# Test-Cost Sensitive Classification Based on Conditioned Loss Functions

Mumin Cebe and Cigdem Gunduz-Demir

Department of Computer Engineering
Bilkent University
Bilkent, Ankara 06800, Turkey
{mumin,gunduz}@cs.bilkent.edu.tr

**Abstract.** We report a novel approach for designing test-cost sensitive classifiers that consider the misclassification cost together with the cost of feature extraction utilizing the consistency behavior for the first time. In this approach, we propose to use a new Bayesian decision theoretical framework in which the loss is conditioned with the current decision and the expected decisions after additional features are extracted as well as the consistency among the current and expected decisions. This approach allows us to force the feature extraction for samples for which the current and expected decisions are inconsistent. On the other hand, it forces not to extract any features in the case of consistency, leading to less costly but equally accurate decisions. In this work, we apply this approach to a medical diagnosis problem and demonstrate that it reduces the overall feature extraction cost up to 47.61 percent without decreasing the accuracy.

## 1 Introduction

In classification, different types of cost have been investigated till date [1]. Among these costs, the most commonly investigated one is the *cost of misclassification errors* [2]. Compared to the misclassification cost, the other types are much less studied. The *cost of computation* includes both static complexity, which arises from the size of a computer program [3], and dynamic complexity, which is incurred during training and testing a classifier [4]. The *cost of feature extraction* arises from the effort of acquiring a feature. This type of cost is especially important in some real-world applications such as medical diagnosis in which one would like to balance the diagnosis accuracy with the cost of medical tests used for acquiring features.

In machine learning literature, a number of studies have investigated the cost of feature extraction [5,6,7,8,9,10,11,12,13,14,15]. The majority of these studies focus on the construction of decision trees in a least costly manner by selecting features based on both their information gain and their extraction cost [5,6,7,8,9]. While the earlier studies [5,6,7] consider only the feature extraction cost, more recent ones [8,9] consider the misclassification cost as well. Another group of studies focuses on the sequential feature selection also based on the information gain of features and their extraction cost [10,11,12]. In these studies, the

J.N. Kok et al. (Eds.): ECML 2007, LNAI 4701, pp. 551–558, 2007.

gain is measured as the difference in the amount of information before and after extracting the features. As the information after feature extraction cannot be known in advance, these studies estimate this information making use of maximum likelihood estimation [10], dynamic Bayesian networks [11], and neural networks [12]. The theoretical aspects of such feature selection are also studied in [13]. The other group of studies considers the feature selection as optimal policy learning and solves it formulating the classification problem as a Markov decision process [14] and a partially observable Markov decision process [15]. All of these studies select features based on the current decision and those obtained after features are extracted. None of them considers the consistency between these decisions.

In this paper, we report a novel cost-sensitive learning approach that takes into consideration the misclassification cost together with the cost of feature extraction utilizing the consistency behavior for the first time. In this approach, we make use of a Bayesian decision theoretical framework in which the loss function is conditioned with the current decision and the estimated decisions after the additional features are extracted in conjunction with the *consistency* among the current and estimated decisions. Using this proposed approach, the system tends to extract features that are expected to change the current decision (i.e., yield inconsistent decisions). It also tends to stop the extraction if all possible features are expected to confirm the current decision (i.e., yield consistent decisions), leading to less costly but equally accurate decisions. In this paper, working with a medical diagnosis problem, we demonstrate that the overall feature extraction cost is reduced up to 47.61% without decreasing the classification accuracy. To the best of our knowledge, this is the first demonstration of the use of conditioned loss functions for the purpose of test-cost sensitive classification.

## 2 Methodology

In our approach, we propose to use a Bayesian decision theoretical framework in which the loss function is conditioned with the current and estimated decisions as well as their consistency. For a given instance, the proposed approach decides whether or not to extract a feature, and in the case of deciding in favor of extraction, which feature to be extracted by using conditional risks computed with the new loss function definition.

In Bayesian decision theory, decision has to be made in favor of the action for which the conditional risk is minimum. For instance $x$, the conditional risk of taking action $\alpha_i$ is defined as

$$R(\alpha_i|x) = \sum_{j=1}^{N} P(C_j|x)\ \lambda(\alpha_i|C_j) \tag{1}$$

where $\{C_1, C_2, ..., C_N\}$ is the set of $N$ possible states of nature and $\lambda(\alpha_i|C_j)$ is the loss incurred for taking action $\alpha_i$ when the actual state of nature is $C_j$. In

**Table 1.** Definition of the conditioned loss function for feature extraction, classification, and reject actions

| | $\texttt{extract}_k$ | $\texttt{classify}$ | $\texttt{reject}$ |
|---|---|---|---|
| **Case 1:** $C_{\texttt{actual}} = C_{\texttt{curr}} = C_{\texttt{est}_k}$ | $\texttt{cost}_k$ | $-\texttt{REWARD}$ | $\texttt{PENALTY}$ |
| **Case 2:** $C_{\texttt{actual}} \neq C_{\texttt{curr}} \neq C_{\texttt{est}_k}$ | $\texttt{cost}_k + \texttt{PENALTY}$ | $\texttt{PENALTY}$ | $-\texttt{REWARD}$ |
| **Case 3:** $C_{\texttt{curr}} = C_{\texttt{est}_k} \neq C_{\texttt{actual}}$ | $\texttt{cost}_k + \texttt{PENALTY}$ | $\texttt{PENALTY}$ | $-\texttt{REWARD}$ |
| **Case 4:** $C_{\texttt{actual}} = C_{\texttt{curr}} \neq C_{\texttt{est}_k}$ | $\texttt{cost}_k + \texttt{PENALTY}$ | $-\texttt{REWARD}$ | $\texttt{PENALTY}$ |
| **Case 5:** $C_{\texttt{actual}} = C_{\texttt{est}_k} \neq C_{\texttt{curr}}$ | $\texttt{cost}_k - \texttt{REWARD}$ | $\texttt{PENALTY}$ | $\texttt{PENALTY}$ |

our approach, we consider $C_j$ as the class that an instance can belong to and $\alpha_i$ as one of the following actions:

(a) $\texttt{extract}_k$: extract feature $F_k$,
(b) `classify`: stop the extraction and classify the instance using the current information, and
(c) `reject`: stop the extraction and reject the classification of the instance.

In the proposed framework, we use a new loss function definition in which the loss is conditioned with the current and estimated decisions along with their consistency. The loss function for each of the aforementioned actions is given in Table 1. In this table, $C_{\texttt{actual}}$ is the actual class, $C_{\texttt{curr}}$ is the class estimated by the current classifier, and $C_{\texttt{est}_k}$ is the estimated class when feature $F_k$ is extracted. Here, $C_{\texttt{actual}}$ and $C_{\texttt{est}_k}$ should be estimated using the current information as it is not possible to know these values in advance.

As shown in Table 1, for a particular action, the loss function takes different values based on the consistency among the actual ($C_{\texttt{actual}}$), current ($C_{\texttt{curr}}$), and estimated ($C_{\texttt{est}_k}$) classes. In this definition, the actions that lead to correct classifications and the action that rejects the classification when the correct classification is not possible are rewarded with an amount of `REWARD` value by adding $-\texttt{REWARD}$ to the loss function. When there are more than one feature that could be extracted, `reject` action is rewarded only if none of the classifiers using each of these features could yield the correct classification. On the contrary, the actions that lead to misclassifications and the action that rejects the classification when the correct classification is possible are penalized with an amount of `PENALTY` value. Additionally, the extraction cost ($\texttt{cost}_k$) is included in the loss function when feature $F_k$ is to be extracted. In this definition of loss function, the only exception that does not follow these rules is the case of $\texttt{extract}_k$ action in Case 1. In this case, although it yields the correct classification, this action is not rewarded since it does not provide any additional information but brings about an extra feature extraction cost. By doing so, for Case 1, we force the algorithm not to extract an additional feature.

For a particular instance $x$, we express the conditional risk of each action using the definition of loss function above. With $\mathcal{C} = \{C_{\texttt{curr}}, C_{\texttt{est}_1}, C_{\texttt{est}_2}, ..., C_{\texttt{est}_M}\}$

being the set of the current class and the classes estimated after extracting each feature, the conditional risk of the $\mathtt{extract}_k$ action is defined as follows.

$$R\ (\ \mathtt{extract}_k|x, \mathcal{C}) = \sum_{j=1}^{N} P(C_{\mathtt{actual}} = j|x)\ \times \tag{2}$$

$$\begin{bmatrix} P(C_{\mathtt{curr}}{=}j|x)\ P(C_{\mathtt{est}_k} = j|x)\ \mathtt{cost}_k\ + \\ P(C_{\mathtt{curr}}{\neq}j|x)\ P(C_{\mathtt{est}_k}{\neq}j|x)\ P(C_{\mathtt{curr}}{=}C_{\mathtt{est}_k}|x)\ [\mathtt{cost}_k + \mathtt{PENALTY}]\ + \\ P(C_{\mathtt{curr}}{\neq}j|x)\ P(C_{\mathtt{est}_k}{\neq}j|x)\ P(C_{\mathtt{curr}}{\neq}C_{\mathtt{est}_k}|x)\ [\mathtt{cost}_k + \mathtt{PENALTY}]\ + \\ P(C_{\mathtt{curr}}{=}j|x)\ P(C_{\mathtt{est}_k}{\neq}j|x)\ [\mathtt{cost}_k + \mathtt{PENALTY}]\ + \\ P(C_{\mathtt{curr}}{\neq}j|x)\ P(C_{\mathtt{est}_k}{=}j|x)\ [\mathtt{cost}_k - \mathtt{REWARD}] \end{bmatrix}$$

$$R\ (\ \mathtt{extract}_k|x, \mathcal{C}) = \sum_{j=1}^{N} P(C_{\mathtt{actual}} = j|x)\ \times \tag{3}$$

$$\left[\begin{array}{lll} P(C_{\mathtt{curr}}{=}j|x) & P(C_{\mathtt{est}_k}{=}j|x) & \mathtt{cost}_k\ + \\ {[1 - P(C_{\mathtt{curr}}{=}j|x)]} & [1 - P(C_{\mathtt{est}_k}{=}j|x)] & [\mathtt{cost}_k + \mathtt{PENALTY}]{+} \\ P(C_{\mathtt{curr}}{=}j|x) & [1 - P(C_{\mathtt{est}_k}{=}j|x)] & [\mathtt{cost}_k + \mathtt{PENALTY}]{+} \\ {[1 - P(C_{\mathtt{curr}}{=}j|x)]} & P(C_{\mathtt{est}_k}{=}j|x) & [\mathtt{cost}_k - \mathtt{REWARD}] \end{array}\right]$$

$$R\ (\ \mathtt{extract}_k|x, \mathcal{C}) = \sum_{j=1}^{N} P(C_{\mathtt{actual}} = j|x)\ \times \tag{4}$$

$$\begin{bmatrix} \mathtt{cost}_k\ + \\ {[1 - P(C_{\mathtt{est}_k} = j|x)]}\ \mathtt{PENALTY}\ + \\ P(C_{\mathtt{est}_k} = j|x)\ [1 - P(C_{\mathtt{curr}} = j)|x]\ [-\mathtt{REWARD}] \end{bmatrix}$$

Equation 4 implies that the extraction of feature $F_k$ requires paying for its cost. It also implies that the $\mathtt{extract}_k$ action is penalized with PENALTY if the class estimated after feature extraction is incorrect and is rewarded with REWARD if this estimated class is correct but it is different than the currently estimated class. Similarly, for a particular instance $x$, we derive the conditional risk of the classify and the reject actions in Equations 5 and 6, respectively.

$$R\ (\ \mathtt{classify}|x, \mathcal{C}) = \sum_{j=1}^{N} P(C_{\mathtt{actual}} = j|x)\ \times \tag{5}$$

$$\big[P(C_{\mathtt{curr}} = j|x)\ [-\mathtt{REWARD}] + [1 - P(C_{\mathtt{curr}} = j|x)]\ \mathtt{PENALTY}\big]$$

$$R\ (\ \mathtt{reject}|x, \mathcal{C}) = \sum_{j=1}^{N} P(C_{\mathtt{actual}} = j|x)\ \times \tag{6}$$

$$\begin{bmatrix} \left[[1 - P(C_{\mathtt{curr}} = j|x)]\ \prod_{m=1}^{M} [1 - P(C_{\mathtt{est}_m} = j|x)]\right]\ [-\mathtt{REWARD}]\ + \\ \left[1 - [1 - P(C_{\mathtt{curr}} = j|x)]\ \prod_{m=1}^{M} [1 - P(C_{\mathtt{est}_m} = j|x)]\right]\ \mathtt{PENALTY}\ + \end{bmatrix}$$

Equation 5 means that classifying the instance with the current classifier (classify action) is rewarded with REWARD if this is a correct classification and is penalized with PENALTY otherwise. Equation 6 means that rejecting the classification is only rewarded with REWARD if neither the estimated classes nor the current class is correct; otherwise, it is penalized with PENALTY.

As given in Equations 4, 5, and 6, the conditional risks are computed using the posterior probabilities. Posterior probabilities $P(C_{\texttt{curr}} = j|x)$ can be calculated by the current classifier before any possible feature extraction, since all of its features are already extracted. On the other hand, posterior probabilities $P(C_{\texttt{est}_k} = j|x)$ could not be known prior to extracting feature $F_k$. Thus, these posteriors should be estimated making use of the currently available information. For that, we use estimators which are trained as follows: First, we learn the parameters of the classifier $Y_k$ that use both the previously extracted features and feature $F_k$ on training samples. Then, for each of these samples, we compute the posterior probabilities using the classifier $Y_k$. Subsequently, we train the estimators to learn these posteriors by using only the previously extracted features. Note that similar posterior probability estimations have been achieved by using linear perceptrons [4] and dynamic Bayesian networks [11].

In the computation of posterior probabilities $P(C_{\texttt{actual}} = j|x)$ in Equation 4, we employ the posteriors computed for the current classifier as well as those estimated for the classifiers whose features are to be extracted. To do so, for each class, we multiply the corresponding posteriors, and then normalize them such that $\sum_{j=1}^{N} P(C_{\texttt{actual}} = j|x) = 1$. For Equations 5 and 6, we only use the posterior probabilities of the current classifier, since the corresponding actions (`classify` and `reject`) require stopping feature extraction, and thus, no additional features are extracted after taking these actions.

In order to dynamically select a subset of features for the classification of a given instance $x$, our algorithm first computes the conditional risk of the `classify` action, the $\texttt{extract}_k$ action for each feature $F_k$ that is not extracted yet, and the `reject` action as given in Equations 4, 5, and 6, and then selects the action for which the conditional risk is minimum. This selection is sequentially conducted until either the `classify` or the `reject` action is selected.

## 3 Experiments

We conduct our experiments on the Thyroid Dataset[1] in which there are three classes (hypothyroid, hyperthyroid, and normal) and 21 features. The first 16 features are based on the answers of the questions that are asked to a patient; thus, we assign no cost to them. The next four features are obtained from the blood tests and the assigned cost of these blood tests is {\$22.78, \$11.41, \$14.51, \$11.41}. The last feature is calculated from the nineteenth and twentieth features; we use the last feature in classification only if these two features are already extracted.

In our experiments, we use decision tree classifiers and Parzen window estimators whose window function defines hypercubes. We train both classifiers and estimators on the training set. For Parzen window estimators, the test set

[1] This dataset is available at the UCI repository [16].

**Table 2.** Confusion matrix for the test set when our test-cost sensitive classification algorithm is used. Here, the reduction in the overall feature extraction cost is 47.61%.

| | | Selected class | | | Reject |
|---|---|---|---|---|---|
| | | Hypothyroid | Hyperthyroid | Normal | cases |
| Actual class | Hypothyroid | 70 | 0 | 0 | 3 |
| | Hyperthyroid | 0 | 173 | 0 | 4 |
| | Normal | 13 | 23 | 3140 | 2 |

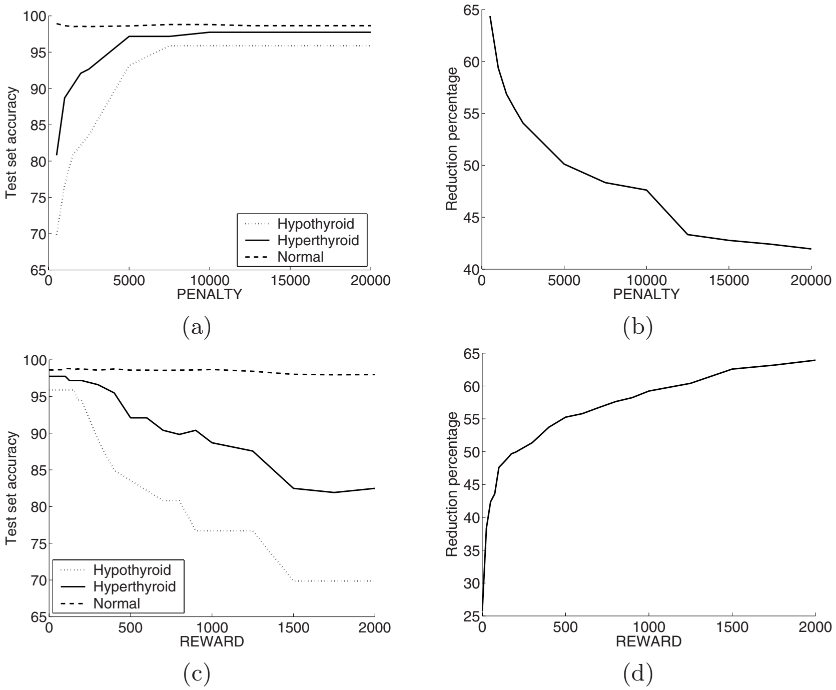
**Table 3.** Confusion matrix for the test set when all features are used in classification

| | | Selected class | | |
|---|---|---|---|---|
| | | Hypothyroid | Hyperthyroid | Normal |
| Actual class | Hypothyroid | 70 | 0 | 3 |
| | Hyperthyroid | 0 | 173 | 4 |
| | Normal | 13 | 29 | 3136 |

includes some samples for which there is no training sample falling in the specified hypercubes. For these samples, we do not penalize any feature extraction since the estimators provide no information and we consider only the posteriors obtained on the current classifier to compute the conditional risks.

In Table 2, we report the test results obtained by our algorithm. In this table, we provide the confusion matrix for the test set, indicating the number of samples for which the `reject` action is taken. These results are obtained when `REWARD` and `PENALTY` values are selected to be 100 and 10000, respectively. For comparison, in Table 3, we also report the confusion matrix for the test set when all features are used in classification; here, we also use a decision tree classifier (herein referred to as *all-feature-classifier*). Tables 2 and 3 demonstrate that, compared to the *all-feature-classifier*, our algorithm yields the same number of correct classifications for hypothyroid and hyperthyroid classes. Moreover, for these classes, our algorithm does not lead to any misclassification. For the samples misclassified by the *all-feature-classifier*, our algorithm takes the `reject` action, reducing the overall misclassification cost. Furthermore, for normal class, our algorithm yields a larger number of correct classifications. For the selected parameters, the decrease in the overall feature extraction cost is 47.61%. This demonstrates that the proposed algorithm significantly decreases the overall feature extraction cost without decreasing the accuracy.

In the proposed algorithm, there are two free model parameters: `REWARD` and `PENALTY`. Next, we investigate the effects of these parameters on the classification accuracy and the reduction in the overall cost of feature extraction. For that, we fix one of these parameters and observe the accuracy and the cost reduction in feature extraction as a function of the other parameter. In Figures 1(a) and 1(b), we present the test set accuracy, for each individual class, and the percentage of the reduction in the overall feature extraction cost as a function of the `PENALTY` value when `REWARD` is set to 100. These figures demonstrate that as the penalty



(a) (b)

(c) (d)

**Fig. 1.** For our test-cost sensitive classification algorithm, (a)-(b) the test set accuracy and the percentage of the cost reduction as a function of PENALTY when REWARD is set to 100, and (c)-(d) the test set accuracy and the percentage of the cost reduction as a function of REWARD when PENALTY is set to 10000.

of misclassifications and selecting the `reject` action increases, the number of correctly classified samples, for especially hypothyroid and hyperthyroid classes, increases too. With the increasing `PENALTY` value, the algorithm tends to extract more number of features not to misclassify the samples, leading to the decrease in the cost reduction. Similarly, in Figures 1(c) and 1(d), we present the test set accuracy and the percentage of the cost reduction as a function of the `REWARD` value when `PENALTY` is set to 10000. These figures demonstrate that the test set accuracy for hypothyroid and hyperthyroid classes decreases with the increasing `REWARD` value. As shown in Equations 4, 5, and 6, as the `REWARD` value increases, the conditional risks decrease. The factor that affects the conditional risks for all actions is $P(C_{\texttt{curr}})$. While this decrease is proportional to $P(C_{\texttt{curr}})$ for the `classify` action, it is proportional to $[1-P(C_{\texttt{curr}})]$ for the $\texttt{extract}_k$ and `reject` actions. This indicates that when $P(C_{\texttt{curr}})$ is just slightly larger than $[1-P(C_{\texttt{curr}})]$ (e.g., 0.51), the decrease in the conditional risk for the `classify` action is larger. Thus, as the `REWARD` value increases, the algorithm tends to classify the samples without extracting additional features. While this decreases the classification accuracy, it increases the cost reduction.

## 4 Conclusion

This work introduces a novel Bayesian decision theoretical framework to incorporate the cost of feature extraction into the cost of misclassification errors utilizing the consistency behavior for the first time. In this framework, the loss function is conditioned with the current decision and the estimated decisions that are to be taken after the feature extraction as well as the consistency among the current and the estimated decisions. By using this framework, we propose a new test-cost sensitive learning algorithm that selects a subset of features, dynamically for each instance. The experiments on a medical diagnosis dataset demonstrate that the proposed algorithm leads to a significant decrease (47.61%) in the feature extraction cost without decreasing the classification accuracy.

## References

1. Turney, P.D.: Types of cost in inductive concept learning. In: Workshop on Cost-Sensitive Learning. ICML 2000, Stanford, CA (2000)
2. Duda, O.R., Hart, E.P., Stork, G.D.: Pattern Classification. Wiley-Interscience, New York (2001)
3. Turney, P.D.: Low size-complexity inductive logic programming: The East-West Challenge considered as a problem in cost-sensitive classification. In: ILP 1995 (1995)
4. Demir, C., Alpaydin, E.: Cost-conscious classifier ensembles. Pattern Recognit Lett. 26, 2206–2214 (2005)
5. Norton, S.W.: Generating better decision trees. In: IJCAI 1989, Detroit, MI (1989)
6. Nunez, M.: The use of background knowledge in decision tree induction. Mach. Learn. 6, 231–250 (1991)
7. Tan, M.: Cost-sensitive learning of classification knowledge and its applications in robotics. Mach. Learn. 13, 7–33 (1993)
8. Turney, P.D.: Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. J. Artif. Intell. Res. 2, 369–409 (1995)
9. Davis, J.V., Ha, J., Rossbach, C.J., Ramadan, H.E., Witchel, E.: Cost-sensitive decision tree learning for forensic classification. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) ECML 2006. LNCS (LNAI), vol. 4212, Springer, Heidelberg (2006)
10. Yang, Q., Ling, C., Chai, X., Pan, R.: Test-cost sensitive classification on data missing values. IEEE T Knowl. Data. En. 18, 626–638 (2006)
11. Zhang, Y., Ji, Q.: Active and dynamic information fusion for multisensor systems with dynamic Bayesian networks. IEEE T. Syst. Man. Cy. B 36 (2006)
12. Gunduz, C.: Value of representation in pattern recognition. M.S. thesis, Bogazici University, Istanbul, Turkey (2001)
13. Greiner, R., Grove, A.J., Roth, D.: Learning cost-sensitive active classifiers. Artif. Intell. 139, 137–174 (2002)
14. Zubek, V.B., Dietterich, T.G.: Pruning improves heuristic search for cost-sensitive learning. In: ICML 2002, San Francisco, CA (2002)
15. Ji, S., Carin, L.: Cost-sensitive feature acquisition and classification. Pattern Recogn 40, 1474–1485 (2007)
16. Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases (1998), Available at http://www.ics.uci.edu/~mlearn/MLRepository.html