerox Research Centre India, Bangalore, India

Abstract

Determining mortality risk is important for critical decisions in Intensive Care Units (ICU). The need for machine learning models that provide accurate patient-specific prediction of mortality is well recognized. We present a new algorithm for ICU mortality prediction that is designed to address the problem of imbalance, which occurs, in the context of binary classification, when one of the two classes is significantly under–represented in the data. We take a fundamentally new approach in exploiting the class imbalance through a feature transformation such that the transformed features are easier to classify. Hypothesis testing is used for classification with a test statistic that follows the distribution of the difference of two χ^2 random variables, for which there are no analytic expressions and we derive an accurate approximation. Experiments on a benchmark dataset of 4000 ICU patients show that our algorithm surpasses the best competing methods for mortality prediction.

Introduction

An Intensive Care Unit (ICU) has the most critically ill patients who are continuously monitored to check for disease progression and potential complications. As the need for ICUs have grown worldwide (Halpern et al. 2013), more ICUs have been created but the availability of resources, both clinical staff and monitoring equipment, remain limited due to many practical constraints. ICU costs have risen amounting to nearly 13% of hospital costs and 5% of the total healthcare cost (Halpern and Pastores 2010).

Continuous monitoring of ICU patients generates a wealth of clinical data presenting opportunities to build predictive models for decision support. The importance of predicting mortality (risk of death) in effective delivery of ICU care is well recognized (Power and Harrison 2014). Identifying high–risk patients can not only aid critical decisions during ICU stay such as interrupting treatments or providing Do–Not–Resuscitate orders but also enable triage, making ICU resources available to other patients in need.

The PhysioNet ICU Mortality Challenge 2012 (Silva et al. 2012) was hosted to encourage the development of new algorithms for ICU mortality prediction. In the challenge itself, the benefits of machine learning algorithms became evident; later algorithmic development on the benchmark dataset has led to up to 170% improvement over traditional risk scoring systems that are used in ICUs today (Johnson et al. 2016). Several aspects of mining clinical data have been examined in these algorithms such as data preprocessing, feature selection, imputation of missing values and the design of new classifiers.

In this paper, we design an algorithm for predicting ICU mortality that addresses the problem of class imbalance. A dataset is called imbalanced if it contains significantly more samples from one class (the *majority* class) than the other class (the *minority* class). Classification on imbalanced datasets is an important problem in clinial data mining (Reddy and Aggarwal 2015). In many datasets, we also find that the sampling distributions of the (training) features *overlap* significantly, that exacerbates the problem of learning from imbalanced data (Denil and Trappenberg 2010).

Our new algorithm for supervised binary classification addresses both the problems of imbalance and overlap in a unique manner. It involves transforming Gaussian random variables into χ^2 random variables where the degree of freedom depends on the mean, variance as well as the class size in the training data. The algorithm exploits the class imbalance in a dataset to achieve a transformation of the features such that the transformed features are well separated. The more the class imbalance, the better the separation we achieve – transforming a disadvantage into an opportunity.

The classification problem is posed as a statistical hypothesis testing problem which involves the distribution of the difference of two χ^2 random variables. However, there is no easily computable analytic expression for the density (PDF) or distribution (CDF) of a linear combination of χ^2–random variables. Approximations exist that can be used in our algorithm and this itself yields better classification accuracy than state–of–the–art methods. Furthermore, we derive new approximations (given in appendix, due to lack of space) to the density and distribution of the difference of two χ^2 variables. These approximations not only improve the classification accuracy further but also possess desirable properties – they are easier to compute, are invertible and are more accurate over a larger range of the random variable – no previous approximation has all these three properties.

To summarize our contributions in this paper,

- We design a new binary classifier consisting of (1) A skewness-based transformation of input features that exploits the class imbalance to achieve better separation (2) Statistical hypothesis tests to obtain the final classification, where the test statistic is a difference between the transformed test feature and skewness measures from training data. We prove that for Gaussian inputs, this test statistic asymptotically follows the distribution of the difference of two independent χ^2 variables (with no known analytic expression).

- We derive new approximations for the PDF, CDF and quantiles of the difference of two χ^2 variables that are more accurate than previous approximations (given in appendix, due to lack of space); when used in our algorithm the approximation further improves predictive accuracy.
- We analyze our algorithm's performance on synthetic data with controlled overlap and imbalance and demonstrate improvement over state-of-the-art methods for ICU mortality prediction on a benchmark dataset of 4000 patients.

Related Work

Mortality Prediction. Several scoring systems for assessing mortality risk in ICUs have been designed such as APACHE (Zimmerman et al. 2006), SAPS (Le Gall, Lemeshow, and Saulnier 1993), and MPM (Lemeshow et al. 1993). They have been developed to assess how care procedures, medications and other clinical factors affect mortality in ICUs. They are mainly adjusted risk models and are not calibrated for patient specific predictions.

In 2012, PhysioNet hosted an ICU Mortality Prediction challenge to encourage the development of new machine learning techniques that can provide patient–specific mortality risk (Silva et al. 2012). Machine learning algorithms designed by the challenge winners surpassed the prediction accuracy of the SAPS risk scoring system that is commonly used in ICUs today. Johnson *et al.* (2014) revisited the problem and through novel data preprocessing techniques further improved the performance of the algorithms.

Design of predictive models using clinical data continues to remain an active research area. Recent literature has examined several new directions in the context of ICU mortality prediction such as computational phenotyping through deep learning (Che et al. 2015) and clinical notes analysis (Ghassemi et al. 2015).

Imbalance and Overlap. The class imbalance problem has been studied extensively – He and Garcia (He and Garcia 2009) and Sun *et al.* (Sun, Wong, and Kamel 2009) provide excellent reviews. Three broad classes of techniques designed for imbalanced–data classification are sampling–based preprocessing techniques, cost–sensitive learning and kernel–based methods. In sampling techniques, the training data is re-sampled, in various ways, to minimize the class imbalance before training the classifier. These include SMOTE (Chawla et al. 2002), where the minority class is inflated by adding synthetic samples that are similar to the data in the feature space, and under–sampling (over-sampling) the majority (minority) class to reduce the imbalance during training. In algorithmic techniques, such as cost–sensitive learning, costs of misclassification or sample weights associated with each class are adjusted in the algorithm itself, e.g. (Breiman, Chen, and Liaw 2004). Several variants of popular classification methods like SVM and ensembles have been proposed to address class imbalance (Akbani, Kwek, and Japkowicz 2004; Galar et al. 2012). Denil and Trappenberg (Denil and Trappenberg 2010; 2011) present a systematic analysis of the overlap problem and its interdependency with class imbalance. Balancing strategies such as SMOTE and undersampling of the majority class have also been used for addressing overlap (Batista, Prati, and Monard 2005).

Our New Classifier

Training. Let A and B be two classes in the context of the given binary classification problem where the training data in class A has n_A observations and training data in class B has n_B observations with $n_A >> n_B$. We denote the training observations in class A as $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_{n_A})$ and the training observations in class B as $\mathbf{y} = (\mathbf{y}_1, ...\mathbf{y}_{n_B})$. Let d be the dimension of each observation. We assume $\mathbf{x}_i$ follows a distribution with mean $\boldsymbol{\mu}_A$ and variance $\boldsymbol{\Sigma}_A$ and $\mathbf{y}_j$ follows a distribution with mean $\boldsymbol{\mu}_B$ and variance $\boldsymbol{\Sigma}_B$ for each i and j. The maximum likelihood estimates (MLE) of the parameters are:

$$\hat{\boldsymbol{\mu}}_A = \frac{1}{n_A}\sum_{i=1}^{n_A}\mathbf{x}_i,\ \hat{\boldsymbol{\mu}}_B = \frac{1}{n_B}\sum_{j=1}^{n_B}\mathbf{y}_j,$$

$$\hat{\boldsymbol{\Sigma}}_A = \frac{1}{n_A}\sum_{i=1}^{n_A}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_A)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_A)^T,$$

$$\hat{\boldsymbol{\Sigma}}_B = \frac{1}{n_B}\sum_{j=1}^{n_B}(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_B)(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_B)^T$$

For each class, from the training observations $\mathbf{x}$ and $\mathbf{y}$, we obtain (scalar) random variables U and V through a cubic-quadratic transformation as given below:

$$U = \sum_{i=1}^{n_A}\sum_{j=1}^{n_A}\left[(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_A)^T\hat{\boldsymbol{\Sigma}}_A^{-1}(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_A)^T\right]^3,$$

$$V = \sum_{i=1}^{n_B}\sum_{j=1}^{n_B}\left[(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_B)^T\hat{\boldsymbol{\Sigma}}_B^{-1}(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_B)\right]^3 \quad (1)$$

Variables U and V are measures of skewness of the distributions of $\mathbf{x}$ and $\mathbf{y}$ (Mardia 1970). For multivariate normal $\mathbf{x}$ and $\mathbf{y}$, the distribution of $\frac{1}{6n_A}U$ and $\frac{1}{6n_B}V$, asymptotically follow the χ^2 distribution with degree of freedom $d(d+1)(d+2)/6$ as shown in (Mardia 1970) [1]:

$$U \sim 6n_A\chi^2_{d(d+1)(d+2)/6},\ \ V \sim 6n_B\chi^2_{d(d+1)(d+2)/6}.$$

To illustrate the transformation, we generate $n_A = 100$ unidimensional observations from the Gaussian $\mathcal{N}(4, 1)$ and $n_B = 25$ unidimensional observations from the Gaussian $\mathcal{N}(6, 1)$. Figure 1 plots the histograms from this data and also shows the distributions of the transformed variables U

[1] This result and theorem 1 below holds for sample sizes $\geq$ 30, from Central Limit Theorem. Larger size will improve the rate of convergence.

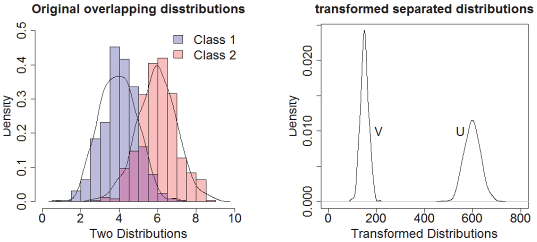

Figure 1: (left) Histogram and fitted densities of simulated Gaussian data, illustrating two overlapping classes, (right) Distributions of random variables U and V obtained through the cubic-quadratic transformations in equation 1.

and V (using equation 1). Notice the overlap in the Gaussians and the distributions of U and V which are well separated. The probability densities are not exact, but calculated from the frequency distribution of the simulated values. Our transformation is similar, in aim but not in technique, to the well–known Linear discriminant analysis (LDA) and Quadratic discriminant analysis (QDA).

Since n_A and n_B are different, the means of U and V that depend on the values of n_A and n_B are well separated. Thus we exploit the imbalance in the data, to achieve a transformation that separates the distributions of U and V. The separation in the distributions is proportional to the difference in the class sizes (assuming moderately large class sizes ≥ 30): the more the difference, the better separation we achieve. The separation is also influenced by the differences in the means and variances of the distributions of $\mathbf{x}$ and $\mathbf{y}$.

We are using the skewness measures of the sampling distributions and not the true distributions. The latter for Gaussian distribution is perfectly symmetric (zero skewness) whereas the former need not be perfectly symmetric. Since the transformations use class sizes (n_A, n_B), the transformed variables follow different χ^2 distributions.

Prediction. Given a test sample Z, we use the same cubic-quadratic transformations to obtain variables Z_1, Z_2 using parameter estimates from class A and B respectively.

$$Z_1 = \frac{1}{6}\left[(Z - \hat{\boldsymbol{\mu}}_A)^T \hat{\boldsymbol{\Sigma}}_A^{-1} (Z - \hat{\boldsymbol{\mu}}_A)\right]^3,$$

$$Z_2 = \frac{1}{6}\left[(Z - \hat{\boldsymbol{\mu}}_B)^T \hat{\boldsymbol{\Sigma}}_B^{-1} (Z - \hat{\boldsymbol{\mu}}_B)\right]^3$$

We pose the classification problem as two hypothesis tests.
1. *Null hypothesis* ($\mathcal{H}_{10}$): Z_1 and $\frac{1}{6n_A}U$ are from the same distribution versus *Alternate hypothesis* ($\mathcal{H}_{11}$): Z_1 and $\frac{1}{6n_A}U$ are from different distributions.
2. *Null hypothesis* ($\mathcal{H}_{20}$): Z_2 and $\frac{1}{6n_B}V$ are from the same distribution versus *Alternate hypothesis* ($\mathcal{H}_{21}$): Z_2 and $\frac{1}{6n_B}V$ are from different distributions.

For hypothesis testing, we need to compute the p–value for the observed value (t) of a test statistic (T). The observed value (t) in our algorithm is $Z_1 - \frac{1}{6n_A}U$ for case 1 and $Z_2 - \frac{1}{6n_B}V$ for case 2 (also see algorithm 1) and the test statistic (T) is a difference of two independent χ^2 variables. The p–value is the probability, under the null hypothesis, of sampling a test statistic at least as extreme as that which was observed, i.e., $P(T > t)$, for positive t. We reject the null hypothesis if the p-value is less than the significance level threshold α. The significance level is the probability of Type I error, i.e. rejection of null hypothesis when it is true. By using a low value of α (≤ 0.05) we can accept the null hypothesis with high confidence at p-values greater than α.

Note that for our classifier, $\mathcal{H}_{10} \implies Z \in A, \mathcal{H}_{11} \implies Z \notin A$. Similarly, $\mathcal{H}_{20} \implies Z \in B, \mathcal{H}_{21} \implies Z \notin B$. Prediction proceeds in the following manner:

- If the p–value in the first test is large ($> \alpha$), then the null hypothesis $\mathcal{H}_{10}$ is accepted, assign label A to Z and stop.
- If the p–value in the first test is small ($< \alpha$), then the null hypothesis $\mathcal{H}_{10}$ is rejected. Since accepting the alternative hypothesis $\mathcal{H}_{11}$, that only implies that $Z \notin A$ does not give us full confidence in assigning label B to Z, we proceed to the second test. If the p–value of the second test is large ($> \alpha$), then the null hypothesis $\mathcal{H}_{20}$ is accepted. Assign label B to Z and stop.
- If the p–value in the first test is small (the null hypothesis $\mathcal{H}_{10}$ is rejected) and the p–value of the second test is also small leading to rejection of the null hypothesis $\mathcal{H}_{20}$ then the tests imply $Z \notin A$ and $Z \notin B$ thereby yielding insufficient confidence in either decision (empirically this occurs in $\sim 5\%$ of the cases). In this case class labels are assigned based on a distance–based rule using the mean squared deviation of Z from $\mathbf{x}$ and $\mathbf{y}$:

$$\text{If } \frac{1}{n_A}\sum_{i=1}^{n_A}(Z - \mathbf{x}_i)^2 < \frac{1}{n_B}\sum_{j=1}^{n_B}(Z - \mathbf{y}_j)^2, \;\; Z \in A \text{ else } Z \in B.$$

Computing p–values. To obtain the p–value, we first need the distribution of our test statistic (T). Let Z^3 denote the component–wise cube of the test sample vector. Let $\widehat{Var_A}(Z^3)$ denote the MLE of the variance of Z^3 based on observations of class A. Assuming that $\widehat{Var_A}(Z^3) = \mathcal{O}(n_A^{-1})$ in probability, theorem 1 states that that our test statistic, $Z_1 - \frac{1}{6n_A}U$, is asymptotically a difference of two independent χ^2 variables (proof in appendix). An equivalent statement holds for $Z_2 - \frac{1}{6n_B}V$. The assumption on $\widehat{Var_A}(Z^3)$ is to ensure that the skewness of the distribution of Z is very low which holds for Gaussian–like distributions.

Theorem 1 *If $\widehat{Var_A}(Z^3) = \mathcal{O}(n_A^{-1})$ in probability, then $Z_1 - \frac{1}{6n_A}U$ asymptotically (as $n_A \to \infty$) follows the distribution of $\chi_1 - \chi_2$ where both χ_1 and χ_2 are independent χ^2 distributions, with degree of freedom $\tilde{d}$ and $d(d+1)(d+2)/6$ respectively. $\tilde{d} = d(d+2)(d+4)/6$ for d even and $\tilde{d} = 2d^2$ for d odd.*

Theorem 1 gives us the distribution of our test statistic but no closed form for the CDF is known, although approximations exist (Press 1966). We derive new approximations for the PDF, CDF and quantiles of the difference of two χ^2 variables (derivations shown in appendix). Using this approximation to compute p–values in our algorithm improves classification accuracy further. The CDF approximations used in our algorithm are strictly necessary when the dimensionality, $d \leq 7$. For $d > 7$, we find that an ensemble–based approach that uses our algorithm on low-dimensional subsets of the data works well in practice (details in appendix). Algorithm 1 shows all the steps in training and prediction.

Algorithm 1 CHISQ Classification Algorithm

```
TRAINING
– Estimate sample means and covariances of classes $A$ and $B$ from the training data:
    $\hat{\boldsymbol{\mu}}_A = \frac{1}{n_A}\sum_{i=1}^{n_A}\mathbf{x}_i, \hat{\boldsymbol{\mu}}_B = \frac{1}{n_B}\sum_{j=1}^{n_B}\mathbf{y}_j$
    $\hat{\boldsymbol{\Sigma}}_A = \frac{1}{n_A}\sum_{i=1}^{n_A}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_A)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_A)^T,$
    $\hat{\boldsymbol{\Sigma}}_B = \frac{1}{n_B}\sum_{j=1}^{n_B}(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_B)(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_B)^T.$
– Compute $U_0$ and $V_0$:
    $U_0 = \frac{1}{6n_A}\sum_{i=1}^{n_A}\sum_{j=1}^{n_A}\left[(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_A)^T\hat{\boldsymbol{\Sigma}}_A^{-1}(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_A)\right]^3,$
    $V_0 = \frac{1}{6n_B}\sum_{i=1}^{n_B}\sum_{j=1}^{n_B}\left[(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_B)^T\hat{\boldsymbol{\Sigma}}_B^{-1}(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_B)\right]^3$
PREDICTION (Given test sample $Z$, significance level $\alpha$)
– Obtain $Z_1$ and $Z_2$:
    $Z_1 = \frac{1}{6}\left[(Z - \hat{\boldsymbol{\mu}}_A)^T\hat{\boldsymbol{\Sigma}}_A^{-1}(Z - \hat{\boldsymbol{\mu}}_A)\right]^3,$
    $Z_2 = \frac{1}{6}\left[(Z - \hat{\boldsymbol{\mu}}_B)^T\hat{\boldsymbol{\Sigma}}_B^{-1}(Z - \hat{\boldsymbol{\mu}}_B)\right]^3.$
  Denote by $T$ the test statistic (difference of two independent $\chi^2$ random variables) whose CDF is evaluated using Press' or our NEW approximation.
– Compute p-value, $p = P(T > Z_1 - U_0)$ where $Z_1 - U_0$ is positive . If $Z_1 - U_0$ is negative, the p-value is given by $p = P(T \leq Z_1 - U_0)$. If $p > \alpha$ assign $Z$ to class $A$ and stop. Else go to the next step.
– Compute p-value, $p = P(T > Z_2 - V_0)$ where $Z_2 - V_0$ is positive . If $Z_2 - V_0$ is negative, the p-value is given by $p = P(T \leq Z_2 - V_0)$. If $p > \alpha$ assign $Z$ to class $B$ and stop. Else go to the next step.
– If $\frac{1}{n_A}\sum_{i=1}^{n_A}(Z - \mathbf{x}_i)^2 < \frac{1}{n_B}\sum_{j=1}^{n_B}(Z - \mathbf{y}_j)^2$ assign $Z$ to class $A$. Else assign $Z$ to class $B$.
```

Computational Complexity. During training, computing U_0 and V_0 takes $\mathcal{O}(d^3 + n_A^2 d^2)$ time which dominates the time complexity. The complexity for prediction is $\mathcal{O}(d^3)$ from the CDF approximation used in computing p–values.

Experiments

Evaluation Metric. We use the Area Under the ROC Curve (AUC) as our evaluation metric. For our algorithm, different operating points on the curve can be obtained by varying the level of significance, α, in hypothesis testing.

Baselines. As baselines we use 6 classifiers: Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and AdaBoost. Additionally we use the following preprocessing techniques all of which are recommended techniques for handling imbalance and overlap (Batista, Prati, and Monard 2005; Denil and Trappenberg 2011): undersampling (UNS) where the majority class is sampled to equal the number of training samples in both classes, SMOTE (Chawla et al. 2002) and cost-sensitive learning (CSL). For CSL, we set the weight in inverse ratio of the number of training samples, i.e., weight of a sample from class $\mathcal{C}_1$ is $n_{\mathcal{C}_2}/n_{\mathcal{C}_1}$ where $n_{\mathcal{C}_1}$ ($n_{\mathcal{C}_2}$) is the number of training samples in class $\mathcal{C}_1$ ($\mathcal{C}_2$). Our classifier with our new approximation and with the approximation of Press are denoted by CHISQ-NEW and CHISQ-PRESS respectively.

Simulated Data. We study the effects of overlap, imbalance and data distributions on classification. We study 4 cases of overlap ($\mathcal{O}$) measured by the common area between the two classes, i.e. $\mathcal{O} = P(\mathcal{U}_{\mathcal{C}_1} \geq \kappa) + P(\mathcal{U}_{\mathcal{C}_2} \leq \kappa)$ where κ is the intersection point of the distribution curves and $\mathcal{U}_{\mathcal{C}_j}$ is the random variable having the same distribution of the j-th class ($j = 1, 2$). See table 1.

	$\mathcal{O}$	λ (Gaussian)	k (Gamma)	I
1	0.617	2	1.5	40:1
2	0.211	3.5	2	20:1
3	0.134	4	3	10:1
4	0.002	7	5	2:1

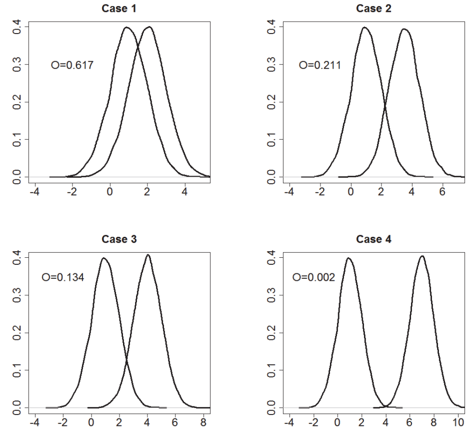

Table 1: Overlap ($\mathcal{O}$) for different values of λ in the Gaussian cluster and k for cluster with gamma distribution $\Gamma(k, \theta)$) and Imbalance Ratio (I) on simulated data. For each value of $\mathcal{O}$, the corresponding value of λ and k are in the same row. Gaussian case in figures below. We simulate 16 datasets by taking all (4 x 4) combinations of $\mathcal{O}$ and I values.

We simulate data from two 3-dimensional Gaussian distributions: $\mathcal{C}_1 \sim N_3(\mathbf{1}, I_3)$ and $\mathcal{C}_2 \sim N_3(\lambda, \Delta_3)$ where I_3 is the 3×3 identity matrix, λ is the mean of the second cluster that controls the cluster overlap ($\mathcal{O}$), and Δ_3 is a covariance matrix with 3×3 dimension having the (i, j)-th component as $0.9^{|i-j|}$. We keep the size of $\mathcal{C}_1$ fixed at

5000, and vary the size of $\mathcal{C}_2$ to create four different imbalance ratios as shown in table 1. To evaluate the performance when there is deviation from model assumptions, we also simulate datasets with two clusters, each with one dimension, where $\mathcal{C}_1$ is from a Gaussian distribution with mean λ (defined above) and variance 1 and $\mathcal{C}_2$ is from a Gamma distribution, $f(x \mid k, \theta) = \theta^{-k}/\Gamma(k)e^{-x/\theta}x^{k-1}$. We take $\theta = 2$ and vary k to create different overlaps as shown in table 1. For each setting, we use 80% of the dataset, chosen randomly, for training and the remaining as test set.

Figure 2 shows the AUC achieved by our classifier (CHISQ) and Random Forest (RF), also applied with preprocessing (RF-UNS, RF-CSL, RF-SMOTE) for the Gaussian case (above) and non-Gaussian case (below). Results of other classifiers are either comparable or worse than RF and are not shown. There are four overlap settings (with different background shades: darker shade for higher overlap) and in each of them four imbalance ratios (increasing imbalance leftwards) as described above.

In the case of Gaussian clusters, in the first 12 out of 16 settings, when there is low to high overlap, CHISQ-NEW outperforms the baselines (and CHISQ-PRESS) in all the imbalance settings. CHISQ-PRESS, that lags behind CHISQ-NEW, outperforms the baselines in 8 out of these 12 settings. With zero overlap (in the last 4 settings), CHISQ does not outperform the others, although the AUC remains above 0.9. For the non-Gaussian case, the performance of all the classifiers is lower, with a noticeable decreasing trend as overlap increases (leftwards in the graph). The performance of CHISQ (with -NEW better than -PRESS) remains higher than all the baselines. With increasing overlap, CHISQ consistently outperforms the baselines at all imbalance ratios.

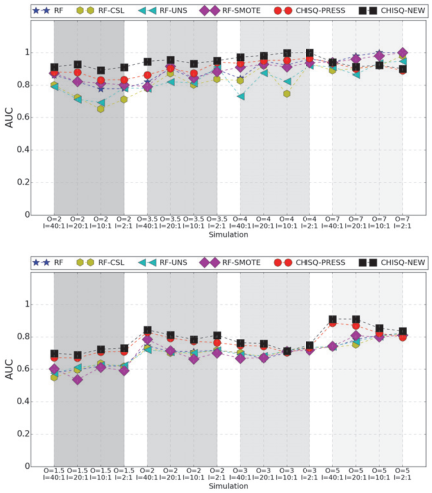

Figure 2: Performance (AUC) of algorithms CHISQ-NEW, CHISQ-PRESS, and RF with preprocessing (-SMOTE, -UNS, -CSL) over different overlaps and imbalance ratios (Above) when both clusters are Gaussian (Above) when one cluster is non-Gaussian (Below). There are 4 overlap settings (darker shade for higher overlap) and in each of them 4 imbalance ratios (decreasing imbalance rightwards)

We conclude that though our algorithm is not the best in the absence of overlap, it does well when there is high overlap even in the presence of high imbalance, the case where other classifiers usually fail. The fact that CHISQ-PRESS, which uses our algorithm but not the new approximation, gives the next best result shows that the improvement is a result of both our algorithm and the new approximation.

ICU Mortality Prediction

Dataset. We use the publicly available labeled dataset of 4000 patients from the Physionet 2012 challenge (Silva et al. 2012) which is from the the MIMIC II ICU database (Goldberger et al. 2000). This dataset, called *Training Set A* in the challenge, is called ICU Mortality dataset in the following. The data for each patient includes age, gender, height, weight, ICU type and 37 time-stamped lab investigations and physiological signal measurements in the first 48 hours of ICU stay. All patients are 16 years or older and had ICU stays of at least 48 hours. Among these, 534 die in the ICU (minority class 1) and 3466 survive (majority class 2).

Preprocessing. We obtain a feature matrix for supervised classification where each row has a single patient's features after variable elimination and feature extraction.

Among the available measurements, 19 of them are absent in more than 35% of the patients; including these would result in a large number of missing values in the feature matrix and so, these measurements are eliminated. We also do not use any measurements that have binary or ordinal datatypes (since our classifier assumes Gaussian inputs). We use the mean values of the measurements averaged over the first 48 hours for each patient, as features. Missing values in the feature matrix are imputed with the column means (from the training fold). Feature selection experiments show that using 6 measurements with the least missing values ($< 2\%$) gives us the best performance.

We briefly outline our feature selection experiments (more details are in the appendix). We select 75% of the data for training and the remaining as test set for these experiments (first fold in the 4-fold CV). The 6 features with the least number of missing values ($< 2\%$) are the mean values of *Mean Arterial Pressure (MAP), heart rate, temperature, sodium level, potassium level, and magnesium level.* With these features CHISQ-NEW achieves a classification accuracy of 90%. The mean values of the remaining 11 continuous–valued features, when added incrementally, do not improve the accuracy obtained by the 6 chosen features. Addition of any of these features individually to the chosen 6 features also does not improve the accuracy. Here we have imputed missing values with column means. With other techniques of missing value imputation or using missing value flags as in Johnson *et al.* (2014) also, we do not observe any increase in accuracy over that achieved by our chosen 6 features.

Other summary statistics such as standard deviation, minimum, maximum and last value in the first 48 hours of patient's ICU stay have been found useful in other studies. In particular, Johnson *et al.* (2014) use 198 such features and with a Random Forest classifier obtain an AUC of nearly 0.84. Use of these features in our classifier (using the

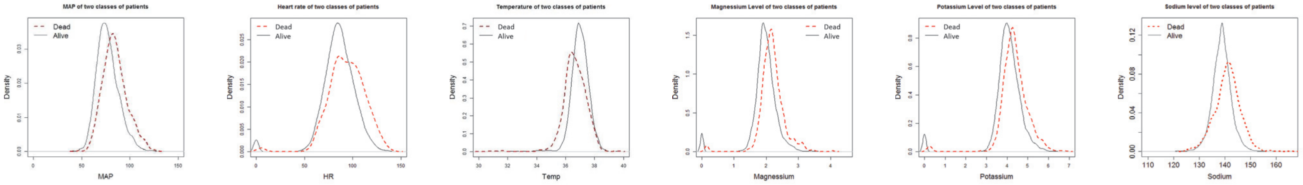

Figure 3: Fitted densities of class 1 (‘dead’: dashed red line) and class 2 (‘alive’: solid black line) in each of the features used. Left to Right: MAP, heart rate, temperature, magnesium, potassium and sodium level.

Algorithm	Score
Johnson *et al.* (2012)	0.48
Citi and Barbieri (2012)	0.47
Vairavan *et al.* (2012)	0.52
CHISQ-PRESS	0.55
CHISQ-NEW	0.60

Table 2: Scores on ICU Mortality dataset (Training Set A) by the top 3 challenge winners and our algorithm CHISQ (-PRESS, using Press’ approximation and -NEW, using our new approximation).

ensemble-based technique for high dimensions described in appendix) does not improve our AUC. The seven most informative features for the Random Forest Classifier (ranked by their Gini importance values) are cumulative sum of urine output, mean, last, minimum and maximum values of BUN, mean value of HCO3 and Age. When these seven features are used with CHISQ-NEW classifier we obtain a classification accuracy of 85% and the AUC remains lesser than 0.867, which is achieved by using our 6 selected features.

Figure 3 shows the fitted densities of the six selected variables that approximately follow Gaussian distribution. An advantage of using only these measurements is that they are part of routine clinical investigations in many patients.

Results

Comparison with 2012 Challenge Winners. The evaluation metric used in the challenge was a score defined as the minimum of recall (R) and precision (P): score $= \min(R, P)$, where $R = TP/(TP + FN)$ and $P = TP/(TP + FP)$; TP, FN and FP denote true positives, false negatives and false positives respectively (positive denotes correct identification of mortality). Table 3 shows the 4-fold cross-validation (CV) scores reported by the top three challenge winners on Training Set A. The mean score obtained by our method with Press’ approximation, CHISQ-PRESS gives a higher score which is further improved by using our new approximation in CHISQ-NEW.

Comparison with other baselines. After the challenge other algorithms were developed for ICU mortality prediction and tested on this dataset. Johnson *et al.* (2014) design new preprocessing techniques and demonstrate improvement over the challenge winners. To our knowledge their method achieves the best reported AUC on this dataset. A deep learning technique using prior-based regularization was also evaluated on this dataset (Che *et al.* 2015). Both these methods use many more features and are not restricted to the 6 features we choose for our method. Table 3 shows the AUC achieved by these methods over 4-fold cross validation. Our algorithm with Press’ approximation, CHISQ-PRESS achieves higher AUC than that of Che *et al.* (2015) and CHISQ-NEW using our new approximation improves the AUC further to outperform that of Johnson *et al.* (2014).

Algorithm	Mean AUC (SD)
Che *et al.* (2015)	0.82 (0.03)
CHISQ-PRESS	0.837 (0.061)
Johnson *et al.* (2014)	0.848 (0.012)
CHISQ-NEW	0.867 (0.031)

Algorithm	Mean AUC (SD)	Pre	Mean AUC (SD)
SVM	0.568 (0.028)	CSL	0.668 (0.017)
RF	0.644 (0.032)	UNS	0.648 (0.029)
LR	0.597 (0.018)	SMOTE	0.599 (0.015)
LDA	0.599 (0.015)	UNS	0.601 (0.016)
QDA	0.66 (0.006)	CSL	0.66 (0.006)
Adaboost	0.541 (0.057)	SMOTE	0.569 (0.015)

Table 3: Results of 4-fold CV on ICU Mortality dataset (Training Set A): our algorithm CHISQ (-PRESS, using Press’ approximation and -NEW, using our new approximation). ABOVE: Mean AUC (Std Dev) comparison with previous algorithms; BELOW: Mean AUC (Std Dev) obtained by 6 baseline classifiers with no preprocessing (2^{nd} column) and with preprocessing (Pre) – best among the 3 techniques (CSL, UNS, SMOTE) shown for each classifier (4^{th} column).

We also test the performance of 6 baseline classifiers using the same 6 features that we choose for our classifier. The aim is to check the improvement with preprocessing techniques like SMOTE, UNS or CSL designed for imbalance. In table 3 we also report the AUC achieved by each of the classifiers before and after preprocessing. We report the highest AUC achieved among the 3 preprocessing techniques for each classifier. None of these classifiers, with or without the preprocessing, are able to achieve an AUC higher than 0.68 which is much lesser than what CHISQ-NEW achieves. Using additional summary statistics (like standard deviation, minimum, maximum of each measurement) as features with RF, SVM and LR improves the AUC up to 0.84 (see Johnson *et al.* 2014) which also is lower than the AUC of CHISQ-NEW.

Conclusion

We present a new binary classification algorithm designed to address the problem of imbalance that is common in clinical datasets. Our algorithm exploits the class imbalance to achieve a unique transformation of the features such that the transformed features are well separated. The transformation results in a difference of χ^2 variables and using approximations for the CDF of the variables, hypothesis testing can be used for classification. We derive new approximations that further improve our algorithm's classification accuracy.

We demonstrate the efficacy of our algorithm on simulated datasets and a large benchmark ICU dataset. An advantage of our method is the use of only six measurements from routine clinical investigations and can easily be obtained from electronic medical records. In comparison other methods or scoring systems use measurements (e.g. Glasgow Coma Scale or an indicator of mechanical ventilation) that may not be available for all patients, may need manual intervention, domain expertise or clinical notes analysis.

References

Akbani, R.; Kwek, S.; and Japkowicz, N. 2004. Applying support vector machines to imbalanced datasets. In *ECML*.

Batista, G. E.; Prati, R. C.; and Monard, M. C. 2005. Balancing strategies and class overlapping. In *Advances in Intelligent Data Analysis VI*. Springer. 24–35.

Breiman, L.; Chen, C.; and Liaw, A. 2004. Using random forest to learn imbalanced data. *Journal of Machine Learning Research* (666).

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 321–357.

Che, Z.; Kale, D.; Li, W.; Bahadori, M. T.; and Liu, Y. 2015. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 507–516. ACM.

Citi, L., and Barbieri, R. 2012. Physionet 2012 challenge: Predicting mortality of ICU patients using a cascaded SVM-GLM paradigm. In *Computers in Cardiology, 2012*. IEEE.

Denil, M., and Trappenberg, T. 2010. Overlap versus imbalance. In *Advances in Artificial Intelligence*. Springer. 220–231.

Denil, M., and Trappenberg, T. 2011. A characterization of the combined effects of overlap and imbalance on the SVM classifier. *arXiv preprint arXiv:1109.3532*.

Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; and Herrera, F. 2012. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 42(4):463–484.

Ghassemi, M.; Pimentel, M. A.; Naumann, T.; Brennan, T.; Clifton, D. A.; Szolovits, P.; and Feng, M. 2015. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *AAAI*.

Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; and Stanley, H. E. 2000. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation* 101(23):e215–e220.

Halpern, N. A., and Pastores, S. M. 2010. Critical care medicine in the United States 2000–2005: An analysis of bed numbers, occupancy rates, payer mix, and costs. *Critical care medicine* 38(1):65–71.

Halpern, N. A.; Pastores, S. M.; Oropello, J. M.; and Kvetan, V. 2013. Critical care medicine in the United States: Addressing the intensivist shortage and image of the specialty. *Critical care medicine* 41(12):2754–2761.

He, H., and Garcia, E. A. 2009. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on* 21(9):1263–1284.

Johnson, A. W.; Dunkley, N.; Mayaud, L.; Tsanas, A.; Kramer, A. A.; and Clifford, G. D. 2012. Patient specific predictions in the intensive care unit using a Bayesian ensemble. In *Computers in Cardiology, 2012*. IEEE.

Johnson, A. E.; Ghassemi, M. M.; Nemati, S.; Niehaus, K. E.; Clifton, D.; and Clifford, G. D. 2016. Machine learning and decision support in critical care. *Proceedings of the IEEE* 104(2):444–466.

Johnson, A. E.; Kramer, A. A.; and Clifford, G. D. 2014. Data preprocessing and mortality prediction: the Physionet/CinC 2012 challenge revisited. In *Computing in Cardiology Conference (CinC), 2014*, 157–160. IEEE.

Le Gall, J.-R.; Lemeshow, S.; and Saulnier, F. 1993. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *The Journal of the Americal Medical Association (JAMA)* 270(24):2957–2963.

Lemeshow, S.; Teres, D.; Klar, J.; Avrunin, J. S.; Gehlbach, S. H.; and Rapoport, J. 1993. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. *The Journal of the Americal Medical Association (JAMA)* 270(20):2478–2486.

Mardia, K. V. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57(3):519–530.

Power, G. S., and Harrison, D. A. 2014. Why try to predict ICU outcomes? *Current opinion in critical care* 20(5):544–549.

Press, S. J. 1966. Linear combinations of non-central chi-square variates. *Annals of Mathematical Statistics* 37:480–487.

Reddy, C. K., and Aggarwal, C. C. 2015. *HealtHcare Data analytics*, volume 36. CRC Press.

Silva, I.; Moody, G.; Scott, D. J.; Celi, L. A.; and Mark, R. G. 2012. Predicting in-hospital mortality of ICU patients: The Physionet/Computing in Cardiology challenge 2012. In *Computers in Cardiology, 2012*. IEEE.

Sun, Y.; Wong, A. K.; and Kamel, M. S. 2009. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence* 23(04):687–719.

Vairavan, S.; Eshelman, L.; Haider, S.; Flower, A.; and Seiver, A. 2012. Prediction of mortality in an intensive care unit using logistic regression and a hidden markov model. In *Computers in Cardiology, 2012*. IEEE.

Zimmerman, J. E.; Kramer, A. A.; McNair, D. S.; and Malila, F. M. 2006. Acute physiology and chronic health evaluation (APACHE IV): Hospital mortality assessment for todays critically ill patients. *Critical care medicine* 34(5):1297–1310.