Genomics & Informatics Vol. 1(2) 80-86, December 2003

Classification of Human Papillomavirus (HPV) Risk Type via Text Mining

Seong-Bae Park*, Sohyun Hwang and Byoung-Tak Zhang

Biointelligence Lab., School of Computer Science and Engineering, Seoul National University, Seoul 151-744, Korea

Abstract

Human Papillomavirus (HPV) infection is known as the main factor for cervical cancer which is a leading cause of cancer deaths in women worldwide. Because there are more than 100 types in HPV, it is critical to discriminate the HPVs related with cervical cancer from those not related with it. In this paper, the risk type of HPVs using their textual explanation. The important issue in this problem is to distinguish false negatives from false positives. That is, we must find high-risk HPVs as many as possible though we may miss some low-risk HPVs. For this purpose, the AdaCost, a cost-sensitive learner is adopted to consider different costs between training examples. The experimental results on the HPV sequence database show that the consideration of costs gives higher performance. The improvement in F-score is higher than that of the accuracy, which implies that the number of high-risk HPVs found is increased.

Keywords: human papillomavirus, cost-sensitive learning, naive Bayes classifier, text classification

Introduction

Cervical cancer is a leading cause of cancer deaths in women worldwide. It, moreover, is the first cause of cancer deaths in Korean women. Since the main etiologic factor for cervical cancer is known as high-risk Human Papillomavirus (HPV) infection (Schiffman et al., 1993), it is now largely a preventable disease. HPV is a double-strand DNA tumor virus that belongs to the papovavirus family. There are more than 100 types of HPV that are specific for epithelial cells including skin, respiratory mucosa, and the genital tract. Genital tract HPV types are classified by their relative malignant potential into low-risk and high-risk types (Janicek *et al.*, 2001). The common, unifying oncogenic feature of the vast majority of cervical cancers is the presence of high-risk HPV. Therefore, the most important thing for diagnosis and therapy is discriminating whether patients have the high-risk HPVs and what HPV types are highly risky.

One way to discriminate the risk types of HPVs is using a text mining technique. Since a great number of research results on HPV have been already reported in biomedical journals (Furumoto and Irahara, 2002; Ishji, 2000), they can be used as a source of discriminating HPV risk types. One problem in discriminating the risk types is that the costs of high-risk HPVs and low-risk HPVs are not identical. This is because high-risk HPVs are seldom while low-risk HPVs are abundant. In addition, in classifying the risk types of HPVs, it is important to distinguish false negatives from false positives. That is, it is not critical to classify the low-risk HPVs as high-risk ones, because they can be investigated by further empirical study. However, it is fatal to classify the high-risk HPVs as low-risk ones. In this case, dangerous HPVs can be missed, and there is no further chance to detect cervical cancer by them.

Most machine learning algorithms for classification problems have focused on minimizing the number of incorrect predictions. However, this kind of learning algorithms ignores the differences between different types of incorrect prediction cost. Thus, recently, there has been considerable interest in cost-sensitive learning (Provost and Fawcett, 1997). Ting and Zheng (1998) proposed two related but different cost-sensitive boosting approaches for tree classification. Their approaches can be applied only to situations where the costs change very often. To apply boosting to situations where misclassification costs are relatively stable, Fan *et al.* (1999) proposed the AdaCost.

In this paper, we propose a cost-sensitive learning method to classify the risk types of HPVs using their textual explanation. In classifying their risk types, we consider the learning costs of each example, because it is far more important to reduce the number of false negatives[1)] than to reduce that of false positives. For this purpose, we adopt

* Corresponding author:
E-mail sbpark@bi.snu.ac.kr, Tel +82-2-880-1847, Fax +82-2-875-2240
Accepted 10 November 2003

1) In this paper, *false negative* implies that high-risk HPV is misclassified as low-risk. Similarly, *false positive* means low-risk HPV that is misclassified as high-risk

AdaCost as a learning algorithm and prove empirically that it shows great performance in classifying the HPV risk types.

One advantage of this work is usefulness for designing the DNA-chip, diagnosing the presence of Human Papillomavirus in cervical cancer patients. Since there are about 100 HPV types, making the DNA chip needs to choose some dangerous ones related with cervical cancer among them. Therefore, this result classifying the risk of HPVs can be a big help to save time to understand information about HPV and cervical cancer from many papers and to choose the HPV types used for DNA chip.

The rest of this paper is organized as follows. Section 2 expresses the problems of normal machine learning algorithms. Section 3 describes the cost-sensitive learning to classify HPV risk types. Section 4 explains how the HPV dataset is generated. Section 5 presents the experimental results. Finally, section 6 draws conclusions.

Problems of normal learning methods

First, let us check what happens unless we consider the cost of each learning example. We classify the risk type of HPVs by their textual explanation given by Los Alamos National Laboratory. The details of this explanation will be explained in Section 4. Table 1 shows the classification result of HPVs according to their risk types. It is classified by the naive Bayes classifier (Lang, 1995) without considering the costs of the examples. This result is obtained when we used only seven HPVs as a training set. Among the seven HPVs, five are high-risk (HPV16, HPV18, HPV31, HPV33, HPV45), and the other two are low-risk(HPV11 and HPV6). Because the risk types of these HPVs are well known (Levy *et al*, 1994), they are chosen to be a training set.

Table 1. Classification of the risk types of HPVs by naive Bayes classifier

Type	Risk	Type	Risk	Type	Risk	Type	Risk
HPV1	Low	HPV2	High	HPV3	Low	HPV4	Low
HPV5	High	HPV7	Low	HPV8	High	HPV9	Low
HPV10	Low	HPV12	Low	HPV13	Low	HPV14	Low
HPV15	Low	HPV17	Low	HPV19	Low	HPV20	Low
HPV21	Low	HPV22	Low	HPV23	Low	HPV24	Low
HPV25	Low	HPV26	High	HPV27	Low	HPV28	Low
HPV29	Low	HPV30	High	HPV32	Low	HPV34	High
HPV35	High	HPV36	Low	HPV37	Low	HPV38	Low
HPV39	High	HPV40	Low	HPV41	Low	HPV42	Low
HPV43	Low	HPV44	Low	HPV47	Low	HPV48	Low
HPV49	Low	HPV50	Low	HPV51	High	HPV52	High
HPV53	High	HPV54	Low	HPV55	Low	HPV56	High
HPV57	High	HPV58	High	HPV59	Low	HPV60	High
HPV61	Low	HPV62	Low	HPV63	High	HPV64	Low
HPV65	Low	HPV66	High	HPV67	High	HPV68	High
HPV69	Low	HPV70	High	HPV72	Low	HPV73	Low
HPV74	Low	HPV75	High	HPV76	High	HPV77	Low
HPV80	Low						

The number of tested HPVs is 69. Assuming that Table 2 below is correct, the risk type for four of 69 HPVs is not known, so that 65 HPVs are evaluated. Twenty among 65 HPVs are classified as high-risk and the remaining 45 are classified as low-risk, while there are only 12 high-risk HPVs in Table 2. Since 53 HPVs are correctly classified, the accuracy is 81.54%.

At first, this accuracy seems reasonable. However, four of 12 misclassified cases are false negative, and 8 are false positive. That is, this is not satisfactory because false negatives are fatal as stated above. The reasons why the method which ignores the cost does not achieve high performance can be summarized into two problems.

The first one is that too many high-risk HPVs are predicted. That is, there are only 12 high-risk ones in the tested HPVs, but the native Bayes predicted 20 HPVs as high-risk. Though we used only two low-risk HPVs, in fact there are far more low-risk HPVs than high-risk HPVs. Therefore, it is required to give a higher cost to high-risk HPVs during training.

The other problem is that there are some HPVs that are difficult to determine their risk types only by their textual explanation. For instance, HPV54 is explained by a single sentence which is "HPV-54 was first isolated from a patient with condyloma acuminata." This problem is inevitable in text classification. Thus, it goes beyond interest of this paper and should be solved by further biomedical experiments.

Materials and Methods

Classifying by cost-sensitive learning adacost algorithm

In order to consider the misclassification cost of HPV risk types, we adopt the AdaCost algorithm (Fan *et al*, 1999). The AdaCost is a variant of AdaBoost (Freund and Schapire, 1996) that uses the cost of misclassifications to update the training distribution on successive boosting rounds.

The algorithm is shown in Fig. 1. *Let* $S = \{(x_1,c_1,y_1), \Lambda,(x_m,c_m,y_m)\}$ be a training set where $c_i \in [0,1]$ is a cost factor and is additionally given to the normal $x_i \in X$ and $y_i \in \{-1,+1\}$. First of all, the distribution of each example is set to $D^1(i) = c_i / \sum_{j=1}^{m} c_j$. When t is an index to show the round of boosting, $Dt(i)$ is the sampling weight given to (x^i, c^i, y^i) at the t-th round. And, $\alpha_t > 0$ is a parameter as a weight for weak learner ht at the t-th round, and its value is given as

$$\alpha_t = \frac{1}{2} \ln \frac{1+r}{1-r}$$

where $\sum_i D(i)y_i h_t(x_i)\beta(i)$. And, $\beta(i)$ is a cost adjustment function with two arguments, sign$(y_i, h_t(x_i))$ and c_i. If $h_t(x_i)$ is correct, then $\beta(i) = -0.5c_i + 0.5$, otherwise $\beta(i) = 0.5c_i + 0.5$.

Input: ▸ $S=\{(x_i,c_i,y_i),\Lambda,(x_m,c_m,y_m)\}$:
$x_i \in X$, $c_i \in [0,1]$, and $y_i \in \{-1,+1\}$
▸ weak learner algorithm **WeakLearn**
▸ integer Tspecifying the number of iterations

Initialize $D_1(i) = c_i / \sum_{j=1}^{m} c_j$ for all i.

For $t = 1, ..., T$:
1. Call **WeakLearn**, providing it with the distribution D_t.
2. Get back a hypothesis h_t: $\mathbf{X} \rightarrow \{-1, +1\}$.
3. Choose $\alpha_t \in R$ and $\beta(i)$, where $\beta(i) = \beta(\text{sign}(y_i h_t(x_i)), c_i)$.
4. Update distribution D_t:

$$D_{t+1}(i) = \frac{1}{Z_t} D_t(i) exp(-\alpha_t y_i h_t(x_i)\beta(i))$$

where Z_t is a normalizing constant.

Output: the final hypothesis:

$$f(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

The main difference between AdaBoost and AdaCost is how the distribution D_t is updated. AdaCost has an additional cost adjustment factor in updating D_t (see step 4 in Fig. 1). As AdaBoost does, the weight of an instance will be increased if it is misclassified. Similarly, its weight will be decreased otherwise. However, the weight change is affected by the value of the cost factor. When an instance has a high cost factor, the weight change will be greater than that with a low cost factor.

Naïve bayes classifier as a weak learner

We have previously proposed the BayesBoost algorithm and showed that it gives great efficiency in text filtering (Kim *et al.*, 2000). It uses naive Bayes classifiers as its weak learner within AdaBoost. Assume that a document d_i is composed of a sequence of words which is $w_{i1}, w_{i2}, \Lambda, w_{i|di|}$, and the words in a document are mutually independent one another and the probability of a word is independent of its position within the document. Though these assumptions are not true in real situations, naive Bayes classifiers showed rather good performance in text classification (McCallum and Nigam, 1998).

Due to the independence assumption, the probability that a document d_i is generated from the class y_j can be expressed as

$$P(d_i|y_j;\hat{\theta}) = P(|d_i|)\prod_{k=1}^{|d_i|} P(w_{dik}|y_j;\hat{\theta})^{N(w_{dik},d_i)},$$

where w_{dik} denotes the k-th word in the document d_i, the weight of word occurring in document d_i, and $|d_i|$ is the number of words in the document. Thus, when assuming $P(|d_i|)$ is uniform, the best class y* of a document di is determined by

$$y^* = \underset{y_j \in \{-1,+1\}}{\text{argmax}}\, P(y_j|d_i;\hat{\theta}),$$

where

$$P(y_j|d_i;\hat{\theta}) = \frac{P(y_j|\hat{\theta})P(d_i|y_j;\hat{\theta})}{P(d_i|\hat{\theta})} = \frac{P(y_j|\hat{\theta})\prod_{k=1}^{|d_i|} P(w_{dik}|y_r;\hat{\theta})^{N(w_{dik},d_i)}}{\sum_{r=1}^{2} P(y_r|\hat{\theta})\prod_{k=1}^{|d_i|} P(w_{dik}|y_r;\hat{\theta})^{N(w_{dik},d_i)}}. \quad (1)$$

In order to calculate this probability, we need to determine $P(w_k|y_j;\hat{\theta})$ and $P(y_j|\hat{\theta})$. These two values can be estimated as

$$P(w_k|y_j;\hat{\theta}) = \frac{1+\sum_{i=1}^{m} N(w_k,d_i)P(y_j|d_i)}{|V|+\sum_{l=1}^{|V|}\sum_{i=1}^{m} N(w_l,d_i)P(y_j|d_i)}$$

$$P(y_j|\hat{\theta}) = \frac{\sum_{i=1}^{m} P(y_j|d_i)}{m}.$$

Here, $|V|$ is the size of vocabulary.

One of the advantages of using naive Bayes classifier as a weak learner is that the naive Bayes utilizes term weights such as term frequency naturally. Moreover, because it is a probabilistic model, it provides a natural measure for calculating confidence ratios in AdaBoost. Thus, in this paper, we also use naive Bayes classifier as a weak learner of AdaCost.

Results and Discussion

Dataset

In general, the research in biomedical domain starts from investigating previous studies in *PubMed* designed to provide access to citations from biomedical literature and available via the NCBI Entrez retrieval system developed by National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) located at the National Institutes of Health (NIH). Most bioinformatics research that handles text information has focused on PubMed as its resource, because it includes most summaries and citations about biomedical literature. However, learning HPV risk types from PubMed is not an

easy work. The difficulties can be summarized with two reasons.

- **The PubMed data are too sparse.**

For example, there are 3,797 articles about HPV and cervical cancer in PubMed, but most of them do not discuss the risk of HPV directly. Thus, it is difficult to capture the risk of HPV from the articles. In addition, the term distribution is totally different according to the interest of the articles.

- **Poor performance of NLP techniques**

The current natural language processing (NLP) techniques are not for text understanding yet. They can not provide even correct syntactic information. The best thing we can expect from NLP techniques is morphological analysis and part-of-speech tagging. Thus, the articles need to be refined for further study.

In this paper, we use *the HPV Sequence Database*[2] in Los Alamos National Laboratory as a dataset. This papillomavirus database is an extension of the HPV compendiums published in 1994, 1995, 1996, and 1997 and provides the complete list of 'papillomavirus types and hosts' and the records for each unique papillomavirus type. An example of the data made from this database is given in Fig. 2. This example is for HPV80 and consists of three parts: definition, source, and comment. The definition indicates the HPV type, the source explains where the information for this HPV is obtained, and the comment gives the explanation for this HPV.

To measure the performance of the results in the experiments below, we manually classified HPV risk types using the 1997 version of Human Papillomaviruses compendium and the comment in the records of HPV types. The classifying procedure is as follows. First, we divided roughly HPV types by the groups in the 1997 version of Human Papillomaviruses compendium. These groups are shown in Fig. 3. This tree, which contains 108 Papilloma Virus (PV) sequences, was computed for the L1 consensus primer region (CPR) using neighbor joining method and a distance matrix calculated with a modified Kimura 2-parameter model (transition/transversion ratio 2.0). Neighbor-joining analysis is a convenient and rapid way to get an initial estimate of branching relationships, especially when a large number of taxa are involved. In the figure, the outermost wide gray arcs show the five PV supergroups (A-E). Each tree branch is labeled with an abbreviated sequence name. For HPVs the 'type' number alone is given in most cases, so the branch labeled 40 is that of HPV40.

2) http://hpv-web.lanl.gov/stdgen/virus/hpv/index.html

<definition>
Human papillomavirus type 80 E6, E7, E1, E2, E4, L2, and L1 genes.
</definition>

<source>
Human papillomavirus type 80.
</source>

<comment>
The DNA genome of HPV80 (HPV15-related) was isolated from histologically normal skin, cloned, and sequenced. HPV80 is most similar to HPV15, and falls within one of the two major branches of the B1 or Cutaneous/EV clade. The E7, E1, and E4 orfs, as well as the URR, of HPV15 and HPV80 share sequence similarities higher than 90%, while in the usually more conservative L1 orf the nucleotide similarity is only 87%. A detailed comparative sequence analysis of HPV80 revealed features characteristic of a truly cutaneous HPV type [362]. Notice in the alignment below that HPV80 compares closely to the cutaneous types HPV15 and HPV49 in the important E7 functional regions CR1, pRb binding site, and CR2. HPV 80 is distinctly different from the high-risk mucosal viruses represented by HPV16. The locus as defined by GenBank is HPVY15176.
</comment>

Fig. 2. An example description of HPV80 from Los Alamos National Labratory.

Second, if the type of the group is skin-related or cutaneous HPV, the members of the group are classified into low-risk type. Third, if the group is known to be high-risk type of cervical cancer-related HPV, the members of the group are classified into high-risk type. Lastly, we used the comment of HPV types to classify some types difficult to be classified. Table 2 shows the summarized classification of HPVs according to its risk.

In the all experiments below, we used only <comment> part. The comment for a HPV type can be considered as a document in text classification. Therefore, each HPV type is represented as a vector of which elements are $tf \cdot idf$ values. In $tf \cdot idf$, $N(w_j, d_i)$ of Equation (1), the weight of a word w_j appeared in the document d_i is given as

$$N(w_j,d_i) = tf_{ij} \cdot log_2 \frac{m}{n} \qquad (2)$$

where tf_{ij} is the frequency of w_j in d_i and n is the number of documents where w_j occurs at least once.

When we stemmed the documents using the Porter' s algorithm (Porter, 1980) and removed words from the stop-

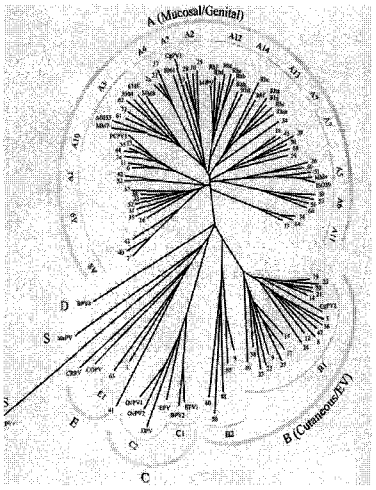

Fig. 3. Neighbor joining phylogenetic tree of 106 PVs based on CPR region of L1.

Table 2. The manually classified risk types of HPVs.

Type	Risk	Type	Risk	Type	Risk	Type	Risk
HPV1	Low	HPV2	Low	HPV3	Low	HPV4	Low
HPV5	Low	HPV6	Low	HPV7	Low	HPV8	Low
HPV9	Low	HPV10	Low	HPV11	Low	HPV12	Low
HPV13	Low	HPV14	Low	HPV15	Low	HPV16	High
HPV17	Low	HPV18	High	HPV19	Low	HPV20	Low
HPV21	Low	HPV22	Low	HPV23	Low	HPV24	Low
HPV25	Low	HPV26	?	HPV27	Low	HPV28	Low
HPV29	Low	HPV30	Low	HPV31	High	HPV32	Low
HPV33	High	HPV34	Low	HPV35	High	HPV36	Low
HPV37	Low	HPV38	Low	HPV39	High	HPV40	Low
HPV41	Low	HPV42	Low	HPV43	Low	HPV44	Low
HPV45	High	HPV47	Low	HPV48	Low	HPV49	Low
HPV50	Low	HPV51	High	HPV52	High	HPV53	Low
HPV54	?	HPV55	Low	HPV56	High	HPV57	?
HPV58	High	HPV59	High	HPV60	Low	HPV61	High
HPV62	High	HPV63	Low	HPV64	Low	HPV65	Low
HPV66	High	HPV67	High	HPV68	High	HPV69	Low
HPV70	?	HPV72	High	HPV73	Low	HPV74	Low
HPV75	Low	HPV76	Low	HPV77	Low	HPV80	Low

list, the size of vocabulary is just 1,434. Thus, each document is represented as a 1,434-dimensional vector.

Experiments

Evaluation measure

Text classification has various measures to evaluate its performance. One of these is the break-even point (Lewis, 1995). However, Schapire et al. (1998) asserted that the break-even points are not very suitable for measuring the performance of classification algorithms.

Table 3. The contingency table to evaluate the classification performance.

	Answer should be *High*	Answer should be *Low*
The Classifier says *High*	*a*	*B*
The Classifier says *Low*	*c*	*d*

In this paper, we evaluate the classification performance using the contingency table method. In this method, recall and precision are defined as follows:

$$recall = \frac{a}{a+c} \cdot 100\%$$

$$precision = \frac{a}{a+c} \cdot 100\% \qquad (3)$$

$$accuracy = \frac{a+b}{a+b+c+d} \cdot 100\%$$

where a, b, c and d are defined in Table 3. The F_β-*score* which combines precision and recall is defined as

$$F_\beta\text{-}score = \frac{(\beta^2+1) \cdot recall \cdot precision}{\beta^2 \cdot recall + precision} \cdot 100\% \quad ,$$

where β is the weight of recall relative to precision. We use β = 1 in all experiments, which corresponds to equal weighting of the two measures.

Experimental Results

Since we have only 74 HPV types and the explanation of each HPV is relatively short, *leave-one-out* (LOO) *cross-validation* is used to determine the performance of the proposed method. We normalized each cost c_i to [0, 1]. That is, the cost for low-risk HPVs is set to 0.1 when the cost for high-risk HPVs is set to 0.9.

Fig. 4 demonstrates the performance of AdaCost. The graphs in this figure show the accuracy and F-score according to the round of AdaCost. Each graph represents the ratio of costs for high-risk and low-risk HPVs. For instance, figure (a) imposes 0.1 on high-risk HPVs and 0.9 on low-risk HPVs. Because the costs in figure (e) are both set to 0.5, it is the performance of the AdaBoost. Figures (a)-(d) plot the performance when lower costs are imposed on high-risk HPVs than those on low-risk HPVs. And, figures (f)-(i) plot the performance when higher costs are imposed on high-risk HPVs.

Generally, when we set different costs to low-risk and high-risk HPVs, higher performance is obtained than AdaBoost shown by figure (e) except the extreme cases represented by figure (a) and (i). Among nine graphs, figure (h) shows the best performance. It implies that 0.8 is the best cost for high-risk HPVs. It is also interesting to see

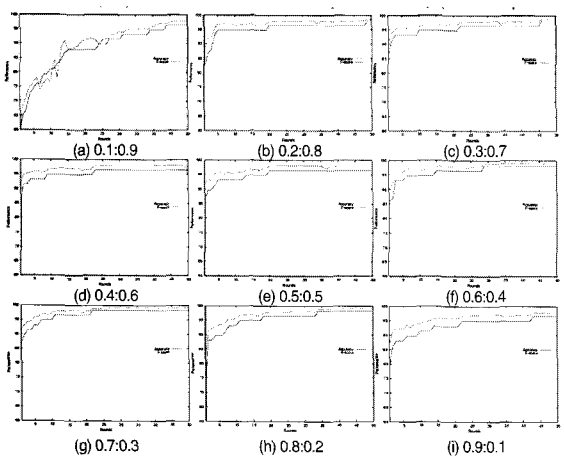

Fig. 4. Performance of AdaCost on HPV risk classification with various costs. The cost ratios are geiven as high vs. Low. For instance, figure (a) imposes 0.1 on high-risk HPVs and 0.9 on lows-risk HPVs.

Table 4. F-score of AdaCost with various cost ratio.

High	Low	F-score	Accuracy (%)
0.1	0.9	96.91±2.1	97.51±1.8
0.2	0.8	98.79±0.9	98.27±0.5
0.3	0.7	98.87±1.5	98.27±0.6
0.4	0.6	97.72±1.5	97.55±0.9
0.5	0.5	98.07±1.2	97.55±1.3
0.6	0.4	98.96±1.8	98.27±1.0
0.7	0.3	98.92±0.3	98.27±0.4
0.8	0.2	99.04±0.2	98.27±1.1
0.9	0.1	97.92±2.6	97.55±1.6

that figure (a) shows the worst performance. That is, in this case AdaCost shows worse performance than AdaBoost. Therefore, if we impose wrong cost, we may obtain worse result.

Table 4 summarizes the graphs in Fig. 4. These results are obtained when 50 weak learners are used in each AdaCost. The accuracy is similar with various costs, but different costs show different performance on F-score. As shown in Equation (3), precision and recall are related with the number of found high-risk HPVs while accuracy is related with the number of correctly predicted HPVs including both low-risk and high-risk HPVs. In our experiments, F-scores are higher than accuracies, which implies that less high-risk HPVs are missed by the proposed method.

Table 5 shows the predicted risk type for the HPV types whose risks are not known exactly. These HPVs are described as '?' in Table 2. According to previous research on HPV (Chan *et al.*, 1997; Favre *et al.*, 1990; Meyer *et al.*, 1998; Nuovo *et al.*, 1988) , HPV70 seems to be misclassified. This is because the comment for HPV70 does not describe its risk but because of its lack of biomedical research it explains only that it is found at the cervix of patients and its sequence is analyzed.

Table 5. The risk type predicted by the proposed method for four HPVs whose risk types are not known exactly.

HPV Types	Risk Types
HPV26	Low
HPV54	Low
HPV57	High
HPV70	Low

Conclusions

This paper proposed a practical method to determine the risk type of Human Papillomavirus. In classifying the risk type, it is important to distinguish false negatives from false positives, where false-negatives are high-risk HPVs that are misclassified as low-risk and false positives are low-risk HPVs misclassified as high-risk.

For this purpose, we set different costs for low-risk and high-risk HPVs. As a learning algorithm, we adopted *AdaCost* and showed empirically that it outperforms AdaBoost which does not consider learning cost. In addition, the experimental results gave higher F-score than accuracy, and it means that more high-risk HPVs are found by AdaCost. This result is important because high-risk HPVs, as stated above, should not be missed. Since HPV is known as the main cause of cervical cancer, high-risk HPVs must be found for further medical investigation of the patients.

Our results can be used as fundamental information to design the DNA-chips for diagnosing the presence of HPV in cervical cancer patients. Because the cost is too high to test all HPV types, the results presented in this paper reduce time and monetary cost to know their relation with cervical cancer.

Acknowledgments

This research was supported by the Korean Ministry of Education under the BK21-IT Program, and by the Korean Ministry of Science and Technology under NRL and BrainTech programs.

References

Chan, S., Chew, S., Egawa. K., Grussendorf-Conen, E., Honda,

Y., Rubben, A., Tan, K., and Bernard, H. (1997). Phylogenetic Analysis of the Human Papillomavirus Type 2 (HPV-2), HPV-27, and HPV-57 Group, Which is Associated with Common Warts. *Virology* 239, 296-302.

Fan, W., Stolfo, S., Zhang, J., and Chan, P. (1999). AdaCost: Misclassification Cost-Sensitive Boosting. In *Proceedings of the 16th International Conference on Machine Learning* 97-105.

Favre, M., Kremsdorf, D., Jablonska, S., Obalek, S., Pehau-Arnaudet, G., Croissant, O., and Orth, G. (1990). Two New Human Papillomavrius Types (HPV54 and 55) Characterized from Genital Tumours Illustrate the Plurality of Genital HPVs. *International Journal of Cancer* 45, 40-46.

Freund, Y. and Schapire, R. (1996). Experiments with a New Boosting Algorithm. In *Proceedings of the 13th International Conference on Machine Learning* 148-156.

Furumoto, H. and Irahara, M. (2002). Human Papillomavirus (HPV) and Cervical Cancer. *The Journal of Medical Investigation*. 49, 124-133.

Ishiji, T. (2000). Molecular Mechanism of Carcinogenesis by Human Papillomavirus-16. *The Journal of Dermatology* 27, 73-86.

Janicek, M. and Averette, H. (2001). Cervical Cancer: Prevention, Diagnosis, and Therapeutics. *Cancer Journal for Clinicians* 51, 92-114.

Kim, Y.-H., Hahn, S.-Y., and Zhang, B.-T. (2000). Text Filtering by Boosting Naive Bayes Classifiers. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 168-175.

Lang, K. (1995). Newsweeder: Learning to Filter Netnews. In *Proceedings of the 12th International Conference on Machine Learning* 331-339.

Levy, J., Fraenkel-Conrat, H., and Owens, R. (1994). *Virology* Prentice Hall.

Lewis, D. (1995). Evaluating and Optimizing Autonomous Text Classification System. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 298-306.

McCallum, A. and Nigam, K. (1998). Employing EM in Pool-Based Active Learning for Text Classification. In *Proceedings of the 15th International Conference on Machine Learning* 350-358.

Meyer, T., Arndt, E., Christophers, E., Beckmann, E., Schroder, S., Gissmann, L., and Stockfleth, E. (1998). Association of Rare Human Papillomavirus Types with Genital Premalignant and Malignant Lesions. *The Journal of Infectious Diseases* 178, 252-255.

Nuovo, G., Crum, C., De Villiers, E., and Silverstein, S. (1988). Isolation of a Novel Human Papillomavirus (Type 51) from a Cervical Condyloma. *Journal of Virology* 62, 1452-1455.

Porter, M. (1980). An Algorithm for Suffix Stripping. *Program* 14, 130-137.

Provost, F. and Fawcett, T. (1997). Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* 43-48.

Schapire, R., Singer, Y., and Singhal, A. (1998). Boosting and Rocchio Applied to Text Filtering. In Proceedings of the 21st *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 215-223.

Schiffman, M., Bauer, H., Hoover, R., Glass, A., Cadell, D., Rush, B., Scott, D., Sherman, M., Kurman, R., and Wacholder, S. (1993). Epidemiologic Evidence Showing That Human Papillomavirus Infection Causes Most Cervical Intraepithelial Neoplasis. *Journal of the National Cancer Institute* 85, 958-964.

Ting, K.-M. and Zheng, T. (1998). Boosting Trees for Cost-Sensitive Classifications. In *Proceedings of the 10th European Conference on Machine Learning* 190-195.