A MAX MARGIN FRAMEWORK ON IMAGE ANNOTATION AND MULTIMODAL IMAGE RETRIEVAL[*]

Zhen Guo, Zhongfei (Mark) Zhang

Computer Science Department
SUNY Binghamton
{zguo,zhongfei}@cs.binghamton.edu

Eric P. Xing, Christos Faloutsos

School of Computer Science
Carnegie Mellon University
{epxing,christos}@cs.cmu.edu

ABSTRACT

This paper presents a max margin framework on image annotation and multimodal image retrieval as a structured prediction model. Following the max margin approach the image retrieval problem is formulated as a quadratic programming problem. By properly selecting joint feature representation between different modalities, our framework captures the dependency information between different modalities and avoids retraining the model from scratch when database undergoes dynamic updates. While this framework is a general approach which can be applied to multimodal information retrieval in any domains, we apply this approach to the Berkeley Drosophila embryo image database for the evaluation purpose. Experimental results show significant performance improvements over a state-of-the-art method.

1. INTRODUCTION

Image retrieval plays an important role in information retrieval due to the overwhelming multimedia data brought by modern technologies, especially the Internet. One of notorious bottleneck in the image retrieval is the semantic gap [1]. Recently, it is reported that this bottleneck may be reduced by the multimodal approach [2, 3] which takes advantage of the fact that in many applications image data typically coexist with other modalities of information such as text. The synergy between different modalities may be exploited to capture the high level concepts.

In this paper, we follow this line of research by proposing the max margin framework on image annotation and image retrieval as a structured prediction model where the input $\mathbf{x}$ and the desired output $\mathbf{y}$ are structures. Our framework is built upon the model proposed by Taskar et al. [4]. Following the max margin approach the image retrieval problem is formulated as a quadratic programming (QP) problem. Given the multimodal information in the image database, the dependency information between different modalities is learned by solving for this QP problem. In this paper we only consider text modality which co-exists with images although our approach can be easily extended for more modalities. Across-modality retrieval (image annotation and word querying) and image retrieval can be done based on dependency information. By properly selecting the joint feature representation between different modalities, our approach captures the dependency information between different modalities which is independent of specific words or specific images. This makes our approach scalable in the sense that it avoids retraining the model from scratch when image database undergoes dynamic updates which include image and word space updates.

While this framework is a general approach which can be applied to multimodal information retrieval in any domains, we apply this approach to the Berkeley Drosophila embryo image database for the evaluation purpose. Experimental results show significant performance improvements over a state-of-the-art method.

2. RELATED WORK

Multimodal approach has recently received substantial attention since Barnard and Duygulu et al. started their pioneering work on image annotation [2, 5]. Recently there have been many studies [6, 7, 3, 8, 9, 10] on multimodal approaches.

The structure model covers many natural learning tasks. There have been many studies on the structure model which include conditional random fields [11], maximum entropy model [12], graph model [13], semi-supervised learning [14] and max margin approach [15, 16, 17, 18]. The max margin principle has received substantial attention since it was used in the support vector machine (SVM) [19]. In addition, the perceptron algorithm is also used to explore the max margin classification [20].

Our main contribution is to develop an effective solution to the image annotation and multimodal image retrieval problem using the max margin approach under a structure model. More importantly, our framework has a great advantage in scalability over many existing image retrieval systems.

3. MAX MARGIN APPROACH

Assume that the training set consists of a set of training instances $S = \{(I^{(i)}, W^{(i)})\}_{i=1}^{L}$, where each instance consists of an image object $I^{(i)}$ and the corresponding annotation word set $W^{(i)}$. We define a block as a subimage of an image such that the image is partitioned into a set of blocks and all the blocks of this image share the same resolution. For each block, we compute the feature representation in the feature space. Since the image database may be large, we apply k-means algorithm to all the feature vectors in the training set.

[*] THIS WORK IS SUPPORTED IN PART BY NSF (IIS-0535162), AFRL (FA8750-05-2-0284), AFOSR (FA9550-06-1-0327), AND A CAREER AWARD TO EPX BY THE NATIONAL SCIENCE FOUNDATION UNDER GRANT NO. DBI-054659.

We define VRep (visual representative) as a representative of a set of all the blocks for all the images in the database that appear visually similar to each other. A VRep is used to represent each cluster and thus is represented as a feature vector in the feature space. Consequently, the training set becomes VRep-annotation pairs $S = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$, where N is the number of clusters, $\mathbf{x}^{(i)}$ is the VRep object and $\mathbf{y}^{(i)}$ is the word annotation set related to this VRep object. We use $\mathcal{Y}$ to represent the whole set of words and $\mathbf{w}_j$ to denote the j-th word in the whole word set. $\mathbf{y}^{(i)}$ is the M-dimensional binary vector ($M = \|\mathcal{Y}\|$) in which the j-th component $\mathbf{y}_j^{(i)}$ is set to 1 if word $\mathbf{w}_j$ appears in $\mathbf{x}^{(i)}$, and 0 otherwise. We use $\mathbf{y}$ to represent an arbitrary M-dimensional binary vector.

We use score function $\mathbf{s}(\mathbf{x}^{(i)}, \mathbf{w}_j)$ to represent the degree of dependency between the specific VRep $\mathbf{x}^{(i)}$ and the specific word $\mathbf{w}_j$. In order to capture the dependency between VReps and words it is helpful to represent it in a joint feature representation $\mathbf{f} : \mathcal{X} \times \mathcal{Y} \to \Re^d$. The feature vector between $\mathbf{x}^{(i)}$ and $\mathbf{w}_j$ can be expressed as $\mathbf{f}(\mathbf{x}^{(i)}, \mathbf{w}_j)$ and the feature vector between $\mathbf{x}^{(i)}$ and word set $\mathbf{y}$ is the sum for all the words: $\mathbf{f}_i(\mathbf{y}) = \mathbf{f}(\mathbf{x}^{(i)}, \mathbf{y}) = \sum_{j=1}^{M} \mathbf{y}_j \mathbf{f}(\mathbf{x}^{(i)}, \mathbf{w}_j)$. In this feature vector, each component may have a different weight in determining the score function. Thus, the score function can be expressed as a weighted combination of a set of features $\alpha^\top \mathbf{f}(\mathbf{x}^{(i)}, \mathbf{w}_j)$, where α is the set of parameters.

The learning task then is to find the optimal weight vector α such that:

$$arg \max_{\mathbf{y} \in \mathcal{Y}^{(i)}} \alpha^\top \mathbf{f}(\mathbf{x}^{(i)}, \mathbf{y}) \approx \mathbf{y}^{(i)} \quad \forall i$$

where $\mathcal{Y}^{(i)} = \{\mathbf{y} | \sum \mathbf{y}_j = \sum \mathbf{y}_j^{(i)}\}$. We define the loss function $l(\mathbf{y}, \mathbf{y}^{(i)})$ as the number of different words between these two sets. In order to make the true structure $\mathbf{y}^{(i)}$ as the optimal solution, the constraint is reduced to:

$$\alpha^\top \mathbf{f}_i(\mathbf{y}^{(i)}) \geq \alpha^\top \mathbf{f}_i(\mathbf{y}) + l(\mathbf{y}, \mathbf{y}^{(i)}) \ \forall i, \forall \mathbf{y} \in \mathcal{Y}^{(i)}$$

We interpret $\frac{1}{\|\alpha\|} \alpha^\top [\mathbf{f}_i(\mathbf{y}^{(i)}) - \mathbf{f}_i(\mathbf{y})]$ as the margin of $\mathbf{y}^{(i)}$ over another $\mathbf{y} \in \mathcal{Y}^{(i)}$. We then rewrite the above constraint as $\frac{1}{\|\alpha\|} \alpha^\top [\mathbf{f}_i(\mathbf{y}^{(i)}) - \mathbf{f}_i(\mathbf{y})] \geq \frac{1}{\|\alpha\|} l(\mathbf{y}, \mathbf{y}^{(i)})$. Thus, minimizing $\|\alpha\|$ maximizes such margin.

The goal now is to solve the optimization problem:

$$\begin{aligned} \min \quad & \|\alpha\|^2 \\ s.t. \quad & \alpha^\top \mathbf{f}_i(\mathbf{y}^{(i)}) \geq \alpha^\top \mathbf{f}_i(\mathbf{y}) + l(\mathbf{y}, \mathbf{y}^{(i)}) \ \forall i, \forall \mathbf{y} \in \mathcal{Y}^{(i)} \end{aligned}$$

3.1. Min-max formulation

The above optimization problem is equivalent to the following optimization problem:

$$\begin{aligned} \min \quad & \|\alpha\|^2 \\ s.t. \quad & \alpha^\top \mathbf{f}_i(\mathbf{y}^{(i)}) \geq \max_{\mathbf{y} \in \mathcal{Y}^{(i)}} (\alpha^\top \mathbf{f}_i(\mathbf{y}) + l(\mathbf{y}, \mathbf{y}^{(i)})) \quad \forall i \end{aligned} \tag{1}$$

We take the approach proposed by Taskar et al. [4] to solve it. We consider the maximization sub-problem contained in the above optimization problem.

We have

$$\begin{aligned} \alpha^\top \mathbf{f}_i(\mathbf{y}) + l(\mathbf{y}, \mathbf{y}^{(i)}) &= \alpha^\top \sum_j \mathbf{y}_j \mathbf{f}(\mathbf{x}^{(i)}, \mathbf{w}_j) + \sum_j \mathbf{y}_j^{(i)} (1 - \mathbf{y}_j) \\ &= \mathbf{d}_i + (\mathbf{F}_i \alpha + \mathbf{c}_i)^\top \mathbf{y} \end{aligned}$$

where $\mathbf{d}_i = \sum_j \mathbf{y}_j^{(i)}$ and $\mathbf{F}_i$ is a matrix in which the j-th row is $\mathbf{f}(\mathbf{x}^{(i)}, \mathbf{w}_j)$; $\mathbf{c}_i$ is the vector in which the j-th component is $-\mathbf{y}_j^{(i)}$.

This maximization sub-problem then becomes:

$$\begin{aligned} \max \quad & \mathbf{d}_i + (\mathbf{F}_i \alpha + \mathbf{c}_i)^\top \mathbf{y} \\ s.t. \quad & \sum_j \mathbf{y}_j = \sum_j \mathbf{y}_j^{(i)} \end{aligned}$$

We map this problem to the following linear programming(LP) problem:

$$\begin{aligned} \max \quad & \mathbf{d}_i + (\mathbf{F}_i \alpha + \mathbf{c}_i)^\top \mathbf{z}_i \\ s.t. \quad & \mathbf{A}_i \mathbf{z}_i \leq \mathbf{b}_i \quad \mathbf{z}_i \geq 0 \end{aligned}$$

for appropriately defined $\mathbf{A}_i, \mathbf{b}_i$, which depend only on $\mathbf{y}, \mathbf{y}^{(i)}$; $\mathbf{z}_i$ is the relaxation for $\mathbf{y}$. It is guaranteed that this LP program has an integral (0/1) solution.

We consider the dual program of this LP program:

$$\begin{aligned} \min \quad & \mathbf{d}_i + \mathbf{b}_i^\top \lambda_i \\ s.t. \quad & \mathbf{A}_i^\top \lambda_i \geq \mathbf{F}_i \alpha + \mathbf{c}_i \quad \lambda_i \geq 0 \end{aligned} \tag{2}$$

Now we can combine (1) and (2) together:

$$\begin{aligned} \min \quad & \|\alpha\|^2 \\ s.t. \quad & \alpha^\top \mathbf{f}_i(\mathbf{y}^{(i)}) \geq \mathbf{d}_i + \mathbf{b}_i^\top \lambda_i \quad \forall i \\ & \mathbf{A}_i^\top \lambda_i \geq \mathbf{F}_i \alpha + \mathbf{c}_i \quad \forall i \end{aligned} \tag{3}$$

This formulation is justified as follows. If (2) is not at the minimum, the constraint is tighter than necessary, leading to a sub-optimal solution α. Nevertheless, the training data are typically hardly separable. In such cases, we need to introduce slack variables ξ_i to allow some constraints violated. The complete optimization problem now becomes a QP problem:

$$\begin{aligned} \min \quad & \|\alpha\|^2 + C \sum_i \xi_i \\ s.t. \quad & \alpha^\top \mathbf{f}_i(\mathbf{y}^{(i)}) \geq \mathbf{d}_i + \mathbf{b}_i^\top \lambda_i - \xi_i \quad \forall i \\ & \mathbf{A}_i^\top \lambda_i \geq \mathbf{F}_i \alpha + \mathbf{c}_i \quad \forall i \\ & \alpha \geq 0 \quad inf > \lambda_i \geq 0 \quad inf > \xi_i \geq 0 \quad \forall i \end{aligned} \tag{4}$$

After this QP program is solved, we have the optimal parameters α. Then we have the dependency information between words and VReps by the score function. For each VRep, we have a ranking-list of words in terms of the score function. Similarly we have a ranking-list of VReps for each word.

3.2. Feature representation

For a specific VRep $\mathbf{x}^{(i)}$ and a specific word $\mathbf{w}_j$, we consider the following feature representation $\mathbf{f}$ between them: $(\frac{\delta_{ij}}{n_j}, \frac{n_j}{N}, \frac{\delta_{ij}}{m_i}, \frac{m_i}{M})$. Here we assume that there are N VReps and M words. n_j denotes the number of VReps in which $\mathbf{w}_j$ appears. m_i denotes the number of words which appear in VRep $\mathbf{x}^{(i)}$. δ_{ij} is an indicator function (1 if $\mathbf{w}_j$ appears in $\mathbf{x}^{(i)}$, and 0 otherwise). Other possible features may depend on the specific word or VRep because some words may be more important than others. We only use the features independent of specific words and specific VReps and we will discuss the advantage later.

3.3. Image Annotation

Given a test image, we partition it into blocks and compute the feature vectors. Then we compute the similarity between feature vectors and VReps in terms of the distance. We return the top n most-relevant VReps. Since for each VRep, we have the ranking-list of words in terms of the score function, we merge these n ranking-lists and sort them to obtain the ranking-list of the whole word set. Finally, we return the top m words as the annotation result.

3.4. Word Query

For a specific word, we have the ranking-list of VReps. we return the top n VReps. For each VRep, we compute the similarity between this VRep and each test image in terms of the distance. For each VRep, we have the ranking-list of test images. Finally, we merge these n ranking-lists and return the top m images as the query results.

3.5. Image Retrieval

Given a query image, we annotate it using the procedure in Sec. 3.3. For each annotation word j, there is a subset of images S_j in which this annotation word appears. Then we have the union set $S = \bigcup S_j$ for all the annotation words.

On the other hand, for each annotation word j, the procedure in Sec. 3.4 is used to obtain the related image subset T_j. Then we have the union set $T = \bigcup T_j$. The final retrieval result is $R = S \bigcap T$.

3.6. Database Updates

Now we consider the case where new images are added to the database. Assume that these new images have annotation words along with them. If they do not, we can annotate them using the procedure in Sec. 3.3. For each newly added image, we partition it into blocks and for each block we compute the nearest VRep in terms of the distance and the VRep-word pairs are updated in the database. This also applies to the case where the newly added images may include new word.

Under the assumption that the newly added images follow the same feature distribution as those in the database, it is reasonable to assume that the optimal parameter α also captures the dependency information between the VReps and the newly added words because the feature representation described in Sec. 3.2 is independent of specific words and specific VReps. Consequently, we do not need to re-train the model from scratch. In fact, the complexity of the update is $O(1)$. As the database scales up, so does the performance due to the incrementally updated data. This is a great advantage over many existing image retrieval systems which are unable to handle new vocabulary at all. The experimental result supports and verifies this analysis.

4. EXPERIMENTAL RESULT

While this approach is a general approach which can be applied to multimodal information retrieval in any domains, we apply this approach to the Berkeley Drosophila embryo image database for the evaluation purpose. We compare the performance of this framework with the state-of-the-art multimodal image annotation and retrieval method MBRM [3].

There are totally 16 stages in the whole embryo image database. We use stages 11 and 12 for the evaluation purpose. There are about 6000 images and 75 words in stages 11 and 12. We split all the images into two parts (one third and two thirds), with the two thirds used as the training set and the one third used as the test set. In order to show the advantage discussed in Sec. 3.6, we use a smaller training subset (110 images) to obtain the optimal parameter α. For these 110 images, there are 35 annotation words. Then we use the test set for evaluation. This experiment result is shown as "Our Framework (1)" in the figures. Then we add the remaining training images to the database and use the test set for evaluations again. This experiment result is shown as "Our Framework (2)" in the figures. When the new images are added to the image database, the new annotation words along with them are also added to the image database.

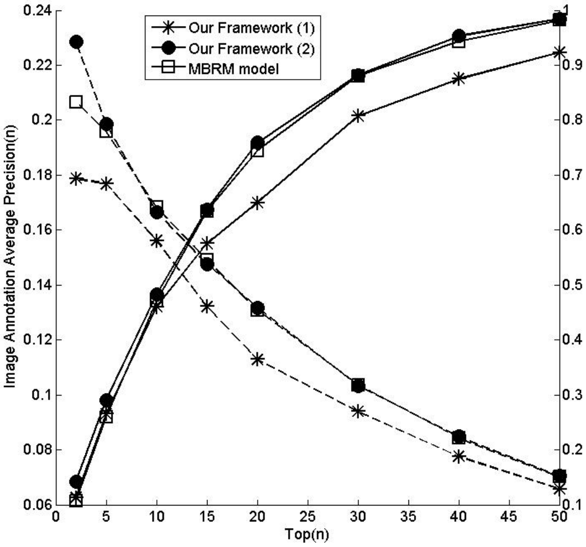

Fig. 1. Evaluation of image annotation between our framework and MBRM model.

In the figures, the dashed lines are for precisions and the solid lines are for recalls. In the image annotation result shown in Fig. 1, the performance becomes better when the new images are added to the image database. This is consistent with the analysis in Sec. 3.6. When the image database scales up to the size as the same as that used by the MBRM model, our framework works slightly better than MBRM. In the word query result shown in Fig. 2, our framework performs significantly better than MBRM. Similarly in the image retrieval performance shown in Fig. 3, our framework works much better than MBRM.

5. CONCLUSION

We present a multimodal framework on image annotation and retrieval based on the max margin approach. The whole problem is mapped to a quadratic programming problem. Our framework is highly scalable in the sense that it takes a constant time to accommodate the database updating without needing to retrain the database from the scratch. The evaluation result shows significant improvements on the performance over a state-of-the-art method.

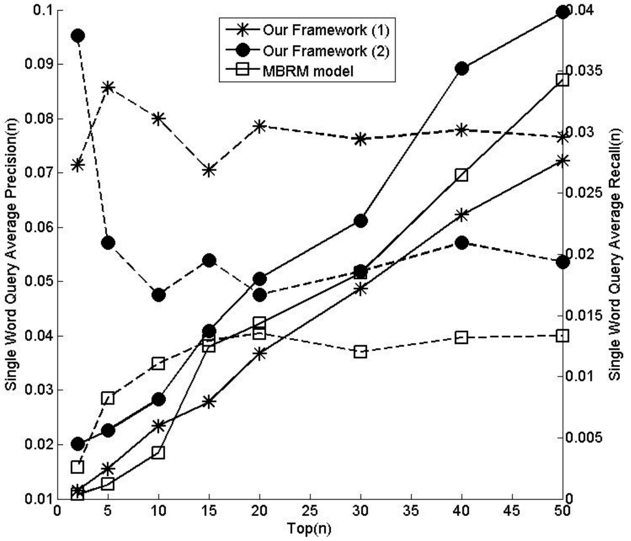

Fig. 2. Evaluation of single word query between our framework and MBRM model.

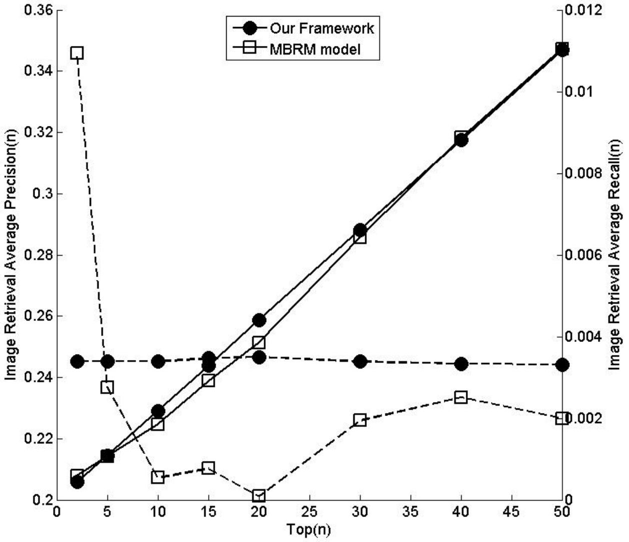

Fig. 3. Evaluation of image retrieval between our framework and MBRM model.

6. REFERENCES

[1] 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 Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349–1380, 2000.

[2] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan, "Matching words and pictures," *Journal of Maching Learning Research*, vol. 3, pp. 1107–1135, 2003.

[3] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *International Conference on Computer Vision and Pattern Recognition*, Washington DC, 2004.

[4] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin, "Learning structured prediction models: A large margin approach," in *Proc. ICML*, Bonn, Germany, 2005.

[5] Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Seventh European Conference on Computer Vision*, 2002, vol. IV, pp. 97–112.

[6] D. Blei and M. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 127–134.

[7] J-Y. Pan, H-J. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery," in *Proceedings of the 10th ACM SIGKDD Conference*, Seattle, WA, 2004.

[8] E. Chang, Kingshy Goh, G. Sychay, and Gang Wu, "Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, pp. 26–38, Jan 2003.

[9] R. Datta, W. Ge, J. Li, and J. Z. Wang, "Toward bridging the annotation-retrieval gap in image search by a generative modeling approach," in *Proc. ACM Multimedia*, Santa Barbara, CA, 2006.

[10] Yi Wu, Edward Y. Chang, and Belle L. Tseng, "Multimodal metadata fusion using causal strength," in *Proc. ACM Multimedia*, Hilton, Singapore, 2005, pp. 872–881.

[11] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001.

[12] Andrew McCallum, Dayne Freitag, and Fernando Pereira, "Maximum entropy markov models for information extraction and segmentation," in *Proc. ICML*, 2000.

[13] W. Chu, Z. Ghahramani, and D. L. Wild, "A graphical model for protein secondary structure prediction," in *Proc. ICML*, Banff, Canada, 2004.

[14] Ulf Brefeld and Tobias Scheffer, "Semi-supervised learning for structured output variables," in *Proc. ICML*, Pittsburgh, PA, 2006.

[15] Hal Daume III and Daniel Marcu, "Learning as search optimization: Approximate large margin methods for structured prediction," in *Proc. ICML*, Bonn, Germany, 2005.

[16] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proc. ICML*, Banff, Canada, 2004.

[17] B. Taskar, C. Guestrin, and D. Koller, "Max-margin markov networks," in *Neural Information Processing Systems Conference*, Vancouver, Canada, 2003.

[18] Yasemin Altun, Ioannis Tsochantaridis, and Thomas Hofmann, "Hidden markov support vector machines," in *Proc. ICML*, Washington DC, 2003.

[19] Vladimir Naumovich Vapnik, *The nature of statistical learning theory*, Springer, 1995.

[20] Yoav Freund and Robert E. Schapire, "Large margin classification using the perceptron algorithm," in *Maching Learning*, 1999, vol. 37.