Cost-Sensitive Access Control for Illegitimate Confidential Access by Insiders

Young-Woo Seo and Katia Sycara

Robotics Institute
Carnegie Mellon University
Pittsburgh PA 15213, USA
{ywseo, katia}@cs.cmu.edu

Abstract. In many organizations, it is common to control access to confidential information based on the need-to-know principle; The requests for access are authorized only if the content of the requested information is relevant to the requester's current information analysis project. We formulate such content-based authorization, i.e. whether to accept or reject access requests as a binary classification problem. In contrast to the conventional error-minimizing classification, we handle this problem in a cost-sensitive learning framework in which the cost caused by incorrect decision is different according to the relative importance of the requested information. In particular, the cost (i.e., damaging effect) for a false positive (i.e., accepting an illegitimate request) is more expensive than that of false negative (i.e., rejecting a valid request). The former is a serious security problem because confidential information, which should not be revealed, can be accessed. From the comparison of the cost-sensitive classifiers with error-minimizing classifiers, we found that the costing with a logistic regression showed the best performance, in terms of the smallest cost paid, the lowest false positive rate, and the relatively low false negative rate.

1 Introduction

Illegitimate access to confidential information by insiders poses a great risk to an organization. Since malicious insiders are well aware of where the valuable information resides and which cause damaging effects, the results of illegitimate confidential access are far more costly. Illegitimate access is difficult to effectively prohibit or detect because malevolent actions are done by already trusted persons.

One of the most common approaches to handle this problem is access control based on the need-to-know principle; The requests for access are authorized only if the content of the requested information is relevant to the requester's project. For example, if an information analyst's current project concerns the development of nuclear weapon by Iran, it would be illegitimate for the analyst to have access to documents on other aspects, e.g., feminist activities in Iran. However, since documents on these different aspects of Iranian politics and welfare are not

S. Mehrotra et al. (Eds.): ISI 2006, LNCS 3975, pp. 117–128, 2006.

necessarily a priori separated in different secured data bases, the issue of allowing access on a need-to-know basis on particular documents is very challenging.

Requests to access the confidential information may occur, for example, when an employee is assigned to a new project and needs to access background knowledge. The project manager will either hand select only those confidential information that he will let the employee see, or completely bar access to the entire collection rather than exposing information that should not be exposed. However this approach is quite inflexible. It does not allow easy adjustment to frequent changes of a user's task assignment. Project assignments for an employee may be changed quite often and hence the employee needs to access confidential information related to the newly assigned project. Alternatively, since the organization wants to make sure that the employee accesses only pertinent information, a set of access control lists (ACL) may be compiled manually to control those requests. Each item of confidential information is associated with an ACL, which ensures a corresponding level of security and can be accessed by anyone who has been authorized. However this approach has a crucial security weakness. Since, for the purpose of indexing and security, confidential information is grouped into containers by project-basis, a user who is authorized to a segment of confidential information in a container is actually able to access the entire container.

As a solution for these problems, we developed a multi-agent system that handles the authorization of requests for confidential information as a binary classification problem [9]. Instead of relying on hand-picked information or coarse-grained ACLs, our system classifies on-the-fly the content of each requested information access as positive or negative with respect to the content of the requester's project and authorizes the request if the requested information is classified as positive to the requester's project. Otherwise the request is rejected because the requester's project description is not similar to the information. Our approach is quite flexible and adaptive to changes of project assignment because only an updated description of newly assigned projects is necessary to re-train the classifiers, instead of re-compiling the ACL on all changing relevant information. Therefore, it is much less expensive, both computationally, and also in terms of human time and effort, than an ACL-based approach.

Although our approach showed a relatively good performance [9], we believe there is room for improvement. Previously we made use of five different error-minimizing classifiers for authorizing the requests to access confidential information. However, in domains where there is differential cost for misclassification of examples, an error-minimizing approach may not give results that reflect the reality of the domain. For example, suppose that there are 100 medical cases that are comprised of 5 cancer cases and 95 flu cases. Without considering the cost for misclassification (e.g., compensation for misdiagnosis), an error-minimizing classifier would simply achieve the lower error rate by ignoring the minority class, even though the actual result of misdiagnosis on cancer is far worse than that of flu. Thus, it is undesirable to use an error-minimizing classification method, which treats all mis-classification costs equally for such a cost-sensitive scenario because primarily it classifies every example as belonging to the most probable class.

In this paper we present our works for testing the effectiveness of cost-sensitive learning for the problem of confidential access control. Section 2 compares cost-sensitive classification with error-minimizing classification in terms of the optimal decision boundary. In addition, it describes two cost-sensitive learning methods for the process of confidential access control. Section 3 describes experimental settings and empirical evaluation of cost-sensitive learners. Section 4 presents related work and section 5 presents conclusion and future work.

2 Cost-Sensitive Classification

A classification method is a decision rule that assigns one of (or more than one) predefined classes to given examples. The optimal decision boundary is a decision criterion that allows a classifier to produce the best performance. Let us consider a hypothetical example in figure 1 which shows two classes with overlapping boundaries due to their intrinsic randomness – their actual values are random variables. In this example, the class-conditional density for each class is a normal distribution, that is, $f_0(x|class = 0) \sim N(\mu_0, \sigma_0^2)$ and $f_1(x|class = 1) \sim N(\mu_1, \sigma_1^2)$ (i.e., $\mu_0 = 0.3500, \sigma_0 = 0.1448, \mu_1 = 0.7000, \sigma_1 = 0.1736$).

Under the equal cost for misclassification, the optimal decision boundary (the solid line in the figure 1) lies in the center of two class distributions. An example randomly generated will be assigned to class 1 if its value is greater than 0.52 (the actual value of the optimal decision boundary in figure 1 is $xe^* = 0.52$). Otherwise it is assigned to class 0. According to the optimal boundary, a classifier can generate four possible classification outcomes for a given example; *a*: true positive, *b*: false positive, *c*: false negative, and *d*: true negative [4]. Table 1 captures this information as well as the cost (λ_{ij}) involved in those four outcomes.

If the cost for misclassification is unequal, where then would be the optimal decision boundary? Let us consider the case that there are text documents belonging to "class 0" and "class 1," and all of them are confidential information of which careless release may have a damaging effect. An employee is newly

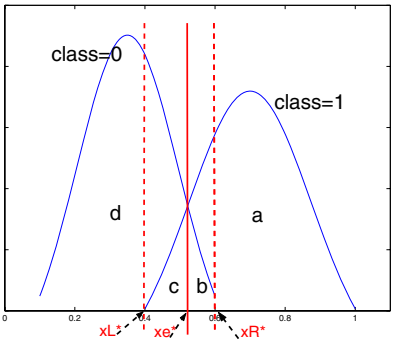

Fig. 1. The optimal decision boundary for a binary classification may vary according to the cost for misclassification

Table 1. A cost matrix represents four possible classification outcomes and their associated costs. In particular, for a given example, x, "a" means true positive, i.e., the example belongs to class 1 and is classified as class 1. " c" is false negative if x is classified as class 0. "d" is true negative if x belongs to class 0 and is classified as class 0. Finally, "b" is false positive if x is classified as positive. λ_{ij} is the cost for classifying an example belonging to j as i.

	true $class = 1$	true $class = 0$
output $class = 1$	a (λ_{11})	b (λ_{10})
output $class = 0$	c (λ_{01})	d (λ_{00})

assigned to a project for which records are labeled as "class 1." He is authorized to access only documents in "class 1" because he needs to know background knowledge of the project. Assuming that a zero cost is assigned to the correct classification [1] (i.e., $\lambda_{11} = \lambda_{00} = 0$), the costs for two types of error should be considered carefully for providing a reliable confidential access control; false negative (λ_{01}) – reject the valid request (e.g., reject the request that the employee asks to access a "class 1" document); false positive (λ_{10}) – accept the invalid request (e.g., accept the request that the employee asks to access a "class 0" document). In particular, a false negative causes the employee to be inconvenienced because he is not able to access need-to-know information. However, not approving valid requests does not cause a serious problem from the security perspective. On the contrary, a false positive is a serious problem because confidential information, which should not be revealed, can be accessed. Therefore, for a need-to-know basis confidential authorization, the cost for false positive (i.e., the damaging effect) is much higher than that of false negative.

Thus the decision boundary for uniform-cost must be re-located, in order to minimize the cost for misclassifications. For example, if the cost of false positive is higher than that of false negative, the decision line should be moved toward the right (e.g., $xe^* \rightarrow xR^*$). Two dashed lines in figure 1 represent the optimal decision boundaries for non-uniform misclassification cost assigned to each example. However a tradeoff must be considered because choosing one of the extremes (e.g., xL^* or xR^*) will not consider the error. In particular, the classifier could reduce the false negative close to zero if we would choose xL^* as a decision line, but with higher false positive. If either of extremes is not the solution, the optimal decision line should be chosen somewhere between extremes by considering the tradeoff.

2.1 Methods for Cost-Sensitive Classification

In the problem of unequal misclassificaiton cost, the goal of cost-sensitive learning is to find the boundary between the regions that divide optimally the example space. Obviously the misclassification cost, particularly a cost table

[1] We assume that the employee is authorized to access if the requested documents is classified by the system as "class 1".

(e.g., table 1), is the dominant factor for the optimal boundaries. That is, the region where class j must be predicted will expand at the expense of the regions of other classes if misclassifying examples of class j is more expensive relative to misclassifying others, even though the class probabilities remain unchanged.

In this paper, we utilize two methods that convert arbitrary error-minimizing classifiers into cost-sensitive ones without modifying classification rules: costing [13] and metacost [2]. These methods make use of sampling techniques that change the original example distribution D to $\hat{D}$ by incorporating the relative misclassification cost of each instance, according to a given cost matrix. This changes the proportion of a certain class (e.g., documents that are "need-to-know" to perform a given project) by re-sampling of the original examples. Then the methods make any cost-insensitive error-minimizing classifiers to perform cost minimization on the newly generated distribution, $\hat{D}$.

Costing. Costing (cost proportionate rejection sampling with aggregation) is a wrapper for cost-sensitive learning that trains a set of error-minimizing classifiers by a distribution, which is the original distribution with the relative cost of each example, and outputs a final classifier by taking the average over all learned classifiers [13]. Costing is comprised of two processes: rejection sampling and bagging. Rejection sampling has been used to generate independently and identically distributed (i.i.d.) samples that are used as a proxy distribution to achieve simulation from the target distribution. Rejection sampling for the costing assigns each example in the original distribution with a relative cost [2] and draws a random number r from a uniform distribution $U(0,1)$. It will keep the example if $r > \frac{c}{Z}$. Otherwise it discards the example and continues sampling until certain criteria are satisfied. The accepted examples are regarded as a realization of the altered distribution, $\hat{D} = \{S_1^{'}, S_2^{'}, ..., S_k^{'}\}$. With the altered distribution, $\hat{D}$, costing trains k different hypotheses, $h_i \equiv Learn(S_i^{'})$, and predicts the label of a test example, $\mathbf{x}$, by combining those hypotheses, $h(\mathbf{x}) = sign\left(\sum_{i=1}^{k} h_i(\mathbf{x})\right)$.

MetaCost. MetaCost is another method for converting an error-minimizing classifier into cost-sensitive classifier by re-sampling [2]. The underlying assumption is that an error-minimizing classifier could learn the optimal decision boundary based on the cost matrix if each training example is relabeled with the cost. MetaCost's learning process is also comprised of two processes: bagging for relabeling and retraining the classifiers with cost. In particular, it generates a set of samples with replacement from the training set and estimates the class of each instance by taking the average of votes over all the trained classifiers. Then MetaCost re-labels each training example with the estimated optimal class and re-trains the classifier to the relabeled training set.

[2] $\hat{x}_i = \frac{c}{Z} \times x_i$, where c is a cost assigned to x_i and Z is a normalization factor, satisfying $\max_{c \in S} c$.

3 Experiments

The scenario which we are particularly interested in is a process of confidential access control based on the need-to-know principle. We model the decision whether to reject or accept the access request as a binary classification. In particular, our system classifies the content of the requested information as positive or negative with respect to the content of the requester's project and authorizes the request if the requested information is classified as positive to the requester's project. Otherwise the request is rejected. To this end, we choose three different classification methods, linear discriminant analysis (LDA), logistic regression (LR), and support vector machines (SVM), because of their relative good performance, particularly in text classification [7], [8], [11].

The purpose of the experiments is two-fold; (1) to find a good classification method that minimizes the cost and the false positive rate while holding the false negative rate reasonably low, (2) to verify that the cost-sensitive learning methods reduce the total cost for misclassification in comparison with error-minimizing classifiers. From these objectives, three performance metrics are primarily used to measure the usefulness of classifiers; false negative, defined as $fn = \frac{c}{a+c}$ by using the values in the table 1, false positive, $fp = \frac{b}{b+d}$, and cost for misclassification. These metrics are better matched to our purpose because we are interested in primarily reducing the error and the cost.

Since there are no datasets available that are comprised of confidential information, we choose the Reuters-21578 document collections for experiments. This data set, which consists of world news stories from 1987, has become a benchmark in text categorization evaluations. It has been partially labelled by human experts with respect to a list of categories. Since our task is a binary classification task where each document must be assigned to either positive or negative, we discarded documents that are assigned no topic or multiple topics. Moreover, classes with fewer than 10 documents are discarded. The resulting data set is comprised of 9,854 documents as a training set and 4,274 documents as a test set with 67 categories.

The experimental setting is as follows. All the documents are regarded as confidential. Documents belonging to the selected category are regarded as confidential information that the requester needs to know. Conversely the rest of test documents are confidential information that should not be revealed. A false positive occurs when a method classifies a document as positive that should have not been revealed whereas a false negative occurs when the method classifies a request as negative that should have been accepted. For both errors, the system pays the cost for misclassification. In the next section, we describe a method for cost assignment.

3.1 Cost Assignment

According to the class assignment – not the original Reuters-21578 category label, but the artificially assigned class label, such as need-to-know confidential or otherwise (simply, positive or negative) – each of the documents in both the

training and testing sets is assigned a cost, ensuring that the mis-classification cost of a need-to-know confidential information is higher than that of the remaining confidential documents (i.e., $\lambda_{10} > \lambda_{01},\ \lambda_{10} > \lambda_{00},\ \lambda_{01} > \lambda_{11}$) [3].

Since the Reuters-21578 document collection does not have cost information, we devised a heuristic for cost assignment. There is a cost involved in incorrect classification. Moreover, a higher cost is assigned to a false positive than a false negative. Particularly, the cost for misclassifying a confidential document, $\mathbf{d}_i$, is computed by:

$$cost(\mathbf{d}_i) = \begin{cases} [s, s + |c_j|] \text{ if } \mathbf{d}_i \in c_j \text{ and } c_j \text{ = positive} \\ \left[0, \frac{\sum_{s \in positive} cost(\mathbf{d}_s)}{|\text{number of negative documents}|}\right] \text{ Otherwise} \end{cases}$$

where $s = \ln\left(\frac{N}{|c_j|}\right) \times 100$, N is the total number of documents and $|c_j|$ is the number of documents belonging to the jth category. The total cost for misclassification is added to the cost of confidential documents misclassed if a classifier is not able to predict any of the positive cases, in order to prevent the case that a low cost is simply achieved by ignoring the class with a low frequency. For example, there are 15 out of 10,000 documents belonging to the positive class. The cost assignment ensures that the total cost for misclassifying those 15 examples should be either equal to or higher than that of the remaining documents [3].

3.2 Experimental Results

From the 67 selected categories of the Reuters-21578 dataset, we choose the five different categories as representative ones according to their category frequencies: small ("livestock" and "corn"), medium ("interest"), and large ("acq" and "earn"). There are 70% of documents in a category (e.g., the "livestock" category) used as "training" and the remaining 30% documents are used for "testing", respectively. There are nine different classifiers tested: LDA, LR, and SVMs, and the combination of those three classifiers with two methods for cost-sensitive learning: metacost and costing. A binary classifier was trained for each of the selected categories by considering the category as positive (i.e., documents that an employee needs to know) with the rest of the data as negative examples. We made use of the LIBSVM[4] and tested three different kernels, such as linear, polynomial, and Gaussian. The Gaussian kernel (i.e., $width = \frac{1}{\text{max feature dimension}}$) was chosen due to its best performance and the different cost factors are assigned [5], $C = 10 \sim 100$. Those values are chosen optimally by 10-fold cross validation.

[3] For this case, the cost for misclassifying a positive document is 650.0789 ($= \ln\left(\frac{9985}{15}\right) \times 100$) and the sum of the cost is 9751.1835 (=650.0789 × 15). Accordingly the cost of misclassification of a negative document is 0.9765 ($= \frac{9751.1835}{9985}$) and the cost sums to 9751.1835 ($= 0.9765 \times 9985$).

[4] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

[5] The cost of constrain violation is set to 100 if there are relatively small amount of positive examples available. Otherwise it is set to about 10.

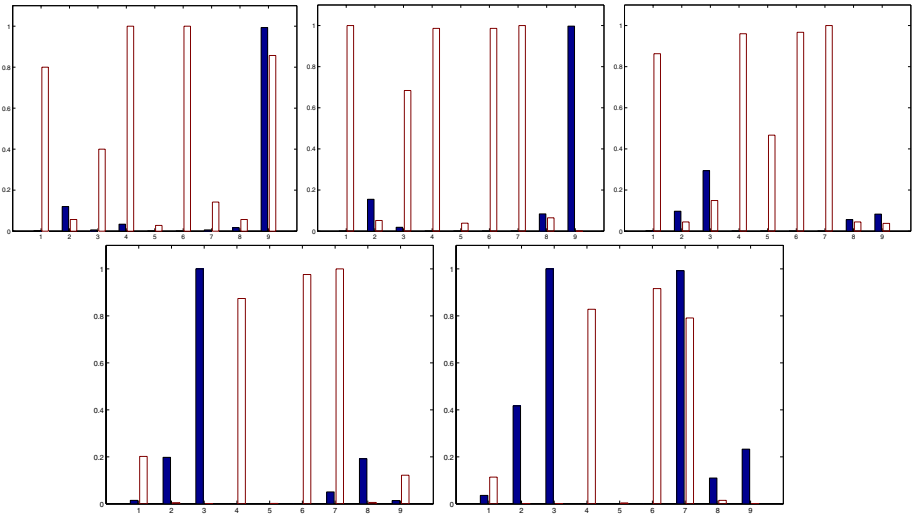

Fig. 2. Pairs of false positive (filled bar) and false negative (empty bar) for three selected categories by nine different classifiers, which are numbered from the left to right: (1) SVM, (2) SVM with costing, (3) SVM with metacost, (4) LR, (5) LR with costing, (6) LR with metacost, (7) LDA, (8) LDA with costing, and (9) LDA with metacost, respectively. From the top left, the results for "livestock," "corn," "interest," "acq," and "earn" are presented

As mentioned earlier, the experimental results are primarily analyzed by "false positive," "false negative," and "cost." The procedure of experiments is as follows: firstly, pick one of five selected categories; secondly, assign the cost to each of documents according to its importance using the heuristic described in section 3.1; then, train each of nine classifiers by training examples with cost; finally compute three performance measure (i.e., false positive, false negative, and cost for incorrect classification). Figure 2 shows pairs of false positive and false negative for the three selected categories by nine different classifiers. Except the "interest" category, LR with the costing showed the best results that minimize false positive while holding false negative low. In particular, for the "livestock" category, LR trained by only 18% training data (i.e., 1,781 out of 9,854 documents) resulted 0% false positive and 2.8% false negative rate. For the costing, we carried out five different sampling trials for each category (i.e., 1, 3, 5, 10, and 15) and represented the trial for the best performance. For this category, a newly generated distribution by 10 rejection sampling trials is used to achieve this result. Each resampled set has only about 178 documents. LDA with the costing showed the smallest error for the "interest" category that is comprised of 5.6% false positive and 4.5% false negative.

Table 2 replicates this trend in terms of the total cost for misclassification. The number in parenthesis next to topic name in table 2 is the total number of text documents belonging to that category. The results reported for the costing

Table 2. The cost for misclassification by nine different classifiers are presented. The values in bold face are the best for corresponding category.

Methods	livestock (114)	corn (253)	interest (513)	acq (2448)	earn (3987)
SVM	13967	66453	54065	83141	108108
SVM (w/costing)	4035 ± 30	8851 ± 52	9058 ± 159	40009 ± 252	96007 ± 331
SVM (w/mc)	7147 ± 50	23596 ± 64	32011 ± 321	194165 ± 451	228612 ± 453
LR	35809	32759	60031	349080	710631
LR (w/costing)	**484 ± 11**	**1333 ± 44**	29614 ± 110	**606 ± 145**	**2521 ± 191**
LR (w/mc)	34980 ± 35	32759 ± 79	60374 ± 154	386859 ± 1185	788819 ± 263
LDA	2638	66453	124733	591300	908690
LDA (w/costing)	1461 ± 28	6092 ± 89	**7301 ± 152**	39354 ± 205	41478 ± 159
LDA (w/mc)	40079 ± 57	45778 ± 71	8955 ± 157	51789 ± 285	54084 ± 244
Cost for base line	42625	79084	139113	591357	1090498

and the metacost are the average of 5 different runs. The bottom line entitled "cost for base line" is the total cost for a category if a classifier classifies all the testing examples incorrectly (e.g., the misclassification cost of a classifier for "livestock" will be 42,625 if the classifier classifies all incorrectly). For the "earn" category, LR with the costing caused only 0.002 out of the total cost (2,521 out of 1,090,498). For the remaining categories, the best-performer paid only less than 0.05 out of the total cost.

From the comparison with error-minimizing classifiers, the costing proved its effectiveness in that it requires relatively small amount of training data for a better performance. For the "corn" category, LR with the costing, which only used 10% of the training data (i.e., 986 out of 9,854 documents) showed the best result in terms of the smallest loss (1,333 out of 79,084), zero false positive, and lower false negative rate (0.039). The LR classifier was trained by a sample set by three rejection sampling trials that is comprised of 458 positive and 528 negative examples. The smallest cost implies that it is expected to pay 1.1% of the total cost caused by incorrect confidential access control (i.e., misclassification). From the false positive perspective (zero false alarm), there is no leaking of confidential information. 39% false negative rate means that there would be 39 out of 1,000 valid requests to the confidential information that are mistakenly rejected. This inconveniences employees because they have to access particular information for their projects, but the system does not authorize their access requests. This trend holds good for the remaining four categories.

4 Related Work

Weippl and Ibrahim [12] proposed content-based management of text document access control. They applied a self-organized map (SOM) to cluster a given collection of text documents into groups which have similar contents. This approach also allowed humans to impose dynamic access control to identified text document groups. However they did not address a potential problem that occurs

when the security policy for individual documents of a cluster does not match with the security policy for that cluster. Giuri and Iglio [5] proposed an approach that determines a user's access to confidential information, which is based on the content of the information and the role of the user. For example, they consider subdividing medical records into several different categories (e.g., pediatrics), and allow that only relevant physicians (e.g., pediatrician) can access them. Since they do not mention automatic techniques in their paper, one is left with the suspicion that they manually categorize content and roles. Aleman-Meza and his colleagues proposed an ontological approach to deal with the legitimate document access problem of insider threat [1]. An attempt to access document is regarded as legitimate if the job assignment of a requester (e.g., an intelligence analyst) has a semantic association with the documents that are accessed. This approach is quite similar to ours in that they enforce the need-to-know principle by using a predefined ontology. A well-defined ontology might be useful to determine the semantic associations between the existing documents and the analysts' assignments, but regular updates are required to accommodate the change of the document collections and the topics of assignments. Symonenko and his colleagues propose a hybrid approach that combines role-based access monitoring, social network analysis, and semantic analysis of insiders' communications, in order to detect inappropriate information exchange [10]. Lee and his colleagues [6] introduced a cost-sensitive framework for the intrusion detection domain and analyzed cost factors in detail. Particularly, they identify the major cost factors (e.g., costs for development, operation, damages and responding to intrusion) and then applied a rule induction learning technique (i.e., RIPPER) to this cost model, in order to maximize security while minimizing costs. However their cost model needs to be changed manually if a system's cost factors are changed.

5 Conclusion and Future Work

In the scenario of confidential access control based on the need-to-know principle, a false positive occurs when the system accepts a request that should not have been accepted whereas a false negative occurs when the system rejects a request that should have been accepted. For both errors, the system pays the cost for misclassification. From the security perspective, it is more tolerable to have an authorization process with a high false negative rather than one with a high false positive rate because the latter is a serious security problem since confidential information, which should not be revealed, can be accessed.

In this paper we test the effectiveness of cost-sensitive learning for confidential access control and improve our previous results by taking into consideration the cost caused by misclassification. To this end, we model the binary decision whether to reject or accept the request in a cost-sensitive learning framework, where the cost caused by incorrect decision for the request is different according to the relative importance of the requested information. In addition, we invented a cost assignment method that ensures that the mis-classification cost of

a need-to-know confidential information is higher than that of the other confidential information. Finally we tested three different error-minimizing classification methods.

From the comparison of the cost-sensitive learning methods with the error-minimizing classification methods, we found that the costing with a logistic regression showed the best performance. In particular, it requires far less training data for much better results, in terms of the smallest cost paid, the lowest false positive rate, and the relatively low false negative rate. The smallest cost implies that it is expected to pay 1.1% of the total cost caused by incorrect confidential access control. The nearly zero false positive rate means that there is no leaking of confidential information. The benefit of smaller training data is two-fold; First, obviously it takes less time to train the classifier; Second, it enables a human administrator to conveniently identify arbitrary subsets of confidential information, in order to train the initial classifier. In other words, through our proposed methods, it becomes easier for a human administrator to define, assign, and enforce an effective access control for a particular subset of confidential information. Although our approach demonstrates a promising result, we believe that such a content-based approach should be used as a complementary tool for a human administrator.

Although to our knowledge, the cost-sensitive learning approach is a novel one for confidential access control, it would be very interesting if we compare the effectiveness of our framework with conventional document management systems (e.g., ACL-based systems) and knowledge-intensive approaches (e.g., ontology-based systems) as future work.

Acknowledgement

This research was partially supported by award BAA-03-02-CMU from Computer Technology Associates to Carnegie Mellon University and in part by AFOSR contract number F49640-01-1-0542.

References

1. Aleman-Meza, B., Burns, P., Eavenson, M., Palaniswami, D., and Sheth, A., An ontological approach to the document access problem of insider threat, In *Proceedings of IEEE International Conference on Intelligence and Security Informatics* (ISI-05), pp. 486-491, 2005.
2. Domingos, P., MetaCost: A general method for making classifiers cost-sensitive, In *Proceedings of International Conference on Knowledge Discovery and Data Mining* (KDD-99), pp. 155-164, 1999.
3. Elkan, C., The foundations of cost-sensitive learning, In *Proceedings of International Joint Conference on Artificial Intelligence* (IJCAI-01), pp. 973-978, 2001.
4. Fawcett, T., ROC graphs: Notes and practical considerations for researchers, HP Lab Palo Alto, *HPL-2003-4*, 2003.
5. Giuri, L. and Iglio, P., Role templates for content-based access control, In *Proceedings of ACM Workshop on Role Based Access Control*, pp. 153-159, 1997.

6. Lee, W., Miller, M., Stolfo, S., Jallad, K., Park, C., Zadok, E., and Prabhakar, V., Toward cost-sensitive modeling for intrusion detection, *ACM Journal of Computer Society*, Vol. 10, No. 1-2, pp. 5-22, 2002.
7. Joachims, T., Text categorization with support vector machines: Learning with many relevant features, In *Proceedings of European Conference on Machine Learning* (ECML-98), 1998.
8. Schutze, H., Hull, D.A., and Pedersen, J.O., A comparison of classifiers and document representations for the routing problem, In *Proceedings of International ACM Conference on Research and Development in Information Retrieval* (SIGIR-95), pp. 229-237, 1995.
9. Seo, Y.-W., Giampapa, J., and Sycara, K., A multi-agent system for enforcing Need-To-Know security policies, In *Proceedings of International Conference on Autonomous Agents and Multi Agent Systems (AAMAS) Workshop on Agent Oriented Information Systems* (AOIS-04), pp. 163-179, 2004.
10. Symonenko, S., Liddy, E.D., Yilmazel, O., Semantic analysis for monitoring insider threats, In *Proceedings of Symposium on Intelligence and Security Informatics*, 2004.
11. Torkkola, T., Linear discriminant analysis in document classification, In *IEEE Workshop on TextMining*, 2001.
12. Weippl, E. and Ibrahim, K., Content-based management of document access control, In *Proceedings of the 14th International Conference on Applications of Prolog*, 2001.
13. Zadrozny, B., Langford, J., and Abe, N., A simple method for cost-sensitive learning, *IBM Tech Report*, 2002.