Minimal Sample Size in Grammatical Inference A Bootstrapping Approach *

Ana L. N. Fred and José M. N. Leitão

Instituto de Telecomunicações, Instituto Superior Técnico,
IST-Torre Norte, Av. Rovisco Pais, 1096 Lisboa Codex, Portugal
eanafred@beta.ist.utl.pt

Abstract. It is well known that the convergence of a grammatical inference method is strongly conditioned by the training data set. Structural completeness is a desired property seldom achieved in real data. The question that naturally arises in these types of problems is: how far is the training data to achieve structural completeness and what is the minimal sample size to use when there is no *a priori* knowledge about the structure of the data. In this paper we propose a simple methodology to give some insight into the later problem. It basically consists of a bootstrapping technique supported on grammars inferred from the existing data. An example of the application of this methodology in the context of automatic sleep analysis is used to illustrate the method.

1 Introduction

Syntactic methods have been applied in a variety of domains [1–5]. Several types of grammars have been used, the most common being finite-state grammars (FSGs) and context-free grammars (CFGs). The formalism of grammars is very powerful in the sense that not only it enables concise modeling of languages, being useful for data compression and providing simple mechanisms for language generation, as they also are suited for classification purposes. An important extension of formal grammars are stochastic grammars [6] where a probabilistic model of language generation is expressed in terms of rules of symbol composition with associated probabilities. This enables the modeling of noisy patterns and string repetition. Classification according to these types of grammars are usually based on Bayesian decision theory or (stochastic) nearest-neighbor techniques, when in the presence of noisy patterns.

The process of automatic definition of rules based on training data is designated as grammatical inference [7, 8]. It is well known that the convergence of a grammatical inference method is strongly conditioned by the training data set. Structural completeness is a desired property seldom achieved in real data. The question that naturally arises in these types of problems is: how far is the

* This work was partially supported by the project PECS/C/SAU/212/95 and PRAXIS 2/2.1/ TIT/1580/95.

training data to achieve structural completeness and what is the minimal sample size to use when there is no *a priori* knowledge about the structure of the data.

The problem of dimensionality has been studied for statistical pattern recognition [9] and rules of thumb have been proposed when no theoretical development has been derived. Concerning syntactic methods much effort has been put on the development and characterization of learning algorithms, the problem of sample size been usually left unspecified. In this paper we propose a simple methodology to give some insight into the later problem.

Section 2 presents notation and definitions used throughout the paper. A simple bootstrapping technique to estimate the dependency of the learning on the training data size is proposed in section 3. The application of this methodology in the context of automatic sleep analysis is summarized in section 4.

2 Notation and Definitions

Let G be a grammar defined by the quadruple (V_N, V_T, R, σ), where V_N, V_T, R and σ are, respectively, a finite set of nonterminals, a finite set of terminals or vocabulary, a finite set of productions and the start symbol. The language described by the grammar G, $L(G)$, is derived from the start symbol σ by successive application of rewriting rules in G. One defines *positive sample set* of a language $L(G)$ as $S^+ \subseteq L(G)$. A set S^+ is *structurally complete* if each rule of G is used in the generation of at least one string in S^+.

The process of automatic definition of productions based on a set of training sentences is called grammatical inference. Typical practical situations include the restriction of data to positive samples and furthermore the number of training patterns is limited. If the number of training sentences is low there is a considerable chance that the set is not structurally complete and the inferred grammar will not cover all the variability of the data. However, since the structure of the data is not usually known *a priori*, it is impossible to know what is the minimal dimension of the training set, assuming that samples are randomly acquired from the language source, to achieve adequate representativity of the language to be modeled. Additionally, the complexity of the structure underlying a language is not linearly related with the length of the sentences involved. How does the particular situation at hand fits with the dimensionality problem? The following section suggests a procedure to tackle this problem.

3 Estimating the Minimal Sample Size

The problem under analysis is the following: given a grammatical inference problem and a training data set, randomly acquired from the language source, one wishes to estimate the minimal number of training samples needed in order to cover the variability and complexity of the language to be modeled. This coverage property reflects in several performance parameters of the resulting syntactic pattern recognition system, namely the probability of error.

Naturally, an optimal solution to this problem is only possible if the exact structure of the language is known (or equivalently, the true grammar) or if one is able to acquire as many training data as desired. However, in most situations such knowledge is inexistent and the samples acquisition is costly or somehow restricted. We therefore assume that an initial finite length training data set is available and it is the only source of information about the language to be modeled. The basic idea to tackle this problem is to use the grammar inferred from this initial data as a model of the complexity of the language to be described. It is true that, if the number of patterns is too small then the true language will be more complex than the language associated with the inferred grammar. Additionally, the inclusion of new sequences should not decrease the complexity of a language; therefore the initial grammar can be seen as exhibiting a lower bound on the true complexity. By using the inferred grammar as a language generator one can replicate arbitrary length sets of sentences, all exhibiting the same structure, which enables the extrapolation of expected performances of the syntactic pattern system above and beyond the number of patterns used in the training. This initial grammar can thus be seen as a generator of bootstrap samples [10] from the training set.

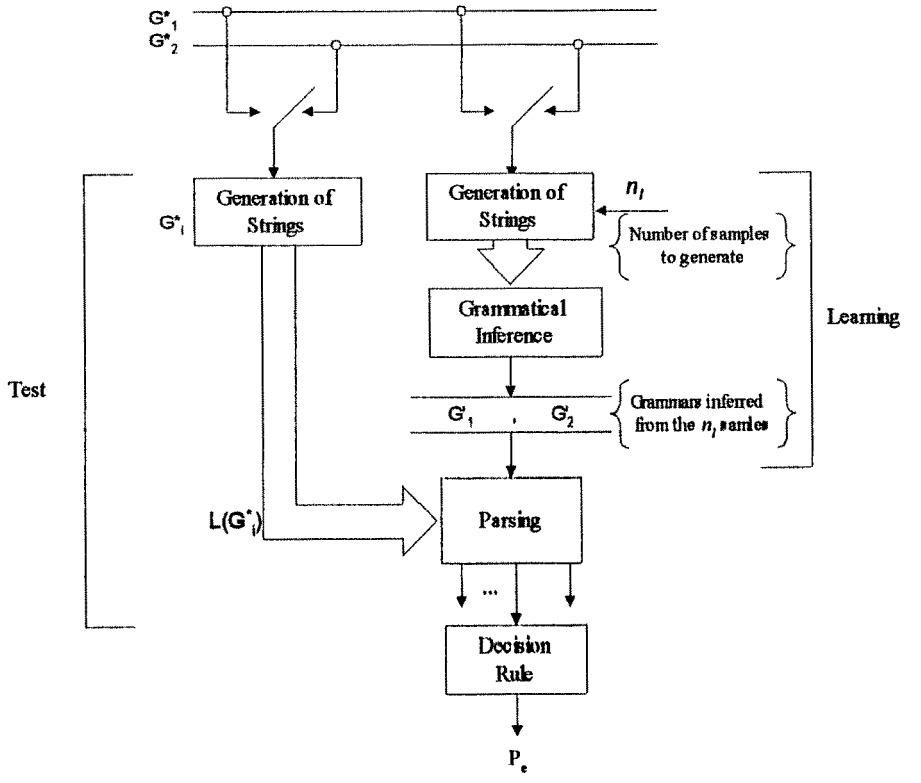

Fig. 1. Estimating the performance of a syntactic pattern recognition system as a function of the training sample size for a two class problem. G_i^* denotes the grammar inferred from the existing data from class i.

The basic idea is therefore the computation of curves of performance as a function of the size of the training set, the later being produced by using the inferred grammars as language generators. The method is schematically described in figure 1 for a two class problem. The grammars G_i^*, inferred using the real data, are used as language generators to produce bootstrap samples of length n_l. For each set, a grammar $G_i^{'}$ is inferred, which will converge to G_i^* for n_l sufficiently large. The grammars $G_i^{'}$ are applied in the recognition and classification of arbitrary samples independently generated from G_i^*. This process is repeated for several values of the n_l parameter, several bootstrap samples being issued for each value. At each step, parameters related to system performance are recorded. Typical variables to be evaluated are: the global probability of error; number of inferred rules; ratio of recognition by the corresponding grammars. The ratio between the average number of inferred rules for each dimension of the training data set and the number of rules in G_i^* gives an idea of the level of structural completeness achieved for that dimension, as far as the estimate G_i^* of the true grammar is concerned. Once again, it should be emphasized that, although G_i^* is not the true grammar, it serves as a model for the corresponding degree of grammatical complexity; as the true grammar should be at least that complex, the results obtained with the model grammar can be seen as estimates of the lower bounds of the true parameter values. The level of structural completeness determines the degree of recognition of sentences according to the true class grammars. This is represented by the last variable in the list above. The global error rate combines both non recognition and erroneous classifications due to the stochastic nature of the grammars.

The graphical representation of the above parameters as a function of the dimension of the bootstrap samples gives some insight into the degree of coverage of the variability of patterns. By selecting a threshold on the desired performance (for instance, the global probability of error) one obtains a lower bound on the dimension of the training data set to be used. The above methodology is illustrated in the following section.

4 Example of Application in the Context of Automatic Sleep Analysis

Sleep is a complex and extremely important dynamic process that takes about one third of human's life. Many disorders are associated with sleep or reflect on the sleep pattern. The global features of sleep dynamics are usually compressed into a sleep stages description (namely: wakefulness; Rapid Eyes Movements - REM - stage; stages 1, 2, 3 and 4NREM), according to the Rechschaffen and Kales (R&K) criteria [11], summarizing typical patterns in physiological variables. The hypnogram is the graphical representation of the evolution of sleep stages along the night. The purposes of automatic hypnogram analysis are: to infer the rules governing its temporal organization and condense them in a model able to reproduce the original structure; to classify hypnogram data according to its specific diagnostic population.

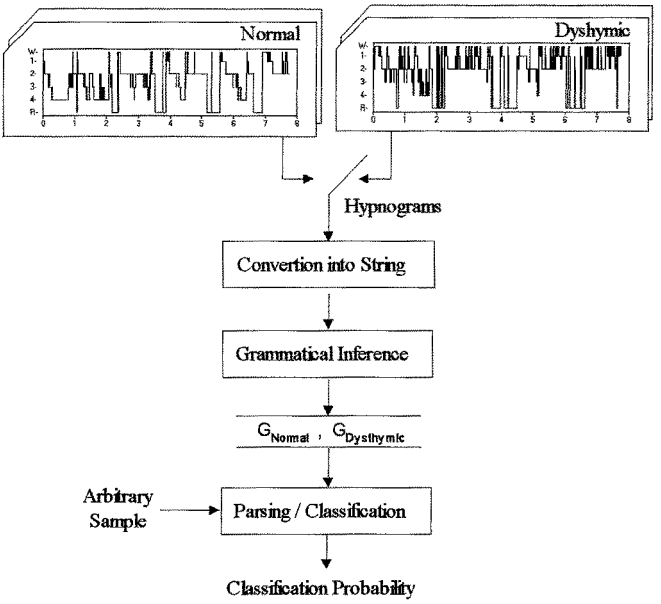

Fig. 2. Syntactic approach to automatic hypnogram analysis.

Figure 2 shows the methodology used. Hypnogram data is converted into string description by simple concatenation of the first symbol of the observed sleep stages. For each population a grammar is inferred using Crespi-reghizzi's method [7,8], no *a priori* information being used except for left-to-right dependency of the symbols. The method of stochastic presentation is used in the estimation of rules probabilities. The inferred grammars are then used for modeling and classification purposes.

The above methodology was applied in the study of a normal population (15 subjects, 26 hynograms) and a population of dysthymic patients (12 subjects, 17 samples) [12,13]. Figure 3 shows the evolution of the recognition rate as a function of the sample size, by using the grammars inferred from the existing data as approximations of the true grammars, while the percentage of uncovered rules is depicted in figure 4. The later parameter is closely related with the structural completeness property of the training data. Each point on the graphics results from averaging over 5 repetitions of the inference/recognition process; the recognition rate at each step is estimated over 100 test samples, randomly generated. It can be seen that the number of samples available for training is clearly insufficient for the order of complexity of the grammars involved. This means that, although high recognition rates can be achieved when resubstitution estimates are used, low recognition rates are to be expected due to no recognition when using estimates based, for instance, on the Jack-knifing method. This is

confirmed by the data where leave-one-out estimates lead to error rates above 50% and simple bootstrap estimates [14] based on 200 bootstrap samples lead to error rates between 24% (for the dysthymic population) and 33% (for the normal population). Examination of these curves suggests the necessity of gathering data sets with more than 40 elements per population in order to achieve recognition rates over 80%. By enlarging the existing data sets for the normal population it was possible to verify that the recognition rate increased over 75% for a sample set of 39 elements.

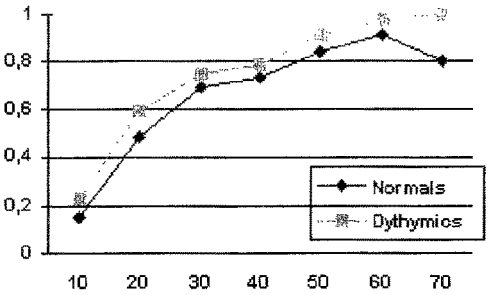

Fig. 3. Recognition rates for the populations normal and dysthymic as a function of the training sample size.

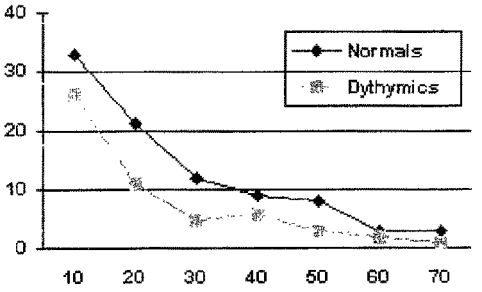

Fig. 4. Percentage of uncovered rules as a function of the training sample size. The number of uncovered rules is counted as the absolute difference between the number of inferred rules from the bootstrap samples (learning phase) and the number of rules of the grammar inferred from the existing real data.

The proposed approach was therefore useful in understanding the reason for the low error rates based on jack-knifing estimates and to give some hint on the minimum desirable size of the training data.

5 Conclusions

The relation between dimensionality and sample size plays an important role in various areas of pattern recognition. In statistical pattern recognition dimensionality concerns the number of features used in the description of patterns; the syntactic equivalent is grammar dimensionality, a concept usually closely related to grammar complexity.

Concerning the syntactic methods, much effort has been put on the development and characterization of learning algorithms, the problem of sample size being usually left unspecified. This paper has presented a simple methodology that gives some insight into this problem. It basically consists on the construction of curves of performance parameters as a function of the training sample size. Parameters such as the global error probability, the recognition ratio according to each class grammar, and the number of inferred rules, can provide some useful information about the degree of structural completeness of the training patterns and of the expected performance of the syntactic pattern recognition system as a function of the dimension of the training data sets. For the computation of these curves, bootstrap samples are generated, not by sampling with replacement from the existing data, but by language generation from the grammar inferred from this data. This grammar can thus be seen as a parameterization of the existing patterns, being a model of the complexity of the underlying structure. This bootstrapping technique enables the extrapolation of the results for dimensions of data sets bellow and above the size of the real data, by means of the generation of arbitrary length sentences, replicating the structure of the original patterns. Analysis of these curves provide some useful information for the understanding of the performance of syntactic pattern recognition systems, being a tool for the estimation of the minimal data set size needed in order to achieve a given performance.

The above methodology has been illustrated and corroborated with an example in automatic sleep analysis using a syntactic approach. In this case the analysis of the curves of recognition rate and of uncovered rules provided a valuable hint on the dimension of the training data sets to gather in this context.

References

1. Gonzalez, R. C., Thomason, M. G. : *Syntactic Pattern Recognition.* Addison-Wesley (1978) 119–142
2. Fu, K. S.: *Syntactic Pattern Recognition and Applications.* Prentice-Hall (1982)
3. Mohr, R., Pavlidis, T., Sanfeliu, A.: *Structural Pattern Analysis.* World Scientific (1990)
4. Bunke, H.: *Advances in Structural and Syntactic Pattern Recognition.* World Scientific (1992)
5. Bunke, H., Sanfeliu, A.: *Syntactic and Structural Pattern Recognition - Theory and Applications.* World Scientific (1990)
6. Thomason, M. G.: Syntactic Pattern Recognition: Stochastic Languages. *Handbook of Pattern Recognition and Image Processing.* Academic Press (1986) 119–142

7. Fu, K. S., Booth, T. L.: Grammatical Inference: Introduction and Survey - Part I and II. *IEEE Trans. Pattern Anal. And Machine Intelligence* 8 (3) (1986) 343–374
8. Miclet, L.: Grammatical Inference. *Syntactic and Structural Pattern Recognition - Theory and Applications.* Scientific Publishing (1990) 237–290
9. Raudys, S., Jain, A. K.: Small Sample Size Effects in Statistical Pattern Recognition. *IEEE Trans. Pattern Anal. And Machine Intelligence* 13 (3) (1991) 252–264
10. Efron, B., Tibshirani, R. J.: *An Introduction to the Bootstrap.* Chapman & Hall, (1993)
11. Rechtschaffen, A., kales, A.: *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects.* U.S. Government Printing Office. Washington DC (1968)
12. Fred, A. L. N.: *Structural Pattern Recognition: Applications in Automatic Sleep Analysis.* PhD Thesis, Technical University of Lisbon (1994)
13. Fred, A. L. N., Leitão, J. M. N.: Use of Stochastic Grammars for Hypnogram Analysis. *Proc. of the 11th IAPR Int'l Conference on Pattern Recognition* (1992) 242–245
14. Jain, A. K.: Bootstrap Techniques for Error Estimation. *IEEE Trans. on Pattern Analysis and machine Intelligence* 9 (5) (1987) 628–633