Estimating False Negatives for Classification Problems with Cluster Structure

György J. Simon Vipin Kumar Zhi-Li Zhang
Department of Computer Science, University of Minnesota
{gsimon,kumar,zhzhang}@cs.umn.edu

Abstract

Estimating the number of false negatives for a classifier when the true outcome of the classification is ascertained only for a limited number of instances is an important problem, with a wide range of applications from epidemiology to computer/network security. The frequently applied method is random sampling. However, when the target (positive) class of the classification is rare, which is often the case with network intrusions and diseases, this simple method results in excessive sampling. In this paper, we propose an approach that exploits the cluster structure of the data to significantly reduce the amount of sampling needed while guaranteeing an estimation accuracy set forth by the user. The basic idea is to cluster the data and divide the clusters into a set of "strata", such that the proportion of positive instances in the stratum is very low, very high or in between. By taking advantage of the different characteristics of the strata, more efficient estimation strategies can be applied, thereby significantly reducing the amount of required sampling. We also develop a computationally efficient clustering algorithm – referred to as class-focused partitioning – which uses the (imperfect) labels predicted by the classifier as additional guidance. We evaluated our method on the KDDCup network intrusion data set. Our method achieved better precision and accuracy with a 5% sample than the best trial of simple random sampling with 40% samples.

Keywords: false negatives estimation, random sampling, clustering

1 Introduction

Determining the performance of a classifier is an important and well-studied problem. While a wealth of metrics (such as precision, recall, F-measure) and techniques (e.g. bootstrapping, cross-validation, ROC curves) have been developed [7], most of them require that the true class labels are known. In many classification problems, however, establishing the true class label of all instances is not feasible. In epidemiology, very reliable classifiers, "gold standard" tests, exist, but they are so intrusive that their application to the generic public without "sufficient indication" of the presence of the disease is deemed unethical and expensive. In case of network intrusion detection, a security expert (say, of an enterprise network) serves as a reliable classifier, who can distinguish malicious traffic behavior (*positive* class) from normal traffic behavior (*negative* class) at high confidence. However, due to the high traffic volume, it is unrealistic to expect a security expert to inspect every single traffic flow in his/her network.

Instead of the "gold standard" test, a less expensive but imperfect classifier, such as an epidemiological screening test or a network intrusion detection system (IDS), is often employed. Positive outcome of this imperfect classifier is considered "sufficient indication" and the true status of these predicted positive instances are investigated.

The goal of this paper is to develop a method for evaluating the performance of these imperfect classifiers. Part of the performance is already known: the true class label for all predicted positive instances is already determined. To obtain a full evaluation, however, *one must also determine the number of false negative instances.*

A frequently used method is to draw a sufficiently large sample of the instances classified as negative, determine the true class label by applying the gold standard test and compute the proportion of the positive instances in it [2]. Since in many applications, *the positive class is typically rare* (e.g., network intrusions and 'diseased' outcomes in a medical screening test), and a (hopefully) large portion of them has been identified by the classifiers, only very few positive instances are left among the (negative) majority of the instances with the true class label not yet determined. Under these conditions, the sample size required to accurately estimate the number of false negatives grows excessively large. We will show later that detecting a certain class of common intrusions within a modest 20% range of error at least at 95% probability in the KDD Cup data set which contains 480,000 instances, would require a sample size of over 100,000 instances.

In this paper, we present a novel approach based on stratified sampling that allows for estimating the number of false negative instances more efficiently, while guaranteeing that the error stays within bounds specified by the user. The basic idea is to exploit the inherent

cluster structure of the classification problem and partition the data into three distinct sets ("strata") of clusters: (1) *Negative stratum* that that is hypothesized to contain clusters with a proportion of false negative (FN) instance close to (or equal to) 0, (2) *Positive stratum* with clusters that are hypothesized to have a FN proportion of close to (or equal to) 1, and (3) *Mixed stratum*, containing the remaining clusters. These strata have different statistical properties which can be exploited to develop a more efficient sampling scheme. The efficiency of our approach stems from two factors. First, in case of simple random sampling (SRS), purer clusters have smaller variability in the proportion of FN instances across possible samples, which translates to smaller sample sizes. Second, when the Positive and Negative clusters are indeed close to pure, a more powerful (in the statistical sense) scheme – the *geometric model* – can be employed to further reduce the sample size.

We evaluate our proposed method on a real-world data set, the KDD Cup intrusion detection dataset. Our proposed method attains a 4-fold improvement in precision and accuracy over SRS with a sample size of 200,000 instances. Our method required a sample of only 25,000 instances.

2 Related Work

In the following, we briefly discuss some of the related work. A more detailed account can be found in the technical report version of this paper [6].

Methods of estimation. In epidemiology, the evaluation of new screening tests is a very important problem [4]. The goal here is to evaluate the performance of an imperfect screening test *without* sampling, based on the predictions of another imperfect classifier. The earliest methods are based on the capture-recapture model [2]. The capture-recapture model in its original form makes the assumption that the classifiers to be evaluated are independent. Although new techniques have been developed (such as latent class analysis and logistic and loglinear modeling, it has been proven that the problem is underdefined [8], and *hence not solvable without incorporating external information.* Mane at al. showed that even moderate deviations from the independence assumption can invalidate the estimate [3]. This recognition led to alternative approaches in the epidemiologic research that utilize multiple tests [8] or two- and multiphase studies.In contrast, our proposed method takes a different approach. The external information we incorporate is the cluster structure of the problem and some limited sampling.

Clustering. Arguably one of the most popular clustering methods is k-means [7]. K-means partitions the set of objects into k disjoint subsets under the constraint that the sum squared error is to be minimized. Semi-supervised clustering [1], aimed at modeling the cluster structure of the problem more accurately, allows the user to specify hints regarding whether two specific points should belong to the same cluster or not. Our problem differs from clustering in that we do not need to discover the *true* cluster structure of the problem. We can break clusters into *partitions* as long as the partitions are pure. While density-based clustering methods with similar goals exist, their computational complexity is prohibitively high for our application.

Semi-Supervised Classification (SSC). Semi-supervised classification [5, 9] is applicable when a large number of unlabeled instances and only a few labeled instances exist. However, a key difference is that while semi-supervised classification aims to *label* the entire data set, we only need to estimate the number of FN instances; we do not need to know *which* instances are FN instances.

3 Problem Formulation

We consider a classification problem, where the instances of a population $\mathcal{D}$ are classified as having either *positive* (+) or *negative* (−) class label. Our goal is to evaluate the performance of a classifier T, when the true labels are not known for all instances. An instance is *predicted positive*, if T predicts it to have positive label; otherwise it is *predicted negative*.

We have an *infallible* expert determine the *true* label of the predicted positive instances. A predicted positive instance is *true positive* (TP) if its true label is positive, otherwise it is *false positive* (FP). Analogously, a predicted negative instance is a *false negative*(FN) (resp., *true negative*) if its true label is positive (resp., negative).

Goal. Our goal is to derive an estimate $\hat{t}$ for the number of FN instances given the predicted labels and the true labels for the predicted positive instances such that

$$\Pr\Big[|\frac{\hat{t}-t}{\hat{t}}| < \varepsilon\Big] \geq 1-\alpha, \tag{3.1}$$

where ε is the margin of error and α is the significance level. Both ε and α are specified by the user. In the experiments, we will use $\alpha = 5\%$ and $\varepsilon = 20\%$.

One way to estimate $\hat{t}$ is to use simple random sampling[1](SRS). When the positive class is rare (i.e., the number of instances with + true labels is very small relative to the total size of the dataset), the required sample size becomes excessively large. In this paper,

[1]By "sampling" we mean drawing a sample and having the expert determine the true class label of the instances in the sample.

we propose an innovative approach that exploits the cluster structure underlying the classification problem to significantly reduce the number of samples that need to be drawn to attain the user-specified accuracy bound.

4 Proposed Approach

First, the classifier is applied to the data. Next, the data is clustered into partitions using our *class-focused partitioning* technique. It aims to cluster the data into *partitions* that are likely *pure*: either a vast majority of the predicted negative instances in the group are positive or negative. Purity, however, cannot be directly observed, so we heuristically approximate it. We consider a partition pure if it is tight and observed-pure. Formal definitions for these properties follow.

PROPERTY 1. (TIGHTNESS) *We call a partition **tight**, if for its mean squared error (MSE), $MSE < minMSE$ holds, where $minMSE$ is a user-defined threshold.*

MSE measures the average dissimilarity of the partition by comparing each instance to the partition mean. MSE does not penalize large partitions as long as their instances are sufficiently similar to each other.

PROPERTY 2. (PURITY) *We call a partition **pure**, if most (or preferably all) instances share the same mode of behavior.*

From the information we have available, we can not measure purity. Instead, *Observed purity* is measured using the imperfect labels. A partition is **observed-pure** if for any of the classifiers $TPR > 0$ and $FPR = 0$; or if $TPR = 0$ and $FPR \geq 0$ [2]. Observed purity does not imply purity (i.e. non-observed purity), however they are correlated.

The Class-Focused Partitioning Algorithm. The algorithm extends bisecting k-means [7]. Initially, it considers the entire data set D as one cluster. Then it iteratively bisects it into two smaller partitions until a stopping conditions is met. Bisection stops when
(1) The class label of all instances is known or
(2) The partition is observed pure or
(3) Further bisection results in no improvement in terms of either purity or tightness.

4.1 Outline of the Estimation.

Once the partitions are created, we form strata on which our stratified-sampling-based estimation is carried out. We shall show that estimation using these strata is correct (i.e. the estimate satisfies the error bound defined in Equation 3.1) and more efficient than just applying simple random sampling (SRS).

We create 3 strata: (i) Positive, which is relatively rich in FN instances, (ii) Negative, which contains relatively few FN instances and (iii) Mixed, containing the rest of the partitions. The Negative stratum is assumed to contain no false negative instances, while the Positive stratum is assumed to contain no true negative instances.

Then the estimation proceeds as follows. Let $\hat{t}$ denote the number of FN instances in a stratum. Since the Negative stratum is assumed to contain no false negatives, we estimate $\hat{t} = 0$. The purpose of sampling here is to prove that there are not too many[3] false negative instances in the stratum w.r.t. the predefined error bounds. We employ the geometric model (described later in Section 4.1.1) to this end. If we happen to find too many false negatives during the sampling, then we revert back to SRS in this stratum. With the Positive stratum, we proceed similarly. This stratum is assumed to contain only positive instances, hence $\hat{t} = u$, where u denotes the number of instances in this stratum with unknown true label. Here we employ the geometric model to prove that indeed there are not too many true negatives in this stratum. If too many true negatives are found, the number of FN instances in this stratum is estimated via SRS. Finally, for the Mixed stratum, we simply employ SRS.

Our approach is correct in the sense that the estimate will be within the error bounds. This is guaranteed for SRS [2]. We only deviate from SRS, when the stratum is observed-pure. Should we discover that it is not sufficiently pure, we revert back to SRS. The worst-case scenario is when the data set is uninformative, i.e. it does not contain relevant clusters. Some partitions could appear (by chance) pure negative, others pure positive. During the application of the geometric model, we will discover (with $1 - \alpha$ probability) that these partitions are not pure and revert to SRS.

The effectiveness of our approach stems from two factors. First, when the partitions are observed pure, the geometric model is applicable[4]. Second, even when the partitions are found to be insufficiently pure, the sample size required to obtain estimates using SRS decreases with increasing purity within a fixed error bound. This is demonstrated in the following examples.

[2] The lack of true and false positive instances can be indicative of the *lack* of FN instances

[3] i.e. the error is within the pre-define error at $1-\alpha$ probability. In the rest of the paper, we will use "too many" in this sense.

[4] SRS is not even applicable when the proportion is 0 or 1. The variance of the estimate becomes 0 and alongside, the sample size becomes 0, too.

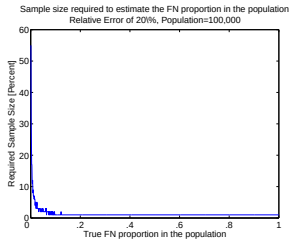

Figure 1: The minimal sample size for estimating within relative error bounds.

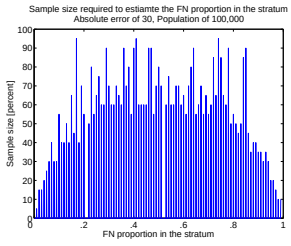

Figure 2: The minimal sample size for estimating within absolute error bounds.

Example 1 – FN estimation in the population. Let us consider a population of 100,000 instances with unknown labels. The goal is to estimate the π^* proportion of FN instances within 20% error at 95% probability. Figure 1 depicts the required sample size as function of π^*. The results are obtained from simulation: for each $\pi^* \in \{1/1000, \ldots, 999/1000\}$, we tested sample sizes of $s \in$\{1%,2%,..., 99% of the population\}. For each π^* and for each s, we ran 100 trials and for each π^* we determined the smallest s such that at least 95 of 100 trials resulted in an estimate $\hat{\pi}^*$ of π^* within 20% error. Figure 1 depicts this minimal s for each π^*.

Figure 1 shows that the sample size required to perform estimation increases as the proportion of FN instances decreases.

□

Example 2 – FN estimation in a stratum. The proportion π^* of FN instances in the population is now fixed. Let $\hat{\pi}^*$ be a fixed estimate of π^*. Let the error ϵ be 20% of $\hat{\pi}^*$; ϵ is also fixed.

Next, we divide the FN population into 3 strata. We distribute the FN instances among these 3 strata arbitrarily. We will concentrate on one of the 3 strata, which we denote by $\mathcal{N}$. In this example, we will control the proportion $\pi_{\mathcal{N}}$ of FN instances in $\mathcal{N}$: $\pi_{\mathcal{N}} = 1/1000, 2/1000, \ldots, 999/1000$. Let us remind the Reader, that the total number of FN instances is fixed, we just change their distribution across the 3 strata.

We also need to distribute the error ϵ among the 3 strata. Arbitrarily, let us assign $\epsilon_{\mathcal{N}} = .3\epsilon$. Let us remind the Reader, that $\epsilon_{\mathcal{N}}$ does not depend on $\hat{\pi}_{\mathcal{N}}$, it only depends on the fixed $\hat{\pi}^*$. Then Figure 2 depicts the sample size required to estimate $\pi_{\mathcal{N}}$ within $\epsilon_{\mathcal{N}}$ error at 95% probability. These sample sizes were determined by simulation identically to Example 1.

Figure 2 shows that the sample size *decreases* with increasing purity (i.e. the increasing $|.5 - \pi_{\mathcal{N}}|$) of the partition. This observation is contrary to the observation in Figure 1. The reason lies in the use of *relative error* versus *absolute error*. In Example 1, ϵ was a *relative* error, it was a function of π^*, the quantity we were estimating. In Example 2, $\epsilon_{\mathcal{N}}$ was an *absolute* error, it did not depend on $\pi^*_{\mathcal{N}}$, the quantity we tried to estimate.

□

Example 2 demonstrates that partitioning the data with the goal of creating purer strata has utility. By discovering pure partitions, we can generate pure strata, which leads to estimation with small samples. On the other hand, Example 2 assumes that an estimate, $\hat{\pi}^*$ exists. Although initial estimates can be obtained, we have to make provisions that these estimates may be incorrect. Incorrect estimates may lead to looser error bounds. With this point in mind, the estimation process will work as follows.

Estimation process.

1. After the discovery of the partitions, the three strata are formed.
2. An initial estimate of the FN proportion $\hat{\pi}^{(0)}$ is computed from the number of True Positive instances and a small sample drawn from the Mixed strata. (This estimate can be viewed as $\hat{\pi}^*$ in Example 2.)
3. A *sampling plan* is created. Given the total allowed error, the sampling plan determines how much of the total allowed error is allocated to each stratum such that the total required sample size is minimal.
4. The sampling plan is executed. The outcome is an updated (more precise) FN proportion estimate ($\hat{\pi}^{(1)}$) and its achieved error. If the achieved error is less than maximal allowed error, then the estimation is complete and the final estimate is $\hat{\pi}^{(1)}$. Otherwise, the estimate needs to be refined. A new total allowed error is computed from $\hat{\pi}^{(1)}$ and the process repeats from Step 3.

Once the proportion $\hat{\pi}$ of the FN instances is obtained, computing the number $\hat{t}$ of FN instances is trivial.

4.1.1 Estimating π in the three strata.

In the Mixed stratum $\mathcal{M}$, the proportion $\hat{\pi}_{\mathcal{M}}$ of FN instances is estimated using SRS. SRS is a standard technique, the Reader is referred to [6] for details.

As discussed before, in the Negative $\mathcal{N}$ and the Positive $\mathcal{P}$ strata, we apply the geometric model. For simplicity let us consider the Negative stratum $\mathcal{N}$; estimation in $\mathcal{P}$ works analogously. In $\mathcal{N}$, all instances are assumed negative. The goal here is to probabilistically prove, that there are no false negative instances in the stratum. More precisely, we shall show that if we draw a sample of size z (as determined by Equation 4.3), and

we find 0 FN instances in that sample, then the estimate $\hat{t}_{\mathcal{N}} = 0$ is correct within an error bound of $\epsilon_{\mathcal{N}}$ at α probability.

N can contain 2 types of instances with unknown labels: (i) predicted negatives that are true negatives and (ii) false negatives that are actually positives. Let us assume that there are x false negatives instances and $N_{\mathcal{N}} - x$ true negative instances. We need to show, that if $x > N_{\mathcal{N}}\epsilon_{\mathcal{N}}$, then by drawing a sample of size z, there will be 0 false negative instances in the sample at most at probability of α.

$$\frac{\binom{N_{\mathcal{N}}-x}{z}}{\binom{N_{\mathcal{N}}}{z}} < \frac{\binom{N_{\mathcal{N}}-N_{\mathcal{N}}\epsilon_{\mathcal{N}}}{z}}{\binom{N_{\mathcal{N}}}{z}} < (1-\epsilon_{\mathcal{N}})^z \quad < \quad \alpha \tag{4.2}$$

Accordingly, in order to probabilistically show that there are not too many FN instances in $\mathcal{N}$, we need to draw a sample of

$$z \geq \frac{\log \alpha}{\log(1-\epsilon_{\mathcal{N}})} \tag{4.3}$$

instances. We expect to find no FN instances in the above sample. If we do find FN instances, we can revert back to SRS.

An analogous estimation can be applied to the $\mathcal{P}$ stratum except we need to find true negatives instead of false negatives.

Once the Negative and Positive strata are proven pure or their FN proportions are determined from SRS, estimating the the total number of FN instances is straightforward.

5 Evaluation

We evaluated our algorithm on the KDD Cup '99 Network Intrusion data set. The data set contains 494,021 instances. The instances belong to the normal class or one of the 20 attack classes. We evaluated a classifier based on Ripper [7], which classified instances as scanner or non-scanner. Out of the 20 attack classes, 4 pertain to scanning activity. The total number of scanning instances is 4,107, which is less than 1% of the total number of instances. Our scan detector detected 86% of all scans.

Evaluating the Estimates. Let $\hat{FN}$ denote the estimate for the number of FN instances; let FN denote the actual number of FN instances (the amount we need to estimate); and let $\mathrm{E}[\hat{FN}]$ and $\mathrm{Var}[\hat{FN}]$ denote the expected value and variance of the estimate, respectively. Then we will evaluate the estimates using the following measures.

$$\begin{aligned} \text{Bias} &= \mathrm{E}[\hat{FN}] - FN \\ \text{Precision} &= \mathrm{Var}[\hat{FN}] = \mathrm{E}[(\hat{FN} - \mathrm{E}[\hat{FN}])^2] \\ \text{Accuracy} &= \mathrm{MSE}[\hat{FN}] = \mathrm{E}[(\hat{FN} - FN)^2] \end{aligned}$$

Bias shows how far the estimate is from the real value and it also shows whether the estimate is an over- or underestimate. Precision is indicative of the variability in the estimate and accuracy shows how close the estimate to the real value is in absolute (squared) terms.

5.1 Estimation via the Proposed Method. We have compared the performance of our method with Simple Random Sampling (SRS). We performed 10 *trial*s for each setting. In each trial, we perform the *entire* estimation process from the partitioning/clustering to the sampling for estimation. For SRS, the sample size of 100,000 was selected, because that is the smallest sample size where the estimates are within the error bounds with (almost) the required probability (See Section 5.1.1). For the proposed scheme, we selected $minMSE = .05$.

Table 1: Comparison of the proposed scheme with SRS

Method	CFP	SRS
Esimate		
Bias	-3.20	2.99
Precision	221.76	1880.80
Accuracy	232.00	1889.80
Lower Qartile	559	549
Median	570	571
Upper Quartile	584	615
Sample Size		
Lower Qartile	24,946	100,000
Median	25,457	100,000
Upper Quartile	25,994	100,000

As Table 1 shows, our proposed method has approximately the same bias as SRS (which is theoretically unbiased), has an 8.48 times better precision and 8.14 times better accuracy. These results were achieved using only a quarter of the the sample size that SRS required.

Before we proceed with analyzing our algorithm further, let us establish a "baseline" performance using Simple Random Sampling.

5.1.1 Baseline: Estimation via Simple Random Sampling (SRS). In this experiment, we took samples of sizes 10,000, 20,000, ..., 200,000 and estimated the number of false negatives from these samples. For each sample size, the estimation was performed on 100 different samples. Figure 3 depicts the estimates. The 100 results for each sample size (along the horizontal axis) are represented by a box plot. The top and bottom of the box denote the upper and lower percentiles and the middle line corresponds to the median. The whiskers extend out to the minimal and maximal estimates.

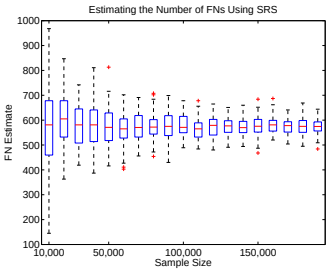

Figure 3: FN Estimation via SRS

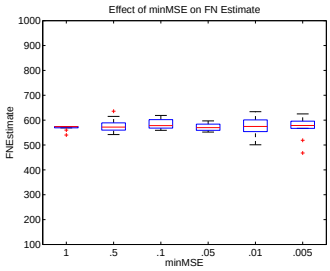

Figure 4: FN Estimation via the proposed method.

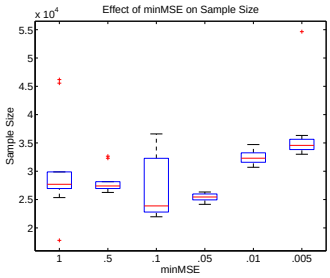

Figure 5: Sample Size Required to Achieve the above Performance

Strictly speaking, the estimates are within 20% error at 95% probability only for a sample size of 200,000, but they already get close (within 20% error at 93% probability) at a sample size of 100,000.

5.1.2 Dependence on $minMSE$. In this section we investigate how $minMSE$ affects the estimation. We ran 10 trials for each of the 6 $minMSE$ values that span almost 3 orders of magnitude.

Figure 4 displays the obtained results. Each boxplot corresponds to the 10 trials with a fixed $minMSE$ value. The estimates are well within the required 20% range.

Due to the guarantees that our sampling scheme must provide, the parameter $minMSE$ appears to have little influence over the estimation precision or accuracy. The difference lies in the sample sizes.

Figure 5 depicts the sample sizes. It shows that a sweet spot exists, where the required sample size is stable and minimal. In our technical report [6], we a describe a method for finding this sweet spot without drawing a single sample. At the sweet spot, our method required only 1/8 of the sample size that needed to be drawn when SRS was used.

6 Summary and Conclusion

In this paper, we considered the problem of estimating the number of false negative instances in a classification task where the true label can be ascertained for only a limited number of instances.

We propose a stratified sampling scheme that exploits the cluster structure of the problem. Clusters are discovered and divided into 3 strata. The clusters in different strata have different statistical properties, hence they give rise to more efficient estimation. Two of the strata are pure: one consisting of mostly positive instances, the other mostly negative instances. The purer the strata, the smaller the sample that is required for estimation within a fix error bound.

We also introduced the *class-focused partitioning* algorithm, which facilitates the efficient discovery of pure clusters. In [6], we have shown that by using our partitions instead of k-means clusters, not only did we achieve reductions in sample sizes and shorter runtimes, but also an approximately 5-fold improvement in estimation precision and accuracy.

Our proposed method (including the class-focused partitioning) has achieved a 4.15-fold improvement in precision, a 3.96-fold improvement in accuracy over SRS with a sample size of 200,000 instances. Our method required a sample of only 25,000 instances.

References

[1] Nizar Grira, Michel Crucinau, and Nozhaa Boujemaa. *Unsupervised and semi-supervised clustering: a brief survey.* Number Report of the MUSCLE European Network of Excellence (FP6). 2005.

[2] Sharon L. Lohr. *Sampling: Design and Analysis.* Duxbury Press, 1999.

[3] Sandeep Mane, Jaideep Srivastava, and San-Yih Hwang. Estimation of missed actual positives using independent classifiers. In *SIGKDD '05*, 2005.

[4] Margaret Sullivan Pepe. *Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford Statistical Science Series. Oxford University Press, 2002.

[5] Matthias Seeger. Learning with labeled and unlabeled data. Technical report, Institute for Adaptive and Neural Computation, 2001.

[6] György J. Simon, Vipin Kumar, and Zhi-Li Zhang. Estimating false negatives for classification problems with cluster structure. `http://www.cs.umn.edu/~gsimon/FNTechRep.pdf`.

[7] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining.* Addison-Wesley, 2005.

[8] Steven D. Walter and L. M. Irwig. Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *Journal of Clinical Epidemiology*, 41(9):923–937, 1988.

[9] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report Computer Sciences 1530, University of Wisconsin – Madison, 2006.