ELSEVIER

Journal of Neuroscience Methods 150 (2006) 105–110

JOURNAL OF NEUROSCIENCE METHODS

www.elsevier.com/locate/jneumeth

Automatic detection of interictal spikes using data mining models

Pablo Valenti [a], Enrique Cazamajou [a], Marcelo Scarpettini [a], Ariel Aizemberg [a], Walter Silva [b], Silvia Kochen [b,*]

[a] *Exact and Natural Sciences Faculty, Buenos Aires University (UBA), Argentina*
[b] *Epilepsy Center, Neurology Division, "Ramos Mejia" Hospital, Institute of Cellular Biology and Neurosciences "Prof Dr Eduardo de Robertis", Faculty of Medicine, Buenos Aires University, Rocamora 4122, 1184 Ciudad de Buenos Aires, Conicet-Cefybo, Argentina*

Received 20 October 2004; received in revised form 21 February 2005; accepted 8 June 2005

Abstract

A prospective candidate for epilepsy surgery is studied both the ictal and interictal spikes (IS) to determine the localization of the epileptogenic zone.

In this work, data mining (DM) classification techniques were utilized to build an automatic detection model. The selected DM algorithms are: Decision Trees (J4.8), and Statistical Bayesian Classifier (naïve model). The main objective was the detection of IS, isolating them from the EEG's base activity. On the other hand, DM has an attractive advantage in such applications, in that the recognition of epileptic discharges does not need a clear definition of spike morphology. Furthermore, previously 'unseen' patterns could be recognized by the DM with proper 'training'.

The results obtained showed that the efficacy of the selected DM algorithms is comparable to the current visual analysis used by the experts. Moreover, DM is faster than the time required for the visual analysis of the EEG. So this tool can assist the experts by facilitating the analysis of a patient's information, and reducing the time and effort required in the process.

Keywords: Data mining models; Detection interictal spikes; Epilepsy; Epilepsy surgery

1. Introduction

In addition to the characteristic electrographic bursts of abnormal activity that are recorded when epileptic patients experience a seizure (ictal episode), the electroencephalogram (EEG) of epileptics will normally display isolated sharp transients or "spikes" in some locations of the brain. These interictal spikes (IS) are a complementary source of information in the diagnosis and localization of epilepsy.

In particular, when a prospective candidate for epilepsy surgery is studied with long-term video/EEG monitoring, both the ictal (electrographic seizures) and interictal (spikes) manifestations of epilepsy are analyzed to determine the localization of epileptogenic zone. The automatic or semiautomatic method for IS detection has been required for several decades for improving the visually analyzing large amounts of data produces fatigue and error.

Numerous attempts have been made to define a reliable spike detection mechanism. However, all of them have faced the lack of a specific characterization of the events to detect. Some approaches have concentrated in measuring the "sharpness" of the EEG signal (Carrie, 1972), which can be expected to soar in the "pointy peak" of a spike. Walker (1996) attempted the detection of spikes through analog computation of the second time derivative (sharpness) of the EEG signals. Smith (1974) attempted a similar form of detection on the digitized EEG signal. This method however, required a minimum duration of the sharp transient to qualify it as a spike. Although these methods involve the duration of the

* Corresponding author. Tel.: +54 11 48633086; fax: +54 11 48633086.
E-mail address: skochen@mail.retina.ar (S. Kochen).

0165-0270/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jneumeth.2005.06.005

transient in a secondary way, they fundamentally consider "sharpness" as a point property, dependent only on the very immediate context of the time of analysis.

The promise shown by that approach has encouraged us to use different data mining (DM) techniques. The DM computational models (Flexer, 2000; Mitchell, 1997) are new information processing techniques that extract previously unknown and potentially useful information from large data series. The philosophy behind these techniques resides on the discovery of global relations and patterns that exist in large databases, but are hidden at first analysis due to actually the huge stored data.

In this paper, the patterns to be discovered are the IS, isolating them from EEG's base activity (Badier and Chauvel, 1995; Bourien and Bellanger, 2004).

2. Methods and materials

Brain electrical activity from intracranial electrodes corresponding to six files (684 spikes) was recorded. It corresponds to three patients, candidates to surgery. The patients were evaluated with a stereo-electroencephalography technique (Talairach et al., 1974). Depth electrodes with 5–15 contacts were implanted in different zones of the brain in relation to the hypothesis of epileptogenic zone of each patients (Chauvel et al., 1987). The EEG signals were recorded using 200 Hz as sample frequency. Expert neurologists analyzed these records in visual form using BioScience Electroencephalography®, with EEG Harmonie Stellate software®.

The signal analysis was approached in three phases, as follows:

a. First, EEG experts neurologists (E1, E2) performed a visual analysis of the EEG signals, identifying and marked the IS, on each one of the input files. These input files, with estimated duration of 20 min, have an average of 241 spikes each one (Fig. 1).
b. Next, we performed the pre-processing of the input signal, with the objective of generating the detection models. The EEG signals were analyzed through the FFT (Dietsch, 1932). Trying to obtain time localization to the properties of the FFT, we added a sliding window to the Fourier formula.

   Based on that, since an IS has an estimated duration of between 20 and 200 (spike 20–70 ms, or sharp wave 70–200 ms (IFSECN, 1974) 200 samples/s × 0.2 s = 40 samples, a sliding window of 64 samples proved to be enough to contain an entire IS signal. It was determined that a shift of eight samples between windows minimizes the quantity of required movements, and maximizes the percentage of coverage of the signal. The FFT assumes a periodic signal in the input, and since we cannot guarantee the periodicity of the sampled signal contained in each sequence of the sliding window, it was necessary to apply windowing techniques to attenuate the borders

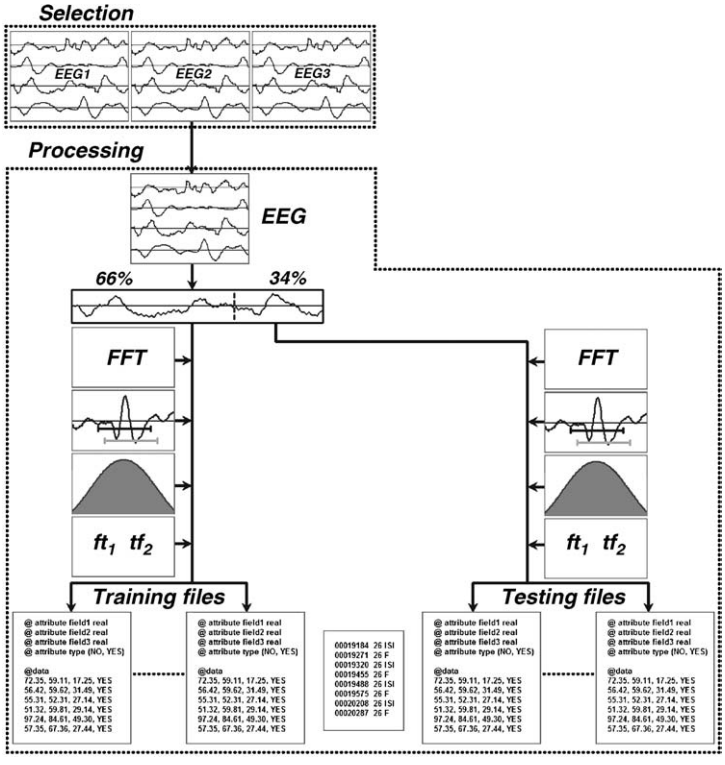


Fig. 1.

of the signal in the window, so as to emulate a periodic signal. Basically this means that the original signal contained in the window is multiplied by a cosinoidal function, which tend to zero on the borders of the window period. Two types of windowing techniques were used: Hanning and Rectangular. The formula used for Hanning is $\mathrm{Hng}(f(n)) = \left(0.5 - 0.5 \times \cos\left(\frac{2\pi n}{T-1}\right)\right) f(n)$, which becomes zero on the borders when "$n = T - 1$", being T the period of the window, and n having values from 0 to $T - 1$. The Rectangular windowing technique is simply leaving the input signal as it is, so the function used was $P(f(n)) = f(n)$.

Finally, two transforming functions (tf_1 and tf_2) were utilized to maximize the inherent characteristics of the IS: $tf_1 = \mathrm{sqrt}(R^2 + I^2)$, $a\tan(I/R)$ and $tf_2 = R^2$, I^2, where R and I are the real and imaginary parts, respectively of the FFT (Fig. 1).

c. The third phase consisted in the generation of the detection models, combining the data mining (DM) classification algorithms with the pre-processing techniques described above. The framework used to build the DM models was an open source package called Weka (Witten and Frank, 2000). And the DM classification algorithms selected for this approach were two, one based on Decision Trees (J4.8), and the second based on Statistical Bayesian Classifier (naïve model):

J4.8: The J4.8 decision tree is an implementation of the C4.5 method (Quinlan and Ross, 1993). A decision tree assigns a class (or output) to a set of input attributes (the instance), filtering those attributes through the decisions (tests) of the tree, going from the root to the leaf nodes. The results obtained at each decision are mutually exclusive and exhaustive (Aha and Breslow, 1997; Breiman et al., 1984).

Naïve model: The Naïve Bayes algorithm implements the statistical Bayes classifier. It uses estimate classes, whose precision values are chosen according to the analysis of training data. Instead of considering a simple normal distribution to model the numerical attributes, an estimation of the kernels was used (Domingos and Pazzani, 1997; Kohavi et al., 1997; Pazzani and Michael, 1997) (Fig. 2).

The combination of the selected DM algorithms (Naïve Bayes and J4.8) with the pre-processing techniques (tf_1 and tf_2, Hanning and Rectangular) led to the generation of eight detection models. Each one of these models applied to a given EEG generates a set of initial detections. In order to provide a way of fine-tuning the considered positive detections from the initial detection set, a precision measure was introduced, with a numerical range from 1 to 8. So, each one of the eight generated set of initial detections combined with the precision measure produced a total of 64 final detection sets (DMT).

This precision "tuning" measure consists in considering a well detected IS, if the detection model provides a positive value to "n" consecutive windows ($1 \leq n \leq 8$). As a shift of eight samples was used in the generation of each file with their detection model, an IS can be shifted to a maximum of eight samples within a given window, that is 4 ms. Thresholds near 1 assigned to the precision measure make the detection very sensitive, and may overgenerate marks. While values near 8 make the detection more accurate but they may lead to an under-detection of IS.

d. The last phase of the process consisted in the validation of the detection models, using new input EEGs that were not used in the training stage. This test approach permitted to evaluate the performance of the models, by simulating a real work scenario were the existence and quantity of IS is unknown. The idea was to compare the result of the IS detection by the DM models, with the visual analysis performed by the experts. E1 and E2 were asked to independently detect all IS in an EEG channel of a patient that had not been used when generating the models. Then, a third expert that had not being involved in any

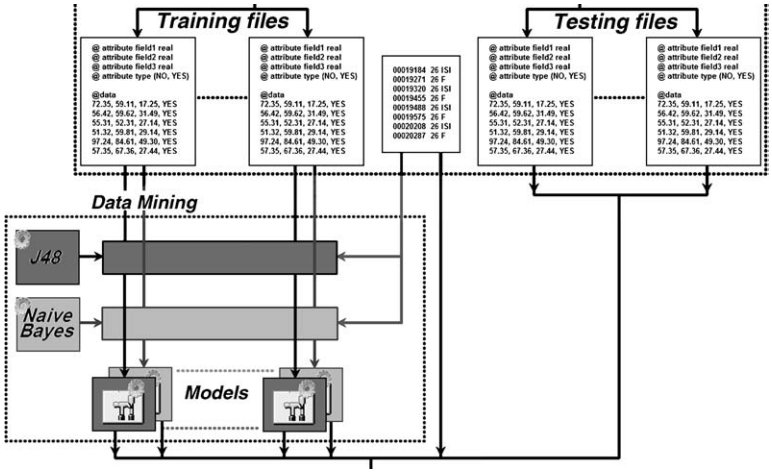


Fig. 2.

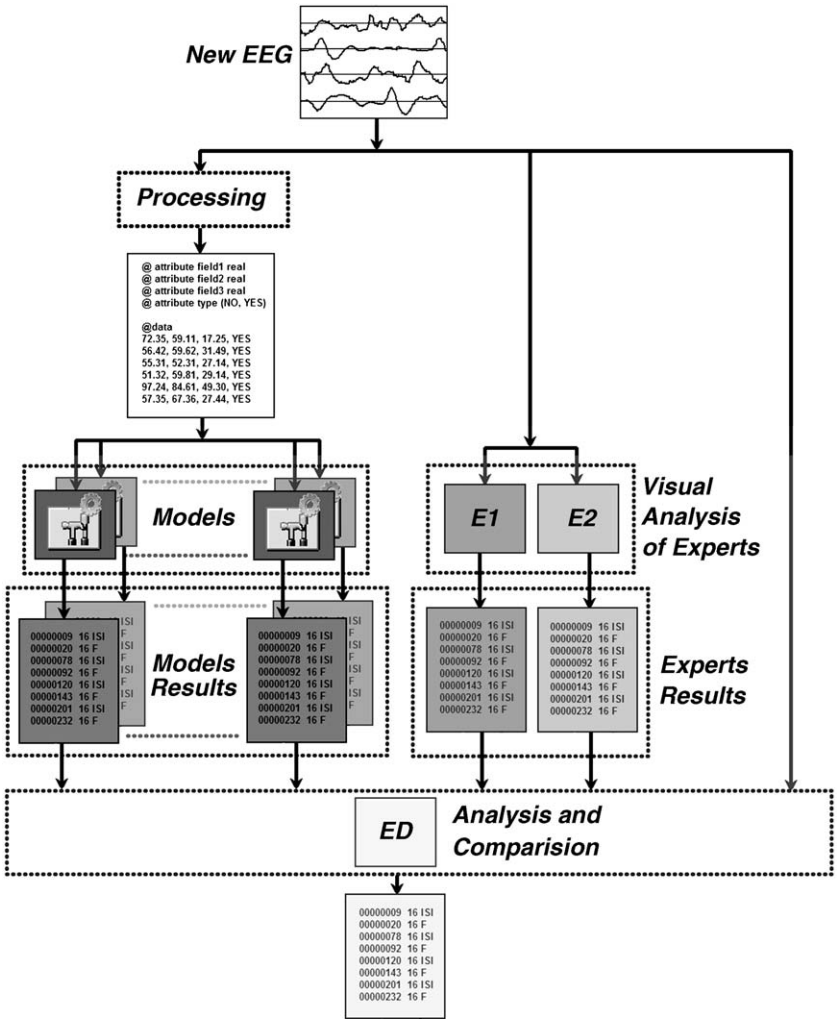


Fig. 3.

of the previous detections (ED), analyzed and compared the positive detections generated by DM, and the positive detections by E1 and E2 (Fig. 3). Consolidating and validating each one of the detections generated by DM, E1 and E2, the expert ED generated a new file containing a set of positive detections, which was considered as the "reference" (R), and was used as the basis for the statistical analysis. On this file, the percentage of detected "matches" and "errors" were calculated, being a "match" (TP, true positive) an IS detected by the DM model and validated by the ED, while an "error" (FP, false positive) is an IS detected by the DM model that was rejected by the ED.

In order to perform the statistical analysis, each one of the 64 DMTs and the two visual detections performed by the experts, E1 and E2, were associated to an ordered pair representing the values of "matches" and "errors" (TP, FP) in the detection. In particular for the R file, these values were (1, 0), since it was used as the basis for the analysis.

3. Results

Fig. 4 shows all the DMTs plotted based on their respective coordinates (TP over the *X*-axis, FP over the *Y*-axis), along with values for E1 and E2. To be able to compare the points plotted in the chart, a "distance" measure to the *R*-coordinates (1, 0) was required. Using the Euclidean distance formula $\sqrt{(1-\mathrm{TP})^2+\mathrm{FP}^2}$, it can be observed that more than 15% of the DMTs are located closer to the R than E2, and almost 50% of the total DMTs are better located than E1. Also, almost 40% of the DMTs detected more TPs (got better performance on TP) than E1 and E2. And 75% of the DMTs got lower FP than E1, and almost 40% were below of E2 FP.

Fig. 5 shows the IS detected in R with the matching detections from E1, E2 and the top-10 ranked DMTs based on their distance to R. The table does not intend to reflect the temporal relation between the IS in the EEG, but to show the 88 detected IS in a sequential order. It is possible observer the top-10 ranked DMTs based on their distance to R, we can see there is no clear winner in terms of DM algorithm, transform function or windowing technique, although

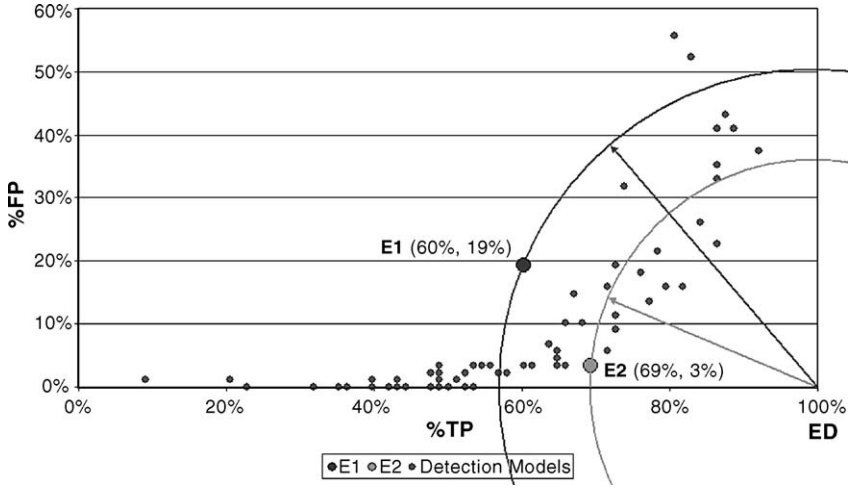
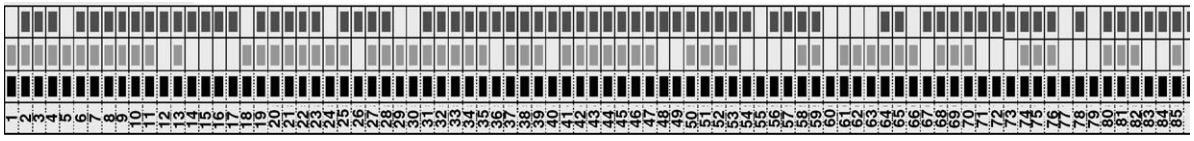


Fig. 4. Chart comparing the performance of the marks generated by E1 and E2, and the positive detections generated by the models. TP, true positives; FP, false positives; ED, expert.



	%VP	%FP	Dist.
NaiveBayes	84%	26%	26.93
E2	69%	3%	27.17
ED	00.%	0%	0.00

Fig. 5. Chart showing the coverage of the marks generated by NaïveBayes and E2, over the "real marks" set generated by ED.

the combination of Naïve Bayes, Hanning and $tf_2 = R^2, I^2$ obtained the shorter distance to R. Furthermore, we can see that within these top-10 DMTs, the DM algorithm Naïve Bayes was used by 6 out of 10 DMTs, and the windowing technique Hanning was used also by 6 out of 10. In terms of transform function, both TF1 and TF2 were used equally (5 and 5) in the top-10 list.

Fig. 2 shows one of the decision tree generated with the DM algorithm J4.8.

4. Conclusions

In this paper we have applied the data mining classification algorithms, Decision Trees (J4.8), and Statistical Bayesian Classifier (naïve model). The DM approach proved to be very useful in the recognition of epileptic discharges.

The obtained results showed that the efficacy of the used DM algorithms is comparable to the current visual analysis used by the experts in their daily work of detecting IS in EEGs. DM offers an advantage in comparison with the classical methods: nonlinear modeling method (Diambra, 2002), wavelets and time–frequency approaches (Senhadji and Wendling, 2002), artificial neural network (Ko and Chung, 2000). And it is possible to modify the characteristics of the algorithms based on the user needs, a DM model can be used that maximizes the percentage of TP, taking the risk of over-detection thus generating more quantity of FP. On the other side, it may be desired to use a model that generate detection of IS under the optimum TP rate, but minimizing the percentage of FP. In this case, the physician will take the risk of "loosing" some IS, but will avoid the time and effort of analyzing the entire EEG in looking for FP.

Another advantage of the methods used in this work, is that the recognition of epileptic discharges does not need a clear definition of spike morphology or the duration, which is certainly necessary in rule-based detection algorithms (Carrie, 1972; Goelz et al., 2000; Kochen et al., 2002; Mitchell, 1997; Smith, 1974). Furthermore, previously 'unseen' patterns could be recognized by the DM with proper 'training'.

On the other hand, the benefit of these computational models is their execution speed, which is enormously faster than the time required for the visual analysis of the EEG. So this tool can assist the experts by facilitating the analysis of patient information, and reducing the time and effort required in the process.

In summary, the above procedure is promising and likely to be useful to the physician as a more sensitive, automated and objective method to help in the localization of the interictal

spike zone of intractable partial seizures. The final output can be visually verified by neurologists. Due to the clinical relevance and demonstrated promise of this method, the consequent implications are on the possible extension to online recognition. Further investigation of this approach is warranted.

References

Aha WD, Breslow LA. Simplifying decision trees: a survey, navy center for applied research in artificial intelligence. Washington, DC: Naval Research Laboratory; 1997.

Badier JM, Chauvel P. Spatio-temporal characteristics of paroxysmal interictal events in human temporal lobe epilepsy. J Physiol Paris 1995;89(4–6):255–64.

Bourien J, Bellanger JJ, Bartolomei F, Chauvel P, Wendling F. Mining reproducible activation patterns in epileptic intracerebral EEG signals: application to interictal activity. IEEE Trans Biomed Eng 2004;51(2).

Breiman, et al. Classification and regression trees. CRC Press LLC; 1984, ISBN 0412048418.

Carrie JR. A hybrid computer technique for detecting sharp EEG transients. Electroencephalogr Clin Neurophysiol 1972;33(3):336–8.

Chauvel P, Buser P, Badier JM, Liegeois-Chauvel C, Marquis P, Bancaud J. The "epileptogenic zone" in humans: representation of intercritical events by spatio-temporal maps. Rev Neurol 1987;143: 443–50.

Diambra L. Detecting epileptic spikes. Epilepsia 2002;43(Suppl. 5): 194–5.

Dietsch G. Fourier Analyse von Elektrenkephalogrammen des Menschen. Pflüger's Arch Ges Physiol 1932;230:106–12.

Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Boston: Department of Information and Computer Science, University of California, ©Kluwer Academic Publishers; 1997.

Flexer A. Data mining and electroencephalography. Stat Meth Med Res 2000;9:395–413.

Goelz H, Jones RD, Bones PJ. Wavelet analysis of transient biomedical signals and its application to detection of epileptiform activity in the EEG. Clin Electroencephalogr 2000;31(4):181–91.

IFSECN, International Federation of Societies for Electroencephalography and Clinical Neurophysiology. A glossary of terms most commonly used by clinical electroencephalographers. Electroencephalogr Clin Neurophysiol 1974; 37:538–53.

Kohavi R, Becker B, Sommerfield D. Improving simple bayes, data mining and visualization group. Silicon Graphics Inc 1997.

Ko CW, Chung HW. Automatic spike detection via an artificial neural network using raw EEG data: effects of data preparation and implications in the limitations of online recognition. Clin Neurophysiol 2000;111(3):477–81.

Kochen S, Giagante B, D'Atellis C, Sirne R, Roitman J. Wavelet analysis preceding seizures. Adv Clin Neurophysiol 2002; 54(Suppl):Cap 67, Elsevier Science B.V.

Mitchell T. Machine learning. The McGraw-Hill Companies, Inc; 1997.

Pazzani, Michael J. Searching for dependencies in Bayesian classifiers. Department of Information and Computer Science, University of California; 1997.

Quinlan, Ross J. C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann Publishers; 1993.

Senhadji L, Wendling F. Epileptic transient detection: wavelets and time–frequency approaches. Neurophysiol Clin Jun 2002;32(3):175–92.

Smith J. Automatic analysis and detection of EEG spikes. IEEE Trans Biomed Eng 1974;BME-21:1–7.

Talairach J, Bancaud J, Szicla G, Bonis A, Geier S. Approche nouvelle de la chirugie de l'epilepsie: methodologie stereotaxique et resultats therapeutiques. Neurochirugie 1974;20(1):1–240.

Walker J. Fast Fourier Transforms (studies in advanced mathematics), CRC Press, (ISBN 0-849-37163-5), 1996.

Witten LH, Frank E. Data mining: practical machine learning tools and techniques with java implementations: San Francisco; Morgan Kaufmann Publishers, 2000.