53

ONE-CLASS LEARNING FOR HUMAN-ROBOT INTERACTION

QingHua Wang
IEETA, University of Aveiro,
Campus Universitário Santiágo, 3810-193, Aveiro, PORTUGAL
qhwang@ieeta.pt

Luis Seabra Lopes
IEETA/Department of Electronics & Telecommunication,
University of Aveiro,
Campus Universitário Santiágo, 3810-193, Aveiro, PORTUGAL
lsl@det.ua.pt

A Suitable learning and classification mechanism is a crucial premise for Human-Robot Interaction. To this purpose, several one-class classification methods have been investigated using wavelet features (parameters of Hidden Markov Tree model) in this paper. Only target class patterns are used to train class models. Good discrimination over outlier (never seen non-target) patterns is still kept based on their distances to class model. Face and non-face classification is used as an example and some promising results are reported.

1. INTRODUCTION

Robots are expected to exist extensively in many areas of our future daily life. So, a basic requirement arises: they must understand what people mean, e.g., instructions from us. To this end, our work focuses on basic natural language concept learning mainly through visual information in the context of Human-Robot interaction (HRI). Let's consider the task of teaching a robot to recognize an object, say, "apple", through its camera in the context of HRI. How can the teaching be conducted? To investigate some state of the art statistical approaches in literature, e.g., Hidden Markov models (Meng, 2000; Zhu & Schwartz 2002), Bayesian networks (Pham *et al*, 2002), naïve Bayes classifier (Schneiderman & Kanade, 2000), PCA based classifier (Turk & Pentland, 1994), and other state of the art methods described in (Yang *et al*, 2002), basically quite a lot of apples and enough non-apples, which is itself an ambiguous concept, must be collected to estimate the class distributions precisely. One might wonder whether these requirements are realistic in the context of HRI.

Different from other learning tasks, the learning of object recognition suitable for HRI has several constraints such as supervised, online, real-time and interactive. To learn a new object (concept), there may be only quite a limited number of samples available. Moreover, the conventional learning mechanisms mentioned above require the preparation of both target and non-target training data in order to learn the correct decision boundary between these two classes. However, in a HRI scenario, typically only positive samples are available. Thus, a so-called one-class classifier might be useful in this situation. These classifiers can learn target class models based on only target patterns, keeping good discrimination for never seen non-target patterns.

Following this idea, an approach based on the combination of the wavelet domain Hidden Markov Tree (HMT) and Kullback-Leibler Distance (KLD) was presented in (Wang & Seabra Lopes, 2004). In this approach, only target samples are used to train an object model in terms of parameters of HMTs. After that, for each sample to be recognized, its KLD to this model is computed. If its KLD is smaller than a certain threshold, obtained during the training session, it is recognized as an instance of the target class; otherwise, it is rejected. This approach has some similarity to the face detection method presented in (*Yang et al*, 2001). There, multimodal density models were used to capture the class (face) distribution only from target patterns. Based on these models, a probability for certain new patterns to be classified as target or non-target class can be computed. The problem of this HMT/KLD based approach is that it can't derive robust object models if there are big in-class variations among the training patterns.

In this paper, some methods described in (Tax, 2001) are investigated to solve this problem. The rest of the paper is organized as follows. The background and motivation to find suitable learning and classification mechanisms for natural language concept grounding for HRI is presented in section 2. In section 3, a brief description of HMTs and one-class learning is provided. The experimental setup and results are given in section 4. Conclusion is provided in section 5 with some discussion and future work.

2. BACKGROUND

Carl, a prototype of an intelligent service robot, designed having in mind such tasks as serving food in a reception or acting as a host in an organization, was developed by our group (Seabra Lopes, 2002) [i]. One of the long-term research goals of this project is to assign robots a capacity for self-development, based on rich sensory and motor capabilities. The idea is to start with minimal initial knowledge, rather than to assign robots rich knowledge in advance (Seabra Lopes & Wang, 2002). After this, robots can learn new skills, explore their environments themselves or under the human guidance.

As a basic requirement of HRI, Carl is able to enter a spoken language conversation with a person. Speech recognition is currently based on IBM NUANCE (Seabra Lopes, 2002). The recognition grammar being used can accept over 12000 different sentences. Carl is able to learn facts about the world through dialog with humans. For instance, Carl can learn that "*Peter is in France*". Later, hearing the question "*Where is Peter?*" Carl can provide an appropriate reply "France". However, Carl

has no idea of what/who France or Peter are. The words (symbols) used in communication to refer to objects still have to be grounded in the robot's own sensory data.

Generally, every word in a conversation should be grounded but, currently, our focus is on grounding of symbols (nouns) that might be a cornerstone to this end, particularly symbols that refer to physical objects in the environment. Therefore, Carl's learning capabilities are being extended to visually recognize objects in a normal office environment. Thus, one can think that we ground the corresponding concepts using visual information or features extracted from visual information. In a learning phase, human tutors teach Carl these concepts. The first stage of grounding turns out to be supervised object learning.

3. LEARNING APPROACH

3.1 One-Class Learning

The design of one-class classifiers is motivated by the fact that patterns from a same class usually cluster regularly together, while patterns from other classes scatter in feature space. One-class learning and classification was first presented in (Moya et al, 1993), but similar ideas also already appeared, including outlier detection (Ritter & Gallegos, 1997), novelty detection (Bishop, 1994), concept learning in the absence of counter-examples (Japkowicz, 1999) and positive-only learning (Muggleton & Firth, 2001). In two-class approaches, information of both target class and non-target class is available. Generally, in multi-class approaches, samples are provided for all classes. Based upon this information, one can precisely capture class descriptions, and therefore find effective decision boundaries between the classes. In contrast, in one-class approaches, only samples of the target class are required. A very natural method for decision-making under this condition is to use some distance-based criterion. If the measurement of an unknown pattern x is smaller than the learned threshold, it can be accepted as the target class pattern; otherwise, it should be rejected. This can be formulated as follows.

$$Class(x) = \begin{cases} target, & if \quad Measurement(x) \le threshold; \\ non\text{–}target, & otherwise. \end{cases} \tag{3}$$

In some sense, this is similar to the famous Bayesian decision rule. The only difference is that, here, the threshold is learned only from target class patterns, while in Bayesian decision rule it's determined based both on target and non-target class patterns. If an appropriate model of the target class (and thus a proper threshold) is found, one can find that most patterns from this target class are accepted and most non-target class patterns are rejected. Of course, the ideal model is one that can accept all target patterns and reject all non-target patterns. But this is usually not easy to find under realistic conditions.

Several methods were proposed to construct models for one-class classification. A very natural method is to generate artificial outlier data (Roberts & Penny, 1996), and thus conventional two-class approaches can be applied. This method severely depends on the quality of artificial data and often works not well.

Some statistical methods were also proposed. One can estimate the density or distribution of the target class, e.g., using Parzen density estimator (Bishop, 1994), Gausian (Parra et al, 1996), multimodal density models (Yang et al, 2001) or wavelet-domain HMTs (Wang & Seabra Lopes, 2004). The requirement of well-sampled training data to precisely capture the density distribution makes this type of methods problematic. In (Moya et al, 1993; Tax, 2001) some boundary-based methods were proposed to avoid density estimation of small or not well-sampled training data. But a well-chosen distance or threshold is needed. Tax provides a systematic description of one-class classification in (Tax, 2001), where the decision criteria are mostly based on the Euclidean distance. The work reported in this paper is based mainly on these methods.

The best results were obtained with SV-DD (Support Vector Data Description (Tax, 2001)), a one-class classification method inspired by Vapnik's Support Vector Machines. SV-DD tries to find a sphere boundary with minimal volume, not a hyperplane as in SVM, containing all or most of objects in a data set. The sphere decision boundary is defined by the so-called support objects or support vectors (Tax, 2001). For classification, objects outside this sphere decision boundary are regarded as outliers (objects from other classes). To obtain a good and compact data description, some remote data points may be discarded although they are not real outliers.

For comparison we also investigate some other one-class classifiers, namely *PCA-DD*, *NN-DD*, *KMEANS-DD*, *KNN-DD* and *GAUSS-DD*. The *PCA-DD* uses subspace composed by Principal Components as data description. The *NN-DD* and *KNN-DD* use simple nearest neighbor and *k*-nearest neighbor data description. The *KMEANS-DD* uses *k*-clusters as data description. In each of these *k* clusters the average distance to its cluster center is minimized. The *GAUSS-DD* assumes data follows simple Gaussian distribution. For more details please refer to (Tax, 2001).

3.2 Hidden Markov Trees

Much work has shown that class distributions can be modeled by modeling the distributions of their wavelet coefficients. Hidden Markov Trees (HMTs) were proposed in (Crouse *et al*, 1998) to precisely characterize these wavelet coefficient distributions, especially the key inter-scale dependencies among parent and children coefficients. In a HMT model, after applying a wavelet transform on an image, any coefficient of it arises from a 2-state zero-mean Gaussian mixture model (Crouse *et al*, 1998). It means the magnitude of a wavelet coefficient is either "large" or "small". Given its state, it is conditionally independent from all other random variables (states of other coefficients).

More specifically, an HMT model for one of three subbands LH, HL, and HH can be fully characterized by the following parameters:

• The pmf (probability mass function) for the root S_1, $\pi = \{\pi_m\}$, $m \in \{1,2\}$. π_m is defined as

$$\pi_m = P(S_1 = m) \tag{1}$$

• The state transition probability matrix A = { $a_{i,\rho(i)}^{rm}$ }, i∈ {2, ... , n}, r, m ∈ {1, 2}. Each $a_{i,\rho(i)}^{rm}$ is the probability when the node *i* is in state m and it parent *ρ(i)* in state *r*. Thus it's defined as

$$a_{i,\rho(i)}^{rm} = P(S_i = m \mid S_{\rho(i)} = r) \tag{2}$$

• The parameters of the zero-mean Gaussian mixture, $\sigma_i^2(m)$, for the state m of each w_i, ∀ i ∈ {1, 2, ..., n}. Here, *n* is the number of coefficients.

To make the HMT model practical, it's assumed that coefficients in each wavelet subband have the same variances (Crouse *et al*, 1998). This assumption, called "tying", reduces HMT parameters to 6 for each subband of each wavelet transform level: 2 for variances and 4 for state transitions. These parameters can be learned using EM algorithm in the sense of maximum likelihood and grouped together as a parametric model, denoted as $\Theta = \{\pi, A_j, \sigma_j^{(m)}\}$, j=1,..., *L* and m=*1, 2*. Here L is the wavelet transform level. For a whole image, three independent HMTs, corresponding to subbands LH, HL and HH respectively, can be used to fully characterize it. For further information on HMTs please see (Choi & Baraniuk, 2001; Crouse *et al*, 1998; Do, 2002; Durand & Gonçalvès, 2002; Fan, 2001; Romberg *et al*, 1999).

4. EXPERIMENTS

4.1 Evaluation approach

The purpose of this paper is to investigate some one-class classifiers available from (Tax, 2001) on HMT features (parameters). The whole evaluation procedure is depicted below in Figure 1 where the specific classifier varies. In a training session, the EM algorithm is applied on target patterns to estimate HMT parameters. Then these HMT parameters are used to train one-class classifiers. For a new pattern, its corresponding HMTs are first estimated and the parameters of these HMTs are fed to learned classifiers for classification.

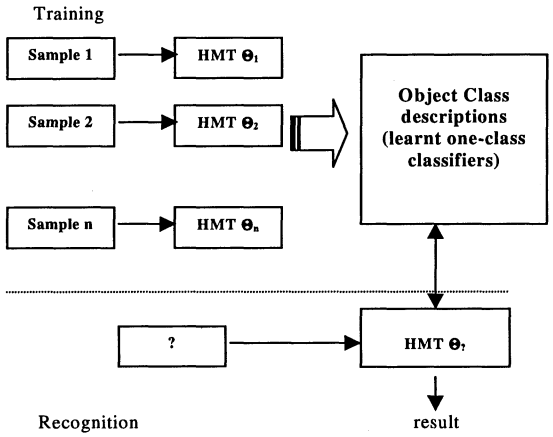

Figure 1-the evaluation scheme

4.2. Experimental Results

All the six one-class classifiers are investigated on the dataset used in (Wang & Seabra Lopes, 2004). This dataset contains two parts. There are 400 pictures from AT&T/ORL face database (AT&T, 2002) and 402 non-face pictures from (Seabra Lopes & Wang, 2002). There are some examples from each part shown in Figure 2 and 3 respectively. It should be noted that all patterns were resized as 32×32. The wavelet transform used in the following experiments is Daub4 (Wang & Seabra Lopes, 2004). Currently, experiments are run into a simulation environment which is based on PRTOOLS (Duin, 2004) and DDTOOLS (Tax, 2001).

Figure 2- Sample images from ORL face database

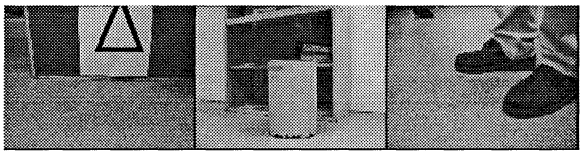

Figure 3-Some samples in our data set

A series of experiments are conducted in which face is the target class and face patterns are used for training. Only a part of face data is used for training, and then the rest of face data and all non-face data is used for independent testing. There is no non-face data introduced into training session. To know how well the amount of

training patterns affects the final classification for each classification method, the number of training patterns is increased from 10% of the face data (40 faces randomly chosen) gradually up to 90% of the face database (360 faces randomly chosen). For a certain amount of face data (10% to 90% of the whole face database), experiments are repeated ten times and the average error rate is used as the final classification score. Thus, a total of 540 experiments were run (6 algorithms, 9 data configurations and 10 trials).

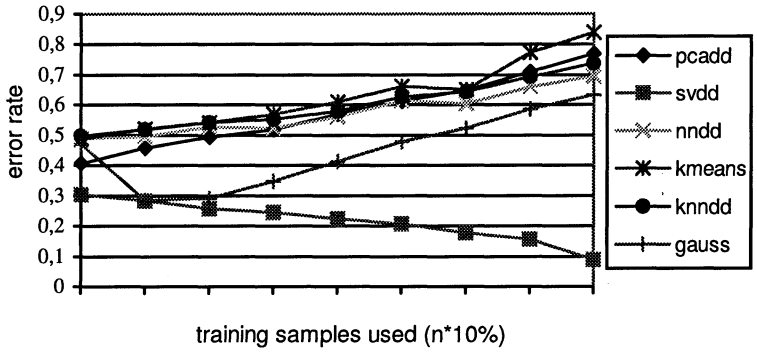

Figure 4-Error rate evaluation of all six one-class classifiers

As one can find from Figure 4, SV-DD always outperforms the other 6 methods in each of those 9 data configurations. The best error rate it can obtain is 8.8% (91.2% accuracy) when 90% faces are used into training. The worst score it achieves is around 30% in error rate when only 10% faces are involved into training. Apparently, for SV-DD can attain good results when more target patterns are involved into training.

The other five one-class classifiers don't show this characteristic. In fact, with more faces used in training, their error rates increase! This looks very unreasonable. With further analysis, some clues are found to explain this behavior. As shown in Table 1, these five classifiers generally work well with face patterns, but they don't work well with non-face patterns. They generally recognize around 80% of non-face patterns as face patterns. Thus, as the number of target patterns used for training increases, the proportion of non-target patterns in the test set also increases, resulting in higher error rates. It can be expected that with some lager data set, with more target class patterns, this might not happen again.

In some sense, several hundreds of training patterns are not easy to prepare in HRI. For the SV-DD method, it shows it can achieve reasonable performances (around 80%) with a small training set (40 to 120 training patterns, 10% to 30%). Since robot learning should be treated as a life-long learning process (Seabra Lopes & Wang, 2002), this moderate performance may form a not so bad starting point.

Finally, a brief comparison is made between the HMT/SV-DD approach and the HMT/KLD based approach described in (Wang & Seabra Lopes, 2004). Applying on the same data set used in this paper, it obtained an average accuracy about 80% on face data. At the same time, it could reject 99% non-face patterns on average although it never sees them before. The main problem of this HMT/KLD based ap-

proach is being very data dependent. Over some part of the ATT&ORL face database, its accuracies were as good as 100%; but over some other part, its accuracies were as low as some 45%. The main cause was the simple average method used to compute the overall class model. When there were great in-class variations, it usually failed since the simple average over parameters of individual HMTs smeared the true class model. When the number of training patterns increased, this problem was more serious. This method only worked well when training patterns were very much similar in pose and scale, as under this assumption the parameters of individual HMTs tended to be similar too, thus the average was closer to the true class model.

Table 1- False Negatives (FN) and False Positives (FP) when 40, 120, 240 and 360 faces were used in training

		40 faces	120 faces	240 faces	360 faces
PCA-DD	FN	0,11	0,02	0,01	0,03
	FP	0,66	0,86	0,86	0,84
SV-DD	FN	0,59	0,39	0,26	0,33
	FP	0,09	0,19	0,23	0,16
NN-DD	FN	0,02	0,02	0,02	0,04
	FP	0,89	0,88	0,84	0,77
KMEANS-DD	FN	0,05	0,01	0,01	0,01
	FP	0,86	0,92	0,92	0,92
KNN-DD	FN	0,02	0,01	0,01	0,00
	FP	0,92	0,91	0,86	0,81
GAUSS-DD	FN	1,00	0,17	0,03	0,04
	FP	0,00	0,40	0,66	0,69

In contrast, the SV-DD based approach tries to find a sphere boundary with minimal volume containing all or most of objects in a data set (Tax, 2001). Basically, when more training patterns are involved into training, this sphere decision boundary is more precise. Experiments have demonstrated this. In all experiments, SV-DD shows very steady performances, deviating in the range of ±5% to the average that is used as the final scores. This feature shows that the SV-DD based approach is less data dependent and more robust than KLD based approach.

5. CONCLUDING REMARKS

In this paper, several one-class classification methods were investigated on HMT features. Some of these classifiers can be further integrated into the context of HRI for natural language concept learning (grounding). Constraints such as only limited (target) training patterns available, interactive and flexible learning (several targets to learn simultaneously) require fast one-class classification methods. This is a very crucial step towards flexible concept learning for robots since it can relieve us from data preparation, especially outlier data. Therefore one may focus more on target class representation. In the reported experiments, only target patterns were used to

train target class models. And the learned models still have good discrimination with respect to outlier (never seen non-target) patterns.

Six one-class classification methods were evaluated on HMT features. Face and non-face classification is used as an example to demonstrate their effectiveness. It can be found from the experiments that some of such one-class classifiers, particularly SV-DD, can attain very nice performance (>90%), at least on the used data. It also provides a kind of fast learning capacity. For example, when 40 faces were used for training and all the other 762 patterns were used for test, the whole process could be done in less than 1 minute in current MATLAB implementation with a laptop having P4 1.6 GHZ CPU and 256 MB RAM. It still has room to improve, e.g., to implement the algorithm in C. All other five one-classifiers performs not well or very badly on our data. It can be concluded that SV-DD forms a promising foundation for developing a learning and classification method suitable for HRI, since not only can it obtain reasonable performance with a (relative) small amount of training patterns, but also it can achieve very good results when a larger amount of training patterns are available. From a viewpoint of lifelong learning in robotics, this potential of SV-DD can be further utilized.

Obviously it's necessary to further study these one-class learning and classification methods, for example, using other data set and/or feature extraction methods. More importantly, it's interesting to integrate some of these methods into a real robot, Carl (Seabra Lopes, 2002). How to precisely connect this kind of SV-DD to certain target class representation, i.e., natural language concept grounding, should be also further studied. Furthermore, the relation between the class representation for certain target class and specific object representation of that class in terms of such SV-DD should also be considered.

6. ACKNOWLEDGEMENTS

This work is funded by IEETA (Instituto de Engenharia Electrónica e Telemática de Aveiro), Universidade de Aveiro, Portugal, under a PhD fellowship to Q. H. Wang. We thank DSP group of Rice University to let us use their HMT source code (partially). We also thank Dr David M. J. Tax of Delft University of Technology for help on using his tools.

7. REFERENCES

1. AT & T, *The Database of Faces,* formerly "The ORL Database of Faces", at http://www.uk.research.att.com/facedatabase.html, 2002.
2. Bishop C. Novelty detection and neural network validation. In: IEE Proc. Vision, Image and Signal Processing, Special Issue on Applications of Neural Networks 1994; 4: 217-222.
3. Choi H. and Baraniuk RG. Multiscale Image Segmentation using Wavelet-Domain Hidden Markov Models. IEEE Transaction on Image Processing 2001; 9:1309-1321.
4. Crouse MS, Nowak RD and Baraniuk RG. Wavelet-Based Statistical Signal Processing using Hidden Markov Models. IEEE Transaction on Signal Processing 1998; 46:886-902.
5. Do MN. Rotation Invariant Texture Characterization and Retrieval using Steerable Wavelet-domain Hidden Markov Models. IEEE Transaction on. Multimedia 2002.
6. Duin R, PRTOOLS 4.0, Delft University of Technology, The Netherlands.

7. Durand JB and Gonçalvès P. Statistical Inference for Hidden Markov Tree Models and Application to Wavelet Trees. IEEE Transaction. On Signal Processing 2002.
8. Fan Guoliang. Wavelet-Domain Statistical Image Modeling and Processing, Ph.D. dissertation, University of Delaware, 2001.
9. Japkowicz, N.: Concept-Learning in the absence of counter-examples: an autoassociation-based approach to classification, Ph D thesis, New Brunswick Rutgers, The State Univ. of New Jersy, 1999.
10. Meng LM. "An Image-based Bayesian Framework for Face detection", In Proc. of IEEE Intl. Conf. On Computer Vision and Pattern Recognition, 2000.
11. Moya M, Koch M and Hostetler L. One-class classifier networks for target recognition applications, In: Proc. World congress on neural networks (1993), pp. 797-801.
12. Muggleton, S. and J. Firth. CProgol4.4: a tutorial introduction. In S. Dzeroski and N. Lavrac, editors, Relational Data Mining, pages 160-188. Springer-Verlag, 2001.
13. Ritter G and Gallegos M. Outliers in statistical pattern recognition and an application to automatic chromosome classification. In: Pattern Recognition Letters 1997; 525-539.
14. Roberts S and Penny W. Novelty, confidence and errors in connectionist systems. Technology report, Imperial College, London, TR-96-1(1996).
15. Romberg JK, Choi H. and Baraniuk RG. "Bayesian Tree-Structured Image Modeling using Wavelet-Domain Hidden Markov Models", In Proc. of SPIE, Denver, CO, vol. 3813, pp. 31--44, 1999.
16. Parra L, Deco G And Miesbach S. Statistical independence and novelty detection with information preserving nonlinear maps. Neural Compiutation 1996; 260-269.
17. Pham TV, Arnold MW and Smeulders WM. Face Detection by aggregated Bayesian network classifiers. Pattern Recognition Letters 2002; 4:451-461.
18. Schneiderman H and Kanade K. "A Statistical Method for 3D Object Detection Applied to Faces and Cars". In Proc. CVPR 2000, pp. 746-751.
19. Seabra Lopes L. "Carl: from Situated Activity to Language-Level Interaction and Learning", In Proc. IEEE Intl. Conf. on Intelligent Robotics & Systems; pp. 890-896, Lausanne, 2002.
20. Seabra Lopes L and Wang QH. "Towards Grounded Human-Robot Communication", In Proc. IEEE Intl. Workshop on RO-MAN, pp. 312-318, Berlin, Germany, 2002.
21. Tax DMJ. One-class classification. Ph D dissertation, Delft University of Technology, The Netherlands, 2001.
22. Turk M and Pentland A. Eigenfaces for recognition. Journal of Cognitive Neuroscience 1994, 1:71-86.
23. Wang QH and Seabra Lopes L An Object Recognition Framework Based on Hidden Markov Trees and Kullback-Leibler Distance. In Proc. 6th Asian Conf. on Computer Vision: pp. 276-281, Korea, 2004.
24. Yang MH, .Kriegman D and Ahuja N. Detecting Faces in Images: A Survey. IEEE Transaction on PAMI 2002.; 1: 34-58.
25. Yang MH, Kriegman D, and Ahuja N. Face Detection Using Multimodal Density Models. Computer Vision and Image Understanding 2001; 284-304 .
26. Zhu Y, Schwartz S. "Efficient Face Detection with Multiscale Sequential Classification". In Proc. IEEE Intl. Conf. on Image Processing, pp. 121-124, Rochester, New York, USA, 2002.

[i] http://www.ieeta.pt/carl