Estimation of False Negatives in Classification *†

Sandeep Mane, Jaideep Srivastava
Department of Computer Science
University of Minnesota
Minneapolis, USA
{smane, srivasta}@cs.umn.edu

San-Yih Hwang
Department of Information Management
National Sun-Yat-Sen University
Kaohsiung, Taiwan
shwang@cs.umn.edu

Jamshid Vayghan
IBM Corporation
Minneapolis, USA
vayghan@us.ibm.com

Abstract

In many classification problems such as spam detection and network intrusion, a large number of unlabeled test instances are predicted negative by the classifier. However, the high costs as well as time constraints on an expert's time prevent further analysis of the "predicted false" class instances in order to segregate the false negatives from the true negatives. A systematic method is thus required to obtain an estimate of the number of false negatives. A capture-recapture based method can be used to obtain an ML-estimate of false negatives when two or more independent classifiers are available. In the case for which independence does not hold, we can apply log-linear models to obtain an estimate of false negatives. However, as shown in this paper, lesser the dependencies among the classifiers, better is the estimate obtained for false negatives. Thus, ideally independent classifiers should be used to estimate the false negatives in an unlabeled dataset. Experimental results on the spam dataset from the UCI Machine Learning Repository are presented.

1 Introduction

Detecting intrusions in a computer network can be considered as a 2-class classification problem. The task is to analyze each network flow and label it as 'suspicious' or 'normal'. [1] There are some unique characteristics of this problem. First, the rate of data generation is very high, e.g. 200,000-300,000 connections per minute. Second, the occurrence of 'intrusions' is much rarer than the occurrence of 'normal' traffic. For such a dataset, a classifier will label relatively very few instances as positive as compared to those labeled negative. The predicted positive instances can be given to an expert who can further analyze them in order to separate the true positives from the false positives. However, the negatively classified instances, being much larger in number, would require an unacceptable amount of time to separate the false negatives from the true negatives. Thus, getting a complete picture of classifier accuracy, e.g. ROC curves, is infeasible. However, since the cost of a false negative may be much higher than of a false positive, e.g. an actual attack being missed, obtaining at least an estimate of false negatives predicted by the classifier is required. This, for example, can be used to estimate false negatives detected by two intrusion detection systems (say SNORT – *http://www.snort.org/* and MINDS – *http://www.cs.umn.edu/research/minds/MINDS.htm*) for an unlabeled dataset, and then comparing their performance.

In the commercial domain, an example of this problem is the estimation of missed opportunities during the sales opportunity analysis process (Vayghan et al. [11]). Here, once a sales opportunity has been classified as negative (not promising) by a human expert (e.g. a business manager), there is no further analysis of that opportunity in order to verify whether it was actually unprofitable or there was a judgment error. A method for estimating the number of false negatives predicted by the decision maker would be useful to estimate the accuracy of the human expert w.r.t. the ground truth (actual outcome). Furthermore, for an individual decision maker, it will help identify strengths and weakness in different domains of opportunities, e.g. the ability to identify 'hardware-selling opportunities' vs. the ability to identify 'software-services opportunities'.

The examples above motivate the need for estimating false negatives for a classifier on an unlabeled dataset. In this paper we present a methodology for obtaining such an estimate for false negatives based on the classical capture-recapture method for parameter estimation in statistics. In addition, we also illustrate a number of important issues

*Sandeep Mane's and Jaideep Srivastava's work was supported in part by NSF Grant ISS-0308264, ARDA Grant F30602-03-C-0243, and a grant from IBM. San-Yih Hwang's work was supported by a Fulbright scholarship. Jamshid Vayghan's work was supported by the IBM Corporation.

†The authors would like to acknowledge comments from Prof. Vipin Kumar and Dr. Philip Yu which led to the formulation of this problem.

[1] Data mining is suitable for detecting novel, i.e. previously unseen, attacks. In such a case, automated techniques can only identify unusual or suspicious behavior. An expert analyst must then examine it to determine if it is truly an intrusion.

IEEE COMPUTER SOCIETY

that need to be explored in making the application of this method practicable. The remainder of this paper is organized as follows: section 2 provides a brief overview of the approach and related work, section 3 presents experimental results, and section 4 concludes future research directions.

2 General approach and related work

Hook and Regal [8] present a survey on false negative estimation in epidemiology using two or more detection methods (classifiers) and the capture-recapture method [4]. Goldberg and Wittes [6] present a generalized approach to false estimation for the multi-class classification problem, which is illustrated using the 2-class case. Consider a labeled dataset which is classified by a $\{True, False\}$-class classifier, whose confusion matrix for the classifier is shown in the Table 1.

		Actual class		
		True	False	Total
Predicted class	True	TP	FP	PP
	False	FN	TN	PN
	Total	AP	AN	N

Table 1: Confusion matrix for a classifier

Here, TP, FP, FN and TN represent the numbers of *true positives*, *false positives*, *false negatives* and *true negatives* respectively. Also, AP, AN, PP and PN are the numbers of *actual positives*, *actual negatives*, *predicted positives* and *predicted negatives* instances, while N is the total number of instances in the dataset. Actual positives are the instances in the dataset whose actual (real) class is $True$. The performance of the classifier can be determined using this confusion matrix.

However, for a skewed-class distribution classifier with a very high data volume, e.g. network intrusion detection, for a given unlabeled dataset only the predicted positive instances are manually classified into true positives and false positives. The predicted negative instances, being very large in number, are not analyzed further by the human expert. Thus, the confusion table for the classifier for the dataset will look as shown in Table 2.

		Actual class	
		True	False
Predicted class	True	TP	FP
	False	FN + TN	

Table 2: Confusion matrix for a rare-class, large-dataset classifier.

The notation used in Table 2 is identical to that in Table 1. Here, only the total (TP+FN) can be obtained. Now, if AP in the dataset is known, then, from the Table 1, FN can be determined. (This is because TP+FN=AP) Thus, the method for estimation of FN is based on the estimation of AP in the dataset.

The main idea behind the method for estimating actual positives using the capture-recapture method can be explained using the following example problem.

Problem: Estimate the number of fish in a pond.
Estimation Method: A two step method, called the 'capture' and 'recapture' steps, is used for this. In step one (capture), let f_1 be the number of fish caught, which are then marked (presumably with an indelible ink) and released in the lake. In the second step (recapture), let f_2 be the number of fish that are caught (presumably after sufficient time to allow the fishes to mix, but not mate and produce more fishes, or even die). Let f_{12} be the number of fish caught in second step, which are found to be marked. Under the stated assumptions, f_{12} will follow a hyper-geometric distribution, since the process is equivalent to 'selection with replacement'. Thus, the estimate for the total number of the fish in the lake is $\left(\frac{f_1 * f_2}{f_{12}}\right)$. Now, if the actual positive instances in the dataset are compared to fish in the lake, then the capture-recapture methodology can be used to estimate the number of actual positives in the dataset, given that the two steps (samplings) are independent of each other. Thus, for applying this technique, there is a need for at least two independent classifiers (detection methods). It should be noted that this method can be extended to the case where more than two independent samplings are available. ■

We now explain the method for estimating the number of actual positives using the capture-recapture method and the classifiers in detail.

		APs detected by classifier 1		
		Yes	No	Total
APs detected by classifier 2	Yes	n_{11}	n_{12}	n_2
	No	n_{21}	n_{22}	n_4
	Total	n_1	n_3	n

Table 3: Contingency table of actual positives for the case of two classifiers

Suppose that two independent classifiers classify the two-class dataset. Let n_1 and n_2 be the number of true positive instances detected by the first and second classifiers, respectively. Let n_{11} be the number of true positives detected by both classifiers. Also, as shown in Table 3, let n_{12} be the actual positive instances classified as $True$ by only the first classifier and let n_{21} be the actual positive instances classified as $True$ by only the second classifier. The value n_{22}, the number of actual positive instances not detected (i.e. classified $False$) by both classifiers, is unknown and needs to be estimated. The sum n of the values in all the cells of the Table 3 is equal to the number of actual positives in the dataset. If the two classifiers are independent, then the ML-estimate for the unknown value n_{22}, as shown by Goldberg and Wittes [6], is : $\mathrm{n}_{22} = \left(\frac{\mathrm{n}_{12} * \mathrm{n}_{21}}{\mathrm{n}_{11}}\right)$.

Wittes et al. [14, 13] discuss the problems arising from decision making in the capture and recapture steps being dependent. If so, i.e. when independence does not hold between the variables in the contingency table, log-linear models (Knoke and Burke [9]) must be used for the contingency table. Fienberg [5] describes a method for constructing log-linear models for the contingency table in such cases and obtaining the best-fitting model. In this approach, the conditional relationship between two or more discrete categorical variables (here, the class labels assigned by the classifiers are discrete categorical variables) is analyzed by taking the natural logarithm of the cell frequencies within a contingency table. For example, for the contingency Table 3, the following model is used to represent the expected frequency of each cell (i,j) in the table –

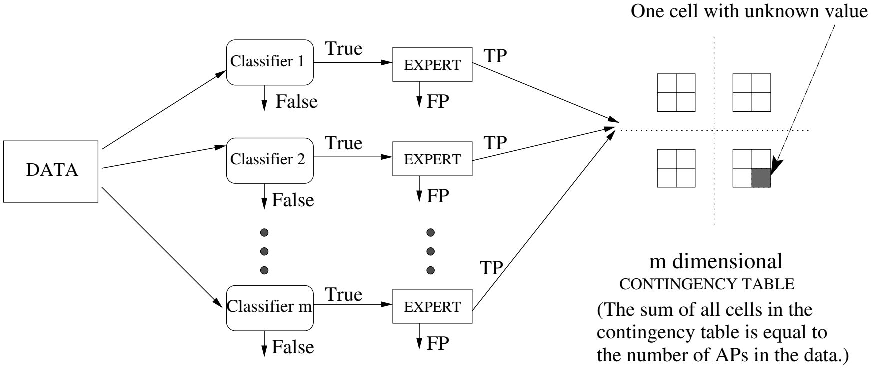
$$\mathrm{Ln}(\mathrm{F_{ij}}) = \mu + \lambda_{\mathrm{i}}^{\mathrm{A}} + \lambda_{\mathrm{j}}^{\mathrm{B}} + \lambda_{\mathrm{ij}}^{\mathrm{AB}}$$

where, $\mathrm{Ln}(\mathrm{F_{ij}})$ is the log of the expected cell frequency of the instances in the cell (i,j) in the contingency table; μ is the overall mean of the natural log of the expected frequencies; A and B are the variables (APs detected by each classifier); i,j refer to the categories within the variables; $\lambda_{\mathrm{i}}^{\mathrm{A}}$ is the main effect of the variable A on the cell frequency; $\lambda_{\mathrm{i}}^{\mathrm{B}}$ is the main effect of the variable B on the cell frequency; and $\lambda_{\mathrm{ij}}^{\mathrm{AB}}$ is the interaction effect of variables A and B on cell frequency.

The basic strategy involves fitting a set of such models to the observed frequencies in all cells of the table. In fitting these models, no distinction is made between independent and dependent variables, i.e. log-linear models demonstrate the general association between variables. Different sets of models depending upon various possible dependencies among the variables are fitted to the table. A log-linear model for the entire table can thus be represented as a set of expected frequencies (which may or may not represent the observed frequencies). Such a model is described in terms of the marginals it fits and the dependencies that are assumed to be present in the data. Iterative computation methods for fitting such a model to a table are described in Christensen [2]. Using deviance measures, e.g. the likelihood ratio or χ^2 measure, as a measure of the goodness-of-fit for a model, the best-fitting, parsimonious (least number of dependencies) model for the table is determined. This model is then used to estimate of the unknown value n_{22}. The purpose of log-linear modeling is thus to choose minimum dependencies in a model for the given cells, while achieving a good goodness-of-fit. This method requires is computationally intensive since models corresponding to all possible dependencies among the variables need to be computed. The disadvantage of this method is that a sufficiently large amount of data (cell values) is required for obtaining a good estimation of the contingency table model. Also, high degrees of association among the variables makes it difficult to comprehend the model. [2]

The Figure 1 illustrates the method of estimating actual positives (and hence false negatives) using m classifiers. Given m different classifiers and a dataset, the number of true positives detected by each of the m classifiers is determined and cross-tabulated in a contingency table. One cell in the contingency table will be unknown, which corresponds to the number of actual positives not detected by all m classifiers. Using ML-estimation technique or log-linear model (depending upon whether independence does or does not hold), an estimate for the unknown cell is obtained. Thus, the total number of actual positives is estimated. The assumption to be noted is that the classifiers used in the capture-recapture method do a good job of keeping the number of false positives low. This helps to keep the number of instances to be manually classified by experts low. Once an estimate of the number of actual positives, $\widehat{AP}$ in dataset has been obtained, the same dataset is classified using a classifier whose performance (accuracy) is to be evaluated. The instances predicted $True$ by the classifier are analyzed manually to separate TP and FP. Next, the estimate $\widehat{AP}$ is used to estimate the false negatives ($\widehat{FN}$) and true negatives ($\widehat{TN}$) detected by the classifier. Using these estimates, the performance (accuracy) of the classifier is evaluated.

Figure 1: Method for estimation of false negatives.

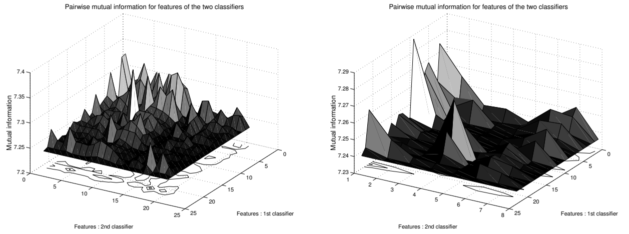

One cell with unknown value
DATA
Classifier 1
Classifier 2
Classifier m
True
False
EXPERT
TP
FP
m dimensional
CONTINGENCY TABLE
(The sum of all cells in the contingency table is equal to the number of APs in the data.)

3 Experimental work

For experimental work, a two-class classification problem using the SPAM email dataset [1] was used. Goldberg and Wittes [6] defined independence of two classifiers as disjoint feature sets. Instead of using disjoint condition as the only criterion for independence, we quantified independence in terms of independence of feature sets, using mutual information [3]. Three disjoint subsets for the dataset were obtained and three different decision tree classifiers A, B and C were trained (using WEKA [12]). As all the features were continuous and due to the limited amount of

[2] The capture-recapture method for false estimation thus requires modeling of concepts for independent, quasi-independent and dependent contingency tables which are summarized by Goodman [7].

Figure 2: Pair-wise mutual information given class 'True' for features used by the classifiers



(a) Classifiers A and B

(b) Classifiers A and C

training data, it was not possible to decide the independence of two feature subsets (in terms of mutual information). In other words, it was not possible to estimate the exact mutual information between two feature subsets, each having sufficiently large number of continuous features. This is an effect of the curse of dimensionality. To overcome this, we instead computed the pair-wise mutual information (MI) [10] between the individual features pairs for each pair of classifiers. The plots of MI for two pairs of classifers, namely (A,B) and (A,C), are shown in Figure 2. [3] It was noted that the feature pairs for the classifiers A and B were on average more pair-wise dependent than feature pairs of classifiers A and C.

The classifiers A, B and C were used to classify the test dataset and the numbers of TPs detected by each classifier were determined. The TPs for each pair of classifiers were cross-tabulated into a contingency table and then the number of APs not detected by all classifiers was estimated using log-linear models. Since the test dataset used was labeled, the number of APs actually missed by all the classifiers was also determined. The results were summarized in the Table 4. The classifiers A, B and C had an approximate training accuracy of 95%, 93% and 74% respectively. Thus, it is noted that even though classifier B had higher accuracy than classifier C, the pair of classifiers A and C gave a better estimate of the total number of actual positives than the pair of classifiers A and B.

Classifiers used	No. of APs not detected by both classifiers (Using labels for test data)	No. of APs not detected by both classifiers (Using log-linear modeling)
A and B	31	4
A and C	5	3
B and C	8	3

Table 4: Cross-tabulation of missed APs and estimated APs.

[3] Note: The range for Z-axis is different for two subfigures. Also light color represents more MI and vice versa. In subfigure (a), the features of the first classifier (A) were plotted along the X-axis and that of the second classifier (B) along the Y-axis. Likewise for subfigure(b).

4 Conclusions

In this paper, a capture-recapture based method for the estimation of false negatives has been presented. The need for having independent classifiers for the method of estimation of false negatives was illustrated using a real-world dataset. Furthermore, it was shown that if the pair-wise MI between the features of a pair of classifiers is low – even if that pair of classifiers has relatively lower accuracy than another pair of classifiers – it may be possible to obtain a better estimate of missed APs using the former pair. Thus, a better estimate for total number of APs, and hence for false negatives, is obtained using independent classifiers. Our current research will address the issues in obtaining *sufficiently accurate* and *independent classifiers* for a given dataset.

References

[1] C. Blake and C. Merz. UCI repository of machine learning databases, 1998.

[2] R. Christensen. *Log-Linear Models and Logistic Regression*. Springer-Verlag Inc, New York, USA, 1997.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York: Wiley, 1991.

[4] J. N. Darroch. The multiple-recapture census: I. estimation of a closed population. *Biometrika*, 45(3/4), 1958.

[5] S. E. Fienberg. An iterative procedure for estimation in contingency tables. *Ann. Math. Stat.*, 41(3), 1970.

[6] J. Goldberg and J. Wittes. The estimation of false negatives in medical screening. *Biometrics*, 34(1):77–86, March 1978.

[7] L. A. Goodman. The analysis of cross-classified data: independence, quasi-independence and interactions in contingency tables with or without missing entries. *J. Amer. Stat. Assn.*, 63(324), 1968.

[8] E. B. Hook and R. R. Regal. Capture-recapture methods in epidemiology: methods and limitations. *Epid. Reviews*, 17(2), 1995.

[9] D. Knoke and P. Burke. *Log-Linear Models*. Sage Publications, Inc. USA, 1980.

[10] R. Moddemeijer. On estimation of entropy and mutual information of continuous distributions. *Sig. Proc.*, 16, 1989.

[11] J. Vayghan, J. Srivastava, S. Mane, P. Yu, and G. Adomavicius. Sales opportunity miner: Data mining for automatic evaluation of sales opportunity. *Book Chapter in New Generation of Data Mining Applications edited by Mehmed Kantardzic and Jozef Zurada*, 2004.

[12] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2000.

[13] J. Wittes. Applications of a multinomial capture-recapture model to epidemiological data. *J. Amer. Stat. Assn.*, 69(345), 1974.

[14] J. Wittes, T. Colton, and V. Sidel. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *J. Chronic Diseases*, 27(1), 1974.