ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/289030311

Extended Naïve Bayes for Group Based Classification

Article *in* Advances in Soft Computing · January 2014

DOI: 10.1007/978-3-319-07692-8_47

CITATIONS
3

READS
317

2 authors:

Noor Azah Samsudin
Universiti Tun Hussein Onn Malaysia
85 PUBLICATIONS 571 CITATIONS
SEE PROFILE

Andrew P Bradley
Queensland University of Technology
226 PUBLICATIONS 13,972 CITATIONS
SEE PROFILE

All content following this page was uploaded by Andrew P Bradley on 08 March 2016.

The user has requested enhancement of the downloaded file.

Extended Naïve Bayes for Group Based Classification

Noor Azah Samsudin[1] and Andrew P. Bradley[2]

[1] Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johor, Malaysia
{azah,mmustafa,shamsulk}@uthm.edu.my
[2] School of Information Technology and Electrical Engineering
The University of Queensland, 4067 QLD, Australia
bradley@itee.uq.edu.au

Abstract. This paper focuses on extending Naive Bayes classifier to address group based classification problem. The group based classification problem requires labeling a group of multiple instances given the prior knowledge that all the instances of the group belong to same unknown class. We present three techniques to extend the Naïve Bayes classifier to label a group of homogenous instances. We then evaluate the extended Naïve Bayes classifier on both synthetic and real data sets and demonstrate that the extended classifiers may be a promising approach in applications where the test data can be arranged into homogenous subsets.

Keywords: group based classification, Naïve Bayes, classification.

1 Introduction

Group based classification (GBC) problem is about labeling a group of multiple instances with the prior knowledge that all the instances in the group belong to same but unknown class [1, 2]. The GBC problem arises in various applications in which there is a need to determine class membership of an object represented by multiple instances such as in cervical cancer screening [3-6] and plant species classification problems discussed in [2].

This paper introduces three techniques to extend Naive Bayes [7] classifier to label a group of instances. We then evaluate the extended Naive Bayes classifiers on both synthetic and real data sets and demonstrate that the extended Bayes classifiers may be a promising approach in applications where the test data can be arranged into homogenous subsets. The performances of the proposed classifiers are compared with the conventional Naive Bayes classifier and another GBC technique, namely F-test based classifier [2].

The rest of the paper is organized as follows. Section 2 formally reviews the property of Naïve Bayes classifier. Section 3 describes the three techniques that we implemented for solving the GBC problem. Section 4 presents the data sets and our experiment methodology. Section 5 presents the results of various techniques and finally provides some concluding remarks.

T. Herawan et al. (eds.), *Recent Advances on Soft Computing and Data Mining SCDM 2014*, Advances in Intelligent Systems and Computing 287,
DOI: 10.1007/978-3-319-07692-8_47, © Springer International Publishing Switzerland 2014

2 Naïve Bayes Classifier

The principal idea of Naïve Bayes classifier originates from Bayes' Theorem [7]. In a classification problem, we are given a data set consisting of N instances and their associated class labels, such as $D = \{(\mathbf{x}^1_, c^1), (\mathbf{x}^2_, c^2),\ldots, (\mathbf{x}^N_, c^N)\}$. Each instance $\mathbf{x}$ is represented by n-dimensional measurements, which are also known as feature vectors—that is, $\mathbf{x} = (f_1, f_2, \ldots, f_n)$. The n-dimensional measurements presented in each instance are obtained from a set of features, $F_1, F_2, \ldots, F_n$. c^N is a class label for $\mathbf{x}^N$, where l belongs to a set of class labels, such that $l=1, \ldots,L, L > 1$.

The aim of a classification problem is to determine the class membership of a single instance, $\mathbf{x}$. Applying Bayes' Theorem, the class membership is determined according to the class that has the highest posterior probability, conditioned on $\mathbf{x}$. Using the Bayes' Theorem, the principal idea of the classification problem can be written as:

$$P(c_l \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c_l)P(c_l)}{P(\mathbf{x})} \tag{1}$$

As $P(\mathbf{x})$ is constant for all classes, $P(\mathbf{x}|c_l)P(c_l)$ therefore needs to be maximised. If the prior probabilities $P(c_l)$ are unknown, the common assumptions are that the classes are equally likely, $P(c_1) = P(c_2) = \ldots = P(c_L)$, and we would therefore maximise $P(\mathbf{x}|c_l)$. That is, the classifier labels $\mathbf{x}$ belongs to class c_i only if:

$$P(c_i \mid \mathbf{x}) > P(c_j \mid \mathbf{x}) \text{ for } 1 \le j \le L, j \ne i$$

Clearly, a class label is determined according to the most likely possible classifications. Note that class prior probabilities can be estimated using the number of instances of class c^l presented in the given data set. A class label c for a single instance $\mathbf{x}$ using Equation (1) is given as:

$$c = \arg\max_{l=1..L} P(c_l)P(\mathbf{x} \mid c_l) \tag{2}$$

where $P(c^l)$ is the prior probability. Due to the naïve independence assumption in Naïve Bayes, all features $F_1, F_2, \ldots, F_n$ are conditionally independent given a class label. Therefore, $P(\mathbf{x}|c^l)$ can be decomposed into a product of n terms, one term for each feature, such as:

$$P(\mathbf{x} \mid c_l) = \prod_{i=1}^{n} P(\mathbf{x} = f_i \mid c_l) \tag{3}$$

Thus, the Naïve Bayes classification rule to determine class label c for an instance $\mathbf{x}$ can be defined as:

$$c = \arg\max_{l=1..L} P(c_l) \prod_{i=1}^{n} P(\mathbf{x} = f_i \mid c_l) \tag{4}$$

Given a data set *D* with *N* instances, if the prior probability $P(c_l)$ is unknown, we can estimate $P(c_1)$ for each class *l*, $\hat{P}(c_l) = \frac{N_l}{N}$, N_l is the number of instances of class *l*. Assuming that the measurements of all *n* features follow the Gaussian distribution, we can estimate the conditional probability for $P(\mathbf{x}= f_i \mid c^l)$ as:

$$P(\mathbf{x} = f_i \mid c_l) = g(f_i \; ; \mu_i, \sigma_i) \text{, e.g.}$$

$$g(f_i \; ; \mu_i, \sigma_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ - \frac{(f-\mu)^2}{2\sigma^2} \right\} \tag{5}$$

The mean μ_i and the standard deviation σ_i are estimated using the measurements of features in the data set *D*.

3 Extended Naïve Bayes Classifier

We propose to extend the Naïve Bayes classifier to address the GBC problem as follows:

$$P(c_l \mid \mathbf{X}_{\mathrm{TE}}) = \frac{P(\mathbf{X}_{\mathrm{TE}} \mid c_l) P(c_l)}{P(\mathbf{X}_{\mathrm{TE}})} \tag{6}$$

given that $\mathbf{X}_{\mathrm{TE}}$ is a group of N_{TE} instances—for example, $\mathbf{X}_{\mathrm{TE}} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^{N_{\mathrm{TE}}}\}$, in which these N_{TE} instances belong to the same unknown class. The GBC approach clearly emphasises that the class label decision is dependent not just on a single instance, but on a group of instances. Note that our GBC aims to determine class label ***c*** for a group of instances. This is different from determining *c* for a single instance **x** presented in Equation (5). Assuming the features of the instances are independent, we propose three approaches to extend the Naïve Bayes classifier to implement the GBC:

1. Naïve Bayes (Voting) (NBV): We estimate posterior probability for every *k*-th instance, $\mathbf{x}^k$ in $\mathbf{X}_{\mathrm{TE}}$. Based on the posterior probability estimation, the class label for every *k*-th instance c_l^k is determined—for example, $c_l^k = \arg\max_{l=1\ldots L} p(c_l \mid \mathbf{x}^k)$. As the Bayes classifier outputs a posterior probability, this voting is actually a threshold at 0.5—that is, for a two-class problem, *i* and *j*, $p(c^i|\mathbf{x}) + p(c^j| \mathbf{x}) = 1$. The total vote of every class label v_l is then determined from the instances in $\mathbf{X}_{\mathrm{TE}}$—for example, $v_l = \sum_{k=1}^{N_{TE}} c_l^k$. Finally, $\mathbf{X}_{\mathrm{TE}}$ is labelled with the class of the majority vote—for example, $\boldsymbol{c}_l = \arg\max_{l=1\ldots L} v_l$.

2. Naïve Bayes (Naive Pooling) (NBP): Like in NBV, the posterior probability is estimated for every instance $\mathbf{x}$ in $\mathbf{X}_{\mathrm{TE}}$. Unlike NBV, these instances are not labelled individually. Instead, combined probabilities are estimated for every class l, $p(C_l)$, using instances in $\mathbf{X}_{\mathrm{TE}}$—for example, $p(C_l) = \dfrac{\prod_{k=1}^{N_{TE}} p(c_l \mid \mathbf{x}^k)}{\prod_{k=1}^{N_{TE}} p(c_l \mid \mathbf{x}^k) + \prod_{k=1}^{N_{TE}} (1 - p(c_l \mid \mathbf{x}^k))}$. Finally, $\mathbf{X}_{\mathrm{TE}}$ is labelled with a class with maximum combined probability—for example, $c_l = \arg\max_{l=1...L} p(C_l)$. The combined probability is related—for example, $p(C_i) = 1\text{-}p(C_j)$, for $i \neq j$.
3. Our naïve assumption in NBP is that each instance in $\mathbf{X}_{\mathrm{TE}}$ is independent from each other, in which the assumption is contrary to our proposed GBC assumption. It is unlikely that individual instances in the same group (e.g. sample petals from the same plant species or cells in a slide) are independent; in fact, the GBC assumes the opposite.
4. Naïve Bayes (Direct Pooling) (DNP): We aim to implement the proposed group-based classifier, $P(c_l \mid \mathbf{X}_{\mathrm{TE}}) = \dfrac{P(\mathbf{X}_{\mathrm{TE}} \mid c_l) P(c_l)}{P(\mathbf{X}_{\mathrm{TE}})}$ presented in Equation (6) directly. That is, we want to use all instances in $\mathbf{X}_{\mathrm{TE}}$ as a group to observe similarity with every class training set $\mathbf{X}_l$. Again in DNP, we make the naïve assumption that features are independent in order to estimate the probability density function (PDF) for every class training set $p(\mathbf{X}_l|c_l)$ for $l = 1,..., L$. We also estimate the PDF for $\mathbf{X}_{\mathrm{TE}}$, $p(\mathbf{X}_{\mathrm{TE}})$. Upon obtaining PDFs for the training sets $p(\mathbf{X}_l|c_l)$ and the test set $p(\mathbf{X}_{\mathrm{TE}})$, the next step is to use these PDFs to estimate the posterior probability for $p(c_l|\mathbf{X}_{\mathrm{TE}})$. Here, we apply the Kolmogorov–Smirnov test (also known as the K–S test) to measure the differences between every class $p(\mathbf{X}_l|c_l)$ and $p(\mathbf{X}_{\mathrm{TE}})$ using the empirical cumulative distribution function (CDF)—for example, $\mathrm{CDF}_l = cdf(p(\mathbf{X}_l|c_l))$ and $\mathrm{CDF}_{\mathrm{XTE}} = cdf(p(\mathbf{X}_{\mathrm{TE}}))$. The K–S test results in a probability of differences between CDF_l and $\mathrm{CDF}_{\mathrm{XTE}}$ e.g. p_l = K-S_test (CDF_l, $\mathrm{CDF}_{\mathrm{XTE}}$). Note that we are estimating the p_l in a naïve manner—that is, one feature at a time. To accommodate n features, we therefore determine the combined probability value for every class l—for example, $\mathbf{p}_l = \prod_{i=1}^{n} p_l^i$. Finally, $\mathbf{X}_{\mathrm{TE}}$ is labelled with a class with minimum combined probability—for example, $c_l = \arg\min_{l=1...L} \mathbf{p}_l$.

Different to NBV, where an individual instance is assigned a class label prior to the group classification stage, in DNP, the class label is assigned directly to the group. Different to NBP, where the instances in the same group are assumed to be independent in contrast to the GBC assumption, in DNP, the assumption of our proposed GBC is not violated.

4 Experiment Methodology

The purpose of our experiments was to initially investigate the efficacy of GBC on both synthetic and real data sets. As all of our experiments involved approximately equal class priors, the error rate was thought to be an appropriate measure of classification performance. This was estimated using 10-fold cross-validation [7, 8]. In the experiments, all instances in the data sets were used, and each cross-validation partition (fold) was randomly selected in order to preserve prior class probability. For every class l, one partition was used as the test data, $\mathbf{X}_{\mathrm{TE}}$, and the remaining partitions as the training data, $\mathbf{X}_l$. However, to plot the classification error rate as a function of group size, all possible subsets of $\mathbf{X}_{\mathrm{TE}}$ larger than size three were evaluated. For every combination size, the error rate was estimated by dividing the number of misclassified instances by the total instances, N.

In all experiments, the performances of the GBC techniques were compared against the Naïve Bayes classifier, which was chosen because the synthetic data sets were normally distributed; thus, it should perform well. Note that for the Naïve Bayes classifier, only one instance from the group is presented to the classifier at a time. Conversely, for the GBC techniques, a group of instances is presented to the respective classifier so that once the group is labelled as belonging to a particular class, all instances in the group are classified as belonging to that class.

4.1 Synthetic Data

We conducted our experiments with commonly used Gaussian data sets—namely the I-I, I-Λ and I-4I data sets originally developed by Fukunaga [9]. Notably, each data set has a different level of 'difficulty', with calculated Bayes error rates of 10 per cent, 1.9 per cent and 9 per cent for I-I, I-Λ and I-4I respectively. Each data set consists of an eight-dimensional data vector with 1,000 samples per class. In these synthetic data sets, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors for class 1 and class 2 respectively. Meanwhile, $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are the corresponding covariance matrices for each class. The values of the $\boldsymbol{\mu}_l$ and $\boldsymbol{\lambda}_l$ are given in Table 1, where I_8 is the 8×8 identity matrix. For the I-Λ data set, $\boldsymbol{\mu}_2$and $\boldsymbol{\lambda}_2$ are provided in Table 2. As we are using 10-fold cross-validation each test partition, $\mathbf{X}_{\mathrm{TE}}$**,** consists of 100 instances per class. For every test partition all possible subsets *(groups)* of size three and above were evaluated. In this way, the proposed GBC techniques determined the class labels for variously sized subsets of the test data. We chose to have subsets of odd-numbered size to avoid tie voting in experiments with NBV approach.

Table 1. Synthetic data sets I-I, I-4I and I-Λ

Data set	μ_1	μ_2	λ_1	λ_2
I-I ($\mu_{1\neq}\ \mu_2$, $\lambda_{1=}\lambda_2$)	0	[2.56, 0,…,0]	I_8	
I-Λ ($\mu_{1\neq}\ \mu_2$, $\lambda_{1\neq}\lambda_2$)		[μ_1,…, μ_8] (Table 2)		[λ_1,…, λ_8] (Table 2)
I-4I ($\mu_{1=}\ \mu_2$, $\lambda_{1\neq}\lambda_2$)		0	I_8	$4I_8$

Table 2. Parameter values of the I-Λ data set

Dimension i	1	2	3	4	5	6	7	8
μ_i	3.86	3.10	0.84	0.84	1.64	1.08	0.26	0.01
λ_i	8.41	12.06	0.12	0.22	1.49	1.77	0.35	2.73

4.2 Iris Data

The iris data set is a collection of plant species from three classes: *Iris setosa*, *Iris versicolor* and *Iris virginica* [10, 11]. There are 50 samples from each class, and each sample is represented by measurements of four features: petal length, petal width, sepal length and sepal width. The data set is obtained from the UCI Machine Learning Repository (http://archive.ics.uci.edu). With 10-fold cross-validation, each test partition, $\mathbf{X}_{TE}$, consists of five (homogenous) instances of unknown class. Therefore, we classified every combination of the test set of size three to five instances.

5 Results

Figures 1–3 show the plots of error rates as a function of group size for each case of synthetic data. In all cases, the error rate for the GBC techniques approach zero as the group size increases. For all three synthetic cases, most of the proposed classifiers outperform the Naïve Bayes when the combination size is larger than seven. Indeed, it is interesting to note that an error rate of zero can be achieved with these data,

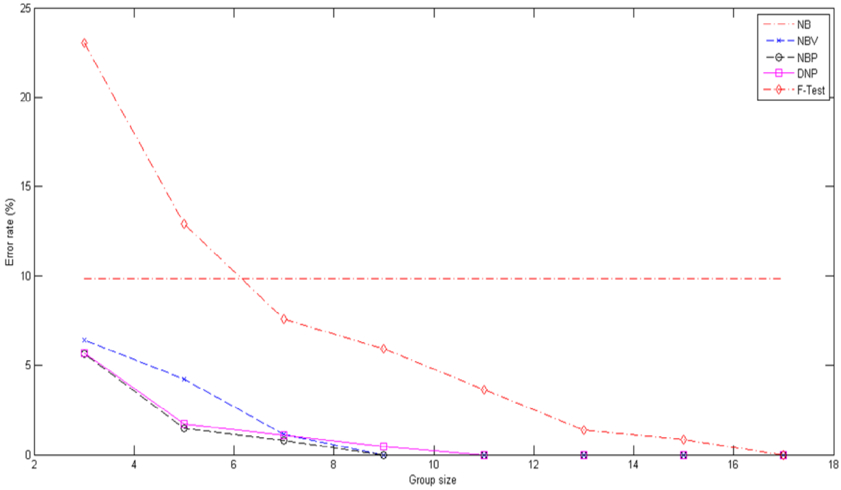

Fig. 1. Comparison of error rate (%): GBC techniques and Naïve Bayes for Case II

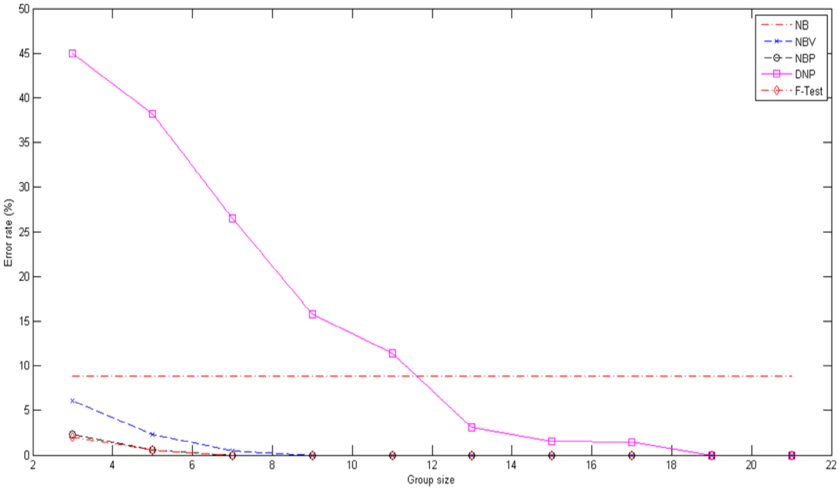

Fig. 2. Comparison of error rate (%): GBC techniques and Naïve Bayes for Case I-4I

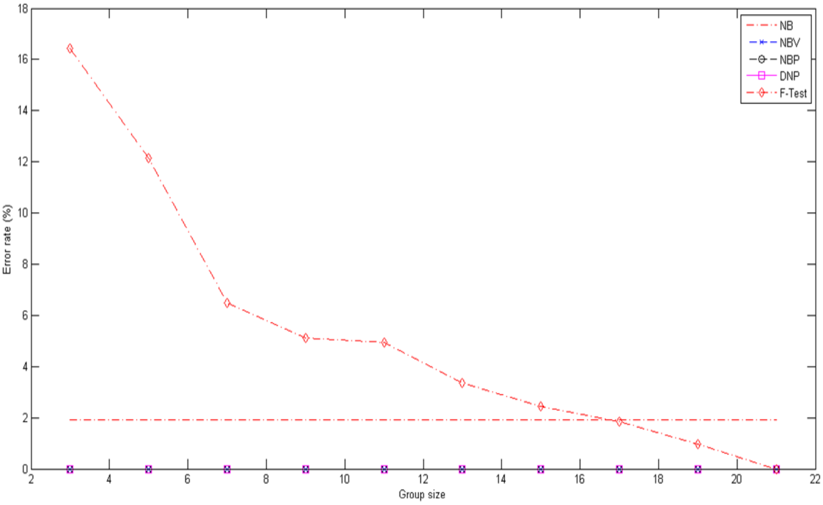

Fig. 3. Comparison of error rate (%): GBC techniques and Naïve Bayes for Case **I-Λ**

which by definition have overlapping class probability distributions and thus a non-zero Bayes error. This indicates the potential benefits of utilising the additional prior knowledge implicit to GBC—that is, that a group of test instances has the same but unknown class membership. The Iris data set is used as one potential practical application of GBC by arranging the test data into homogenous subgroups. The results in Figure 4 suggest that GBC outperforms the Naïve Bayes classifier when the group size is three or more. Clearly, GBC is benefitting from the prior knowledge that all test samples in the sub-group are homogeneous and should be given the same class label.

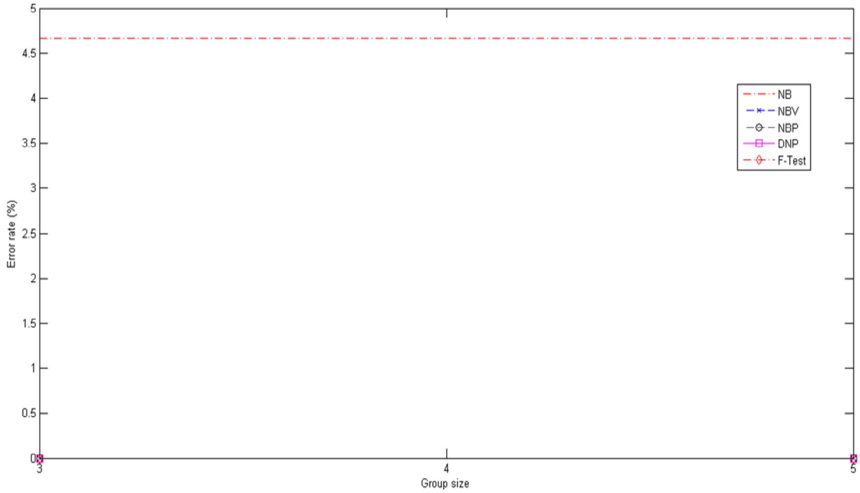

Fig. 4. Comparison of error rate (%): GBC techniques and Naïve Bayes for the Iris data set

6 Conclusions

In this paper, we presented the underlying concepts and rationale behind GBC. We developed and evaluated four GBC techniques, comparing them to individual based classification technique, the conventional Naïve Bayes classifier. In particular, three different ways of extending the Naive Bayes classifier to accumulate information about a group of test samples were presented: voting, naive pooling and direct pooling. The extended Naive Bayes classifiers were then evaluated for a variety of group sizes using both synthetic and real-world data, and their performances were evaluated in terms of average error rate. The results indicate that the proposed GBC techniques have the potential to outperform the Naive Bayes classification technique, especially as the (group) size of the test set increases. Clearly, these results indicate that the additional prior knowledge that a group of test samples belongs to the same but unknown class label can be effectively utilised to reduce misclassification problems.

References

1. Samsudin, N.A., Bradley, A.P.: Nearest neighbour group-based classification. Pattern Recognition 43, 3458–3467 (2010)
2. Samsudin, N.A., Bradley, A.P.: Group-based meta-classification. In: 19th International Conference on Pattern Recognition, IEEE, Tampa (2008)
3. Moshavegh, R., Bejnordi, B.E., Mehnert, A., Sujathan, K., Malm, P., Bengtsson, E.: Automated segmentation of free-lying cell nuclei in Pap smears for malignancy associated change analysis. In: 34th Annual International Conference of the IEEE EMBS, pp. 5372–5375 (2012)

4. Bengtsson, E.: Computerized cell image analysis: Past, present, and future. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 395–407. Springer, Heidelberg (2003)
5. Nordin, B., Bengtsson, E.: Specimen analysis by rare event, cell population, and/or contextual evaluation. In: Grohs, H.K., Husain, O.A.N. (eds.) Automated Cervical Cancer Screening, pp. 44–51. IGAKU-SHOIN Medical Publishers, New York (1994)
6. Mehnert, A.J.H.: Image analysis for the study of chromatic distribution in cell nuclei with application to cervical cancer screening. School of Information Technology and Electrical Engineering, vol. Phd. The University of Queensland, Australia (2003)
7. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, New York (2001)
8. Alpaydin, E.: Introduction to machine learning. The MIT Press, London (2004)
9. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Press (1990)
10. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugenics, 179–188 (1936)
11. Duda, R.O., Hart, P.E.: Pattern classification and scene analysis. John Wiley & Sons (1973)

View publication stats