Risk Clearance with Guaranteed Precision

Ryan McBride* Ke Wang* Viswanadh Nekkanti* Wenyuan Li†

Abstract

In real life applications, we often face the following risk clearance problem: given a set of instances with a known numeric outcome Y (e.g., a toxicity level), we want to learn a model to identify new instances that have a low risk, i.e., the probability of the Y value exceeding a certain maximum MAX is less than some threshold t. This problem guarantees that the cleared instances have the minimum precision of $1-t$ for $Y \leq MAX$. By clearing such low risk instances, we can allocate costly resources to the remaining high risk instances. In this work, we formulate this problem as Risk Clearance with a goal of maximizing the clearance of low risk instances. Existing classification models fail to solve Risk Clearance adequately, so we develop algorithms designed specifically for this problem. We then validate that our approach improves on existing work via experiments on an industrial case study.

1 Introduction

1.1 Motivation In many real life applications we encounter the following *Risk Clearance* problem: given a training set T of observed instances with independent attributes $A_1, \cdots, A_l$ and one dependent numeric attribute Y called the class, we want to partition the attribute space of $A_1, \cdots, A_l$ into *cleared* regions and *non-cleared* regions. A region is a description of the independent attributes for a collection of instances, such as "$Year\_Made = 1975$ and $Manufacturer = M_1$". A region R_i is cleared if the probability that an independently and identically distributed (i.i.d.) random instance sampled from the space of R_i has a class value $Y > MAX$ is less than t. This condition, called the *clearance criterion*, is represented by the expression

$$\Pr[Y > MAX \mid R_i] < t$$

where the MAX and threshold t are a tolerance level of risk. $\Pr[Y > MAX \mid R_i]$ is called the *risk probability* with respect to MAX. We define this problem as:

Definition 1 *(Risk Clearance):* Given a training set T over independent attributes $A_1, \cdots, A_l$ and a dependent numeric class Y, and parameters MAX and t, find a partition of the attribute space for $A_1, \cdots, A_l$ into regions, $\{R_1, \cdots R_m\}$, where each region R_i is marked as either cleared or non-cleared according to the clearance criterion, such that the number of future instances (with unknown class values) that belong to a cleared region is maximized.

The above problem was motivated by a task of identifying toxic transformers for the electrical company British Columbia Hydro. Until 1986, polychlorinated biphenyls (PCBs) based oil mixtures were used in many high voltage transformers; unfortunately, PCBs are carcinogenic and if the PCB concentration exceeds 50 ppm, the transformer's oil must be replaced [1]. We therefore want to infer which transformers likely exceed $MAX = 50$ since directly testing toxicity can cost tens of thousands of dollars per transformer. This problem can be modeled as Risk Clearance with a set of historically tested transformers as the training set. For example, setting $MAX = 50$ and $t = 0.01$ ensures that an unknown toxicity transformer is cleared if its probability of exceeding 50 ppm is less than 1%. The learnt model could then infer, for example, that instances in the region described by "$Year\_Made = 1975$ and $Manufacturer = M_1$" would fulfill this clearance condition.

Risk Clearance finds applications in risk management [2] where instances that have a large risk probability of $Y > MAX$ are handled first. By clearing a large majority of low risk instances, engineers can focus limited resources on a small number of non-cleared, therefore, high risk, instances. The potentially high cost of clearing a high risk instance requires each cleared instance to have some minimum guarantee of low risk. This is expressed by our clearance criterion $\Pr[Y > MAX \mid R_i] < t$: if an instance is cleared, its probability of having $Y \leq MAX$ is at least $1-t$. This *minimum precision* requirement on clearance differs from existing supervised learning that aims to maximize some aggregated measures of recall and precision (e.g. the F-measure and precision-recall curves), and other related objectives in the literature, discussed more

*Computing Science, Simon Fraser University, Burnaby, BC, Canada. {rom2,wangk,vnekkant}@sfu.ca

†The State Key Laboratory of Power Transmission Equipment & System Security and New Technology, Chongqing University, Chongqing, China. wenyuanl.li@ieee.org. Wenyuan is also a former Principal Engineer at BC Hydro

in Section 2. In general, MAX and t can be specified by following regulatory guidelines and industrial standards. For example, many communities in India are devastated by arsenic in ground water wells since chronic exposure of arsenic over a maximum of 10 micrograms per litre causes brain damage, heart disease, and cancer [3]. We could clear most wells with $\Pr[Arsenic > 10 \mid R_i] < t$ for some threshold t by modeling it as a Risk Clearance problem.

It can be shown that all of the following criteria can be converted to the clearance criterion $\Pr[Y > MAX|R_i] < t$: (1) $\Pr[Y \leq MAX|R_i] \geq t$, (2) $\Pr[Y \geq MAX|R_i] \geq t$, and (3) $\Pr[Y < MAX|R_i] < t$, by introducing a new threshold $t' = 1 - t$ and/or a new class label $Y' = MAX - Y$. Assuming that a large Y value represents "badness" or high risks, (1) specifies a condition for good regions, like $\Pr[Y > MAX|R_i] < t$, while (2) and (3) specify conditions for bad regions. Thus, a solution to Risk Clearance also provides a solution for identifying bad regions.

1.2 Contributions The contributions of this work are summarized as follows.

- In Section 1, we motivate and formulate a new learning problem called Risk Clearance that aims to clear as many instances as possible for a given clearance criterion $\Pr[Y > MAX \mid R_i] < t$. The novelty of this problem is considering the actual distribution of the numeric class values Y relative to the cutoff value MAX, instead of the binary class of $Y > MAX$ and $Y \leq MAX$, thus allowing a more informative estimation of the risk probability and clearing many low risk instances with a guaranteed precision (i.e., $1 - t$). Indeed, our survey of related work in Section 2 suggests that existing supervised learning solutions do not address Risk Clearance's objective.
- We present an efficient solution to the Risk Clearance problem in Section 3. This solution assumes a method for estimating the risk probability $\Pr[Y > MAX|R_i]$ for a given region R_i.
- We present a procedure for estimating the risk probability $\Pr[Y > MAX|R_i]$ in Section 4 that fulfills three essential requirements: it exploits exact numeric values of Y, is statistical in that $\Pr[Y > MAX|R_i]$ is valid for the whole population in R_i, and requires minimal assumptions on the Y value distribution.
- In Section 5, we conduct experiments to evaluate the effectiveness of our solution in meeting the clearance criterion while achieving the goal of clearance maximization.

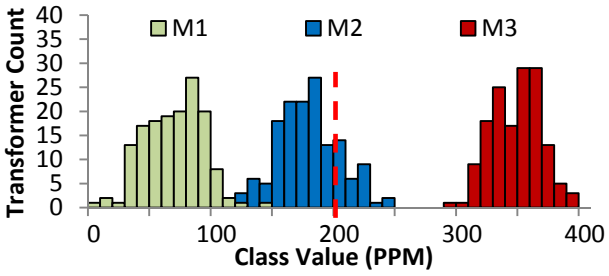

Figure 1: $t = 10\%$ and $MAX = 200$. Risk Clearance will clear the transformers produced by manufacturers M1 and M2 by grouping them into one region. The regression method will clear only the transformers produced by M1 by producing one group for each manufacturer to minimize the sum of squared error.

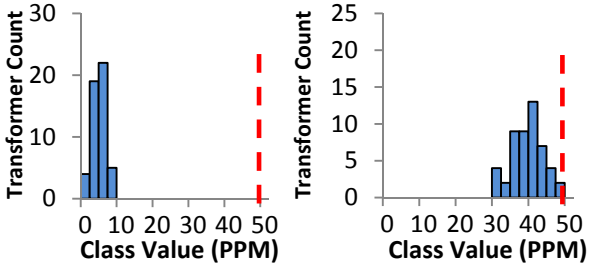

Figure 2: CUT Classification cannot distinguish these two cases with a class value ≤ 50, though the left distribution is less likely to exceed the $MAX = 50$ ppm than the right distribution.

2 Related Work

Existing supervised learning solutions in literature have not addressed the design goal or motivation of Risk Clearance. Below, we survey several well-studied problems in this regard.

Regression. Given a set of training examples $X_1 \ldots X_n$ with numeric class values of $Y_1 \ldots Y_n$, regression methods aim to predict Y_{n+1} for a new example X_{n+1}. To maximize the prediction accuracy, Regression Trees [4] search for a partition of the attribute space to minimize the sum of squared errors in each region. This objective does not address our goal of maximizing clearance with respect to the MAX and t. For example, to minimize the sum of squared error, Regression Trees would partition the 150 transformers in Figure 1 into three regions corresponding to three manufacturers M1, M2, and M3, where the horizontal axis represents the PCB level. For $MAX = 200$ and $t = 10\%$, only the M1 region can be cleared because the risk probability of each of M2 and M3 is greater than t. In contrast, Risk Clearance can clear all the instances in M1 or M2 by combining them as a single region that has a risk probability of no more than t.

CUT Classification [5], motivated by a similar transformer toxicity problem, considers a clear-

ance problem for binary classes (Positive and Negative) where an instance is cleared if its probability of being the negative class is no more than t. To solve Risk Clearance, one can first relabel each training instance as Negative or Positive according to whether $Y > MAX$, and then apply CUT Classification to this binary data set. Unfortunately, the actual distribution of numeric class values is lost in the binary class. For example, the binary class can no longer distinguish the two cases in Figure 2 because both have 50 positive ($\leq MAX$) instances in the training set. However, the actual distribution implies that the future instances from the left distribution are much less likely to exceed MAX than those from the right distribution.

Cost-Sensitive Models [6] recognize the asymmetry of misclassification by specifying the relative cost of predicting an instance as having a class value of $\hat{Y}$ when its true class value is Y (see [7] for one example of this formulation), and trying to minimize the overall cost of all predictions. In practice, such detailed cost information may be unavailable or hard to obtain [5]. Even if these costs were estimated, minimizing the overall estimated cost cannot guarantee that our clearance criterion $\Pr[Y > MAX \mid R_i] < t$ is satisfied.

Classification for Imbalanced Data [8] aims to bias classifiers towards good identification of the minority class (e.g. $Y > MAX$ cases in our context) by, for example, oversampling the minority class cases then running a regular classifier on this rebalanced data. This modeling ignores the clearance threshold t and the actual distribution of the numeric Y values.

Uplift Modeling [9] searches for the instances that benefit the most from a treatment action (e.g. replacing a transformer) in comparison to not intervening. This corresponds to searching for instances with a large change $\Pr[Y \leq MAX|Treated] - \Pr[Y \leq MAX|Not\ Treated]$. In the example of the transformer problem, $\Pr[Y \leq MAX|Treated]$ after replacing contaminated oil is always 100%, thus, Uplift Modeling becomes equivalent to searching for instances minimizing $\Pr[Y \leq MAX|Not\ Treated]$, which fails to address the threshold requirement for the probability specified by t.

In summary, Risk Clearance encodes a new risk prioritization problem with exact numeric class values and a specification of the clearance criterion in terms of MAX and t. Existing work either fails to consider this clearance criterion or fails to benefit from exploiting the actual distribution of exact numeric class values in clearance maximization.

3 *RiskClear:* The Algorithm

This section describes our solution, RiskClear, that partitions the attribute space into regions $\{R_1, \cdots, R_m\}$ where a region R_i is cleared if $\Pr[Y > MAX|R_i] < t$. The goal of this partitioning is to maximize the number of future instances that are assigned to a cleared region. We describe the main algorithm structure in this section then discuss how to estimate the risk probability $\Pr[Y > MAX|R_i]$ in Section 4.

We present RiskClear in two steps. In Section 3.1, we present a region partitioning procedure to clear a high percentage of training instances based on $\Pr[Y > MAX|R_i] < t$. With $\Pr[Y > MAX|R_i]$ being estimated for the whole population in the region R_i, not just for the observed training instances in R_i, this solution implies that a similar proportion of instances will be cleared when applied to future instances. This procedure is called ClearTree. In Section 3.2, we discuss clearing more instances by iteratively applying ClearTree to the training instances that have not been cleared by any previous iteration (i.e., the attribute space of all currently non-cleared regions).

3.1 ClearTree: One Iteration Given a region R containing training instances T, the cutoff class value MAX, and the clearance threshold t, $ClearTree$ in Algorithm 1 searches for a tree where the leaf nodes represent a partitioning of R into cleared and non-cleared regions. The algorithm first checks if R satisfies the clearance criterion, i.e., $\Pr[Y > MAX|R] < t$, and labels it as *cleared* if this condition is met (in lines 1-2); otherwise, the algorithm checks if the stopping condition $Stopping(R, MAX, t)$ (defined later) is satisfied and labels R as *non-cleared* if true (in line 3-4). If the stopping condition does not hold then in line 5 the algorithm checks every candidate partitioning of R, returned by $Candidate(R)$, and chooses the one that assigns the most training instances to a cleared region R_i as determined by $\Pr[Y > MAX|R_i] < t$. If no partitioning can clear any instances, at line 9 the algorithm chooses the partitioning with the highest "potential" of clearing future instances measured by $\Sigma_{i=1}^{m} Score(R, R_i, MAX, t)$. $Best$ denotes the chosen partitioning. $ClearTree$ is repeated on each region R_i in $Best$ (in lines 10-11). We do not prune the resulting tree because the estimation of the probability $\Pr[Y > MAX|R_i]$ is statistical (i.e., holds for the whole population in R_i) and our stopping condition (see below) helps prevent regions with too small sample sizes from being explored.

Next, we define $Stopping$, $Candidate$, and $Score$.

Stopping($\mathbf{R}, \mathbf{MAX}, \mathbf{t}$): We choose to stop further partitioning if the number of training instances in R is below 20, based on the rule of thumb that statistical tests with fewer than 20 or 30 samples do not generally work effectively.

Candidate($\mathbf{R}$): This function enumerates all can-

Algorithm 1 $ClearTree(R, MAX, t)$

Require: A training set in a region R, the class cutoff value MAX, and the clearance threshold t.

1: **if** $\Pr[Y > MAX|R] < t$ **then**
2: label R as *cleared* and return
3: **if** $Stopping(R, MAX, t)$ is true **then**
4: label R as *non-cleared* and return
5: let $Max\_Cleared$ be the maximum count of training instances cleared by any candidate partitioning in $Candidate(R)$
6: **if** $Max\_Cleared > 0$ **then**
7: let $Best = \{R_1, \cdots, R_m\}$ be the candidate partitioning that clears $Max\_Cleared$ training instances
8: **else**
9: let $Best = \{R_1, \cdots, R_m\}$ be the candidate partitioning maximizing $\sum_{i=1}^{m} Score(R, R_i, MAX, t)$
10: **for all** R_i **in** $Best$ **do**
11: call $ClearTree(R_i, MAX, t)$

Algorithm 2 $RiskClear(R, MAX, t)$

Require: A training set in a region R, the class cutoff value MAX, and the clearance threshold t.

1: **repeat**
2: call $ClearTree(R, MAX, t)$
3: Update R to the set of training instances in R not cleared
4: **until** R does not change

didate partitions of R from partitions related to each of R's attributes. For a numeric attribute A_j we consider the binary partitioning $\{R_1, R_2\}$ for each distinct split value v, the midpoint of two consecutive distinct values of A_j in the training set, where R_1 contains all instances in R having their A_j value $\leq v$ and R_2 contains all other instances in R. For a categorical attribute A_j we consider all "one-vs-rest" partitions where each partitioning has two regions $\{R_1, R_2\}$ such that R_1 contains instances in R matching one categorical value of the attribute and R_2 contains all other instances in R.

Score$(\mathbf{R}, \mathbf{R_i}, \mathbf{MAX}, \mathbf{t})$: This function computes a score measuring the potential clearance of a region R_i. If $\Pr[Y > MAX|R_i < t$, i.e., R_i is cleared, then its score is $|R_i|$, i.e., the number of training instances in R_i, otherwise, if $\Pr[Y > MAX|R_i]$ is greater than $\Pr[Y > MAX|R]$, R_i makes clearance harder, so we set the score for R_i to zero; otherwise, R_i improves the probability to clear over its parent region R and we prefer an R_i that covers many instances with a probability closer to the threshold t, i.e., a larger $\mathbf{\Delta} = |R_i| \cdot [1 - (\Pr[Y > MAX|R_i] - t)]$. Our overall scoring function is thus:

$$Score(R, R_i, MAX, t) = \begin{cases} |R_i| & \text{if } \Pr[Y > MAX|R_i] < t \\ 0 & \text{if } \Pr[Y > MAX|R_i] \\ & \quad > \Pr[Y > MAX|R] \\ \mathbf{\Delta} & \text{otherwise} \end{cases}$$

3.2 RiskClear: Iterative Algorithm The method $ClearTree(R, MAX, t)$ produces a set of regions $R_1, \cdots, R_m$ represented by the tree's leaf nodes, where each region is labeled as either *cleared* or *non-cleared*. To maximize clearance, we repeatedly apply $ClearTree()$ to the set of training instances belonging to the non-cleared regions. This algorithm is given in Algorithm 2, which repeatedly calls $ClearTree(R, MAX, t)$ in iteration, where R contains the training instances not cleared by any previous iteration. The iterative process terminates when R remains unchanged in two consecutive iterations.

3.3 Discussion Next, we will discuss how to clear future instances as well as the method's efficiency/novelty.

Clearing future instances: The output of RiskClear is a sequence $CR_1, CR_2, \cdots, CR_k$, where CR_i denotes the set of cleared regions produced by the ith call of $ClearTree()$ in $RiskClear(R, MAX, t)$. To determine whether a future instance I with known independent attribute values but an unknown class value can be cleared, we examine CR_i in the order of i and find the first CR_i that has a cleared region matching the values of independent variables of I. If this CR_i is found, I is declared as cleared. If no such CR_i is found then I is declared as non-cleared.

Efficiency: $ClearTree$ has a similar complexity to Regression Trees since it also heuristically splits a parent region into several child regions by considering a similar search space. This method thus has a well-documented efficiency [4]. $RiskClear$ is therefore also efficient in practice since there are only a few $ClearTree$ calls due to the stopping condition used by $ClearTree$ or, if necessary, a maximum number of calls/iterations can be specified.

Novelty: While the top-down search strategy of ClearTree is similar to existing tree-based methods such as CART [4], we argue that our method is innovative: as explained in Sections 1 and 2, Risk Clearance differs from existing work due to its clearance criterion and clearance maximization goal, implying existing solutions cannot be adopted to exactly address our objective. Specifically, the clearance objective must consider (1) the exact numeric Y values, instead of the binary classes of $Y > MAX$ and $Y \leq MAX$, in the estimation of the risk probability $\Pr[Y > MAX|R_i]$, and (2) the clearance criterion specified by MAX and

t that imposes a minimum precision requirement. As we shall see in the next section, estimating $\Pr[Y > MAX|R_i]$ requires careful consideration because existing approaches, such as assuming Y is from a normal distribution, could too easily clear many instances then violate the minimum precision requirement.

4 Estimating $\Pr[\mathbf{Y > MAX|R_i}]$

In this section, we consider the remaining question of how to estimate the risk probability $\Pr[Y > MAX|R_i]$ using the training instances in the region R_i. We consider the following three properties to be essential in this estimation:

- *Numeric-centric:* The exact numeric Y values of training instances must be explored to better characterize the chance to exceed MAX.
- *Statistical:* The estimated probability $\Pr[Y > MAX|R_i]$ should hold for the whole population from which the training instances are drawn, not just on the training instances, because our goal is to clear future i.i.d. samples from the population.
- *No assumption on the Y value distribution:* Since typically little is known about the distribution of Y values, an estimation method that makes as few assumptions on the distribution of Y values as possible is preferred.

The numeric-centric requirement excludes all methods that treat Risk Clearance as binary classification (e.g. CUT Classification [5]). The statistical requirement implies that the training set should be treated as a sample of R_i's population and the sample size should be considered in the estimation. The "no assumptions on Y distribution" is necessary since an assumption on a particular distribution (e.g., the Y values are normally distributed [10]) can lead to unacceptable clearance decisions. For example, suppose that the top figure in Figure 3 is the *actual* PCB concentrations of transformers, and the bottom figure shows the *assumed* normal distribution with the same sample mean and variance. With the assumed normal distribution, the estimated $\Pr[Y > MAX|R_i]$ is roughly 19%, which is vastly below the observed 43% on the actual distribution. Therefore, if t is between 19% and 43% then the normality assumption will incorrectly clear all these transformers.

One solution may be to apply a statistical normality test (e.g. Shapiro-Wilk) then only assume the normal distribution if normality is not rejected. However, this approach is problematic: [11] pointed out that, for a small set of samples, the Shapiro-Wilk test can easily report a non-normal distribution as normal. In addition, this approach does not offer a solution if normality is rejected and running normality tests for every candidate partitioning is computationally expensive because of the large number of candidates. Similar discussions apply to assuming other distribution forms.

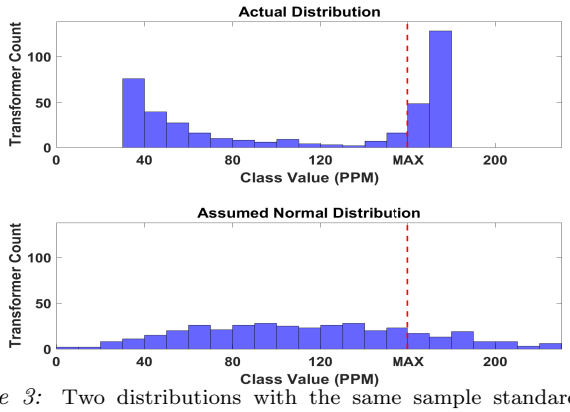

Figure 3: Two distributions with the same sample standard deviation (60) and sample mean (110).

There are many other works dedicated to estimating probabilities, such as Kernel Density Estimation, a non-parametric method to estimate the probability density function of a random variable [12]. However, we found these approaches not suitable for our problem because their estimates are built blind to the clearance criterion parameters, i.e., MAX and t, and the fact that the hazard/cost of incorrect clearance is usually very high. For example, running Matlab's default Kernel Density function on the right distribution in Figure 2 returns an estimated probability to exceed MAX as virtually zero (because every observed instance is under MAX) even though the distribution's dispersion and low sample size intuitively imply that there is a far higher risk that future instances will exceed MAX.

In the rest of this section, we develop two new estimates for $\Pr[Y > MAX|R_i]$ that fulfill our listed requirements. Table 1 summarizes our notation. The first one, called the Cantelli Estimate, makes no assumption on the distribution of Y values, and the second one, called the NormCan Estimate, benefits from the normality assumption opportunistically.

4.1 The Cantelli Estimate

Our first estimate is based on Cantelli's Concentration Inequality [13], that uses Markov's Inequality to state that, if $\lambda > 0$,

$$\Pr[Z - \mu \geq \lambda] \leq \frac{\sigma^2}{\sigma^2 + \lambda^2} \tag{4.1}$$

where Z is a real-valued random variable, μ is the expected value of Z, and σ^2 is the variance of Z. This inequality makes no assumption about the underlying distribution, thus, can be applied for all distributions. With the numeric class Y being the random variable Z, sampled from the underlying distribution $D_Y(R_i)$ with

a mean μ and variance σ^2, and λ being $MAX - \mu$, we derive

$$\Pr[Y > MAX|R_i] \leq \frac{\sigma^2}{\sigma^2 + (MAX - \mu)^2} \quad (4.2)$$

assuming that MAX is greater than μ. If $\mu \geq MAX$, we set $\Pr[Y > MAX|R_i]$ to 1 so that no instance will be cleared. If the true mean μ and true variance σ^2 are not available, applying the unbiased estimator approach [10] we can substitute them with the sample mean $\hat{\mu}$ and the sample variance $\hat{\sigma}^2$:

$$\Pr[Y > MAX|R_i] \leq \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + (MAX - \hat{\mu})^2} \quad (4.3)$$

However, this approach ignores the impact of the sample size $|R_i|$. To fix this issue, we adjust the estimate by applying the Central Limit Theorem to take sample size into account. For a given confidence level CF, this theorem says that the true μ is no more than μ_{ub} in $CF\%$ of cases, where μ_{ub} is given by

$$\mu_{ub} = \hat{\mu} + z\frac{\hat{\sigma}}{\sqrt{|R_i|}} \quad (4.4)$$

z is the critical value for the one-sided t-distribution with $|R_i| - 1$ degrees of freedom[1] [10]. For example, with $CF = 99.5\%$, $z \approx 2.6259$ if $|R_i| - 1 = 100$, and $z \approx 3.1692$ if $|R_i| - 1 = 10$. Our first estimate, called the Cantelli estimate, replaces μ with the upper bound $\mu \leq \mu_{ub}$: if $\mu_{ub} \leq MAX$,

$$\Pr_{Cantelli}[Y > MAX|R_i] = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + (MAX - \mu_{ub})^2} \quad (4.5)$$

If $\mu_{ub} > MAX$, we set $\Pr_{Cantelli}[Y > MAX|R_i]$ to 1.

With enough samples, the Cantelli estimate approaches the bound in Eqn 4.3, which approaches the bound in Eqn 4.2. These trends make a region R_i easier to clear if its class values have a lower dispersion (i.e. a small sample variance $\hat{\sigma}^2$), are further below MAX (i.e. a small sample mean $\hat{\mu}$), and its sample size is large (i.e. that μ_{ub} is close to $\hat{\mu}$).

In Section 2, we commented that CUT Classification [5] cannot distinguish the two distributions of the binary class for >50 and ≤50 in Figure 2, which can miss opportunities to clear more instances. Indeed, for the distribution of >50 and ≤50 in Figure 2, the Wilson interval employed by CUT Classification gives an estimated 15.5% probability of Y exceeding 50. In contrast, applying Eqn 4.5 to the actual numeric class distribution on the left of Figure 2 with $CF = 99.5\%$ achieves a nearly zero estimated probability to exceed 50 because of the very low average and dispersion. Thus, the Cantelli estimate can clear this region even for a low threshold t close to zero, but CUT Classification cannot for any threshold t less than 15.5%.

[1] We preferred the t-distribution's critical value over the normal distribution's because it is larger and therefore more conservative with fewer samples.

Table 1: Notation

Notation	Definition
Y	The random variable for class values
t and MAX	The threshold and numeric cutoff
R_i	A region of attribute space
$D_Y(R_i)$	R_i's true class value distribution
$\{R_1, \cdots, R_m\}$	A set/partition of disjoint regions
$\|R_i\|$	The count of training instances in R_i
$\Pr[Y > MAX\|R_i]$	The probability that a class value sampled from $D_Y(R_i)$ exceeds MAX
$\hat{p}(R_i)$	The observed proportion of $>MAX$ training instances in R_i
$\Phi(MAX\|\mu, \sigma^2)$	$\Pr[Y \leq MAX\|R_i]$ when $D_Y(R_i)$ is normal with mean/variance=μ/σ^2
$\hat{\mu}, \hat{\sigma}^2$	$D_Y(R_i)$'s sample mean and variance
μ_{ub}	μ's confidence interval upper bound.

4.2 The NormCan Estimate The Cantelli estimate does not benefit from a normal distribution that may be present in a data set. In this case, the Cantelli approach could over-estimate the probability $\Pr[Y > MAX|R_i]$, and thus fail to clear R_i that in fact should be cleared. Instead of never considering the normal distribution assumption, our second estimate restricts when the normality assumption is appropriate based on the observed proportion of training instances in R_i with a class value $>MAX$.

Let $\Phi(MAX|\mu, \sigma^2)$ denote the probability that the random class variable Y does not exceed MAX, assuming that Y follows the normal distribution with the mean μ and the variance σ^2. Replacing the mean/variance with the estimated $\hat{\mu}$ and $\hat{\sigma}^2$, we have

$$\Pr_{Norm}[Y > MAX|R_i] = 1 - \Phi(MAX|\hat{\mu}, \hat{\sigma}^2) \quad (4.6)$$

Adjusting this estimate to consider sample size, we use the upper bound μ_{ub} in Eqn 4.4 instead of the sample mean to derive a more conservative estimate:

$$\Pr_{AdjustedNorm} = 1 - \Phi(MAX|\mu_{ub}, \hat{\sigma}^2) \quad (4.7)$$

However, the estimate given by Eqn 4.7 may still be too optimistic (i.e., under-estimate the probability) because of the normality assumption. To fix this, we propose using the pessimistic Cantelli estimate (Eqn 4.5) if $\hat{p}(R_i) \geq t$, where $\hat{p}(R_i)$ denotes the observed proportion of training instances in R_i with a class value $>MAX$; otherwise, we use the adjusted normal estimate (Eqn 4.7):

$$\Pr_{NormCan} = \begin{cases} \Pr_{Cantelli} & \text{if } \hat{p}(R_i) \geq t \\ \Pr_{AdjustedNorm} & \text{otherwise} \end{cases}$$

In other words, the NormCan estimate opportunistically benefits from the normality assumption: when the observed data has a too high proportion of $Y > MAX$ instances it takes more precaution by adopting the pessimistic Cantelli estimate, and adopts the optimistic AdjustedNorm estimate otherwise. For example, with $t = 30\%$, $\hat{p}(R_i) \geq t$ for Figure 3's top distribution, the pessimistic Cantelli estimate is used, giving the estimated probability 63%, which correctly prevents this distribution from being cleared. In contrast, for Figure 3's bottom distribution, $\hat{p}(R_i) < t$, which allows the optimistic $\Pr_{AdjustedNorm}$ to be applied, giving the estimated probability 26% for $CF = 99.5\%$; thus, this distribution is correctly cleared for $t = 30\%$.

5 Experiments

In this section we evaluate the effectiveness of RiskClear on solving Risk Clearance problems[2].

5.1 Methodology Given an observed data set T, the numeric class cutoff MAX, and the clearance threshold t, we partition T into a training set T_1 and a testing set T_2 using k-fold cross validation. For each fold, we build a model using T_1 and apply this model to clear the instances in T_2. The performance is measured by the recall and precision averaged over the k folds. A true positive (TP) is an instance in T_2 that has a class value $Y \leq MAX$ and is cleared by the model; a false positive (FP) is an instance in T_2 that has a class value $Y > MAX$ and is cleared by the model; a true negative (TN) is an instance in T_2 that has a class value $Y > MAX$ and is non-cleared by the model; a false negative (FN) is an instance in T_2 that has a class value $Y \leq MAX$ and is non-cleared by the model. The *recall* and *precision* are:

$$Recall = \frac{TP}{TP + FN} \qquad Precision = \frac{TP}{TP + FP}$$

Our clearance criterion $\Pr[Y > MAX|R_i] < t$ is equivalent to $\Pr[Y \leq MAX|R_i] \geq 1 - t$, that is, for a cleared instance, its probability of $Y \leq MAX$ is at least $1 - t$. This imposes the minimum precision requirement $Precision \geq 1 - t$ for $Y \leq MAX$ on all cleared instances. Note that this requirement *must* be satisfied, and only if this requirement is satisfied, we evaluate recall. This minimum precision requirement distinguishes our evaluation from other evaluators that measure the trade-off between precision and recall, such as the F-Measure and recall-precision curve evaluators.

[2] Additional experiments and code are available at: `https://www2.cs.sfu.ca/~wangk/software/RiskClear/`

5.2 Compared Methods We evaluate two variants of the **RiskClear** algorithm proposed in Section 3: **RiskClear(Cantelli)** refers to RiskClear with $\Pr_{Cantelli}$ as the probability estimate and **RiskClear(NormCan)** refers to RiskClear with $\Pr_{NormCan}$ as the probability estimate. In both solutions, the confidence level CF is set to 99.5% due to the importance of not clearing toxic transformers in our case study. We compare these methods with:

CUT Classification. This algorithm proposed in [5] aims to clear instances with binary class values. To apply this algorithm to our Risk Clearance problem with numeric classes, we first relabel all instances with $Y > MAX$ as *Negative* and all instances with $Y \leq MAX$ as *Positive*. We set CF to 99.5% and use the Pure Potential variant of CUT Classification because it was shown to outperform other variants. We do not consider other binary classification approaches (e.g., Decision Trees) because [5] shows that CUT Classification outperforms such approaches.

Regression. This baseline represents the regression approach that seeks a partitioning to minimize the sum of squared error rather than maximize clearance. We use Regression Trees from [4]. Our implementation uses Weka's Reduced Error Pruning variant with the default parameters (http://www.cs.waikato.ac.nz/ml/weka/). To label a leaf node as cleared or non-cleared, we consider two estimates for $\Pr[Y > MAX|R_i]$ at a leaf node R_i. **REGRESSION(Normal)** uses $\Pr_{Norm}$, and **REGRESSION(Cantelli)** uses $\Pr_{Cantelli}$. Since this method use non-binary categorical splits while the other methods use binary splits, for a fair comparison we convert each categorical attribute into multiple binary attributes.

We evaluate the below motivating claims:

- *Claim 1*: RiskClear meets the clearance criterion by satisfying $Precision > 1 - t$.
- *Claim 2*: RiskClear achieves higher recall than the regression baseline.
- *Claim 3*: RiskClear achieves a higher recall than CUT Classification by considering the actual distribution of numeric class values.
- *Claim 4*: The normal distribution assumption, i.e., $\Pr_{Norm}$, can too easily clear instances, thus, fails to satisfy $Precision > 1 - t$.

5.3 Data Sets We report the results on the BC Hydro data set, provided by our partner BC Hydro, that is composed of records for roughly one thousand observed oil-filled transformer parts. Each part is described by 10 independent variables and one numeric class Y representing the PCB toxicity in parts per

million (ppm), described in Table 2. We use two settings of MAX based on the industrial standards of [1], $MAX = 400$ and $MAX = 50$. For each setting, $\hat{p}$ denotes the percentage of instances in the data set that have the class value $Y > MAX$. For $MAX = 400$, $\hat{p}_1 < 5\%$ and for $MAX = 50$, $\hat{p}_2 < 25\%$. Due to confidentiality, we cannot divulge the actual toxicities or the exact $\hat{p}$ values.

Table 2: Transformer Part Attributes

Attribute Type	Relevant Attributes
Toxicity Class	Parts per million of PCBs (ppm)
Numeric	Current, Oil Volume, Voltage.
Categorical	Area, Subarea, Manufacturer, Type, Model, Substation, Bushing Position

As discussed in Section 5.1, $1 - t$ corresponds to the minimum precision required by the Risk Clearance problem, where t is the clearance threshold. Typically, this minimum precision $1 - t$ is much higher than the proportion of instances with $Y \leq MAX$ in the training set, i.e., $1 - \hat{p}$, since clearing an instance having $Y > MAX$ has a much worse consequence than not clearing an instance that should be cleared. Thus, t should be significantly smaller than $\hat{p}$. To ensure this, we test 50 settings for t: for $i = 1, \cdots, 50$, $t = i \cdot \frac{\hat{p}/2}{50}$. In particular, for $MAX = 400$, $\hat{p}_1 < 5\%$, so the minimum precision $1 - t$ for a t chosen above is higher than 97.5%, and for $MAX = 50$, $\hat{p}_2 < 25\%$, the minimum precision $1 - t$ for a t chosen above is higher than 87.5%.

We use 3-fold cross validation to ensure a sufficient number of $Y > MAX$ cases in the testing set and use stratified sampling to ensure that each fold's testing set gets roughly the same count of $Y > MAX$ cases. The reported precision and recall are the mean over the folds.

5.4 MAX=400 ppm Results Figure 4 shows the precision and recall for $MAX = 400$. This minimum precision $1 - t$ is represented by the dashed line for various settings of t. For every t examined, the two RiskClear variants (i.e., dark blue and dark red) meet the minimum precision requirement by staying above the dashed line, supporting *Claim 1*.

Claim 2 is also supported: RiskClear(Cantelli), in dark blue, has a far higher recall than REGRESSION(Cantelli), the light blue line, and RiskClear(NormCan), the dark red line, has a higher recall than REGRESSION(Normal), the light red line. Also, RiskClear(NormCan) has a higher recall than RiskClear(Cantelli) due to adopting the adjusted normal distribution assumption when appropriate.

Claim 3 is supported by CUT Classification clearing zero training instances, as shown with the 0% recall. Recall that CUT Classification converts the numeric Y values into a binary class $>MAX$ and $\leq MAX$, and

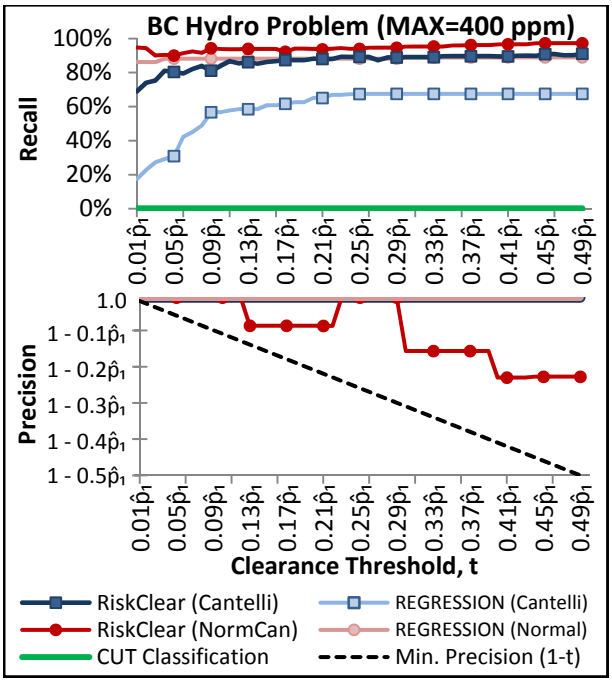

Figure 4: For $MAX = 400$ ppm, RiskClear(NormCan) (the dark red line) is the best performer in terms of our clearance maximization objective: it has the highest recall (the upper figure) while attaining a precision above the minimum $1 - t$ indicated by the dashed line. CUT Classification does not clear any instance so it has 0% recall with an undefined precision. The two REGRESSION methods and RiskClear(Cantelli) attain 100% precision, but their recall is not as high as RiskClear(NormCan)'s. Every method's standard error is no more than 3.5% for recall and no more than 0.3% for precision.

estimates the true proportion of the $>MAX$ class by a confidence interval that tends to be large for low sample sizes. This makes it difficult for the method's estimate, this interval's upper limit, to be under the low threshold t (particularly for $MAX = 400$ where t is a fraction of $\hat{p}_1 < 5\%$). Consequently, CUT Classification could not even clear a single instance even if a region contains only uncontaminated cases for most tested ts. In contrast, RiskClear clears more instances by finding many regions similar to the left distribution in Figure 2 where the observed toxicity values are concentrated far away from MAX. This study clearly suggests that replacing the numeric class with a binary class can be detrimental.

Claim 4: To evaluate this claim, we run RiskClear with $\Pr_{Norm}$; this optimistic estimate under the normal distribution assumption allows all instances to be cleared, giving 100% recall and the precision of $1 - \hat{p}_1$, far below the required minimum precision $1 - t$ since t is a fraction of $\hat{p}_1$. This implies that the normal distribution assumption can too easily clear instances. In contrast, RiskClear(NormCan) uses the more conservative Cantelli estimate in general and adopts the normal distribution assumption only under certain conditions,

which ensures a precision over the required $1-t$. A similar case occurs for $MAX = 50$, implying *Claim 4*.

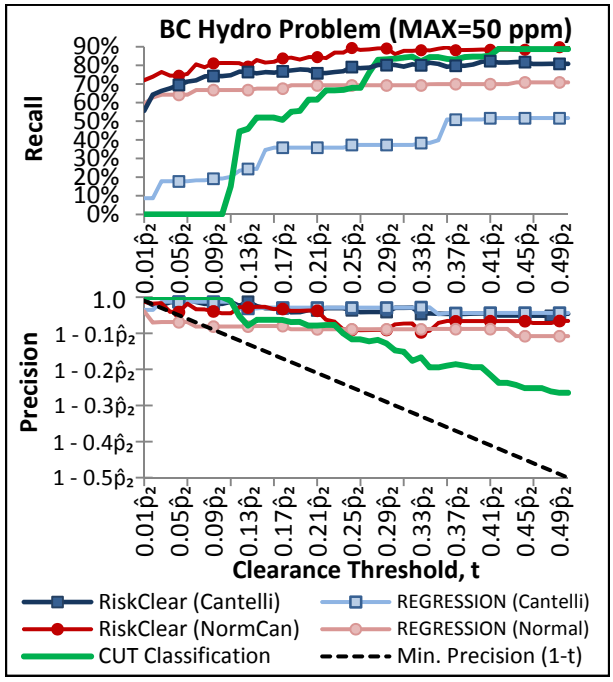

Figure 5: RiskClear(NormCan) is the best model for $MAX = 50$: it has the highest recall while meeting the precision requirement. For all algorithms the standard error is no more than 5.1% for recall and no more than 2.4% for precision.

5.5 MAX=50 ppm Results Figure 5 shows the results for $MAX = 50$ with $\hat{p}_2 < 25\%$. While the general performance is similar to the $MAX = 400$ case, we highlight some key points. First, the RiskClear algorithms stay above the dashed line, thus, supporting *Claim 1*. *Claim 2* is also supported: RiskClear(Cantelli) has a far higher recall than REGRESSION(Cantelli) and RiskClear(NormCan) has a higher recall than REGRESSION(Normal) by roughly 25%. We found that REGRESSION's goal of minimizing the sum of squared errors tends to produce mostly small and non-cleared regions (e.g. splitting by "Model"), while RiskClear instead prioritizes partitions that maximize clearance (e.g. a large and cleared region based on manufacturers). Also, CUT Classification performs badly for a low t. In practice, t is small because a high precision is needed to avoid costly false positives, validating *Claim 3*.

5.6 Recommendation RiskClear(NormCan) is the overall winner: it fulfills the minimum precision requirement while clearing more than competitors by leveraging exact numeric values, the strength of the normal distribution assumption in generating high recall, and the Cantelli estimate's prevention of invalid clearance.

6 Conclusion

This work presented a new supervised learning approach, Risk Clearance, to clear future low risk instances that have a less than t probability for a numeric risk indicator Y (e.g., a toxicity level) to exceed a cutoff value MAX. This problem applies to a wide range of domains where risk prioritization is desired. The novelty of Risk Clearance is that, unlike cost-sensitive learning in the literature, it does not require the user to specify detailed cost metric information; instead, it leverages sample size and the distribution of the numeric Y values to estimate the risk and clear low risk instances.

Acknowledgments. We thank BC Hydro for sponsorship and the Natural Sciences and Engineering Research Council of Canada for a Graduate Scholarship and a Collaborative Research Development Grant (NSERC Project CRDPJ/445210).

References

[1] United Nations Environment Programme Chemicals. "PCB Transformers and Capacitors from Management to Reclassification and Disposal." 2002.

[2] D. Hubbard, *The Failure of Risk Management: Why It's Broken and How to Fix It.* John Wiley & Sons., 2009.

[3] K. Daigle, "Arsenic: A Growing Plague in the World's Drinking Water," *Scientific American*, Jan. 2016.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Statistics/Probability Series. Belmont, California, U.S.A.: Wadsworth Publishing Company, 1984.

[5] R. McBride, K. Wang, and W. Li, "Classification by CUT: Clearance Under Threshold," in ICDM 2014, Shenzhen, China, pp. 410–419, 2014.

[6] C. Elkan, "The Foundations of Cost-Sensitive Learning," in *IJCAI*, pp. 973–978, 2001.

[7] J. Hernández-Orallo, "Probabilistic Reframing for Cost-Sensitive Regression," *ACM Trans. Knowl. Discov. Data*, vol. 8, no. 4, pp. 17:1–17:55, Aug. 2014.

[8] O. Maimon and L. Rokach, "Data Mining for Imbalanced Datasets: An Overview," in *Data Mining and Knowledge Discovery Handbook*, 2005.

[9] M. Jaskowski and S. Jaroszewicz, "Uplift Modeling for Clinical Trial Data," in *ICML Workshop on Clinical Data Analysis*, 2012.

[10] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference.* Springer, 2010.

[11] J. Rochon, M. Gondan, and M. Kieser, "To Test or not to Test: Preliminary Assessment of Normality when Comparing Two Independent Samples," *BMC Medical Research Methodology*, vol. 12, no. 1, 2012.

[12] B. W. Silverman, *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall, 1986.

[13] R. Savage, "Probability Inequalities of the Tchebycheff Type." *Journal of Research of the National Bureau of Standards - B Mathematics and Mathematical Physics*, vol. 65B, 1961.