Mining the Risk Types of Human Papillomavirus (HPV) by AdaCost

S.-B. Park, S. Hwang, and B.-T. Zhang

School of Computer Science and Engineering
Seoul National University
151-742 Seoul, Korea
{sbpark,shhwang,btzhang}@bi.snu.ac.kr

Abstract. Human Papillomavirus (HPV) infection is known as the main factor for cervical cancer, where cervical cancer 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, we classify 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 out high-risk HPVs 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 considering costs gives higher performance. The F-score is higher than the accuracy, which implies that most high-risk HPVs are found.

1 Introduction

Human papillomavirus (HPV) is a double-strand DNA tumor virus that belongs to the papovavirus family, and there are more than 100 types in HPV that are specific for epithelial cells including skin, respiratory mucosa, and the genital tract. Especially, the genital tract HPV types are classified by their relative malignant potential into low-, and high-risk types [6]. 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 what HPV types are high-risk.

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 [4,5], they can be used as a source of discriminating HPV risk types. One problem in discriminating the risk types is that 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.

V. Mařík et al. (Eds.): DEXA 2003, LNCS 2736, pp. 403–412, 2003.

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 [11]. Ting and Zheng proposed two related but different cost-sensitive boosting approaches for tree classification [13]. 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. proposed the AdaCost algorithm [2].

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

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

2 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, most bioinformatics research on text mining 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. 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* in Los Alamos National Laboratory as a dataset. This papillomavirus database is an extension of the HPV compendiums published in 1994 – 1997, and provides the complete list of 'papillomavirus types and hosts' and the records for each unique papillomavirus

[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.

<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. 1. An example description of HPV80.

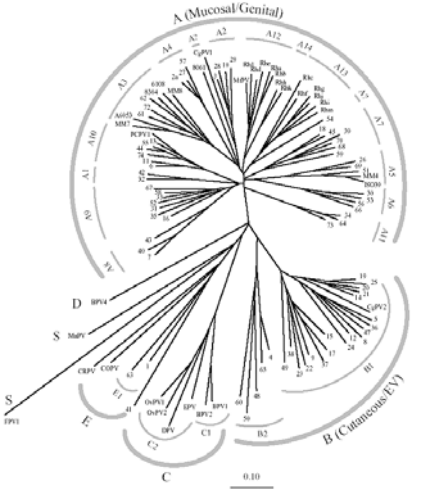

Fig. 2. Neighbor joining phylogenetic tree of 106 PVs based on CPR region of L1.

type. An example of the data from this database is given in Figure 1. This 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 HPV compendium

Table 1. The manually classified risk types of each HPV.

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	Don't Know	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	Don't Know	HPV55	Low	HPV56	High	HPV57	Don't Know
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	Don't Know	HPV72	High	HPV73	Low	HPV74	Low
HPV75	Low	HPV76	Low	HPV77	Low	HPV80	Low

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 HPV compendium. These groups are shown in Figure 2. This tree, which contains 108 Papillomavirus (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.

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 1 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 $tf \cdot idf$, the weight of a word w_j appeared in the document d_i is given as

$$N(w_j, d_i) = tf_{ij} \cdot \log_2 \frac{m}{n}, \tag{1}$$

where tf_{ij} is the frequency of w_j in d_i, m is the total number of documents, and n is the number of documents where w_j occurs at least once. When we stemmed the documents using the Porter's algorithm and removed words from the stop-list, the size of vocabulary is just 1,434. Thus, each document is represented as a 1,434-dimensional vector.

3 Classifying by Cost-Sensitive Learning

3.1 AdaCost Algorithm

In order to consider the misclassification cost of HPV risk types, we adopt the AdaCost algorithm [2]. Let $S = \{(x_1, c_1, y_1) \ldots, (x_m, c_m, y_m)\}$ be a training set where $c_i \in \mathcal{R}^+$ is a cost factor and is additionally given to the normal $x_i \in \mathcal{X}$ 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, $D_t(i)$ is the 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 h_t at the t-th round, and its value is given as

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

where $r = \sum_i D(i) y_i h_t(x_i) \beta(i)$. 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 $\beta(i) = -0.5c_i + 0.5$.

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. 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.

3.2 Naive Bayes Classifier as a Weak Learner

Kim et al. proposed the BayesBoost algorithm and showed that it gives great efficiency in text filtering [7]. 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 $w_{i1}, w_{i2}, \ldots, w_{i|d_i|}$, 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 [8].

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

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

Table 2. 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

where $w_{d_{ik}}$ denotes the k-th word in the document d_i, $N(w_{d_{ik}}, d_i)$ given by Equation (1) denotes the weight of word $w_{d_{ik}}$ 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 d_i is determined by

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

where

$$\begin{aligned} P(y_j|d_i;\hat{\theta}) &= \frac{P(y_j|\hat{\theta})P(d_i|y_j;\hat{\theta})}{P(d_i|\hat{\theta})} \\ &= \frac{P(y_j|\hat{\theta})\prod_{k=1}^{|d_i|} P(w_{d_{ik}}|y_j;\hat{\theta})^{N(w_{d_{ik}},d_i)}}{\sum_{r=1}^{2} P(y_r|\hat{\theta})\prod_{k=1}^{|d_i|} P(w_{d_{ik}}|y_r;\hat{\theta})^{N(w_{d_{ik}},d_i)}}. \end{aligned} \tag{2}$$

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

$$\begin{aligned} P(w_k|y_j;\hat{\theta}) &= \frac{1+\sum_{i=1}^{m} N(w_k,d_i)P(y_j|d_i)}{|V|+\sum_{k=1}^{|V|}\sum_{i=1}^{m} N(w_k,d_i)P(y_j|d_i)}, \\ P(y_j|\hat{\theta}) &= \frac{\sum_{i=1}^{m} P(y_j|d_i)}{m}. \end{aligned}$$

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.

4 Experiments

4.1 Evaluation Measure

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

$$\begin{aligned} recall &= \frac{a}{a+c} \cdot 100\% \\ precision &= \frac{a}{a+b} \cdot 100\% \\ accuracy &= \frac{a+d}{a+b+c+d} \cdot 100\%, \end{aligned} \tag{3}$$

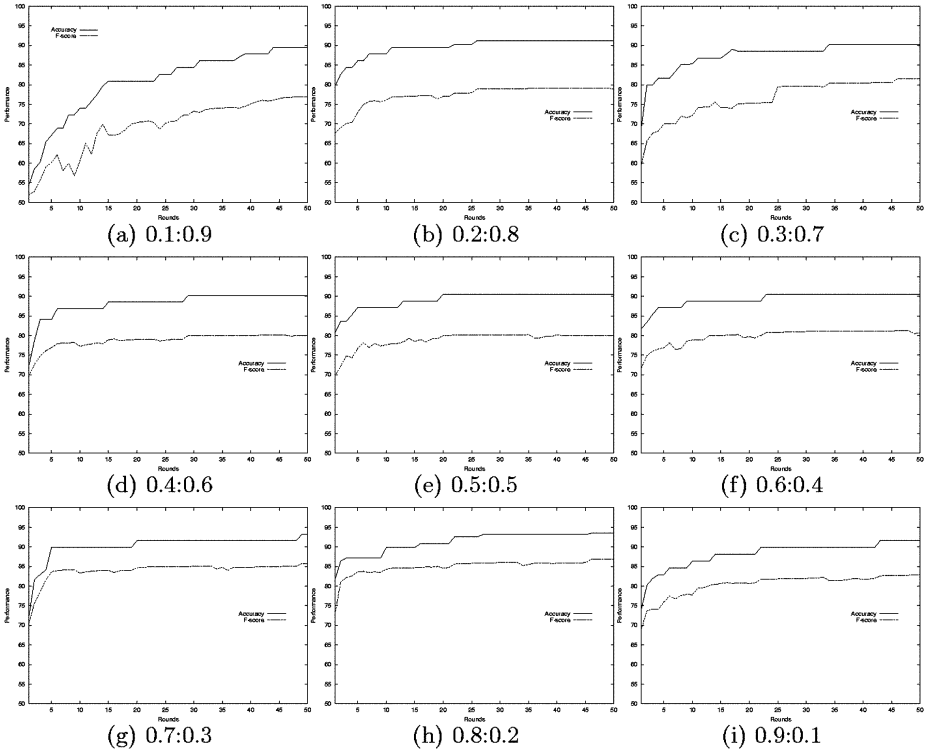

Fig. 3. Performance of AdaCost on HPV risk classification with various costs.

where a, b, c and d are defined in Table 2. The F_β-*score* which combines precision and recall is defined as

$$F_\beta = \frac{(\beta^2 + 1) \cdot recall \cdot precision}{\beta^2 \cdot recall + precision},$$

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

4.2 Experimental Results

Since we have only 72 HPV types except "Don't Know"s 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.

Figure 3 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

Table 3. The performance comparison of AdaCost and AdaBoost on HPV risk classification.

	AdaCost	AdaBoost	naive Bayes
Accuracy (%)	93.05	90.55	81.94
F-score	86.49	80.08	63.64

(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 higher cost to high-risk HPVs than to low-risk HPVs, we obtained higher performance than AdaBoost shown by figure (e). When we impose lower cost to high-risk HPVs than to high-risk HPVs, the performance gets lower than AdaBoost except figure (c). These results coincide with the intuition that we should set higher costs to high-risk HPVs. It is also interesting to see that figure (a) shows the worst performance. Therefore, if we impose wrong cost, we may obtain worse result. Among nine graphs, figure (h) shows the best performance. It implies that 0.8 is the best cost for high-risk HPVs.

The final classification performance is given in Table 3. It compares three learning methods: AdaCost, AdaBoost, and naive Bayes classifier which is used as a weak learner in AdaCost and AdaBoost. AdaCost shows 93.05% of accuracy and 86.49 of F-score, while AdaBoost gives only 90.55% of accuracy and 80.08 of F-score. Especially, naive Bayes classifier reported 26 high-risk HPVs. Among 26 high-risk HPVs reported by naive Bayes classifier, only fourteen are correctly predicted. Thus, it shows only 81.94% of accuracy and 63.64 of F-score. As shown in Equation (3), F-score is closely 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 previous study, we showed that even AdaBoost has an implicit ability of cost learning [12]. That is, AdaBoost can show higher F-score than naive Bayes classifier. In our experiments, the F-score of AdaCost and AdaBoost is actually far higher than that of naive Bayes classifier. And, this result supports our goal to reduce false negatives. In addition, when we strongly pose cost factors as in AdaCost, it shows higher F-score than AdaBoost. The difference in accuracy between AdaCost and AdaBoost is just 2.5%, but the difference in F-score is 6.41. This implies that more high-risk HPVs found by AdaCost than by AdaBoost.

This can be found also in Figure 4 which depicts the receiver operating characteristic (ROC) curves of AdaCost. In this figure, the dotted line plots the ROC curve of AdaBoost, while the thick line plots that of AdaCost when the cost of 0.8 is imposed on high-risk HPVs. Since two curves do not intersect and the curve of AdaCost is above that of AdaBoost, the performance of AdaCost is superior under all relative weightings of true positive and false positive rates.

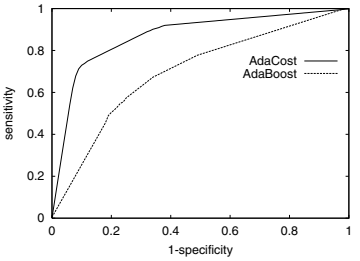

Fig. 4. ROC curve of AdaCost.

Table 4. The risk type predicted by the proposed method for four HPVs whose risks are not known exactly.

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

Table 4 shows the predicted risk type for the HPV types whose risks are not known exactly. These HPVs are described as "Don't Know" in Table 1. According to previous research on HPV [3,1,9,10], only 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.

5 Conclusions

This paper proposed a practical method to determine the risk type of human papillomaviruses. In classifying their 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 improvement on F-score than that on 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 not be missed 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.

Acknowledgements. 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 BrainTech and NRL programs.

References

1. S. Chan, S. Chew, K. Egawa, E. Grussendorf-Conen, Y. Honda, A. Rubben, K. Tan, and H. Bernard, "Phylogenetic Analysis of the Human Papillomavirus Type 2 (HPV-2), HPV-27, and HPV-57 Group, Which is Associated with Common Warts," *Virology*, Vol. 239, pp. 296–302, 1997.
2. W. Fan, S. Stolfo, J. Zhang, and P. Chan, "AdaCost: Misclassification Cost-Sensitive Boosting," In *Proceedings of the 16th International Conference on Machine Learning*, pp. 97–105, 1999.
3. M. Favre, D. Kremsdorf, S. Jablonska, S. Obalek, G. Pehau-Arnaudet, O. Croissant, and G. Orth, "Two New Human Papillomavirus Types (HPV54 and 55) Characterized from Genital Tumours Illustrate the Plurality of Genital HPVs," *International Journal of Cancer*, Vol. 45, pp. 40–46, 1990.
4. H. Furumoto and M. Irahara, "Human Papilloma Virus (HPV) and Cervical Cancer," *The Jounral of Medical Investigation*, Vol. 49, No. 3–4, pp. 124–133, 2002.
5. T. Ishiji, "Molecular Mechanism of Carcinogenesis by Human Papillomavirus-16," *The Journal of Dermatology*, Vol. 27, No. 2, pp. 73–86, 2000.
6. M. Janicek and H. Averette, "Cervical Cancer: Prevention, Diagnosis, and Therapeutics," *Cancer Journal for Clinicians*, Vol. 51, pp. 92–114, 2001.
7. Y.-H. Kim, S.-Y. Hahn, and B.-T. Zhang, "Text Filtering by Boosting Naive Bayes Classifiers", In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 168–175, 2000.
8. A. McCallum and K. Nigam, "Empolying EM in Pool-based Active Learning for Text Classification," In *Proceedings of the 15th International Conference on Machine Learning*, pp. 350–358, 1998.
9. T. Meyer, R. Arndt, E. Christophers, E. Beckmann, S. Schroder, L. Gissmann, and E. Stockfleth, "Association of Rare Human Papillomavirus Types with Genital Premalignant and Malignant Lesions," *The Journal of Infectious Diseases*, Vol. 178, pp. 252–255, 1998.
10. G. Nuovo, C. Crum, E. De Villiers, R. Levine, and S. Silverstein, "Isolation of a Novel Human Papillomavirus (Type 51) from a Cervical Condyloma," *Journal of Virology*, Vol. 62, pp. 1452–1455, 1988.
11. F. Provost and T. Fawcett, "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*, pp. 43–48, 1997.
12. S.-B. Park and B.-T. Zhang, "A Boosted Maximum Entropy Model for Learning Text Chunking," In *Proceedings of the 19th Internatinal Conference on Machine Learning*, pp. 482–489, 2002.
13. K.-M. Ting and Z. Zheng, "Boosting Trees for Cost-Sensitive Classifications," In *Proceedings of the 10th European Conference on Machine Learning*, pp. 190–195, 1998.