Evaluation of Learning from Screened Positive Examples

Keywords: machine learning, evaluation metrics, one-class learning, privacy, electronic information

Marina Sokolova MSOKOLOVA@EHEALTHINFORMATION.CA
CHEO Research Institute, 401 Smyth Rd., Ottawa, ON K1H 8L1 CANADA

Khaled El Emam KELEMAM@UOTTAWA.CA
CHEO Research Institute, 401 Smyth Rd., Ottawa, ON K1H 8L1 CANADA
University of Ottawa, 800 King Edward Av, Ottawa, ON K1N 6N5 CANADA

Abstract

In this work, we propose metrics to assess the classification performance of an algorithm that learns from a very small number of examples of one class. The metrics, *True Detection Probability* and *False Referral Probability*, have been used in medicine to evaluate classification performance from examples screened to be positive. Necessity of such metrics comes from a real life application: we are developing a tool to detect Personal Health Information (PHI) in files obtained from peer-to-peer file sharing networks. On the data available to us, the files with PHI represent a small portion of all files (1%). However, the detection and prevention of PHI leaks is important: such inadvertent disclosure of PHI increases opportunities for privacy breaches.

1. Appraisal of Machine Learning metrics

In supervised Machine Learning (ML), application of common evaluation measures implies that a classifier's performance is evaluated on *all* examples. With data labels always assumed to be correct, examples on which the classifier's and data labels coincide are called correctly classified. For binary classification, the following three measures represent three popular approaches to evaluate the classifier's performance: *Accuracy* estimates overall performance, without differentiation between positive and negative classes; *Fscore* focuses on positive class classification; *Area Under Curve* separates performance on positive and negative classes.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}, \quad (1)$$

$$Fscore = \frac{(\beta^2 + 1)tp}{(\beta^2 + 1)tp + \beta^2 fn + fp} \quad (2)$$

$$Area\ Under\ Curve = \frac{1}{2}(\frac{tp}{tp + fn} + \frac{tn}{tn + fp}) \quad (3)$$

where tp is the number of correctly labeled positive examples, fp – examples incorrectly labeled by the classifier as positive, tn – correctly labeled negative examples, fn – examples incorrectly labeled by the classifier as negative. To use these metrics, we need to know positive and negative labels `provided with data` and their correspondence with positive and negative labels `output by the classifier`.

However, knowing correspondence between data and classifier labels for positive `and` negative examples can be a fundamental problem in many practical applications. These applications include, but not restricted to, situations when full classification of examples labeled as negatives by the first phase of an algorithm is either unethical (e.g., cancer testing) or economically prohibitive (e.g., full security audit). In this work, we propose performance evaluation restricted to full classification of examples which are labeled as positives (i.e., *screened positive*) by the algorithm's first phase. Further, Sections 2 and 3 discuss an application where classification of only screen positive examples is financially feasible. Section 4 presents a learning algorithm. Section 5 introduces evaluation measures *True Detection Probability* and *False Referral Probability*. Example, related work, conclusions and intended future work complete the paper.

Appearing in *Proceedings of the 3rd EMML Workshop co-joint with the ICML 2008*, Helsinki, Finland, 2008.

2. Motivation for the use of new metrics

We are developing a tool to learn the characteristics of *Personal Health Information* (PHI). PHI consists of *personally identifying information* (PII), e.g., person names, and *health information* (HI), e.g., disease symptoms. Indicators of both types of information have to be present in a file in order to label it as a PHI file. This tool will be used to detect inadvertent PHI leaks from peer-to-peer file sharing networks (El Emam et al., 2008).

To evaluate the tool's performance in realistic conditions, we need to run it on a large number of files that contain PHI (henceforth, the PHI files). PHI files differ in content, language style, and vocabulary, therefore, it is necessary to maximize the number of files included in the evaluation. Our initial empirical evidence showed that PHI files constitute 0.98% of 407 Canadian files randomly chosen from a peer-to-peer network and only 0.44% of 452 US files. At such a low prevalence, to adequately test the performance of the tool we will have to process tens of thousands of files. Because of the sheer number of to-be-processed files, the financial resources to manually label each of these files become prohibitive very quickly.

We propose an evaluation procedure where we only manually verify and label the files classified as PII by our tool, and ignore all the other files that are not classified as PII. Our procedure in a nutshell: eighteen PII characteristics are listed in the US Health Insurance Portability and Accountability Act (HIPAA). Our tool screens all in-coming files for these characteristics (Step 1). Only the files that exhibit one of the characteristics are selected for manual verification and labeling as PHI (Step 2); Figure 1 shows the system sketch.

Step 1 makes standard evaluation metrics not applicable for the algorithm's assessment. Due to the file pre-selection, the contents of two cells of a standard confusion matrix are missing; see Table 1.

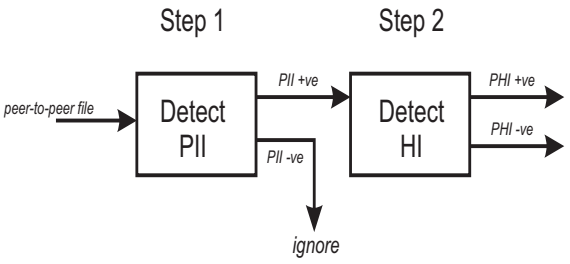

Figure 1. The system design for the PHI detection tool.

Table 1. Comparison of confusion matrices. PII is the target screened concept. $n^+(n^-)$ is the number of potentially positive(negative) examples. PHI denotes verified labels. 1 denotes presence of indicators, 0 - their absence.

Classification of screened positive examples

	PII =1	PII =0
PHI=1	n^+_{PHI}	?
PHI=0	$n^+_{\overline{PHI}}$	?
	n^+	n^-

Classification of positive and negative examples.

	PII =1	PII =0
PHI=1	n^{PII}_{PHI}	$n^{\overline{PII}}_{PHI}$
PHI=0	$n^{PII}_{\overline{PHI}}$	$n^{\overline{PII}}_{\overline{PHI}}$
	n^+	n^-

To obtain meaningful performance estimation, we apply measures that only evaluate classification of examples which promise to be positive. The measures, *True Detection Probability* and *False Referral Probability*, have been used in medicine (Pepe, 2004). Specifically, they are used to evaluate a diagnostic test when it is unethical to use an obtrusive or high risk verification procedure on a patient who is not likely to have a disease. Therefore, the true disease status is only established for those patients who screen positive on an earlier test. To the best of our knowledge, this is the first application of the measures to a machine learning study. Previously, analysis of evaluation measures showed that metrics used in medical tests can be successfully applied in machine learning settings (Sokolova et al., 2006).

3. Personal Health Information

With the increasing adoption of electronic medical records and personal health records, patients, medical professionals and government agencies are concerned about unauthorized disclosure and use of PHI. For example, a recent study has shown that significant PHI is inadvertently leaking from second hand disk drives (El Emam et al., 2007).

According to HIPAA, there are 18 categories of identifying information: person names, addresses, IDs, phone numbers, and so on (Wojcik et al., 2007). Some researchers have expanded this list by including medical practitioner names (Uzuner et al., 2007). We have added the names of organizations (e.g., schools, aid centres) which can act as substitute addresses. The latter agrees with privacy protection recommendations issued by the Canadian Judicial Council (2005); see Table 2.

Empirical evidence (Section 2) showed that 4 patients can potentially be identified from the PHI-labelled Canadian files randomly chosen from a peer-to-peer file sharing network. The PHI texts contained person names, their home or temporal addresses, ID num-

Table 2. Examples of indicators of *personally identifying information* and *health information*.

PII	HI
Personal name	Physical condition
Street Address	Disease symptoms
Health Care providers	Behavioral state
Phone numbers	Mental state

bers, and medical diagnoses. Additionally to the patients, two files identified a family member. At the current stage, we seek cases when these characteristics are found in the same file and belong to the same person. We leave in-width co-reference resolution for future work.

4. Learning PHI

A PHI text class can be identified through a restricted number of PII characteristics and their combinations. Hence, it is reasonable to assume that an algorithm will be able to learn from this class only (Angluin, 1980). This ability is especially important because non-PHI files are too heterogeneous, making it difficult to specify a narrow set of characteristics for them. For example, non-PHI files downloaded from the peer-to-peer file sharing network contained ebooks, photos in doc and pdf formats, business forms, personal non-identifying communications, and drafts of various documents.

To learn the PHI file class, we describe the target concepts through all the PHI characteristics and store the prototypes into memory. As a result, we build the one-class learning algorithm as an instance-based learner. We define the instance-based inductive bias as follows:
I To define the distance between examples, we look at the PHI prototype structure. A PHI prototype text has to have two sets of characteristics:

$$PHI = IdentifyingInformation \wedge HealthInformation \quad (4)$$

$IdentifyingInformation = PersonName \wedge (ID \vee Venue)$
$PersonName = FirstName \wedge LastName$,
$ID = SIN \vee HealthCardN \vee PassportN \vee DriverLicenseN$,
$Venue = OrganizationName \vee StreetAddress \vee HealthProfessionalName \vee PhoneNumber \vee Email$.
The search for *IdentifyingInformation* indicators is related to the *Name Entity Recognition* problem. A few of PII characteristics are enough to make the file sensitive; their total number is restricted. We modeled them by a set of Regular Expressions and partially applied statistical *Name Entity Resolution*. We are planning to build the *HealthInformation* part by extracting information from available electronic versions of medical dictionaries (Hecht & Shiel, 2003). We expect a large variety of HI characteristics because they are often written in natural language and involve a large vocabulary. We are planning to use *bag-of-terms* to represent the HI characteristics and apply k-Nearest Neighbor to label examples.

The difference between PII and HI characteristics implies that the tool can use a two-fold distance, each fold computed separately on Step 1 and Step 2 of the PHI detection process (Figure 1). *Hamming Distance*, the number of bits which differ between two binary strings A and B of length n, is applied to compute the difference between the prototype and an incoming data entry on PII attributes:

$$Hamming\ Distance = \sum_{1}^{n} |PII(A)_i - PII(B)_i| \quad (5)$$

For Step 2, the HI indicators learning, *Eucledian Distance* will be used on the document level to compute similarities between the prototypes and new examples.
II During the testing phase, we employed screening for the PII characteristics. The tool first looks for the presence of the PII characteristics, then we manually search for HI characteristics in that subset of files which have PII. Files that are classified as PHI are manually verified. The *True Detection Probability* (TDP) and *False Referral Probability* (FRP) were used to measure the performance of the tool. The measures we used in verification are suitable when verification can only be done on screened positive files (Pepe, 2004). We discuss the measures in Section 5.

5. Evaluation restricted to screened positives

The inability to verify all cases happens in many settings of medical testing when it is unethical or very costly to further test patients (to verify their disease status using a gold stanard test) whose initial results were negative (Pepe, 2004). Final results, thus, depend on the algorithm performance and the predictive power of a "screened for" characteristic. Consequently, we need measures that evaluate the predictive ability of both, the characteristics and the algorithm's performance.

In learning from potentially positive examples, the standard conditional probabilities do not work. Consider two measures:

$$True\ Positive\ Fraction = P[PII = 1|PHI = 1] \quad (6)$$

$$False\ Positive\ Fraction = P[PII = 1|PHI = 0] \quad (7)$$

With only positive examples screened, they will be changed as follows:

$$\textit{True Positive Fraction} = \frac{n^+_{PHI}}{n^+_{PHI}+?} \quad (8)$$

$$\textit{False Positive Fraction} = \frac{n^+_{\overline{PHI}}}{n^+_{\overline{PHI}}+?} \quad (9)$$

As a result, both measures will be undetermined.

To estimate the classification performance, we switch from conditional to joint probabilities. We compute the probability of PII and HI happening together, *True Detection Probability*(TDP), and the probability of HI absence and the presence of PII, *False Referral Probability*(FRP):

$$TDP = P[PHI = 1, PII = 1] \quad (10)$$

$$FRP = P[PHI = 0, PII = 1] \quad (11)$$

Using the notations of Table 1, we approximate the two measures as:

$$\hat{TDP} = \frac{n^+_{PHI}}{n^+ + n^-} \quad (12)$$

$$\hat{FRP} = \frac{n^+_{\overline{PHI}}}{n^+ + n^-}. \quad (13)$$

TDP shows the proportion of files an algorithm marked as having an PII indicator `and` containing PHI. It should be as high as possible, ideally converging to $P[PHI = 1]$. FRP shows the proportion of files the algorithm marked as having the PII indicator but `not` containing PHI. It should be as low as possible, ideally converging to 0. Both measures depend on the precision of the pre-defined characteristic and on the classifier's ability to discriminate between positive and negative examples, e.g., the PHI and non-PHI files.

6. Example

Although our tool will be applied to prevent the PHI leakage from peer-to-peer networks, our initial tests are on data from second-hand hard drives (El Emam et al., 2007). The total number of files was $2,579,425$ of them verified to be the PHI files. On this data, the proportion of PHI files on the same hard drive varied from 1.30% to 20.00%, with potential identification of 197 patients from the files found on only one hard drive (three rounds of independent manual evaluation). The non-PHI files contained advertisements, business correspondence, drafts of various documents and school assignments.

Table 3. Confusion matrix for classification of 123 second-hand hard drive files. Only PII positive files were verified for PHI; 1 denotes presence of indicators, 0 - their absence.

Organization names positive.

	PII=1	PII=0
PHI=1	71	?
PHI=0	4	?
	75	48

Street addresses positive.

	PII=1	PII=0
PHI=1	75	?
PHI=0	3	?
	78	45

Table 4. Performance evaluation for PII characteristics.

Measure	OrgName	StrAddr
$\hat{TDP}$	0.577	0.610
$\hat{FRP}$	0.033	0.024

We illustrate the measures' properties by computing them for the "PHI vs non-PHI" binary classification of 123 files obtained from second-hand disk drives. Here we assume that PII characteristics include only *OrganizationNames* and *StreetAddress*. In the first set of experiments, the files were screened for *OrganizationNames*, in the second set – for *StreetAddress*. Table 3 reports the confusion matrices. Table 4 reports the obtained measure values. With higher $\hat{TDP}$ and lower $\hat{FRP}$, *StreetAddress* is a better predictor of a PHI file on this data.

Note that both measures evaluate the characteristic coverage of the files. In the case of second-hand hard drive files, the *OrganizationNames* characteristic covers only 0.610% of the files, *StreetAddress* – 0.634% of the files. Among the remaining files, 6 contained other potentially identifying information, either phone numbers or email. Only 18 files did not contain identifying information, except for person names. The latter files were non-PHI.

7. Related work

Current tools that detect PHI for the purpose of anonymization focus on the quality of PII detection (Uzuner et al., 2007). While showing F-score 98% on supplied data, the tools perform poor on previously unseen sets, e.g. they do not recognize names written without preceding titles *Mrs., Mr* (Uzuner et al., 2007). The tools' learning component depends on document types and may not be applicable when PHI texts are buried among many diverse documents which do not contain PHI. We showed that PHI text characteristics can be learned, first, by representing texts through designated features, and second, by applying an instance-based algorithm on these representations.

The aim of a one-class learner is to establish a set of rules that discriminate between examples of a known

class and all other examples. One-class learning is often used for anomaly detection in the medical domain (Glickman et al., 2005) and security applications (Liu et al., 2006). In anomaly detection, examples from a known class usually outnumber the unknown classes. As a result, the negative-selection procedure is implemented: algorithms are trained to detect anomalies, which do not correspond to the established profile of the known class. We, however, solve an opposite problem where the algorithm seeks examples corresponding to the PHI profile.

When learning from a small positive class, some methods model the class with probabilistic classifiers (Li et al., 2007). In contrast, we concentrate only on potentially positive examples that contain pre-defined PII characteristics. Text classification from only positive examples was used in a news monitoring system (Zizka et al., 2006). The authors built a prototype-based algorithm which evaluates closeness of incoming examples to stored prototypes. The examples are represented by bag-of-words. The unlabeled examples are ranked according to the computed cosine similarity measure. The bag-of-words representation and cosine measure treat all features equally. In contrast, we represent the known class example by two sets of features which are treated separately on the testing phase. Our algorithm computes *Hamming Distance* between examples' PII attributes and *Euclidean Distance* between examples' HI attributes. None of the above work considered evaluation measures restricted to potentially positive examples. Our work, in contrast, presents two measures that are specifically designed to assess classification performance on such examples.

8. Conclusions and future work

In this paper, we showed limitations of common ML evaluation measures that restrict their use in some practical applications. Specifically, we showed that a measure's dependence on classification of positive and negative examples prohibits its use when full classification of negative examples is either un-feasible or non-desirable. We presented two measures, *True Detection Probability* and *False Referral Probability*, that can be used when only examples screened positive on the first stage become candidates for full classification.

In empirical support of our claim, we presented a real-life situation (i.e., prevention of PHI leaks) where such measures can be applied and the advantages of their application (i.e., freeing manual and financial resources). We introduced an algorithm that learns from screened positive examples. We discussed the PII characteristics and showed that their restricted number allows learning from only a class of screened positive examples. We applied *True Detection Probability* and *False Referral Probability* to evaluate the predictive ability of the PII characteristics and, consequently, the algorithm's performance.

We plan to continue work on the algorithm and its performance evaluation. Implementation of *HealthInformation* will be our next step, followed by completing the statistical part of *Name Entity Resolution*. Peer-to-peer file sharing networks and second-hand disk drives are confirmed to contain PHI files. There are other contexts where a large volume of files need to be scanned to detect a few PHI files, for example, enforcing organizational security policies for email whereby PHI leaks through email attachments need to be detected. Outside PHI detection, *True Detection Probability* and *False Referral Probability* can be used in security audits where learning focuses on screened positive examples (i.e., only examples that exhibit certain characteristics are kept for further check-ups).

Acknowledgements

This work is supported by Natural Sciences and Engineering Research Council of Canada and Ontario Centre of Excellence research grants. The authors thank the workshop organizers for many helpful comments.

References

Angluin, D. (1980). Inductive inference of formal languages from positive data. *Information and Control*, *45*, 117–135.

Canadian Judicial Council (2005). Use of Personal Information in Judgments and Recommended Protocol. http://www.cjc-ccm.gc.ca/cmslib/general/news_pub_techissues_UseProtocol_2005_en.pdf. Accessed June 13, 2008.

El Emam, K., Neri, E., & Jonker, E. (2007). An evaluation of personal health information remnants in second hand personal computer disk drives. *Journal of Medical Internet Research*, *9*, e24.

El Emam, K., Neri, E., Sokolova, M., Jonker, E., & Peyton, L. (2008). *The inadvertent disclosure of personal health information through peer-to-peer file sharing programs* (Protocol). Children's Hospital of Eastern Ontarion Research Institute, Ottawa, Ontario, Canada.

Glickman, M., Balthrop, J., & Forrest, S. (2005).

A machine learning evaluation of an artificial immune system. *Evolutionary Computation Journal*, *13*, 179–212.

Hecht, F., & Shiel, W. (2003). *Medical dictionary*. Wiley Publishing.

Li, X., Liu, B., & Ng, S.-K. (2007). Learning to classify documents with only a small positive training set. *Proceedings of European Conference on Machine Learning* (pp. 201–213).

Liu, Z., Zhu, M., & Fan, K. (2006). One-class learning based algorithm for the freeway automatic incident detection. *Journal of Computer Science and Network Security,*, *6*, 289–293.

Pepe, M. (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. *Proceedings of the ACS Australian Joint Conference on Artificial Intelligence* (pp. 1015–1021).

Uzuner, O., Luo, Y., & Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-indentification. *Journal of the American Medical Informatics Association*, *14*, 550–563.

Wojcik, R., Hauenstein, L., Sniegoski, C., & Holtry, R. (2007). Obtaining the data. In J. Lombardo and D. Buckeridge (Eds.), *Disease surveillance*, 91 – 142. Wiley.

Zizka, J., Hroza, J., Pouliquen, B., Ignat, C., & Steinberger, R. (2006). The selection of electronic text documents supported by only positive examples. *Proceedings of JADT 2006*.