Imbalanced Learning in Relevance Feedback with Biased Minimax Probability Machine for Image Retrieval Tasks

Xiang Peng and Irwin King

Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{xpeng, king}@cse.cuhk.edu.hk

Abstract. In recent years, Minimax Probability Machine (MPM) have demonstrated excellent performance in a variety of pattern recognition problems. At the same time various machine learning methods have been used on relevance feedback tasks in Content-based Image Retrieval (CBIR). One of the problems in typical techniques for relevance feedback is that they treat the relevant feedback and irrelevant feedback equally. In other words, the negative instances largely outnumber the positive instances. Hence, the assumption that they are balanced is incorrect. In this paper we study how MPM can be applied to image retrieval, more precisely, Biased MPM during the relevance feedback iterations. We formulate the relevance feedback based on a modified MPM called Biased Minimax Probability Machine (BMPM). Different from previous methods, this model directly controls the accuracy of classification of the future data to build up biased classifiers. Hence, it provides a rigorous treatment on imbalanced data. Mathematical formulation and explanations are provided for showing the advantages. Experiments are conducted to evaluate the performance of our proposed framework, in which encouraging and promising experimental results are obtained.

1 Introduction

With the recent progress of devices for capturing and storing image data, Content-based Image Retrieval (CBIR) has attracted a lot of research interests in the past decade [10]. However two semantically similar image may be located far from each other in the feature space, while two absolutely different images may lie close to each other [9]. This is known as the problem of semantic gap between low-level features and high-level concepts [11]. Relevance feedback has been shown to be a powerful tool to address this problem and improve retrieval performance in CBIR [3]. Recently, researchers proposed a number of classification techniques to attack relevance feedback tasks including some state-of-the-art models such as Support Vector Machines (SVMs) [3]. However most of the classification techniques treat the relevance feedback problem as a strict binary classification problem and they do not consider the imbalanced dataset problem, which means the number of irrelevant images are significantly larger than the number of relevant images. This

I. King et al. (Eds.): ICONIP 2006, Part I, LNCS 4232, pp. 342–351, 2006.

imbalanced dataset problem would lead the positive data (relevant images) be overwhelmed by the negative data (irrelevant images).

Minimax Probability Machine (MPM) has been used as a very important tool to perform classification tasks [7]. Compared with SVMs, it has promising performance. In order to attack the problem of imbalance dataset in CBIR, we propose to use a modified Minimax Probability Machine, called Biased Minimax Probability Machine (BMPM) which can better model the relevance feedback problem and reduce the performance degradation caused by the imbalanced dataset problem.

The rest of this paper is organized as follows. We first review some related research efforts on relevance feedback and MPM in Section II. In Section III, we present and formulate the BMPM which is derived from MPM. We then formulate the relevance feedback technique employing Biased MPM and show the benefits compared with the conventional techniques. Experiments, performance evaluation, and discussions are given in Section IV and Section V concludes our work.

2 Related Work

Here we will give a brief introduction about the related work in the research area of relevance feedback in CBIR and the theory of MPM.

Relevance Feedback for CBIR. Relevance feedback techniques have been used as a powerful tool for Content-based Image Retrieval [13]. There are various methodologies involving in that research area such as using Self-organizing Map (SOM) [1], Decision Tree [6], Artificial Neural Network [2], and Bayesian Learning Network [12], etc. Moreover many newly proposed popular classification techniques have been proposed to tackle the relevance feedback, such as Bayesian classifiers and Support Vector Machines (SVMs) [10], etc. Among them, SVM-based techniques are the quite promising and effective techniques to solving the relevance feedback tasks.

Minimax Probability Machine. We here introduce the basic concept of Minimax Probability Machine [7], which translate a classification problem into an optimization problem. In pattern classification problems, MPM provides very good generalization performance in empirical applications.

Let us illustrate MPM in a binary classification case. Suppose two random n-Dimensional vectors, $\mathbf{x}$ and $\mathbf{y}$ represent two classes of data, where $\mathbf{x}$ belongs to the family of distributions with a given mean $\overline{\mathbf{x}}$ and a covariance $\Sigma_{\mathbf{x}}$, denoted as $\mathbf{x} \sim (\overline{\mathbf{x}}, \Sigma_{\mathbf{x}})$; similarly, $\mathbf{y}$ belongs to the family of distributions with a given mean $\overline{\mathbf{y}}$ and a covariance $\Sigma_{\mathbf{y}}$, denoted as $\mathbf{y} \sim (\overline{\mathbf{y}}, \Sigma_{\mathbf{y}})$. Here $\mathbf{x}$, $\mathbf{y}$, $\overline{\mathbf{x}}$, $\overline{\mathbf{y}} \in R^n$, and $\Sigma_{\mathbf{x}}$, $\Sigma_{\mathbf{y}} \in R^{n \times n}$. In this paper, class $\mathbf{x}$ represents the relevance image class and $\mathbf{y}$ represents the irrelevance image class.

The Minimax Probability Machine attempts to determine the hyperplane $\mathbf{a}^T\mathbf{z} = b$ ($\mathbf{a} \in R^n$, $\mathbf{z} \in R^n$, $b \in R^n$) which can separate two classes of data

with maximal probability. The formulation for the original MPM model [8] is written as follows:

$$\begin{aligned} \max_{\alpha,\mathbf{a}\neq\mathbf{0},b} \quad & \alpha \quad s.t. \\ & \inf_{\mathbf{x}\sim(\bar{\mathbf{x}},\Sigma_{\mathbf{x}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq a\} \geq \alpha, \\ & \inf_{\mathbf{y}\sim(\bar{\mathbf{y}},\Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \alpha, \end{aligned} \tag{1}$$

where α represents the lower bound of the accuracy for future data. Future points $\mathbf{z}$ for which $\mathbf{a}^T\mathbf{z} \geq \alpha$, are then classified as the class $\mathbf{x}$; otherwise they are judged as the class $\mathbf{y}$.

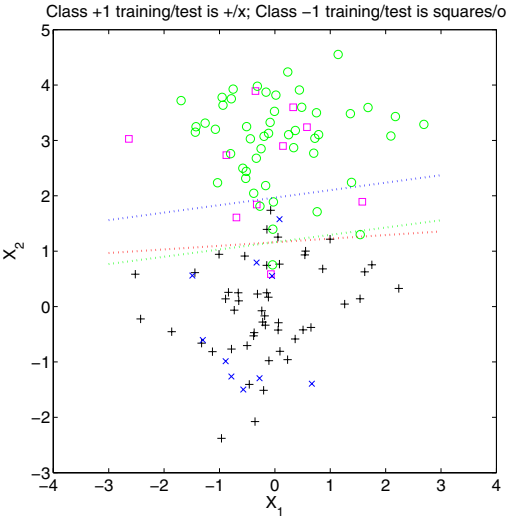


Fig. 1. Decision lines comparison: MPM decision line (dotted red line), BMPM decision line (dotted green line), SVM decision line (dotted blue line)

Later, Huang et al. [4] improved the model by removing away the assumption that these two classes have the same importance, and furthermore adding a bias to the more important class. As we could observe from the above formulation, this model actually assumes that two classes have the same importance. Hence it makes the worst-case accuracies for two classes the same. However, in real applications, especially in relevance feedback of content-based image retrieval, two classes of data are usually biased, i.e., the relevant class is often more important than the irrelevance class and the quantities of both datasets are imbalanced. Therefore it is more appropriate to take the inherited bias into account in this context. In the following section, we will introduce Huang's development, an extensions of MPM, i.e., the Biased Minimax Probability Machine (BMPM), which is used in our proposed framework.

3 Relevance Feedback Using Biased MPM

In this section, we first introduce the model definition of BMPM. Next, we show the benefits by applying BMPM in Relevance Feedback in Content-Based Image Retrieval. We then in Section 3.3, present and formulate our proposed Biased MPM methodology, applying to Relevance Feedback.

3.1 Model Definition

We assume two random vectors $\mathbf{x}$ and $\mathbf{y}$ represent two classes of data with means and covariance matrices as $\{\overline{x}, \Sigma x\}$ and $\{\overline{y}, \Sigma y\}$, respectively in a two-category classification task.

With given reliable $\{\overline{x}, \Sigma x\}$, $\{\overline{y}, \Sigma y\}$ for two classes of data, we try to find a hyperplane $\mathbf{a}^T\mathbf{z} = b(\mathbf{a} \neq 0, \mathbf{z} \in R^n, b \in R)$ with $\mathbf{a}^T\mathbf{z} > b$ being considered as class $\mathbf{x}$ and $\mathbf{a}^T\mathbf{z} < b$ being judged as class $\mathbf{y}$ to separate the important class of data ($\mathbf{x}$) with a maximal probability while keeping the accuracy of less important class of data ($\mathbf{y}$) acceptable.[1] We formulate this objective as follows:

$$\begin{aligned} \max_{\alpha,\beta,b,\mathbf{a}\neq\mathbf{0}} \quad & \alpha \quad s.t. \\ & \inf_{\mathbf{x}\sim(\overline{\mathbf{x}},\Sigma_{\mathbf{X}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha, \\ & \inf_{\mathbf{y}\sim(\overline{\mathbf{y}},\Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta, \\ & \beta \geq \beta_0, \end{aligned} \tag{2}$$

where α represents the lower bound of the accuracy for the classification, or the worst-case accuracy of future data points $\mathbf{x}$; the same for β. The parameter β_0 is a pre-specified positive constant, which represents an acceptable accuracy level for the less important class $\mathbf{y}$.

The above formulation is derived from MPM, which requires the probabilities of correct classification for both classes to be an equal value α. Through this formulation, the BMPM model can handle the imbalanced classification in a direct way by changing the value of $\beta_\mathbf{0}$. This model provides a different treatment on different classes, i.e., the hyperplane $\mathbf{a}_*^T\mathbf{z} = b_*$ given by the solution of this optimization will favor the classification of the important class $\mathbf{x}$ over the less important class $\mathbf{y}$. Furthermore, the derived decision hyperplane is directly associated with two real accuracy indicators of classification of the future data, i.e., α and β, for each class.

3.2 Advantages of BMPM in Relevance Feedback

From the above formulations, one may see that the optimization in BMPM is similar to the one in the MPM, which is in SOCP format and could be efficient solved by SeduMi or Mosek [4]. Now, we show the mathematical differences and the advantages of our proposed BMPM framework from an analytical perspective for solving the relevance feedback problems compared with MPMs and SVMs.

[1] The reader may refer to [5] for a more detailed and complete description.

Obviously we see that BMPM is with the following constraint, contrast to the one of MPM in [8] and the one of MEMPM in [5]

$$\begin{aligned} \inf_{\mathbf{x}\sim(\bar{\mathbf{x}},\Sigma_{\mathbf{x}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha, \\ \inf_{\mathbf{y}\sim(\bar{\mathbf{y}},\Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta, \\ \beta \geq \beta_0, \end{aligned} \tag{3}$$

The difference indicates that our proposed BMPM framework tries to improve the accuracy of relevant images while maintaining an acceptable specificity of irrelevant ones. This methodology provides a rigorous way to handle the relevance feedback problem by directly control the accuracy and this can be useful for solving this imbalanced dataset problem. However, SVMs and Neural Networks treat the two classes without any bias or direct control, which is not effective to model and solve the relevance feedback problem.

3.3 Modified Relevance Feedback Framework by BMPM

Here, we describe how to formulate the relevance feedback algorithm by employing the BMPM technique. Applying BMPM-based technique in relevance feedback is similar to the conventional classification tasks. However, the relevance feedback needs to construct an iterative function to produce the value of the retrieval results. The following is our proposed methodology for retrieval task in CBIR.

Framework 1. BMPM-based Relevance Feedback

Input: $\mathbf{Q}_{im}$ (query image, represented by its features)
Output: $\mathbf{R}_{im}$ (images belongs to the same class with similar semantic content)

F$_q \Leftarrow$ **Q**$_{im}$ /*Feature extraction for query image*/
F$_q \Leftarrow$ **x**/**y** /*Assign labels to query image*/
For i = 1: MaxIt
    R$_{im} \Leftarrow$ **R**$^i_{im}$ /*Update*/
    Involve feedback information using BMPM
        R$^1_{im}$, **R**$^2_{im} \Leftarrow$ **R**$_{im}$ /*Separate returned images into two sets*/
        R$^1_{im}$, **R**$^2_{im} \Leftarrow \{\mathbf{x}, \mathbf{y}\}$ /*Assign labels to classes by experts*/
        Classification task by BMPM
    $i \Leftarrow i + 1$
end For
Return R$_{im}$

After a certain number iterations of relevance feedback finished, our proposed strategy returns the *Top-k* most relevant images and also learn a more reasonable classifier to classify the imbalanced image datasets.

4 Evaluation and Experiment Results

We implement this machine learning scheme and apply to relevance feedback in Content-based Image Retrieval. In this section, we describe the iterative framework and show the experimental results. Moreover we compare the performance of some different algorithms for relevance feedback: Minimax Probability Machine (MPM), Support Vector Machines (SVMs) and our proposed BMPM-based approach. The experiments are evaluated on both a synthetic dataset and a real-world image dataset. All our works are done on a 3.2GHz machine with Intel Pentium 4 processor having 1Gb RAM.

4.1 Experiment Datasets

Synthetic Dataset. We generate a synthetic dataset to simulate the real-world image dataset. The dataset consists 10 categories, 9 of which contains 100 data points randomly generated by Gaussian Distribution with different means and covariances. The remaining class contains 30 instances generated by another mean and covariance, denoting the relevant samples.

Real-world Datasets. The dataset are chosen from the COREL image CDs. We organize one dataset which contain various images with different semantic meanings, such as *bird*, *pyramid*, *model*, *autumn*, *dog*, and *glacier*, etc. It is with 6 categories (6-Bird). The class of *bird* which we recognize as the positive class contains 50 images while the other 5 categories each includes 100 instances belonging to the negative classes.

Image Representation. Here we extract three different features to represent the images: *color*, *shape*, and *texture*. The color feature employed is the color histograms. We quantize the number of pixels into 10 bins for each color channel (H, S, and V) respectively. Then we could obtain a 30-dimensional color

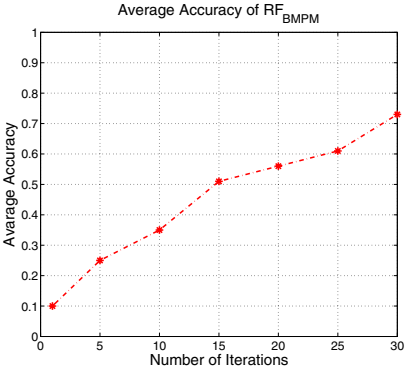


Fig. 2. Average accuracy of proposed algorithm for Relevance Feedback in Synthetic Dataset

histogram. We use edge direction histogram as shape feature to represent an image. First we compute the edge images by Canny edge detector and obtain the edge direction histogram by quantizing the results into 15 bins of 20 degrees. Therefore a 15-dimensional edge direction histogram is used as the edge feature. We apply the wavelet-based texture in our experiments. Gabor Wavelet Decomposition is first performed and we compute the features for each Gabor filter output afterwards. Following this approach we obtain a 16-dimensional vector to represent the texture information for each image.

4.2 Results on Synthetic Dataset

In the experiments, the relevant category is first picked from the dataset. Our framework then improves retrieval results by relevance feedbacks from the user. We select 10 points from the dataset and label them as either relevant or irrelevant based on the ground truth of the dataset in each iteration. Three positive instances and eight negative points are randomly picked for the first iteration, and the machine learning scheme of BMPM is applied with this initial set. Our proposed framework selects 10 points from the dataset for the iterations afterward and record the number of the points in the positive and negative regions during the retrieval process. More specifically, the whole process is repeated for 30 times before calculating the average performance measurements.

4.3 Results on Realworld Dataset

In the following, we present the experimental results by the algorithm on real-world images. The metric of evaluation is the *Average Precision* which is defined as the average ratio of the number of relevant images of the returned images over the total number of the returned images.

In the real-world dataset experiments, the iteration is similar to synthetic dataset, except that we need to first perform feature extraction for the query images and image database. In each iteration of the feedback process, 10 images are picked from the database and labelled as either relevant or irrelevant based on the ground truth of the database. The precision is then recorded, and the whole process is repeated for 10 times to produce the average precision in each iteration for the proposed method.

The framework implemented in our experiments is based on modifying the codes in the BMPM tool box [5]. To enable an objective measure of performance, we choose the same parameters for all the settings such as α and β_0 et al.

The first evaluation is about the effect of iterations in our proposed method. Fig. 4 shows the evaluation results of returned images with no feedback and 10-iterations. We can observe that our method efficiently improves the retrieval result.

4.4 Comparison of SVMs, MPM and BMPM

Here are some experimental results that we have gathered from our framework to validate some simple assumptions and demonstrate the effectiveness of our

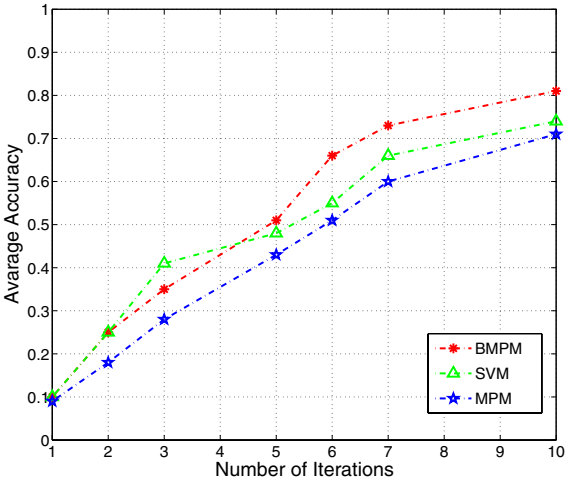
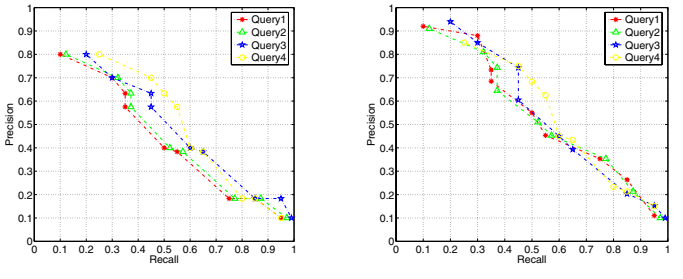


Fig. 3. Performance comparison of three algorithms for Relevance Feedback

proposed learning scheme. Because we are interested in examining how much the proposed method improves the retrieval results with the user's feedback, we conduct the evaluations of these three different algorithms over the image dataset. Furthermore, we also test the number of relevant images returned for a fixed iteration number.



(a) No Relevance Feedback **(b)** After 10 Iterations of Feedback

Fig. 4. Algorithm performance over COREL Images Dataset

Fig. 3 shows the evaluation results on the 6-Bird dataset. From the results on the real-world dataset, we can observe that our proposed BMPM-based methodology outperforms other approaches such as MPM and SVMs. From here we could see how the bias works. In order to know the detailed comparison of the three methods after a set number of iterations, we list the retrieval results in Table 1. From the results, we can also see similar results that verify our hypothesis.

Table 1. Number of relevant images in Top 50 Returns

Different	Number of Iterations				
Algorithms	0	1	3	7	10
BMPM	5	**9**	22	**35**	**42**
MPM	5	6	20	32	39
SVMs	**6**	8	**26**	33	40

5 Conclusion and Future Work

In this paper, we address the problem of imbalanced classification needed with the relevance feedback in CBIR and present a novel learning framework, the Biased Minimax Probability Machine (BMPM)-based approach, to treat this problem more precisely. In contrast to the traditional methods, the BMPM does not adopt an indirect approach, but directly controls the worst-case classification accuracy in order to impose a certain bias in favor of the relevant images. This provides a more admirable way to handle imbalanced classification problems. We evaluate the performance of the BMPM-based on the synthetic dataset and the COREL Image Dataset and compare it with two popular classifiers: Minimax Probability Machine (MPM), and Support Vector Machines (SVMs). The results on both synthetic and real-world datasets show that the BMPM outperforms the other two models on the problem of relevance feedback.

Although we could observe that our proposed learning framework is more precise than other state-of-art techniques, in some pattern recognition tasks, especially in information retrieval, effectiveness is sometimes more important than efficiency, in which expensive time-cost presents one of the main bottlenecks of the BMPM learning model. To solve such problem, a possible direction is to propose other methods to solve this optimization problem. Undoubtedly, a new learning scheme for image retrieval will be still a highly active research topic in the future.

Acknowledgment

The work described in this paper is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4235/04E) and is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

References

1. I. J. Cox, M. L. Miller, T. P. Minka, and P. N. Yianilos. An optimized interaction strategy for bayesian relevance feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1998.
2. G. Giacinto and F. Roli. Bayesian relevance feedback for content-based image retrieval. *Pattern Recognition*, 2004.

3. P. Hong, Q. Tian, and T. S. Huang. Incorporate support vector machines to content-based image retrieval with relevance feedback. In *IEEE International Conference on Image Processing*, 2000.
4. K. Huang, H. Yang, I. King, and M. R. Lyu. Imbalanced learning with biased minimax probability machine. *IEEE Transactions on System, Man, and Cybernetics*, 2005.
5. K. Huang, H. Yang, I. King, M. R. Lyu, and L. Chan. The minimum error minimax probability machine. *Journal of Machine Learning Research*, 5:1253–1286, 2004.
6. A. K. Jain and A. Vailaya. Shape-based retrieval: a case study with trademark image database. *Pattern Recognition*, 9:1369–1390, 1998.
7. G. Lanckriet, L. Ghaoui, C. Bhattacharyya, and M. I. Jordan. Minimax probability machine. In *Advances in Neural Infonation Processing Systems*, 2002.
8. G. Lanckriet, L. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2003.
9. Y. Rui, T. S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in mars. In *IEEE International Conference on Image Processing*, 1997.
10. Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content–based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.
11. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
12. Z. Su, H. Zhang, and S. Ma. Relevance feedback using a bayesian classifier in content-based image retrieval. In *SPIE Electronic Imaging*, 2001.
13. R. Yan, A. G. Hauptmann, and R. Jin. Negative pseudo-relevance feedback in content-based video retrieval. In *ACM Multimedia*, 2003.