Introduction to the concept of structural HMM: application to mining customers' preferences in automotive design

Bouchaffra, D.; Tan, J.;
Dept. of Comput. Sci. & Eng., Oakland Univ.,
Rochester, MI, USA

This paper appears in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on
Issue Date : 23-26 Aug. 2004
Volume : 2
On page(s): 493 - 496 Vol.2
ISSN : 1051-4651
Print ISBN: 0-7695-2128-2
INSPEC Accession Number: 8280150
Digital Object Identifier : 10.1109/ICPR.2004.1334274
Date of Current Version : 20 September 2004

ABSTRACT

We have introduced in this paper the concept of structural hidden Markov models (SHMM). This new paradigm adds the syntactical (or structural) component to the traditional HMM. SHMM introduce relationships between the visible observations of a sequence. These observations are related because they are viewed as evidences of a same conclusion in a rule of inference. We have applied this novel concept to predict customer's preferences for automotive designs. SHMM has outperformed both the k-nearest neighbors and the neural network classifiers with an additional 12% increase in accuracy.

Introduction to the Concept of Structural HMM: Application to Mining Customers' Preferences in Automotive Design

Djamel Bouchaffra and Jun Tan
School of Engineering and Computer Science
Oakland University
Michigan, USA

Abstract

We have introduced in this paper the concept of structural hidden Markov models (SHMM's). This new paradigm adds the syntactical (or structural) component to the traditional HMM's. SHMM's introduce relationships between the visible observations of a sequence. These observations are related because they are viewed as evidences of a same conclusion in a rule of inference. We have applied this novel concept to predict customer's preferences for automotive designs. SHMM has outperformed both the k-nearest neighbors and the neural network classifiers with an additional 12% increase in accuracy.

1. Introduction

Almost all system modeling techniques include two simple relationships: the *classification* relationship and the *componential* relationship. The classification relationship is the means by which the human mind generalizes experience so that the class stars is filled with all those shiny dots that we see in the sky of a summer night. The componential relationship is the means by which we organize the whole made up of many parts that seems to be an inherent quality of all patterns, from stars to automobiles to people to sand. Therefore, statistics and structure are always driving humans in a decision problem in pattern recognition (PR) [1]. Ideally, researchers would have liked a solution to a PR problem consists of the following stages: (i) find a feature vector x, (ii) train a system using a set of training patterns whose classification is *a-priori known*, and (iii) classify unknown incoming patterns. Unfortunately, for most practical problems, this approach is not feasible because the precise feature vector is not obvious and thus training becomes impossible. Therefore, the analytical approaches which process the patterns only on a quantitative basis but ignore the interrelationships between the components of the patterns quite often fail. The truth is that a pattern contains some relational and structural information from which it is difficult and sometimes impossible to derive an appropriate feature vector. Syntactical pattern recognition [3], Bayesian belief networks [7], and Hidden Markov models (HMM's) [8] are some of the techniques that can handle statistical and structural data but *separately*. However, there is a growing need for developing mathematical paradigms that embed both statistics and syntax at the same time. The main goal in this paper is to merge statistics and syntax in a seamless way within a novel concept that we called "structural hidden Markov models". This paper is organized as follows: section 2 covers the mathematical description of a structural hidden Markov model. The application and the experiments are presented in section 3 and the conclusion and future work are laid in section 4.

2. Fusion of Statistics and Syntax: The Concept of Structural HMM

In this section, we build a mathematical model that merges statistical and structural information together. This model that we called SHMM goes beyond the traditional HMM since it emphasizes the structure (or syntax) of the visible sequence of observation. It provides information about the structure formed by the visible sequence of observations. Let $O = (v_1 v_2 \ldots v_T)$ be the observation sequence of length T and $q = (q_1 q_2 \ldots q_T)$ be the state sequence where q_1 is the initial state, given model λ, we can write:

$$P(O \mid \lambda) = \sum_{all\ q} P(O, q \mid \lambda) = \sum_{all\ q} \left[P(O \mid q, \lambda) \times P(q \mid \lambda) \right],$$

and using state conditional independence, we obtain:

$$P(O \mid q, \lambda) = \prod_{t=1}^{T} P(v_t \mid q_t, \lambda).$$

However, there are several scenarios where the conditional independence assumption doesn't hold. For example, while standard HMM's perform well in recognizing amino acids and consequent construction of proteins from the first level structure of DNA sequences [5], they are inadequate for predicting the secondary structure of a protein. The reason for the inadequacy comes from the fact that the same order of amino acid sequences have different folding modes in natural circumstances. Therefore, there is a need to balance the loss incurred by this state conditional independence assumption. Our idea is to create syntactical rules that possess these observation sequences as evidences. These rules show how the secondary structure of a protein is constructed: they represent the structural information.

In the SHMM framework, the observation sequence O is not only one sequence in which all observations are conditionally independent, but a sequence that is divided into a series of subsequences $O_i = (v_1 v_2 \ldots v_{r_i})$ $(1 \leqslant i \leqslant s)$, where s is the number of subsequences. The observations in a subsequence are related in the sense that they represent evidences for a same conclusion of a rule such as $\mathbf{C}_{R_i} \stackrel{R_i}{\Longleftarrow} v_1 \wedge v_2 \wedge \ldots \wedge v_{r_i}$. The length of each subsequence is r_i $(1 \leqslant i \leqslant s)$ and $T = \sum_{i=1}^{s} r_i$. The structural information in this model is expressed through the activation of the rules set by the experts. The whole sequence of observations can be written directly as:

$$O = (v_1 v_2 \cdots v_{r_1}, C_{R_1}, v_{r_1+1} v_{r_1+2} \cdots v_{r_1+r_2}, C_{R_2}, \cdots, v_{r_1+r_2 \cdots r_s}, C_{R_s}) = (O_1 C_{R_1}, O_2 C_{R_2}, \cdots, O_s C_{R_s}).$$

Therefore, we can define a Structural HMM as:

Definition 2.1 *A structural hidden Markov model is a quintuple $\lambda = (\pi, \mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$, where:*

- *π is the initial state probability vector, $\pi = \{\pi_i\}$, where $\pi_i = P(q_1 = i)$ and $1 \leqslant i \leqslant N$, $\sum_i \pi_i = 1$. N is the number of states in the model. We label the individual states as 1, 2, ..., N, and denote the state at time t as q_t.*
- *$\mathcal{A}$ is the state transition probability matrix, $A = \{a_{ij}\}$, where $a_{ij} = P(q_{t+1} = j \mid q_t = i)$ and $1 \leqslant i, j \leqslant N$, $\sum_j a_{ij} = 1$.*
- *$\mathcal{B}$ is the state conditional probability matrix of the visible observations, $B = \{b_j(k)\}$, in which $b_j(k) = P(v_k \mid q_j)$, $1 \leqslant k \leqslant M$ and $1 \leqslant j \leqslant N$. M is the number of distinct observations in one state. We use a symbol to represent each observation, and the set of symbols is denoted as $V = \{v_1, v_2, \ldots, v_M\}$.*
- *$\mathcal{C}$ is the posterior probability matrix of a conclusion given a sequence of observations, $\mathcal{C} = \{c_j(i)\}$, where $c_j(i) = P(C_{R_i} \mid O_j)$, $\sum_i c_j(i) = 1$. We denote the conclusion assigned to O_j via rule R_i as C_{R_i}. This can be depicted as: $\mathbf{C}_{R_i} \stackrel{R_i}{\Longleftarrow} v_{j_1} \wedge v_{j_2} \wedge \ldots \wedge v_{j_k}$, where j_k is the number of evidences (observations) in the tail of the rule. The meaning of R_i depends on the applications at hand. For example, a protein's type can be expressed as a conjunction of amino acids.*
- *$\mathcal{D}$ is the conclusion transition probability matrix, $D = \{d_{ij}\}$, where $d_{ij} = P(C_{R_i} \mid C_{R_j})$, $\sum_j d_{ij} = 1$.*

An example of the interaction between sequences of observations and their corresponding rule conclusions can be illustrated by Figure 1. The choice of the topology of the network in the figure depends on the information we have regarding a particular application. We now define the problems that are involved in a structural hidden Markov model.

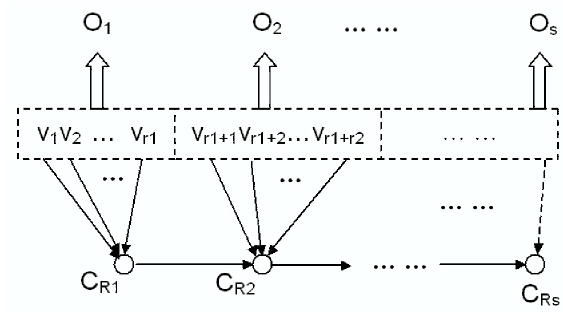

Figure 1. Structural HMM topology.

2.1. Problems assigned to a Structural HMM

There are four problems that are assigned to a SHMM: (i) Probability evaluation, (ii) Statistical decoding, (iii) Structural decoding and (iv) Parameter estimation.

2.1.1 Problem 1: Probability Evaluation

The evaluation problem in SHMM is to compute:

$$P(O \mid \lambda) = P(O_1 C_{R_1}, O_2 C_{R_2}, \cdots, O_s C_{R_s} \mid \lambda)$$

$$= \prod_{i=1}^{s} P(O_i C_{R_i} \mid O_{i-1} C_{R_{i-1}}, \ldots, O_1 C_{R_1}, \lambda). \quad (1)$$

Because C_{R_i} is conditionally dependent on $C_{R_{i-1}}$ and O_i is independent of $C_{R_{i-1}}$, Equation 1 can be expressed as:

$$P(O \mid \lambda) = \prod_{i=1}^{s} P(O_i C_{R_i} \mid C_{R_{i-1}}, \lambda),$$

$$= \prod_{i=1}^{s} \left[P(C_{R_i} \mid O_i, C_{R_{i-1}}, \lambda) \times P(O_i \mid C_{R_{i-1}}, \lambda) \right]$$

$$= \prod_{i=1}^{s} \left[P(C_{R_i} \mid O_i, C_{R_{i-1}}, \lambda) \times P(O_i \mid \lambda) \right]$$

$$= \prod_{i=1}^{s} P(C_{R_i} \mid O_i, C_{R_{i-1}}, \lambda) \times \prod_{i=1}^{s} P(O_i \mid \lambda).$$

We assume for now that C_{R_i} is independent of $C_{R_{i-1}}$, finally, this provides:

$$P(O \mid \lambda) = \prod_{i=1}^{s} c_i(i) \times \sum_{q_1, q_2, \ldots, q_{r_i}} \pi_{q_1} b_{q_1}(v_1) a_{q_1 q_2} b_{q_2}(v_2) \ldots a_{q_{(r_i-1)} q_{r_i}} b_{q_{ri}}(v_{r_i}).$$

2.1.2 Problem 2: Statistical Decoding

The statistical decoding problem consists of determining the optimal state sequence $q^* = \underset{q}{argmax}(P(O, q \mid \lambda))$ that best "explains" the sequence of observations . It can be computed using Viterbi algorithm as in traditional HMM's.

2.1.3 Problem 3: Structural Decoding

The structural decoding problem consists of determining the optimal rule conclusion sequence $\mathcal{C}^* = < C^*_{R_1} C^*_{R_2} \dots C^*_{R_t} >$ such that: $\mathcal{C}^* = \underset{\mathcal{C}}{argmax}(P(O,\mathcal{C} \mid \lambda))$.

We define: $\delta_t(i) = \underset{C_{R_1} C_{R_2} \dots C_{R_t}}{max} P(O, C_{R_1} C_{R_2} \dots C_{R_t} = i \mid \lambda)$ that is, $\delta_t(i)$ is the highest probability along a single path, at time t, which accounts for the first t observations and ends in rule conclusion i. Then, we estimate the following by induction: $\delta_{t+1}(j) = \underset{i}{max}\, \delta_t(i) d_{ik} \quad b_j(v_{t+1})$. Similarly, this can be computed using *Viterbi* algorithm. However, we estimate δ in each step *through conclusion transition probability matrix instead of state transition probability matrix*. This optimal sequence of conclusions describes the structural pattern piecewise.

2.1.4 Problem 4: Parameter Estimation

The re-estimation phase of the parameters $\{\pi_i\}$, $\{a_{ij}\}$, $\{b_j(k)\}$ and $\{d_{ij}\}$ is conducted as in traditional HMM's, using the Baum-Welch optimization technique. However, *the most difficult problem* is the estimation of $c_j(i)$. There are two types of uncertainty that can be expressed using first-order logical rule: *statistical uncertainty*, where we are uncertain of the distribution of conclusions across properties, and *propositional uncertainty*, where we are uncertain of the truth of logical sentences [6].

We define both of the uncertainties in the following:

- **Statistical Uncertainty:** In this method, the uncertainty on the conclusion is expressed as a posterior probability. Using *naive Bayes' rule*, we make the following estimation:

$$P(C_{R_i} | v_{j1} v_{j2} \dots v_{jk}) \approx \frac{\prod_{j=1}^{k} P(v_{ij}|C_{R_i}) \times P(C_{R_i})}{\sum_{C_{R_i}} \prod_{j=1}^{k} P(v_{ij}|C_{R_i}) \times P(C_{R_i})}. \tag{2}$$

  The term $P(v_{ij} \mid C_{R_i})$ is estimated using the ML criterion. The prior distribution $P(C_{R_i})$ is assumed to be uniform.

- **Propositional Uncertainty:** Nilsson was among the first to consider the problem of representing propositional uncertainty, i.e., uncertainty regarding the truth of logical sentences. The implication rule $C_{R_i} \overset{R_i}{\Longleftarrow} v_{j1} v_{j2} \dots v_{jk}$ is viewed as an entailment between a tail and a head predicate. We transform the chain relation into predicates:
  $v_{j_1} \wedge v_{j_2} \wedge \dots \wedge v_{j_k} \Longrightarrow C_{R_i}$
  $A_1 \wedge A_2 \wedge \dots \wedge A_k \Longrightarrow C_{R_i}$
  where $A_k = v_{j_k}$. If $A = A_1 \wedge A_2 \wedge \dots \wedge A_k$ then the problem consists of determining the probability assigned to C_{R_i} given the probability of the entailment (that depends on our expertise in the application at hand) and the probability assigned to A (estimated). The truth of logical sentences is defined in term of *possible worlds*. A probability distribution over *possible worlds* is built. An agent's world model express its degree of belief that any possible world is the actual world, and can be used to compute the degree of belief (sentence probability) of a sentence. Given a probabilistic knowledge $P(A)$ that expresses our propositional uncertainty, we would like to compute the degree of belief for $P(C_{R_i})$. The random worlds formulation allows us to reason under propositional uncertainty, given a world model. However, we are immediately faced with identifiability: in general, our probabilistic knowledge base $P(A)$ can be compatible with infinitely many possible world models. We can either accept this indeterminacy or introduce an additional criterion such as the "Jaynes" [4] maximum entropy that eliminates it. In this probabilistic logic framework, the probability of an observation is $p(v_{j_k}) = p(v_{j_k} = true)$, which is the probability that a predicate is *true*. The truth table of predicates of the rule $C_{R_i} \overset{R_i}{\Longleftarrow} v_{j1} v_{j2} \dots v_{jk}$ is illustrated in Table 1. To estimate the probability of $P(C_{R_i})$, we need to determine the vector $\mathcal{W} = P(w_i)$ such that $C \cdot \mathcal{W} = \Pi$, where C is the consistent logical truth table, $\mathcal{W}$ is the probability vector assigned to possible worlds, and Π is the probability vector assigned to predicates. In order to determine a unique solution to this problem, we maximize the entropy assigned to the possible worlds distribution [4].

	Possible worlds: $w_1\ w_2 \cdots\ w_{k-1}\ w_k$
v_{j_1}	0 0 0 0 $\cdots$ 1 1 1 1
v_{j_2}	0 0 1 1 $\cdots$ 0 0 1 1
...	...
v_{j_k}	0 0 0 0 $\cdots$ 0 0 1 1
A	0 0 0 0 $\cdots$ 0 0 1 1
$A \Longrightarrow C_{R_i}$	1 1 1 1 $\cdots$ 1 1 0 1
C_{R_i}	0 0 0 0 $\cdots$ 0 0 0 1

Table 1. Logical truth table assigned to predicates

3. Application and Experiments

We have applied this research in order to aid automotive design engineers in predicting customers' perceptions on particular car makes before these cars are put into making. This enables automotive companies to save money by data mining customers preferences. We collected 228 images of regular cars with their three views (front, size and rear, i.e., 684 images). During our survey, we extracted the contour of the three views of the whole car (thus removed the influence of colors on a student's opinion), then presented these contours to 100 university students. The students were asked to give their opinions on the three views of a car viewed separately as well as their opinions on the car as a whole. Opinions are adjectives that express their feelings of the car view at first sight. Thus we obtained 912 adjectives clustered with synonymy using the online lexical database WordNet [2]. Each centroid of a cluster is called a *perception* which is a conclusion in SHMM modeling. Each respondent's opinion (adjective, such as beautiful, sporty, etc.) belongs to one and only one perception.

Therefore, we extracted the contour of "front (f)" and represented it as $O_f = (v_1 v_2 \ldots v_{r_f})$, where v_i are bits representing the chain code directions of the contour and r_f is the length of O_f. The customer's opinion assigned to this view is represented by C_{R_f}, where C_{R_f} is the conclusion assigned to rule R_f that defines how the opinion of this view is obtained from the chain code description of its contour. An example of such a rule is: "$attractive$" $\Longleftarrow$ $3017432 \ldots 12$, which means that the contour of the view represented by the chain code string is tagged as "attractive". We did similar task on the "side (s)" and the "rear (r)" views and obtained O_s, C_{R_s}, O_r and C_{R_r} respectively.

The k-nearest neighbors and neural networks classifiers have also been experimented in order to compare them with SHMM. We used the statistical uncertainty discussed in section 2.1.4 for classification. Preliminary performance results are depicted in Table 2. If our predicted conclusion (or category) is C_p and the true conclusion obtained from survey is C_t, then our precision is defined as:

$$Precision = \frac{\sum \delta(C_p - C_t)}{\mid input \;\; patterns \mid} \tag{3}$$

where $\delta(x - a)$ is the Kronecker symbol which is "1" if x=a, and "0" otherwise, the denominator $\mid input\, patterns \mid$ represents the total number of patterns (external contours of a car). SHMM outperformed the two traditional classifiers since its accuracy is 90%. This optimal prediction of user perceptions is fed to the design engineer before the car is put into making.

Precision (%) **Sample Size**	k-NN	NN	SHMM
140 cars	52.1	54.2	66.7
228 cars	73.2	78.6	90

Table 2. Performances obtained using the k-nearest neighbors, the neural network and the SHMM classifiers.

4. Conclusion and Future Work

We have introduced a novel mathematical paradigm that is capable of exploiting both statistical and syntactical information at the same time. We believe that the concept of structural hidden Markov model will bridge the gap between statistical and syntactical researchers within the PR community. Our approach relates visible observations through their contribution to a same logical conclusion of a syntactic rule. We have seen that the structural decoding can have two different interpretations. We have used the statistical uncertainty approach to answer the "structural decoding problem". We have obtained promising results in the automobile application described above. However, this research is still ongoing, more data need to be collected in order to measure the real contribution of SHMM's. We also need to test the propositional uncertainty approach and apply SHMM's in other areas.

References

[1] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, New York, 2001.

[2] C. Fellbaum. *WordNet: an Electronic Lexical Database*. Bradford Book, 1998.

[3] K. Fu. *Syntactic Pattern Recognition and Applications*. Prentice-Hall, Englewood Cliffs, N.J., 1982.

[4] E. T. Jaynes. On the rationale of maximum entropy methods. *Proceedings of the IEEE*, 70:939–952, 1982.

[5] A. Krogh, M. Brown, I. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.*, 235(1994):1501–1531.

[6] N. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–87, 1986.

[7] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, 1988.

[8] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257 – 286, 1989.