TR-97–8

Multiple Attribute Learning with Canonical Correlation Analysis and EM algorithm

Shotaro Akaho* Satoru Hayamizu†
Osamu Hasegawa† Takashi Yoshimura†
Hideki Asoh*

Received 11 April 1997

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba,
Ibaraki, 305, JAPAN

Abstract This paper presents a new framework of learning pattern recognition, called "multiple attribute learning". In usual setting of pattern recognition, target patterns have several attribute such as color, size, shape, and we can classify the patterns in several ways by their color, by their size, or by their shape. in normal pattern recognition problem, an attribute or a mode of classification is chosen in advance and the problem is simplified. To the contrary, the problem considered in this paper is to make the learning system solve multiple classification problems at once. That is, a mixture of learning data set for multiple classification problems are given to the learning system at once and the system learn multiple classification rules from the data. We propose a method to solve this problem using canonical correlation analysis and EM algorithm. The effectiveness of the method is demonstrated by experiments.

*Information Science Division
†Machine Intelligence Division

1 Introduction

Recently, research aiming to software agent based multimodal interface system become popular. With this interface system, user communicate with human-like agents on screen using spoken language, visual cues etc. We also have been building a prototype of the system [2, 10, 4]

Since it is difficult to program the complete knowledge of the complicated real world environment and various users, in order to realize flexible agents which can naturally communicate with users in real world setting, learning ability is necessary. The agent should learn about the environment through interaction with users.

To maintain the easiness to use, the interaction between users and systems should not be strongly controlled. This means the learning processes of the agents should be also informal and should not be controlled strongly.

Let us imagine that you will teach the agent about your favorite mug cup. The mug cup may be white large one. You show the cup to the system and may tell to the agent that "This is white." Or you may tell to the agent that "This is large." The visual image of the mug cup has at least three attributes color, size, and shape. This means there are at least three ways of classification of the visual image, classification by color, by size, and by shape.

In normal pattern recognition learning, a way of classification is chosen in advance. However, in the real situation of concept learning of interface agent, learning data for multiple classification problems may be given to the system at once. The system should treat such uncontrolled data set and learn multiple classification rules from the data. Note that at the same time the system discovers the attributes such as color, size, shape. Once the system discovered the attribute, adding new category to the attribute, for example, adding "pink" to color attribute, may become easier task.

In this paper, we consider this problem of multiple attribute learning. User shows an image of an object and teaches one of the attributes of the object by speech. The system does not know which attribute the speech represents. After learning proceeded, when a human shows a new image of an object, the system is expected to play back a set of spoken words each of which corresponds to an attribute of the object (see table 1).

Our approach is basically a memory-based approach which does not include any explicit pattern recognition or any symbolic representation. In this task, even if the pattern recognition succeeded completely, the categorization of attributes is still a complicated problem. For example, suppose there are two attributes for each object and the pattern recognition can be done completely. If training samples $(A.a)$, $(A.b)$, and $(B.c)$ are given, where a pair $(X.y)$ denotes a pair of image X and speech y, we find that a and b represent different attributes (attribute 1 and attribute 2 respectively). However, we cannot determine which attribute c belongs to. If we get an additional sample $(B.a)$, we find that c belongs to the attribute 2, but the situation does not change if we get an additional sample $(B.d)$, and we need other samples except for A and B. In this manner, this task

Table 1: An example of the task

Example of training samples:

Image X	Speech Y
[White cup]	"white"
[Blue pen]	"pen"
...	...
[White cup]	"cup"
[Red book]	"red"

After learning:

Image X		Speech Y
[White pen]	$\rightarrow$	"white" + "pen"
[Blue book]	$\rightarrow$	"blue" + "book"
...		...
[Red cup]	$\rightarrow$	"red" + "cup"

includes combinatorial complexity, and moreover the pattern recognition is not always perfect (teacher may include some errors/noise).

For such a noisy and complicated task, statistical approach is often suitable because it gives a solution which is robust against noise and reduces time complexity when a rough solution is allowed.

Since it is not known which attribute of an object the speech represents, we apply the EM (expectation-maximization) algorithm, which is a statistical estimation method for the problem including hidden variables. As a preprocess, we apply canonical correlational analysis (CCA) in order to reduce the dimensions of image and speech.

This paper is organized as follows. First, we explain our framework including CCA and the EM algorithm. We then show our experimental results using real images and speeches. Finally, we discuss the extension of our framework and consider advantages and disadvantages of our approach.

2 Attribute learning from two information sources

In this section, we formalize the task described in the previous section.

Suppose $\boldsymbol{x}$ and $\boldsymbol{y}$ are data provided from two information sources X and Y respectively. In this paper, $\boldsymbol{x}$ represents an image and $\boldsymbol{y}$ represents a spoken description. $\boldsymbol{x}$ has K kinds of attributes and $\boldsymbol{y}$ represents one of those attributes. The number of attributes K is known but it is not known which attribute $\boldsymbol{y}$ represents. In an experiment discussed later, we consider color, shape, and size as attributes ($K = 3$).

In the learning phase, a lot of pairs of $\boldsymbol{x}$ and $\boldsymbol{y}$ are given to the system as training samples. After learning, all attributes $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K$ corresponding to newly given $\boldsymbol{x}$ are

expected to be played back.

3 Reduction of dimensions

In real world application, the dimensional characteristics of $\boldsymbol{x}$ and $\boldsymbol{y}$ are too large for the system to learn their relation. Furthermore, the dimensionality of speech data changes as the length of speech changes. This problem is called the curse of dimensionality.

To deal with it, the first thing to do is to extract features from raw data. However, by design, feature extraction is general-purpose, and extracted features are still redundant in order to find the relation of two information sources.

For this purpose, we extract only information which is necessary to find the relation, using canonical correlation analysis (CCA). CCA is a statistical technique where $\boldsymbol{x}$ and $\boldsymbol{y}$ are mapped into the space in which the correlation is maximized. In this paper, a linear map is used as the map because of its simplicity, but it can be extended to nonlinear map easily[1]. In the case of the linear map, the map is given as a solution of an eigen value problem.

4 Finite Mixture and the EM algorithm

In this section, we explain the method of finding one-to-many relationships between two information sources whose dimensions are reduced by CCA. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote data with reduced dimensions below for the purpose of simplification.

4.1 Finite Mixture

Let us consider the k-th attribute $(1 \leq k \leq K)$ $\boldsymbol{y}_k$ of $\boldsymbol{x}$, and suppose the relation between $\boldsymbol{y}_k$ and $\boldsymbol{x}$ is modeled by a certain parametric conditional probability distribution $f_k(\boldsymbol{y}_k \mid \boldsymbol{x}; \theta_k)$. Since k is not known by the system in the current framework, k is a hidden variable. Thus if k is randomly selected in probability ξ_k, the distribution of the observed variable is given by

$$f(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \xi_k f_k(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}_k), \tag{1}$$

which is called the finite mixture model. Recently this model has been attracting much attention because it can be applied to a lot of problems such as multi-agent learning and modular learning[5].

In this paper, a linear map with gaussian noise is used as f_k because of its simplicity.

4.2 EM Algorithm

The EM algorithm is a recurrent algorithm to solve the maximum likelihood estimation problem when hidden variables are included[3]. When f_k is a linear map, each step of the

EM algorithm is given as follows: Let the conditional probability distribution of hidden variable k be

$$q(k \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}^{(t)}) = \frac{f_k(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}_k^{(t)})}{f(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}^{(t)})}, \tag{2}$$

where $\boldsymbol{\theta}^{(t)}$ is a parameter obtained in the t-th step. Then parameters of the $t+1$-th step is given as follows.

The weight probabilities are given by

$$\xi_k^{(t+1)} = \frac{1}{N} \sum_{(\boldsymbol{x},\boldsymbol{y})} q(k \mid \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}^{(t)}). \tag{3}$$

The parameters of linear mapping are given by the regression parameters where each pair $(\boldsymbol{x}, \boldsymbol{y})$ is weighted by $q(k \mid \boldsymbol{y}, \boldsymbol{x})$.

The time complexity of each step of the EM algorithm is given by $O(K \times N)$ where N is the number of training samples.

4.3 Parameter Initialization

Since the EM algorithm converges to a local optimum and the finite mixture has a lot of local optima, it is important to find a good initial solution.

In our problem, the solution satisfies the following condition: if there are two such pairs of $(\boldsymbol{x}, \boldsymbol{y})$ where the $\boldsymbol{x}$'s are the same or similar and the $\boldsymbol{y}$'s are different, the $\boldsymbol{y}$'s represent different attributes.

The following value roughly reflects the above property,

$$\sum_{i<j} \frac{\boldsymbol{c}'\boldsymbol{y}_i - \boldsymbol{c}'\boldsymbol{y}_j}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\| + \epsilon}, \quad \|\boldsymbol{c}\| = 1, \tag{4}$$

hence we find the projection axis $\boldsymbol{c}$ which maximizes the above criterion and divides a sample set into K subgroups on the projected axis. The projection axis is given as a solution of an eigen value problem.

5 Experiment

We applied the learning framework above to real images and spoken descriptions.

5.1 Collected Data

We recorded images of colored plastic number blocks with a camera, and recorded attributes with a microphone. The images and spoken descriptions are collected separately and the training pairs are generated afterward.

There are 10 different shapes of blocks each assigned a number from 0 to 9, and for each shape there are 4 colors (red, blue, yellow, and green). Sizes (small, medium

and large) are changed by adjusting the camera. Each image is captured directly into a computer from the camera and stored as a full color YIQ image with 320×240 pixels (but the Y element is discarded). For each category, we took 29 images by adding a small perturbation to the position and the direction.

From raw images, we generated 3 resolution images (1/1, 1/4, 1/16) by thinning out, and extracted 25 dimensional local higher-order autocorrelation features[7] for each generated image. A total of 150 dimensional image features (2 colors, 3 resolutions, and 25 correlation features) were extracted.

Dialogue describing the corresponding attributes were recorded by 2 male speakers. 17 Japanese were recorded corresponding to 4 colors, 10 shapes, and 3 sizes. First we extracted the melcepstrum features of each speech (12 degrees for each time step), and then in order to obtain the fixed dimensional features, we normalized them into 10 frames by thinning out. For example, if there are 30 frames, we selected the 1st, 4th, 7th, ..., 28th frames. A total of 120 dimensional speech features were extracted.

5.2 Experiment process

First, we prepared training samples. From the extracted features, we randomly generated 200 pairs of training samples for each attribute.

Second, we applied CCA to the generated training samples. The purpose of CCA is to discard redundant dimensions and leave essential information, thus it is important to choose an appropriate number of dimensions. If we choose too small dimension, it is impossible to categorize, while if we choose too large dimension, it is difficult to learn the relation between image and speech. We selected 10 dimensional spaces where the correlation coefficients of images and speeches are dominant (more than about 0.5).

Thirdly, the EM algorithm was applied to find a one-to-many map in the space generated by CCA.

Finally, we evaluated the results by showing test images which were not included in training samples. Each attribute is determined by the nearest neighbor (1-NN) method based on the categorized samples. The correct answer rate is calculated by counting the correctly answered attributes.

5.3 Result

Fig. 1 and 2 shows the performance of categorization. Each bar indicates the true attributes and each color indicates the frequency of categorized attributes. If the categorization is perfect, each bar would be colored by a separate color. Fig. 1 is the initial state of the EM algorithm and fig. 2 is the state after 30 steps of the EM algorithm.

We tried such experiments for 50 different seeds of random numbers, the correct rate was between 65 % and 90 %, and 75.8 % in average after 30 steps of the EM algorithm. This would be a satisfactory level as preliminary experiments, since we utilize only linear

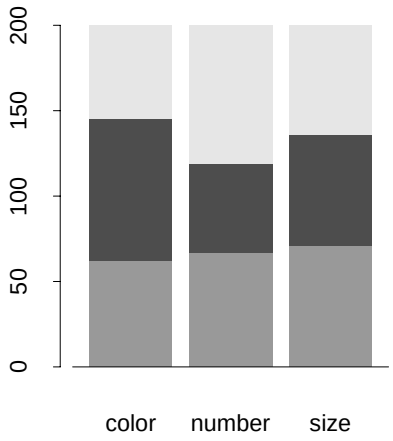

Figure 1: Distribution of true category (initial state of the EM algorithm)

models.

6 Discussion

In this section, we would like to mention about problems and improvement of our framework described in the preceding sections.

The number of attributes Although the system needs to know the number of attributes to apply the EM algorithm, it is not practical in some situations. One simple idea is to try several numbers of attributes and to choose the best number among them. A lot of criteria to choose the best model are proposed in statistics and information science society, for instance, MDL (Minimum Description Length principle) and AIC (Akaike's Information Criterion). However, we must take care of using those criteria, since they are not so robust under real world settings.

More than two information sources In our framework, the number of information sources is assumed to be two, and one source (image) has several attributes while another (speech) has just one attribute.

In general, the number of information sources would be more than two and each source would have multiple attributes. We have to extend our framework in order to do that. CCA for more than two information sources is studied in statistics society and is known as generalized CCA, but it is not still an established method[6]. If each information source has multiple attributes, we have to consider many-to-many relation between the sources. Such an extention is possible, but it enlarges the number of hidden variables in the EM algorithm, which will make the learning difficult.

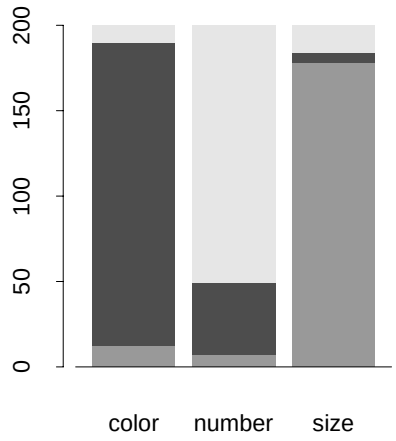

Figure 2: Distribution of true category (after 30 steps of the EM algorithm)

Hierarchical structure of concepts Although we treated concepts(attributes) without any structures, some concepts are related to each other and they form a hierarchy in general[8, 9]. For example, "birds" and "fish" are concepts by themselves, but they are also members of a concept "animals".

In order to deal with hierarchical structure of attributes, we need to use a hierarchical model in the EM algorithm, which is easy to implement. However, it also enlarges the number of hidden variables, and the learning might be difficult.

Symbols are not necessary at all? The answer is "Yes, we need symbols". As we wrote at the beginning, our framework does not include any symbolic representation. However, that does not mean our approach is contradictory to symbols. Our framework should be considered as a level under acquiring symbols.

Symbols are very useful tool for reasoning and thinking, but it is not impossible and not necessary to define all symbols in advance. Our framework will help to acquire the concept of symbols only from superficial data like images and speeches. It is easy to associate the acquired concepts with the corresponding symbols.

7 Conclusion

We have proposed a framework of learning in the multimodal interaction system through the task of concept acquisition. It aims at reducing the participation of humans to a symbolic level as much as possible and the system is expected to learn the environment merely by using the superficial interaction with humans.

References

[1] Asoh, H. and Takechi, O.: An approximation of nonlinear canonical correlation analysis by multilayer perceptrons. In *Proc. of ICANN'94*, pp. 713–716, 1994.

[2] Bolt, R.: Put that there : Voice and gesture at the graphics interface. *Computer Graphics*, Vol. 4, No. 3, pp. 262–270, 1980.

[3] Dempster, A., Laird, N., and Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* B, Vol. 39, pp. 1–38, 1977.

[4] Hasegawa, O. *et al*: Active agent oriented multimodal interface system. In *IJCAI-95*, pp. 82–87, 1995.

[5] Jordan, M. I. and Jacobs, R. A.: Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, Vol. 6, pp. 181–214, 1994.

[6] Kettering, J.: Canonical analysis of several sets of variables. *Biometrica*, 1971.

[7] Kurita, T., Otsu, N., and Sato, T.: A face recognition method using higher order local autocorrelation and multivariate analysis. In *Proc. of 11th International Conf. on Pattern Recognition*, volume II, pp. 213–216, The Hague, 1992.

[8] Mervis, C. and Rosch, E.: Categorization of natural objects. In Rosenz, M. and Potter, L. (eds.), *Annual review of psychology*, volume 21. Annual Reviews, 1981.

[9] Smith, E. and Medin, D.: *Categories and Concepts*. Harvard University Press, 1981.

[10] Vo, M. and Wabel, A.: Multimodal human-computer interaction. In *Proc. Int. Symp. on Spoken Dialogue*, pp. 95–101, 1993.