

Lecture Notes in Statistics

89

**P. Cheeseman
R. W. Oldford (Eds.)**

Selecting Models from Data

Artificial Intelligence and Statistics IV



Springer-Verlag

Contents

Preface	v
I Overviews: Model Selection	1
1 Statistical strategy: step 1 D.J.Hand	3
2 Rational Learning: Finding a Balance Between Utility and Efficiency Jonathan Gratch, Gerald DeJong and Yuhong Yang	11
3 A new criterion for selecting models from partially observed data Hidetoshi Shimodaira	21
4 Small-sample and large-sample statistical model selection criteria S. L. Sclove	31
5 On the choice of penalty term in generalized FPE criterion Ping Zhang	41
6 Cross-Validation, Stacking and Bi-Level Stacking: Meta-Methods for Classification Learning Cullen Schaffer	51
7 Probabilistic approach to model selection: comparison with unstructured data set Victor L. Brailovsky	61
8 Detecting and Explaining Dependencies in Execution Traces Adele E. Howe and Paul R. Cohen	71
9 A method for the dynamic selection of models under time constraints Geoffrey Rutledge and Ross Shachter	79
II Graphical Models	89
10 Strategies for Graphical Model Selection David Madigan, Adrian E. Raftery, Jeremy C. York, Jeffrey M. Bradshaw, and Russell G. Almond	91
11 Conditional dependence in probabilistic networks Remco R. Bouckaert	101

12 Reuse and sharing of graphical belief network components	113
Russell Almond, Jeffrey Bradshaw, and David Madigan	
13 Bayesian Graphical Models for Predicting Errors in Databases	123
David Madigan, Jeremy C. York, Jeffrey M. Bradshaw, and Russell G. Almond	
14 Model Selection for Diagnosis and Treatment Using Temporal Influence Diagrams	133
Gregory M. Provan	
15 Diagnostic systems by model selection: a case study	143
S. L. Lauritzen, B. Thiesson and D. J. Spiegelhalter	
16 A Survey of Sampling Methods for Inference on Directed Graphs	153
Andrew Runnalls	
17 Minimizing decision table sizes in influence diagrams: dimension shrinking	163
Nevin Lianwen Zhang, Runping Qi, and David Poole	
18 Models from Data for Various Types of Reasoning	173
Raj Bhatnagar and Laveen N Kanal	
 III Causal Models	 181
19 Causal inference in artificial intelligence	183
Michael E. Sobel	
20 Inferring causal structure among unmeasured variables	197
Richard Scheines	
21 When can association graphs admit a causal interpretation?	205
Judea Pearl and Nanny Wermuth	
22 Inference, Intervention, and Prediction	215
Peter Spirtes and Clark Glymour	
23 Attitude Formation Models: Insights from TETRAD	223
Sanjay Mishra and Prakash P. Shenoy	
24 Discovering Probabilistic Causal Relationships: A Comparison Between Two Methods	233
Floriana Esposito, Donato Malerba, and Giovanni Semeraro	
25 Path Analysis Models of an Autonomous Agent in a Complex Environment	243
Paul R. Cohen, David M. Hart, Robert St. Amant, Lisa A. Ballesteros and Adam Carlson	

IV Particular Models	253
26 A Parallel Constructor of Markov Networks Randy Mechling and Marco Valtorta	255
27 Capturing observations in a nonstationary hidden Markov model Djamel Bouchaffra and Jacques Rouault	263
28 Extrapolating Definite Integral Information Scott D. Goodwin, Eric Neufeld, and André Trudel	273
29 The Software Reliability Consultant George J. Knafl and Andrej Semrl	283
30 Statistical Reasoning to Enhance User Modelling in Consulting Systems Paula Hietala	293
31 Selecting a frailty model for longitudinal breast cancer data D. Moreira dos Santos and R. B. Davies	299
32 Optimal design of reflective sensors using probabilistic analysis Aaron Wallack and Edward Nicolson	309
V Similarity-Based Models	319
33 Learning to Catch: Applying Nearest Neighbor Algorithms to Dynamic Control Tasks David W. Aha and Steven L. Salzberg	321
34 Dynamic Recursive Model Class Selection for Classifier Construction Carla E. Brodley and Paul E. Utgoff	329
35 Minimizing the expected costs of classifying patterns by sequential costly inspections Louis Anthony Cox, Jr. and Yuping Qiu	339
36 Combining a knowledge-based system and a clustering method for a construction of models in ill-structured domains Karina Gibert and Ulises Cortés	351
37 Clustering of Symbolically Described Events for Prediction of Numeric Attributes Bradley L. Whitehall and David J. Sirag, Jr.	361
38 Symbolic Classifiers: Conditions to Have Good Accuracy Performance C. Feng, R. King, A. Sutherland, S. Muggleton, and R. Henery	371

VI Regression and Other Statistical Models	381
39 Statistical and neural network techniques for nonparametric regression Vladimir Cherkassky and Filip Mulier	383
40 Multicollinearity: A tale of two nonparametric regressions Richard D. De Veaux and Lyle H. Ungar	393
41 Choice of Order in Regression Strategy Julian J. Faraway	403
42 Modelling response models in software D.G. Anglin and R.W. Oldford	413
43 Principal components and model selection Beat E. Neuenschwander and Bernard D. Flury	425
VII Algorithms and Tools	433
44 Algorithmic speedups in growing classification trees by using an additive split criterion David Lubinsky	435
45 Markov Chain Monte Carlo Methods for Hierarchical Bayesian Expert Systems Jeremy C. York and David Madigan	445
46 Simulated annealing in the construction of near-optimal decision trees James F. Lutsko and Bart Kuijpers	453
47 SA/GA : Survival of the Fittest in Alaska Kris Dockx and James F. Lutsko	463
48 A Tool for Model Generation and Knowledge Acquisition Sally Jo Cunningham and Paul Denize	471
49 Using knowledge-assisted discriminant analysis to generate new comparative terms Bing Leng and Bruce G. Buchanan	479

Capturing observations in a nonstationary hidden Markov model

Djamel Bouchaffra and Jacques Rouault

Cristal-Gresec - Université Stendhal
B.P. 25 - 38040 Grenoble Cedex 9 - France

ABSTRACT This paper is concerned with the problem of morphological ambiguities using a Markov process. The problem here is to estimate interferent solutions that might be derived from a morphological analysis. We start by using a Markov chain with one long sequence of transitions. In this model the states are the morphological features and a sequence corresponds to a transition from one feature to another. After having observed an inadequacy of this model, one will explore a nonstationary hidden Markov process. Among the main advantages of this latter model we have the possibility to assign a type to a text, given some training samples. Therefore, a recognition of "style" or a creation of a new one might be developed.

27.1 Introduction

27.1.1 Automatic analysis of natural language

This work lies within a textual analysis system in natural language discourse (French in our case). In most systems used today, the analysis process is divided into *levels*, starting from morphology (first level) through syntax, and semantics to pragmatics. These levels are sequentially activated, without backtracking, originating in the morphological phase and ending in the pragmatic one. Therefore, the *i*-th level knows only the results of preceding levels. This means that, at the morphological level, each word in the text (*a form*) is analyzed autonomously out of context. Hence, for each form, one is obliged to consider all possible analysis.

Example : let's consider the sequence of the two forms *cut* and *down* :

- *cut* can be given 3 analyses : verb, noun, adjective ;
- *down* can be a verb, an adverb or a noun.

The number of possible combinations based upon the independance of the analysis of one form in relation with the others implies that the phrase *cut down* is liable to *nine* interpretations, independently on the context.

These multiple solutions are transmitted to syntactic parsing which doesn't eliminate them either. In fact, as a syntactic parser generates its own interferent analyses, often from interferent morphology analysis, the problems with which we are confronted are far from being solved. In order to provide a solution to these problems, we have recourse to statistical methods. Thus the result of the morphological analysis is filtered when using a Markov model.

¹ *Selecting Models from Data: AI and Statistics IV*. Edited by P. Cheeseman and R.W. Oldford. © 1994 Springer-Verlag.

27.1.2 Morphological analysis

A morphological analyser must be able to cut up a word form into smaller components and to interpret this action. The easiest segmentation of a word form consists in separating word terminations (inflexional endings) from the rest of the word form called *basis*. We have then got a *inflexional morphology*. A more accurate cutting up consists in splitting up the basis into affixes (*prefixes, suffixes*) and *root*. This is then called *derivational morphology*.

The interpretation consists in associating the segmentation of a word form with a set of informations, particularly including :

- the general morphological class : verb, noun-adjective, preposition, ...
- the values of relevant morphological variables : number, gender, tense, ...

Therefore, an interpretation is a class plus values of variables ; such a combination is called a *feature*. Note that a word form is associated with several features in case where there are multiple solutions.

27.1.3 Why statistical procedures?

Because of the independance of the analysis levels, it is difficult to provide contextual linguistic rules. This is one of the reasons why we fall back on statistical methods. These latter method possess another advantage : they reflect simultaneously language properties, e.g. the impossibility to obtain a determinant followed directly by a verb, and properties of the analysed corpus, e.g. a large number of nominal phrases.

Some researchers used Bayesian approaches to solve the problem of morphological ambiguities. However, these methods have a clear conceptual framework and powerful representations, but must still be knowledge-engineered, rather than trained. Very often in the application of these methods, researchers have a good observation of the individuals of the population, *because the observation is a relative notion*. Therefore, we have difficulty in observing possible transitions of the individuals. The way of "capturing" the individuals depends on the environment encountered.

27.2 A morphological features Markov chain

27.2.1 The semantic of the model

Let m be the number of states, T the length of state sequence and $\{f_i/1 \leq i \leq m\}$ the states or morphological features ; we have only one individual ($n = 1$) for each transition time $t = 1, 2, \dots, T$. A first order m -states Markov chain is defined by an $m \times m$ state transition matrix P , an $m \times 1$ initial probability vector Π , where :

$$P = (P_{f_i, f_j})$$

$$i, j = 1, 2, \dots, m$$

$$P_{f_i, f_j} = \text{Prob}[e_{t+1} = f_j / e_t = f_i]$$

$$\Pi_{f_i} = \text{Prob}[e_1 = f_i]$$

$$i = 1, 2, \dots, m$$

By definition, we have :

$$\sum_{j=1}^{j=m} P_{f_i, f_j} = 1 \quad \text{pour } i = 1, 2, \dots, m$$

$$\sum_{k=1}^{k=m} \Pi_{f_k} = 1$$

The probability associated to a realization E of this Markov chain is :

$$Prob[E/P, \Pi] = \Pi_{e_1} \times \prod_{t=2}^{t=T} P_{e_{t-1}, e_t}$$

27.2.2 Estimation of transition probabilities

As pointed out by Bartlett in Anderson and Goodman [AG57] the asymptotic theory must be considered with respect to the variable *number of times of observing the word form in a single sequence of transitions*, instead of the variable *number of individuals in a state when T is fixed*. However, this asymptotic theory was considered because the number of times of observing the word form increases ($T \rightarrow +\infty$). Furthermore, we cannot investigate the stationary properties of the Markov process, since we only have one word form (one individual) at each transition time. Therefore, we assumed stationarity. Thus, if N_{f_i, f_j} is the number of times that the observed word form was in the feature f_i at time $t-1$ and in the feature f_j at time t , for $t \in \{1, 2, \dots, T\}$, then the estimates of the transition probabilities are :

$$\hat{P}_{f_i, f_j} = \frac{N_{f_i, f_j}}{N_{f_i+}}$$

where N_{f_i+} is the number of times that the word form was in state f_i . The estimated transition probabilities are evaluated on one training sample. We removed the morphological ambiguities by choosing the sequence E of higher probability.

27.3 A Markov model with hidden states and observations

The inadequacy of the previous model to remove certain morphological ambiguities has led us to believe that some unknown hidden states govern the distribution of the morphological features. Instead of passing from one morphological feature to another, we focused only on the surface of one random sample, i.e. an observation was a morphological feature. As pointed out in [ROU88], this latter entity cannot be extracted without a context effect in a sample. In order to consider this context effect, we have chosen criteria like *the nature of the feature, its successor feature, its position in a sentence, the position of the sentence in the text*. An observation o_i is then a *known hidden vector* whose components are values of the criteria presented here. Of course, one can explore other criteria.

Définition 1 A hidden Markov model (HMM) is a Markov chain whose states cannot be observed directly but only through a sequence of observation vectors.

A HMM is represented by the state transition probability P , the initial state probability vector Π and a $T \times K$ matrix V (K is the number of states); the elements of V are the conditional densities $v_i(o_t) = \text{density of observation } o_t \text{ given } e_t = i$. Our aim is the determination of the optimal model estimate $\mathcal{V}^* = (\Pi^*, P^*, V^*)$ given a certain number of samples: this is the training problem.

Theoreme 1 The probability of a sample $S = \{o_1, o_2, \dots, o_T\}$ given a model \mathcal{V} can be written as:

$$\text{Prob}(S/\mathcal{V}) = \sum_E \Pi_{e_1} v_{e_1}(o_1) \times \prod_{t=2}^{t=T} P_{e_{t-1}, e_t} v_{e_t}(o_t)$$

Proof: For a fixed state sequence $E = (e_1, e_2, \dots, e_T)$, the probability of the observation sequence $S = \{o_1, o_2, \dots, o_T\}$ is:

$$\text{Prob}(S/E, \mathcal{V}) = v_{e_1}(o_1) \times v_{e_2}(o_2) \times \dots \times v_{e_T}(o_T)$$

The probability of a state sequence is:

$$\text{Prob}(E/\mathcal{V}) = \Pi_{e_1} \times P_{e_1, e_2} \times P_{e_2, e_3} \times \dots \times P_{e_{T-1}, e_T}$$

Using the formula:

$$\text{Prob}(S, E/\mathcal{V}) = \text{Prob}(S/E, \mathcal{V}) \times \text{Prob}(E/\mathcal{V})$$

and summing this joint probability over all possible states sequences E , one demonstrates the theorem.

The interpretation of the previous equation is: initially at time $t = 1$, the system is in state e_1 with probability Π_1 and we observe o_1 with probability $v_{e_1}(o_1)$. The system then makes a transition to state e_2 with probability P_{e_1, e_2} and we observe o_2 with probability $v_{e_2}(o_2)$. This process continues until the last transition from state e_{T-1} to state e_T with probability P_{e_{T-1}, e_T} and then we observe o_T with probability $v_{e_T}(o_T)$.

In order to determine one of the estimate of the model $\mathcal{V} = (\Pi, P, V)$, one can use the maximum likelihood criterion (or a max entropy) for a certain family S_i where $i \in \{1, 2, \dots, L\}$ of training samples. Some methods of choosing representative samples of fixed length are presented in [BOU92]. The problem is expressed mathematically as:

$$\max_{v_i} f(S_1, S_2, \dots, S_L/\mathcal{V}) = \max_{v_i} \left\{ \prod_{j=1}^{j=L} \left[\sum_E \Pi_{e_1} \times v_{e_1}(o_1^j) \times \prod_{t=2}^{t=T} P_{e_{t-1}, e_t} v_{e_t}(o_t^j) \right] \right\}$$

There is no known method to solve this problem analytically, that is the reason why we use iterative procedures. We start by determining first the optimal path for each sample. An optimal path E^* is the one which is associated to the higher probability of the sample. Using the well-known Viterbi algorithm, one can determine this optimal path. The different steps for finding the single best state sequence in the Viterbi algorithm are:

step 1: initialization

$$\delta_1(i) = \Pi_i v_i(o_1) \quad (1 \leq i \leq K)$$

$$\psi_1(i) = 0$$

step 2 : recursion for $2 \leq t \leq T$ and $1 \leq j \leq K$:

$$\delta_t(j) = \max_{1 \leq i \leq K} [\delta_{t-1}(i) P_{i,j}] v_j(o_t)$$

$$\psi_t(j) = \arg \max_{1 \leq i \leq K} [\delta_{t-1}(i) P_{i,j}]$$

step 3 : termination

$$P^* = \max_{1 \leq i \leq K} [\delta_T(i)]$$

$$e_T^* = \arg \max_{1 \leq i \leq K} [\delta_T(i)]$$

state sequence backtracking for $t = T-1, T-2, \dots, 1$:

$$e_t^* = \psi_{t+1} e_{t+1}^*$$

P^* is the state-optimized likelihood function and $E^* = \{e_1^*, e_2^*, \dots, e_T^*\}$ is the optimal state sequence. Instead of tracking all possible paths, one successively tracks only the optimal paths E_t^* of all samples. Thus, this can be written as :

$$g(o_1, o_2, \dots, o_T; E^*, \mathcal{V}) = \max_E \{ \Pi_{e_1} \times v_{e_1}(o_1) \times \prod_{t=2}^{t=T} P_{e_{t-1}, e_t} v_{e_t}(o_t) \}$$

This computation has to be done for all the samples. Among all the v_i ($i \in \{1, 2, \dots, L\}$) associated to optimal paths, we decide to choose as best model estimate the one which maximizes the probability associated to a sample. It can be written as :

$$\mathcal{V}^* = \arg \{ \max_{v_i} g(o_1^i, o_2^i, \dots, o_T^i; E^*, \mathcal{V}_i) \}$$

$$i \in \{1, 2, \dots, L\}$$

27.4 The different steps of the method

We present an interactive method which enables us to obtain an estimator of the model V . This method is suitable for direct computation.

First step : one has to cluster the sample with respect to the chosen criteria. two possibilities are offered : a *classification* or a *segmentation*. In this latter procedure, the user may structure the states ; operating in this way, the states appear like unknown hidden states. However, in a classification the system structures its own states according to a suitable norm. Thus, the states appear like unknown hidden ones. The clusters formed by one of the two procedures represent the first states of the model, they form *the first training path*.

Second step : one estimate the transition probabilities using the following equations and the probability of each training vector for each state $v_i(o_t)$. This is the first model \mathcal{V}_1 . Let $i, i \in \{1, 2, \dots, K\}$ and $t \in \{1, 2, \dots, T\}$.

- Let $Nb(o_1, i)$ be the number of times the observation o_1 belongs to the state i and Nbp the number of training paths, then :

$$\hat{\Pi}_i = \frac{Nb(o_1, i)}{Nbp}$$

- Let $Nb(o_{t-1}, i; o_t, j)$ be the number of times the observation o_{t-1} belongs to the state i and the observation o_t belongs to the state j , then :

$$\hat{P}_{i,j}(t) = \frac{Nb(o_{t-1}, i; o_t, j)}{Nb(o_{t-1}, i)}$$

- The previous estimation formula can be written as :

$$\hat{P}_{i,j}(t) = \frac{N_{i,j}(t)}{N_i(t-1)} = \frac{N_{i,j}(t)}{N_{i+}(t)}$$

where $N_{i,j}(t)$ is the number of transitions from state i at time $t-1$ to state j at time t and $N_i(t-1)$ the number of times the state i is visited at time $t-1$.

- Let $Nbez(o_1, i)$ be the expected number of times of being in state i and observing o_1 and $Nbez(i)$ the expected number of times of being in state i , then :

$$\hat{v}_i(o_1) = \frac{Nbez(o_1, i)}{Nbez(i)}$$

Third step : one computes $f(o_1, o_2, \dots, o_T; \mathcal{V}_1)$ and determines the next training path, or clustering, necessary to increase $f(o_1, o_2, \dots, o_T; \mathcal{V}_1)$. We apply the second step to this training path. The procedure is repeated until we reach the maximum value of the previous function. At this optimal value, we have E_1^* and v_1 of the first sample. This step uses Viterbi algorithm.

This algorithm is applied to a family of samples of the same text, so we obtain a family of E_i^* and \mathcal{V}_i . As mentioned previously, one decides reasonably to choose the model \mathcal{V}^* whose probability associated to a sample is maximum. This last model makes the sample the most representative, i.e. *we have a good observation in some sense*. This optimal model estimate is considered as *a type of the text processed*.

27.5 Test for first-order stationarity

As outlined by Anderson and Goodman [AG57] the following test can be used to determine whether the Markov chain is first-order stationary, or not. Thus, we have to test the null hypothesis (H) :

$$P_{i,j}(t) = P_{i,j} \quad (t = \{1, 2, \dots, T\})$$

The likelihood ratio with respect to the null and alternate hypothesis is :

$$\lambda = \prod_{t=1}^{t=T} \prod_{i=1}^{i=K} \prod_{j=1}^{j=K} \frac{P_{i,j}^{N_{i,j}(t)}}{P_{i,j}^{N_{i,j}(t)}(t)}$$

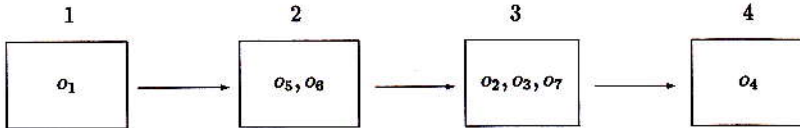
We now determine the confidence region of the test. In fact, the expression $-2 \log \lambda$ is distributed as a Chi-square distribution with $(T-1) \times K \times (K-1)$ degrees of freedom when the null hypothesis is true. As the distribution of the statistic $S = -2 \log \lambda$ is χ_2 , one can compute a β point ($\beta = 95, 99.95$ %, etc.) as the threshold S_β . The test is formulated as :

If $S < S_\beta$, the null hypothesis is accepted, i.e. the Markov chain is first-order stationary. Otherwise, the null hypothesis is rejected at $100\% - \beta$ level of significance, i.e. the chain is not a first order stationary and one decides in favour of the nonstationary model.

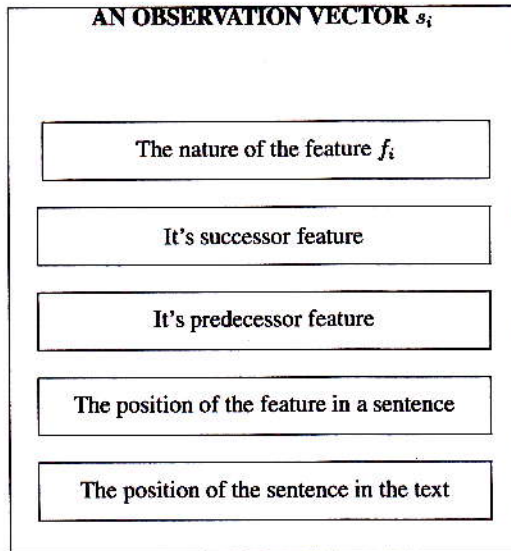
27.6 How to solve the morphological ambiguities

This is the most important phase of our application. Let's consider an example of nine possible paths encountered in a test. Among these paths, the system has to choose the most likely according to the probability measure (see the third figure). Our decision of choosing the most likely path comes from the optimal model \mathcal{V}^* obtained in the training phase. We show in this example how to remove the morphological ambiguities.

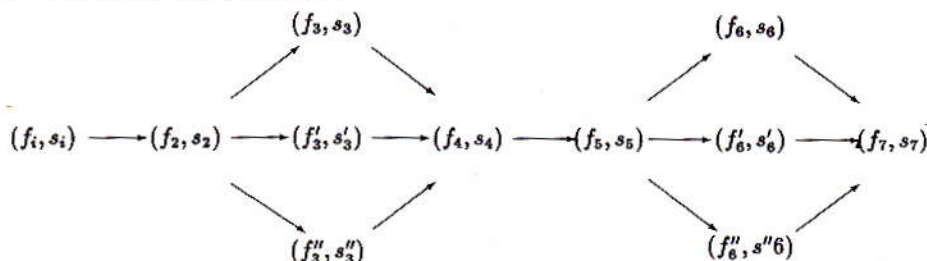
If the optimal state sequence obtained in the training phase is the one which corresponds to the figure :



K=4 : the optimal state sequence is 1, 3, 3, 4, 2, 2, 3



the one for example can choose between the two following paths of the figure :



Path 1 $s_1 \ s_2 \ s_3 \ s_4 \ s_5 \ s_6 \ s_7$

Path 2 $s_1 \ s_2 \ s'_3 \ s_4 \ s_5 \ s'_6 \ s_7$

One compute the probabilities of these two realizations of the observations o_i ($i = 1, 2, \dots, 7$) using the formula :

$$Prob(o_1, o_2, \dots, o_7 / \mathcal{V}^*) = \Pi_{e_1} v_{e_1}(o_1) \times \prod_{t=2}^{t=7} P_{e_{t-1}, e_t} v_{e_t}(o_t)$$

The first figure shows that each s_i belongs to a state e_i and, using the optimal model $\mathcal{V}^* = (\Pi, P, V)$ one can compute the probability of a path. Our decision to remove the morphological ambiguities is to choose the path with the highest probability.

27.7 Conclusion

We have presented a new approach for solving the morphological ambiguities using a hidden Markov model. This method may also be applied to other analysis levels as syntax. The main advantage of the method is the possibility to assign many different classes of criteria (fuzzy or completely known) to the training vectors and investigating many samples. Furthermore, we can define a "distance" between any sample and a family of type of texts called models. One can choose the model which gives the higher probability of this sample and conclude that the sample belongs to the specific type of texts. We can also develop a proximity measure between two models \mathcal{V}_1^* and \mathcal{V}_2^* through representative samples. However, some precautions must be taken in the choice of the distance used between the training vectors in the cluster process. In fact, the value of the probability associated to a sample may depend on this norm and, therefore, the choice of the best model estimated can be affected.

So far, we supposed that the criteria described the observations and the states are completely known (hard observation). Very often, when we want to make deep investigations, fuzziness or uncertainty due to some criteria or states are encountered, what should be done in this case? How can we cluster the observations according to some uncertainty measure? What is the optimal path and the best estimate model according to the uncertainty measure? We are working in order to propose solutions to those questions in the case of a probabilistic [BOU91, BOU] and fuzzy logic.

27.8 REFERENCES

[AG57] T.W. ANDERSON and L. A. GOODMAN. Statistical inference about markov chains.

Annals of Mathematical Statistics, 28, 1957.

- [BOU] D. BOUCHAFFRA. *A relation between isometrics and the relative consistency concept in probabilistic logic*. Elsevier, North Holland.
- [BOU91] D. BOUCHAFFRA. *A relation between isometrics and the relative consistency concept in probabilistic logic*. In *13th IMACS World Congress on Computational and Applied Mathematics*, juillet 1991.
- [BOU92] D. BOUCHAFFRA. *Echantillonnage multivarié de textes pour les processus de Markov et introduction au raisonnement incertain dans le Traitement Automatique de la Langue naturelle*. PhD thesis, Université des Sciences Sociales, novembre 1992.
- [ROU88] J. ROUAULT. *Linguistique automatique : applications documentaires*. P. Lang, 1988.