ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Content-based image retrieval with hierarchical Gaussian Process bandits with self-organizing maps

Ksenia Konyushkova and Dorota Głowacka *

Helsinki Institute for Information Technology,
Department of Computer Science, University of Helsinki, Finland

Abstract. A content-based image retrieval system based on relevance feedback is proposed. The system relies on an interactive search paradigm where at each round a user is presented with k images and selects the one closest to her target. The approach based on hierarchical Gaussian Process (GP) bandits is used to trade exploration and exploitation in presenting the images in each round. Experimental results show that the new approach compares favorably with previous work.

1 Introduction

We consider content-based image retrieval in the case when the user is unable to specify the required content through tags or other image properties. Instead, the system must extract information from the user through limited feedback. We consider a protocol that operates through a sequence of rounds in each of which a set of k images is displayed and the user must indicate which image is closest to their ideal target image. We assume that there is a hypothetical target image in the user's mind and the user's likelihood of choosing one of the displayed images is proportional to a polynomially decaying function of the distance between the displayed images and the target. While this problem has been studied before (e.g. [1]), we propose a novel computationally efficient approach based on a 2-level hierarchical Gaussian Process bandits. At the first stage, we select a cluster containing the most promising set of images based on the user feedback and at the next stage we select an image from that cluster to present to the user. At each iteration of the search, the procedure is repeated k times to obtain k images to present to the user. The main advantage of this approach is that it is more efficient in terms of time complexity than its competitors and thus more applicable to on-line retrieval systems.

Many traditional image retrieval systems, e.g. Google Image Search, AltaVista or TinEye utilize image metadata, such as captions and tags. However, it is not always possible to tag new images in a dataset quickly and efficiently. Content Based Image Retrieval (CBIR) systems, on the other hand, rely only on features extracted directly from images. Many CBIR systems, e.g. CIRES

*The work has been partly supported by the Academy of Finland under the Finnish Center of Excellence in Computational Inference Research (COIN), by the Finnish Funding Agency for Technology and Innovation under project D2I, and by the IST Programme of the European Community under the PASCAL Network of Excellence.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

[2] or SuperFish (www.superfish.com) employ elaborate computer vision techniques in order to find images similar to the target. Pre-clustering can also be incorporated to take data distribution into account during the search [3]. There are attempts to make the image search systems scalable as, for instance, in SALSAS [4], where the complexity is almost constant for any database size thanks to locality sensitivity hashing. The problem with these approaches is that their success is highly dependent on the availability of a good example of the target image. To resolve this issue a lot of work has been done recently in an attempt to incorporate the user feedback into the search (e.g. clustering-based CBIR with relevance feedback [5] or PinView [1]). However, systems of this kind often face problems with the usability due to the running time of each iteration of the algorithm, i.e. they often perform advanced optimization tasks which take a long time to execute thus inconveniencing the user, or do not scale for large sets of images.

2 Related algorithms

There is a class of algorithms often employed CBIR systems that trade off between exploration and exploitation, i.e. they propose to the user images that are most likely to be of interest based on previous user feedback, while at the same time gaining more knowledge about users' preferences to make better suggestions in future search rounds. In this section, we discuss two algorithms against which we benchmark our system: the LinRel algorithm [6], which forms an integral part of the PinView system, and the simple GP bandit algorithm [7].

2.1 LinRel

In each iteration i, LinRel calculates estimated relevance score μ_i of each image:

$$s_i = X_p \cdot (X_{s_i}^\top \cdot X_{s_i} + \lambda I)^{-1} X_{s_i}^\top \tag{1}$$

$$\mu_i = s_i \cdot f_i, \tag{2}$$

where X_p is a matrix, where each row is a kernelized representation of images we want to estimate; X_{s_i} is a matrix with images shown up to the ith iteration and f_i is a vector of relevance feedback obtained so far. The kernel function is the Euclidean distance between images. After each iteration, LinRel selects for presentation images with the highest relevance score.

2.2 Gaussian Process Bandits

Another policy used for balancing exploration and exploitation is Gaussian Process Bandits Upper Confidence Bound (GP-UCB) algorithm [7]. The general idea is to maximize the upper confidence bound which is a combination of the predicted mean and variance of an image [8]. At each iteration i, we present to the user the image that maximizes $argmax\{\mu_i + \sqrt{\beta} \cdot \sigma_i\}$, where μ_i is a predicted mean of the relevance score, σ_i is a standard deviation and β is a constant or

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

some function of time to adjust the confidence level. In GP-UCB, we define μ as

$$\mu = K_* K^{-1} r, \tag{3}$$

and variance as

$$K_{**} - K_* K^{-1} K_*^T, \tag{4}$$

where r is the relevance feedback, and K, K_* and K_{**} are parts of the kernel matrix. K corresponds to a pairwise kernel function between all shown images so far, K_* between shown images and those whose relevance we need to predict, and K_{**} between the images whose relevance we need to predict [9].

3 The GP-SOM algorithm

The algorithms discussed in the previous section have proven theoretical regret bounds. However, in order to be applied in real-life systems they also need to be time-efficient and easily scalable to large datasets. In order to tackle this problem, we employ hierarchical GP bandits [10], where Self-Organizing Maps (SOM) [11] of image features are used as layers in the bandit hierarchy. The SOM is a discretization of input space topology that represents a non-linear projection of high dimensional space into a lower dimension. We call our algorithm GP-SOM.

3.1 Preprocessing

In order to save computation time in the on-line retrieval system, we precompute the SOMs of images. SOM is an unsupervised method for reducing dimensionality of input space by constructing an artificial neural network of instances that reflects their topological order. SOM provides the so-called model vectors that are treated in our algorithm as discretization of the input space. One of the most popular ways to obtain SOM is through Expectation-Maximization [11]. In the maximization step, when recalculating model vectors, which correspond to centroids in traditional clustering, images assigned to other model vectors are also considered and the influence they have depend on the neighborhood function between model vectors. The neighborhood function chosen in our implementation is the Gaussian kernel. The expectation step is similar to the classic K-means algorithm, i.e. images are assigned to the closest model vectors. The preprocessing step results in an objects hierarchy which serves as an input to the hierarchical GP bandits algorithm.

3.2 Hierarchical GP UCB Bandits

We apply a 2-layer bandit settings. First, we select a model vector and then we sample an image from among the images associated with a particular model vector. Thus, in the first layer, the arms are considered to be model vectors and we select one model vector. In the next step, the arms are images associated with the chosen model vector and we select one image. We repeat the selection

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

procedure k times in order to obtain k images to present to the user. In order to avoid presenting the same images to the user, we exclude the images presented so far from the temporary tree structure used for selection. At each level of the hierarchy, we apply the GP-UCB algorithm as defined above.

4 Complexity analysis

Most theoretical analysis of bandit-style algorithms concentrate on regret bounds, but the time required for computing every prediction is neglected. In this section, we analyze the running time of an on-line iteration of LinRel, GP-UCB and GP-SOM. Let N denote the number of images in the dataset. At each iteration, we present k images. Let us consider the $i+1$th iteration. We assume that the search can only last as long as all images have been displayed, so that $i \cdot k < N$. Let us denote the constant in multiplication complexity as c_m, summation as c_s and inversion as c_i. We will consider the basic linear algebra operations, where matrix multiplication of $[m \times n]$ by $[n \times p]$ takes $c_m \cdot \mathcal{O}(mnp)$, summation of two matrices of the size $[m \times n]$ takes $c_s \cdot \mathcal{O}(mn)$ and inversion of an $[n \times n]$ matrix takes $c_i \cdot \mathcal{O}(n^3)$.

Thus, the complexity of each step of LinRel does not depend on the number of iterations, but on the number of images in the dataset as $\mathcal{O}(N^3)$. One iteration of GP-UCB takes only $\mathcal{O}(Ni^2k^2)$ compared to $\mathcal{O}(N^3)$ of LinRel. At the same time, we avoid inverting a huge $[N \times N]$ matrix and do it only with a smaller $[ik \times ik]$ matrix. In GP-SOM, when building the Self-Organizing Map, we choose the number of points it contains, which means that the number of model vectors in the map is chosen to be approximately $\sqrt{N}$. Thus, the complexity of GP-SOM is $2 \cdot k \cdot \mathcal{O}(\sqrt{N}i^2k^2)$, and $2 \cdot k$ is much smaller than $\sqrt{N}$ in any realistic situation. Moreover, the Self-Organizing Map approach can be generalized into an l level hierarchical bandits by introducing additional levels in the map and increasing the complexity only by a scalar factor. If we fix the number of arms in each run to be P and allow an l level hierarchy, we can process P^l images with the complexity $l \cdot \mathcal{O}(Pi^2k^3)$.

5 Experimental Results

In the previous section, we have shown the advantage of the proposed algorithm in terms of time complexity. In order to compare the performance of the three algorithms, we ran a set of simulation experiments. We used the MIRFLICKR-25000 dataset [12] containing 25000 Flickr images consisting of 3 sets of visual descriptors: texture, shape and color. We consider 3 visual aspects of each image when constructing a map and the closest image is determined as a harmonic mean of similarities in visual aspects

$$1/(1/d_t + 1/d_s + 1/d_c), \tag{5}$$

where d_t is distance from a datapoint to the model vector in texture space, d_s in shape space, and d_c in color space.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

5.1 The User Model

We assume that the choice of one of the presented images is a random process, where more relevant images are more likely to be chosen. In our simulation experiments, we will rely on the user model proposed in [13], which has been shown to be a close approximation of real user behavior. We assume a similarity measure $S(\mathbf{x}_1, \mathbf{x}_2)$ between images $\mathbf{x}_1, \mathbf{x}_2$, which also measures the relevance of an image $\mathbf{x}$ compared to an ideal target image $\mathbf{t}$ by $S(\mathbf{x}, \mathbf{t})$. Let $0 \leq \lambda \leq 1$ be the uniform noise in the user's choice. The probability of choosing image $\mathbf{x}_{i,j}$ is given by:

$$D\{\mathbf{x}_i^* = \mathbf{x}_{i,j} \mid \mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,k}; \mathbf{t}\} = (1-\lambda)\frac{S(\mathbf{x}_{i,j}, \mathbf{t})}{\sum_{j=1}^{k} S(\mathbf{x}_{i,j}, \mathbf{t})} + \frac{\lambda}{k}. \tag{6}$$

Assuming a distance function $d(\cdot, \cdot)$, a possible choice for the similarity measure is $S(\mathbf{x}, \mathbf{t}) = d(\mathbf{x}, \mathbf{t})^{-a}$ with parameter $a > 0$. With the polynomial similarity measure, the user's response depends on the relative size of the image distances to the ideal target image. We use Euclidean norm as the distance measure between image $\mathbf{x}$ and the target image $\mathbf{t}$. In all the experiments, the values of a and λ were kept constant at 4 and 0.1, respectively (the optimal values based on [13]).

5.2 Experiments

All the reported results are averaged over 100 searches for randomly selected target images from the dataset. We also tested the influence of k, i.e. the number of images displayed at each iteration, on the performance of the algorithms. We measured the average and standard deviation of the number of iterations required to find the target. The results are summarized in Table 1. The first entry in each cell is the average and the second is the standard deviation.

Algorithm	$k = 5$	$k = 10$	$k = 20$
LR	37.25; 32.52	20.43; 16.07	9.56; 6.22
GP	33.15; 31.14	20.9; 15.98	13.23; 8.19
GP SOM	34.5; 31.09	20.48; 15.68	14.0; 8.54

Table 1: Comparison of the performance of LinRel, GB-UCB and GB-SOM.

There is no significant difference between simple GP bandits and GP-SOM, however GP-SOM is more efficient in terms of computational complexity. For small values of k, both GP algorithms outperform LinRel but for large values of k, LinRel performs slightly better. However, we must bear in mind that LinRel is much slower than both GP algorithms and it does not scale up to large datasets of images.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

6 Conclusion and discussions

We proposed an algorithm for content based image retrieval that combines hierarchical GP Bandits with SOM and reduces the running time of each iteration of the algorithm compared to simple GP bandits while preserving the accuracy of the original algorithms. Instead of searching the entire database, we utilize SOM to precompute efficient discretization of the image space and then perform a sequential Gaussian Process Bandits search. After testing the performance of the algorithm in simulations, we are ready to test the system on real users with large database of images. The next step in the development of our system is building a user interface and run extensive user studies to tune the parameters in the GP-SOM algorithm and demonstrate its benefits.

References

[1] P. Auer, Z. Hussain, S. Kaski, A. Klami, J. Kujala, J. Laaksonen, A.P. Leung, K. Pasupa, and J. Shawe-Taylor. Pinview: Implicit feedback in content-based image retrieval. *JMLR: Workshop and Conference Proceedings: Workshop on Applications of Pattern Analysis*, 11:51 – 57, 2010.

[2] Q. Iqbal and J. K. Aggarwal. Cires: A system for content-based retrieval in digital image libraries. In *Invited session on Content Based Image Retrieval: Techniques and Applications International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 205–210, Singapore, 2002.

[3] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79. ACM, 2004.

[4] David Gorisse, Matthieu Cord, and F Precioso. Salsas: Sub-linear active learning strategy with approximate k-nn search. *Pattern Recognition*, 44(10):2343–2357, 2011.

[5] I. Mironica and C. Vertan. An adaptive hierarchical clustering approach for relevance feedback in content-based image retrieval systems. In *Signals, Circuits and Systems (ISSCS), 2011 10th International Symposium on*, pages 1 –4, 30 2011-july 1 2011.

[6] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397 – 422, 2002.

[7] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process bandits without regret: An experimental design approach. *CoRR*, abs/0912.3995, 2009.

[8] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[9] C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[10] S. Pandey, D. Agarwal, D. Chakrabarti, and V. Josifovski. Bandits for taxonomies: A model-based approach. In *SIAM Intl. Conf. on Data Mining (SDM)*, 2007.

[11] T. Kohonen. *Self-organizing maps*, volume 30. Springer Verlag, 2001.

[12] B. Thomee Mark J. Huiskes and Michael S. Lew. New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In *MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval*, pages 527–536, 2010.

[13] P. Auer, D. Głowacka, A. Leung, S. Hong Lim, A. Medlar, and J. Shawe-Taylor. Study of exploration-exploitation trade-offs with delayed feedback, fp7–216529 pinview. Technical report, European Community's Seventh Framework Programme, 2011.