

A SVM-HMM Based Online Classifier for Handwritten Chemical Symbols

Yang Zhang, Guangshun Shi, Kai Wang

Institute of Machine Intelligence
College of Information Technical Science, Nankai University
Tianjin, China
{yangzhang, gssshi, kaiwang}@imi.nankai.edu.cn

Abstract—This paper presents a novel double-stage classifier for handwritten chemical symbols recognition task. The first stage is rough classification, SVM method is used to distinguish non-ring structure (NRS) and organic ring structure (ORS) symbols, while HMM method is used for fine recognition at second stage. A point-sequence-reordering algorithm is proposed to improve the recognition accuracy of ORS symbols. Our test data set contains 101 chemical symbols, 9090 training samples and 3232 test samples. Finally, we obtained top-1 accuracy of 93.10% and top-3 accuracy of 98.08% based on the test data set.

Keywords—online recognition; handwritten chemical symbols; double-stage classifier; stroke-order independent algorithm;

I. INTRODUCTION

With the popularity of pen-based interaction, and also the complexity of the entry for chemical symbols, applications of online handwritten recognition in chemical field become more and more popular. Ming[2] proposed a framework for recognizing handwritten chemical expressions and obtained 85.9% top-1 accuracy. TY[3] proposed a pen-based interaction system to interpret hand-drawn organic structural diagrams. Xin[4] proposed an approach to understand handwritten chemical formulas. In our preliminary work[1] and [5], we proposed the system architecture for handwritten chemical formulas recognition, and used hmm-based method to recognize inorganic symbols and achieved 89.5% top-1 accuracy.

By now, existing methods usually tended to recognize handwritten chemical symbols through single stage classifier. But those are not reliable for ORS symbols with arbitrary writing and more strokes. To achieve stable, accurate recognition purpose, we design a two-stage classifier based on svm and hmm method. For rough classification, svm classifier based on radial basis function(RBF) kernel is used to classify chemical symbols into ORS symbols and NRS symbols. For fine classification, we choose multiple combinations with variable states and Gaussians to train hmm classifiers. 8-states and 12-Gaussians is the best combination for both NRS and ORS. In particular, to improve the accuracy of ORS in the fine classification, we design a PSR algorithm to preprocess the

sample and the PSR algorithm increases the accuracy of ORS significantly.

This paper is organized as below:

Section 2 gives the methodology which shows a skeleton of our method; Section 3 presents the designed classifier and selected features used in our experiment; Section 4 presents the collection of data set and all the experimental results at first classification stage and second classification stage; Section 5 concludes this paper and puts forward future direction of this work; Section 6 acknowledges.

II. METHODOLOGY

In this paper, chemical symbol set is comprised of ORS symbols and NRS symbols. Because of the distinction between ORS symbols and NRS symbols, we can't use a single classifier to make classification of all the symbols. Hence, we design a double-stage classifier which roughly distinguishes NRS symbols and ORS symbols, then performs fine classification on NRS and ORS symbols respectively.

SVM is an accurate and stable model targeting two classes and is suitable to classify chemical symbols into ORS symbols and NRS symbols. Hence, in rough classification stage, we construct three svm classifiers based on polynomial kernel, RBF kernel and Sigmoid kernel respectively to compare, the classifier based on RBF kernel achieve the best 99.88% accuracy on test set (see Section 4.2).

As a stochastic model, HMM are very suitable for handwritten symbols and characters recognition [9]. Based on the first-stage classification result, we mainly adopt the method in our previous work [5] to build hmm classifiers and expand it to the classification of ORS Symbols. To improve the accuracy of ORS symbols, we design a PSR algorithm to preprocess samples before feature extraction.

Finally, we obtained 91.91% top-1 accuracy of NRS and 97.53% top-1 accuracy of ORS on test set.

The process flow chart is as follows:

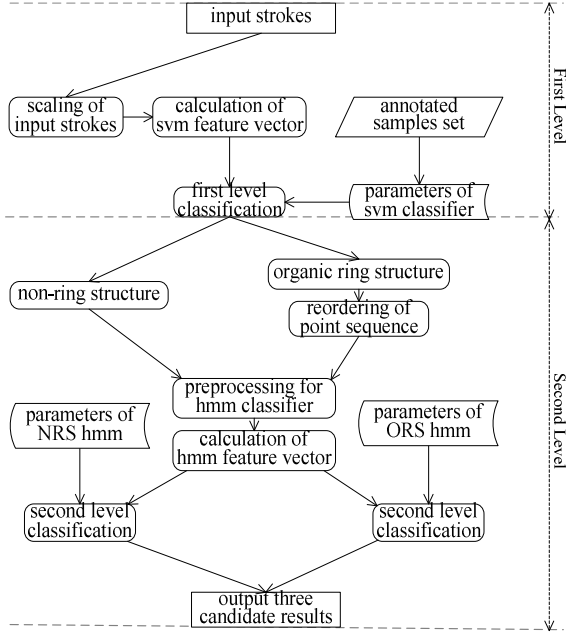


Figure1 Flow Chart of Recognition

III. CLASSIFIER AND FEATURE EXTRACTION

A. Svm Classifier and Its Features

For rough classification, we construct three svm classifiers based on three different kernels. We defined ORS symbols as positive sample and NRS as negative sample, then use the famous algorithm [7] to train the SVM parameters. Experiment shows that the RBF kernel achieved the best result (see Section IV Part B). The decision function we used in svm classifier is:

$$f(\vec{x}) = \text{sgn} \left\{ \sum_{i \in SV} y_i \alpha_i^* \langle \vec{x}_i, \vec{x} \rangle + b^* \right\}$$

58-dimensional feature vector is defined for each scaled sample as bellow:

$$\vec{v} = \{ \bar{m}, \bar{o}, \bar{p}, \bar{a} \}, \text{ where}$$

1) \bar{m} (mesh): divide the bounding rectangle into $M \times N$ grids, then calculate the ratio of points in each grid to the total points of the symbol:

$$m_i = \frac{\text{\#dots in } i\text{th grid}}{\text{\#total dots}} \quad (i = 1, \dots, M \times N)$$

In this paper, we specify 4x4 grid, obtain 16D features.

2) \bar{o} (outline): from each edge of the bounding rectangle, draw K equi-distance scan lines, and start scanning along the scan line until encounter the first black point or the mid-axis perpendicular to the scan line, note the distance of scanning and normalize it between 0-1:

$$o_i = \frac{\text{distance of scan line}}{\text{distance from edge to mid-axis}} \quad (i = 1, \dots, K)$$

In this paper, we specify K as 5, obtain 20D features.

3) \bar{p} (projection): from each edge of the bounding rectangle, segment L equi-distance bins and calculated the ratio of points to total points of the symbol:

$$p_i = \frac{\text{\#dots in } i\text{th bin}}{\text{\#total dots}} \quad (i=1, \dots, L)$$

In this paper, we specify L as 5, obtain 20D features.

4) \bar{a} (aspect):

$$a_1 = \frac{\text{height of boundrect}}{\text{height} + \text{width}} \quad a_2 = \frac{\text{width of boundrect}}{\text{height} + \text{width}}$$

We randomly selected two sets of samples (2 per symbol), extracted partial features and calculated the average for each dimension in NRS and ORS sets respectively. From the scatter plots in Fig 2, we can see that the features have a good distinguishing performance.

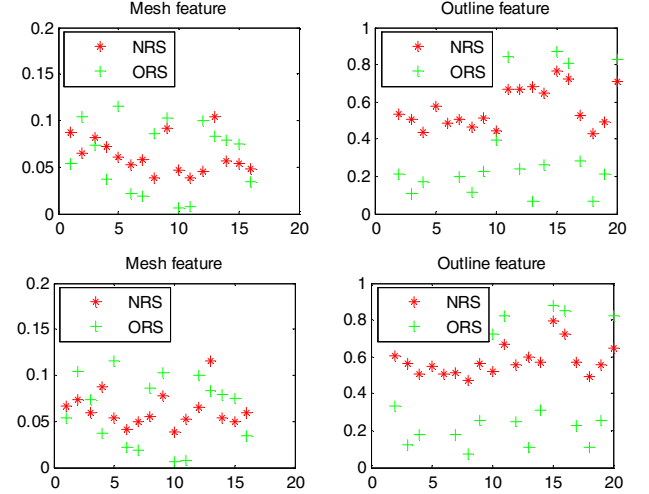


Figure 2 Mesh & Outline Scatter Plot

B. Hmm Classifier and Its Features

As a continuous HMM, the left-right HMM [8] has been successfully used in speech recognition and handwritten recognition. Therefore, we choose left-right hmm to build the second-stage classifier (one per symbol) and use the famous Baum [8] algorithm to train HMM parameters. Then we choose multiple combinations with variable states and Gaussians to compare their performance (see Section IV Part B). In the end, we obtain two groups of optimal parameters in NRS set and ORS set respectively.

We mainly adopt the preprocessing and features used in our previous work [5]. Especially, to improve the accuracy of ORS, we use PSR to preprocess the samples of ORS before feature extraction.

C. PSR Algorithm

ORS has a special two-dimensional hexagonal structure, which results in the uncertainty of samples in number of strokes and stroke order. Compared with NRS, this uncertainty is so serious that it directly affects HMM to model ORS in the fine classification. Hence, we proposed the point sequence reordering algorithm (PSR in brief) to eliminate the uncertainty, making HMM independent of number of strokes and stroke order.

Given the closed ring structure, we adopt a counter-clockwise scanning, we adopt a counter-clockwise scanning to reorder the sequence of points of ORS as A in Figure 3:

In NRS fine classification, Top-1 results indicate that 8 states and 12 Gaussians is the best combination of parameters. Most misclassified samples usually occurred in the confusing symbols, such as uppercase “O”, digit “0” and lowercase “o”; uppercase “C”, lowercase “c” and brace “(”.

In addition, samples with poor writing quality are often misclassified. We can provide more candidate results or build discriminatory classifiers targeting confusing symbols to improve the recognition accuracy.

In ORS fine classification, Top-1 results indicate that 8 state and 12 Gaussians is the best combination of parameters. Most misclassified samples usually have poor writing quality. This misclassification can be reduced by fine pre-processing.

TABLE 4 BEFORE AND AFTER PSR

4-STATE 3-GAUSSIAN	ACCURACY IN TEST (1216)		
	Top-1	Top-2	Top-3
Before PSR	34.95%	49.84%	60.44%
After PSR	95.23%	98.36%	98.85%

The data in Table 4 indicate that PSR has greatly improved the accuracy of ORS. In addition, PSR is not dependent on the ORS and can also be used for other multi-stroke, closed symbols.

Finally, we construct a two-stage classifier based on the best svm classifier and best hmm classifier to recognize all the 101 chemical symbols as below:

TABLE 5 ACCURACY FOR SVM-HMM CLASSIFIER

8-STATE 12-GAUSSIAN	TRAIN (9090)		TEST (3232)
	Top-1	Top-2	Top-3
Train Set	97.66 %	99.65%	99.87%
Test Set	93.10%	97.09%	98.08%

V. CONCLUSION AND OUTLOOK

In the future, we will combine symbol recognition with segmentation and structure analysis of chemical expressions to form a systematic mechanism to process and analyze chemical expression.

We will also focus on reducing the consumption of system resource to enhance the friendly human-computer interaction.

ACKNOWLEDGMENT

This work is funded by Microsoft Research Asia and TJNSFC Grant #09JCZDJC26000.

REFERENCES

- [1] J.F. Yang, G.S. Shi, Q.R. Wang, “A Study of On-line Handwritten Chemical Expressions Recognition”, In Proc. Of 19th Intl. Conf. on Pattern Recognition, 2008.
- [2] M. Chang, S. Han, D.M. Zhang, “A Unified Framework for Recognizing Handwritten Chemical Expressions”, In Proc. Of 10th Intl. Conf. on Document Analysis Recognition, 2009.
- [3] T.Y. Ouyang, R. Davis, “Recognition of Hand Drawn Chemical Diagrams”, Master Thesis in MIT, 2007.
- [4] X. Wang, G.S. Shi, J.F. Yang, “The Understanding and Structure Analyzing for Online Handwritten Chemical Formulas”, In Proc. 10th Intl. Conf. on Document Analysis Recognition, 2009.
- [5] Y. Zhang, G.S. Shi, J.F. Yang, “HMM-based Online Recognition of Handwritten Chemical Symbols”, In Proc. Of 10th Intl. Conf. on Document Analysis Recognition, 2009.
- [6] N. Cristianini, J. Shawe-Taylor, “Support Vector Machines”, In Cambridge University Press, 2000.
- [7] T. Joachims, “Making Large-Scale SVM Learning Practical”, In Advances in Kernel Methods-Support Vector Learning, MIT Press, 1999.
- [8] L. Rabiner, “A Tutorial of Hidden Markov Models and Selected Applications in Speech Recognition”, Proc. IEEE, 77:257-286, 1989.
- [9] J. Tokuno, “Context-dependent Substroke Model for HMM-based Online Handwriting Recognition”, 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR), pp. 78-83. 2002.