World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:6, No:4, 2012

An Overview of the Application of Fuzzy Inference System for the Automation of Breast Cancer Grading with Spectral Data

Shabbar Naqvi, Jonathan M. Garibaldi

***Abstract*—**Breast cancer is one of the most frequent occurring cancers in women throughout the world including U.K. The grading of this cancer plays a vital role in the prognosis of the disease. In this paper we present an overview of the use of advanced computational method of fuzzy inference system as a tool for the automation of breast cancer grading. A new spectral data set obtained from Fourier Transform Infrared Spectroscopy (FTIR) of cancer patients has been used for this study. The future work outlines the potential areas of fuzzy systems that can be used for the automation of breast cancer grading.

***Keywords*—**Breast cancer, FTIR, fuzzy inference system, principal component analysis

I. Introduction

BREAST cancer is a disease that is one of the most common causes of cancer deaths in women throughout the world including U.K [1]. Current cancer diagnostic procedures involve inspection of sample under a microscope by pathologist. This procedure is time consuming as well as it involves more chances of errors resulting in incorrect conclusions [2]. Therefore, there is need to develop new automated methods that provide accurate results helping pathologists in real clinical practice.

One of the relatively new approaches is the use of spectral data sets to be used in cancer diagnosis and prognosis. FTIR is one of the techniques that have been frequently used for various cancer diagnosis [3]. In FTIR, infrared radiation is passed through the sample and resulting spectra is signature of that sample different from spectra of other samples [3]. In our previous papers, we have shown the potential of FTIR for the automation of breast cancer grading with unsupervised computational method of clustering on the basis of widely accepted criteria of Nottingham Grading System (NGS) and have also discussed the complications involved in the analysis [4, 5].

In this paper we present an overview on application of supervised learning method of fuzzy systems on FTIR spectral data set to be used for the automation of breast cancer grading. A new high dimensional spectral data set has been used for the experiments. There is no significant reported work on automation of breast cancer grading using NGS criteria with the help of fuzzy inference system with FTIR spectral data sets.

Shabbar Naqvi is a PhD student in the Intelligent Modelling and Analysis (IMA) research group, School of Computer Science, University of Nottingham (phone +44-115-9514299 e-mail: szn@cs.nott.ac.uk)

Jonathan M. Garibaldi is an Associate Professor in the Intelligent Modelling and Analysis (IMA) research group, School of Computer Science, University of Nottingham (phone +44-115-9514216 e-mail: jmg@cs.nott.ac.uk)

The rest of this paper describes the background of the fuzzy inference system (FIS) alongwith previous work done on spectral data sets with fuzzy systems, detailed description of the data set alongwith pre-processing, development of FIS, our initial experiments, discussion on results and the future work describing potential areas that can be further investigated for development of a novel method.

II. Background

A. Breast Cancer Grading

The NGS is widely used around the world for grading the breast cancer tumours and is also a component of the Nottingham Prognostic Index (NPI). It is based on clinical parameters of mitotic count, tubule formation and nuclear pleomorphism. These parameters are calculated on the basis of the visual inspection of tumour under microscope by expert pathologists. The grades are ranked from 1 to 3. Grade 1 indicates less aggressive tumour with more chances of survival, grade 2 indicates medium aggressive tumour and moderate chances of survival and grade 3 indicates highly aggressive tumour and less chances of survival. The correct identification of grade is very important as it plays a vital role in deciding about the future treatment and prognosis of disease [6, 7].

B. Fuzzy Inference System (FIS)

Fuzzy inference system (FIS) is also called Fuzzy Expert System (FES) and Fuzzy Logic Controller (FLC) depending upon the area of its application. It is a rule-based system that uses fuzzy logic, rather than Boolean logic, to reason about data.

Its basic structure has four main components

i.A Fuzzifier which translates real - valued inputs into fuzzy values.

ii.An inference engine that applies a fuzzy reasoning mechanism to obtain a fuzzy output.

iii.Defuzzifier, Which translates this latter into a crisp value.

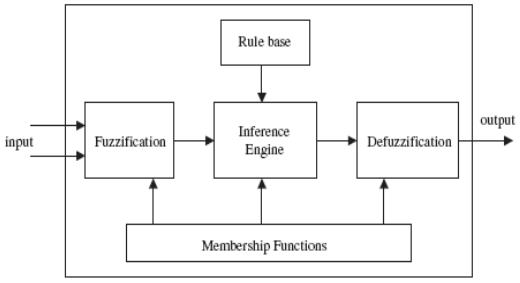

Fig. 1 Block diagram of FIS [9]

iv.Knowledge base, which contains fuzzy rules

The two most common fuzzy inference methods are Mamdani's fuzzy inference (Mamdani) method and Takagi-Sugeno (Sugeno) [8]. A membership function contains the value between 0 and 1 for a given input. A basic structure of FIS has been shown in Fig.1.

C. Work Done on Spectral Data with FIS

Fuzzy systems have been used in literature with spectral data sets on various application areas. Castanys et al [10] have described a three phase case-based reasoning system(CBR) to identify unknown materials by means of the automatic recognition of their Raman spectra. The first phase consists of dimensionality reduction by means of PCA. The second phase consists of defining similarity measures to objectively quantify the spectral similarity with a final value obtained by the fuzzy logic system. The final phase consists of revision and validation of the results. The total number of rules developed for the fuzzy logic system was 4. Evsukoff et al [11] have presented a frame work for intelligent data analysis of spectral data in classification and regression problems. In this frame work, the number of interpolation function is computed using spectral analysis. Each function is then associated with a symbol to generate fuzzy rule. Each symbol is related with a prototype that can be computed using a clustering algorithm. A rule induction algorithm determines the minimum number of rules for the frame work. Cernuda et al [12] suggested the use of specific fuzzy system called Takagi-Sugeno fuzzy system for calibrating the chemometric models. This fuzzy system was used to model the non-linearity contained in the production process of polytheracrylat (PEA). Mahmoodabadi et al [13] have presented a fully automated system in order to analyse and classify magnetic resonance spectroscopy (MRS) signals of patients with metabolic brain diseases. The authors proposed novel fuzzy membership functions that consider the human subjectivity in decision making by using a fuzzy classifier to categorize the metabolic brain diseases. Similarly Zhengmao Ye [14], Pueyo et al [15]and Kong et al [16] have also developed fuzzy systems for spectral data sets. The literature review reveals that use of fuzzy systems has produced good results in these areas and it is worth applying this method for the automation of breast cancer grading.

III. Data set

The data set used for the current work is called BR804. It was obtained with the collaboration from Prof. Rohit Bhargava from University of Illinois at Urbana Champaign, USA [17]. It consists of 80 cores of 40 cases of paired breast invasive cancer and matched normal adjacent tissue with single core per class. Fig. 2 shows the Tissue Microarray (TMA) slide of the data set and Fig. 3 explains the pair wise categorization of the cores. It consists of 10 x 8 Matrix with dark blue shades cores indicating malignant tumours and light ones indicates the corresponding normal cells that are mentioned as NAT (Not a Tumour).

The cancer grades break down of the data set is described in Table I. These grades were calculated by expert histopathologists with NGS criteria. Table I shows that there are 6 cases whose grade could not be determined by the normal histopathological procedure and are undefined.

TABLE I
BREAK DOWN OF CASES IN TERMS OF GRADE

Grade I (G1)	Grade 2 (G2)	Grade 3 (G3)	Undefined
2	6	26	6

The FTIR data set for this TMA slide is of size 85.4GB. It has been calculated between (722 – 4000) cm-1 for every alternative wave number. It consists of three parameters.

X: Samples =4062

Y: Bands = 3420

For each (X, Y) pair there is an absorbance ‘Z’ that consists of 1641 wave numbers. Each point is of type float (4 bytes).Therefore, the total data size is calculated as.

Total data set size = X xYxZx4 = 85.4GB.

It indicates that data set is of extremely large size. As FTIR spectra had been calculated for the whole TMA slide, therefore, it results in such a high data size.

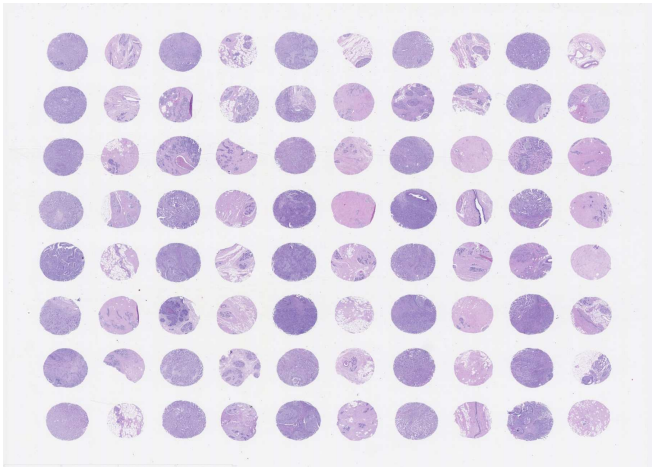

Fig. 2 TMA slide of data set

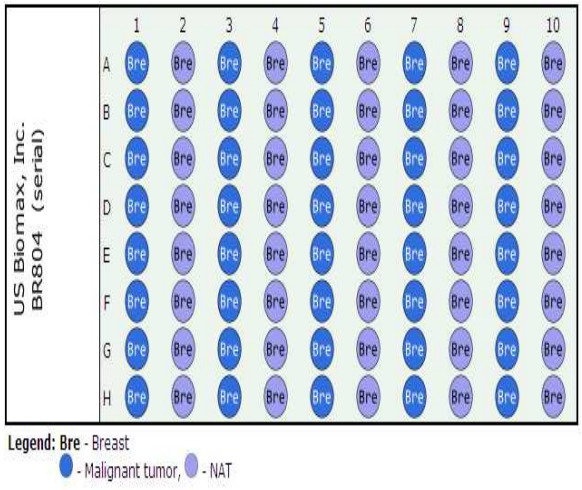

Fig. 3 Pair wise categorization of cores

IV. METHODS

A. Preprocessing

As it is very difficult as well as computationally expansive to process the whole data set, therefore, for our initial experiments ten areas for each grade were selected from the data set. Each area consisted of 10x10 (X, Y) square. In this way 100 spectra for each square were extracted. These squares were selected nearly from centre of the core as it is expected to contain the maximum absorbance in the core and their FTIR spectra were used. An example of such a square has been presented in Fig. 4. 1000 spectra for each grade were obtained in this manner. The total data size was 3000 x1641 absorbance values. All the spectra were base line corrected and normalised using a script written in Matlab ®. The mean spectra before and after pre-processing is shown in Fig. 5 and 6 respectively. It can be seen that after preprocessing spectra are generally more closed to each other removing abnormalities in spectra.

B. Dimension Reduction

It is very beneficial to reduce the dimensions of the data as it makes it computationally efficient. In literature, principal component analysis (PCA) has been widely used for the reduction of dimensions keeping most of the important information inherited in the data intact. A few principal components (PCs) carry most of the data variance thus substantially reducing the size of the data set [18]. We have used first 3 principal components (PCs) for our work as they contain 63.2% of the variance of the data set. The final data set selected was of size 3000 x 10 PCs.

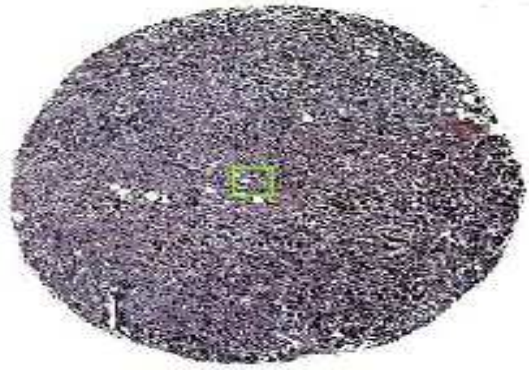

Fig. 4 The dark green square indicates the approximate position of 10x10 square of a core

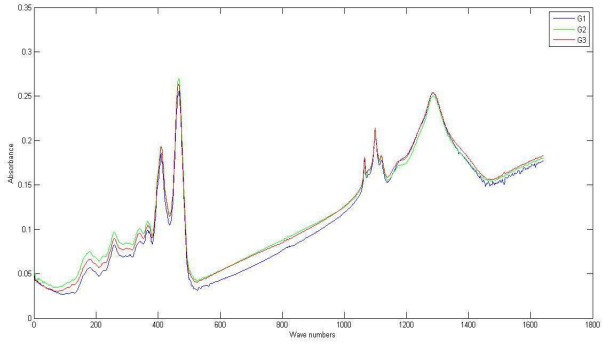

Fig. 5 Mean spectra before pre-processing

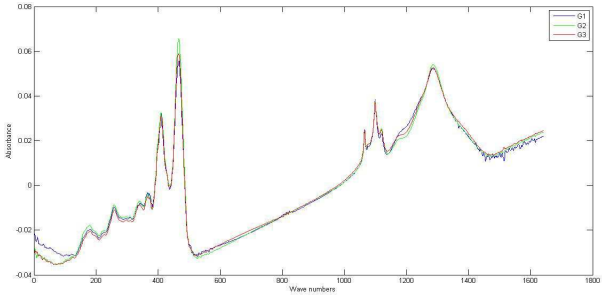

Fig. 6 Mean spectra after pre-processing

C. Development of FIS

The data set of 3000 x3 PCs was used for our experiments. All experiments were carried out using Matlab ® using fuzzy logic tool box. We have used Sugeno type fuzzy system for our experiments with a single output. 50% of the data set was used for the training purpose. In this way the total training data was of size 1500 x3 PCs. The rest of the 50% data was used for testing the system. During the training, output of the system was the classified value of grade as 1, 2 or 3. Subclustering function of Matlab ® was used to create three clusters each representing one grade. Three membership functions (mfs) were defined for each input PC belonging to three clusters. These membership functions were simply chosen by trial and error method. Membership function used was of triangular type. This selection was made as it is one of the most common types of membership function used in the literature. Membership functions for each principle component have been shown in Fig. 7-9 respectively. Initially, 27 rules were created for the system showing all possible combinations of the membership functions with clustering result. Table II shows these rules.

These rules were joined by the fuzzy AND operator resulting in a single output as grade. Majority vote was selected as criteria for the final output as shown in Table II. Out of 27, 6 rules were rejected as they did not provide any meaningful result because they assigned each membership function to a different cluster and majority vote was not able to allocate any grade in such case. Therefore, total number of rules finally selected was 21.

TABLE II
FUZZY RULE SET FOR FIS

Rule	mf1 cluster	mf2 cluster	mf3 cluster	Output cluster
1	1	1	1	1
2	1	1	2	1
3	1	1	3	1
4	1	2	1	1
5	1	2	2	2
6	1	2	3	Rejected
7	1	3	1	1
8	1	3	2	Rejected
9	1	3	3	3
10	2	1	1	1
11	2	1	2	2
12	2	1	3	Rejected
13	2	2	1	2
14	2	2	2	2
15	2	2	3	2
16	2	3	1	Rejected
17	2	3	2	2
18	2	3	3	3
19	3	1	1	1
20	3	1	2	Rejected
21	3	1	3	3
22	3	2	1	Rejected
23	3	2	2	2
24	3	2	3	3
25	3	3	1	3
26	3	3	2	3
27	3	3	3	3

V. RESULTS

Results on testing data have been shown in Table III.

TABLE III
TESTING DATA RESULTS

Grade	Number of spectra	Correct percentage
Grade 1	500	65.4%
Grade 2	500	53.4%
Grade 3	500	64%

It shows that the FIS did not perform well on the data. Only 53.4% data of grade 2 was correctly classified. Though grade 1 and 3 percentage was higher (65.4% and 64% respectively) but still it was not statistically significant. These results indicate that there needs to be more work done in order to get appreciable results out of the FIS. The discussion section highlights some of the areas that need to be addressed for the future work.

VI. DISCUSSION

Our system did not perform well on the data set. It may be because membership functions did not perform well and we need to try various membership functions in future to find the best suited membership functions. Also increasing the training data may result in better performance.

TABLE IV
SPECTRA WISE BREAK DOWN

Grade	Grade 1	Grade 2	Grade 3
Grade 1	327	100	73
Grade 2	34	267	199
Grade 3	40	140	320

Table IV shows spectra wise breakdown of data. It can be seen that in case of grade 1, only 327 spectra were identified correctly. Rest of the spectra were misclassified as either of grade 2 (100) or grade 3 (73). In case of grade 2 misclassified spectra were largely categorized as grade 3 (199) and for grade 3 misclassified spectra mainly belonged to grade 2 (140). The incorrect classification indicates that there is more uncertainty involved in spectra and more complex methods are required to find the correct classification.

Another reason may be use of the first 3PCs for the system as they only contained 63.2% variance of data and it may be the case of losing important information. In future we anticipate looking at higher number of PCs to see whether they provide better results. We did not use higher number of PCs for our initial experiments as they were resulting in higher number of rules whereas with only 3 PCs we had 21 rules. We also need to find a mechanism for rule reduction for our future work. We have used whole data set as an initial input to our method. From Fig.6, it can be seen that using specific areas of spectral data instead of using the whole data set may provide better results. Although these results do not properly classify the data but they indicate that FIS is a strong candidate as an advanced computational method to be further investigated for the automation of the breast cancer grading.

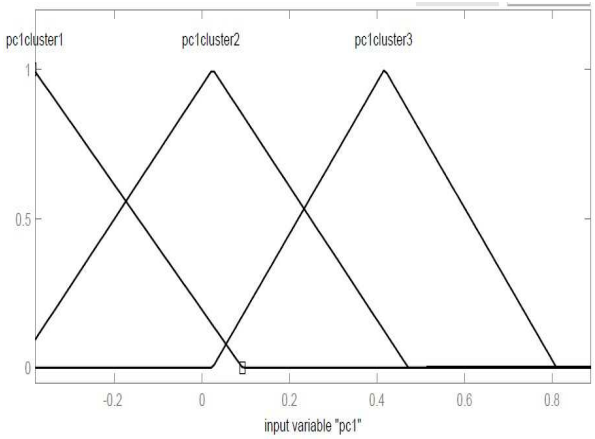

Fig. 7 Membership functions for PC1

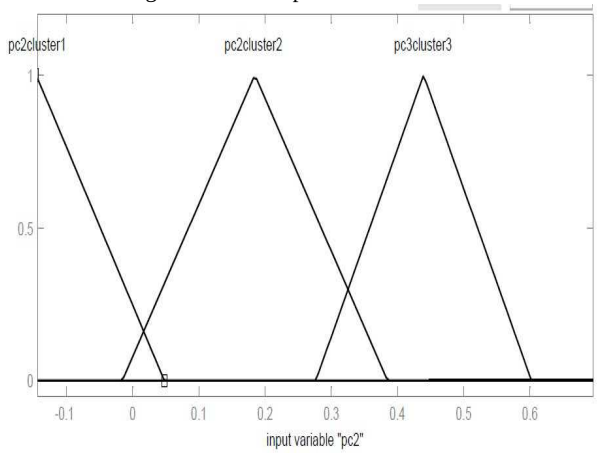

Fig. 8 Membership functions for PC2

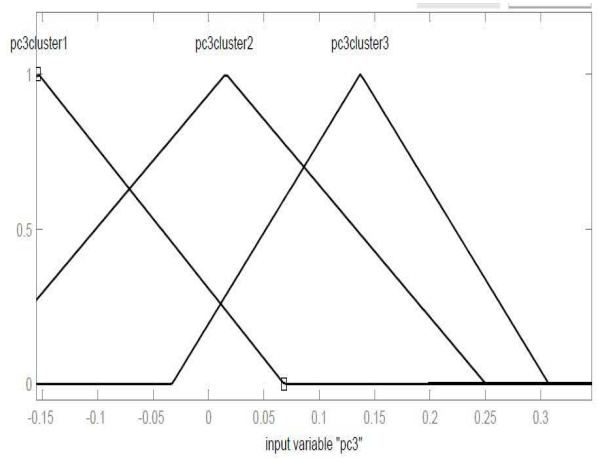

Fig. 9 Membership functions for PC3

FUTURE WORK

In future, we intend to develop novel advanced computational methods using fuzzy systems that can be used for cancer spectral data sets in the area of breast cancer for the automation of breast cancer grading. Although our initial focus is on developing simple FIS but in future we shall be investigating more complex type II fuzzy logic based system as it is more suitable to cases where more uncertainty is involved. The ultimate aim of this research is to develop novel frame work with the help of novel advanced computational methods with FIS which provide an automated system for the grading of breast cancer. Such a frame work will be very beneficial for the clinicians helping them in real clinical practice obtaining better prognosis for the cancer patients increasing their long term survival.

REFERENCES

[1] A. C. Society, "Cancer Facts and Figures " 2012.
[2] J. Backhaus, R. Mueller, N. Formanski, N. Szlama, H.-G. Meerpohl, M. Eidt, and P. Bugert, "Diagnosis of breast cancer with infrared spectroscopy from serum samples," Vibrational Spectroscopy, vol. 52, pp. 173-177, 2010.
[3] X. Y. Wang, "Fuzzy Clustering in the Analysis of Fourier Transform Infrared Spectra for Cancer Diagnosis," in School of Computer Science. PhD: University of Nottingham, 2006.
[4] S. Naqvi and J. M. Garibaldi, "The complexities involved in the analysis of Fourier Transform Infrared Spectroscopy of breast cancer data with clustering algorithms," in Computer Science and Electronic Engineering Conference (CEEC), 2011 3rd, 2011, pp. 80-85.
[5] S. Naqvi and J. Garibaldi, "An Investigation into the use of Fuzzy C-Means Clustering of Fourier Transform Infrared Microscopic Data for the Automation of Breast Cancer Grading," in Proceedings of the 9th Annual Workshop on Computational Intelligence (UKCI 2009) Nottingham, UK 2009.
[6] E. A. Rakha, M. E. El-Sayed, A. H. S. Lee, C. W. Elston, M. J. Grainge, Z. Hodi, R. W. Blamey, and I. O. Ellis, "Prognostic Significance of Nottingham Histologic Grade in Invasive Breast Carcinoma," J Clin Oncol, vol. 26, pp. 3153-3158, July 1, 2008 2008.
[7] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in Biomedical Imaging: From Nano to Macro, ISBI 2008. , 2008, pp. 284-287.
[8] S. N. Sivanandam, S. Sumathi, and S. N. Deepa, Introduction to Fuzzy Logic using MATLAB vol. 15: Springer, 2007.
[9] H. Hamdan and J. M. Garibaldi, "Adaptive neuro-fuzzy inference system (ANFIS) in modelling breast cancer survival," in Fuzzy Systems (FUZZ), 2010 IEEE International Conference on, pp. 1-8.
[10] M. Castanys, R. Perez-Pueyo, M. J. Soneira, E. Golobardes, and A. Fornells, "Identification of Raman spectra through a case-based reasoning system: application to artistic pigments," Journal of Raman Spectroscopy, vol. 42, pp. 1553-1561, 2011.
[11] A. G. Evsukoff, A. C. S. Branco, and S. Galichet, "Intelligent data analysis and model interpretation with spectral analysis fuzzy symbolic modeling," International Journal of Approximate Reasoning, vol. 52, pp. 728-750, 2011.
[12] C. Cernuda, E. Lughofer, W. MÃ¤rzinger, and J. r. Kasberger, "NIR-based quantification of process parameters in polyetheracrylat (PEA) production using flexible non-linear fuzzy systems," Chemometrics and Intelligent Laboratory Systems, vol. 109, pp. 22-33, 2011.
[13] S. Z. Mahmoodabadi, J. Alirezaie, P. Babyn, A. Kassner, and E. Widjaja, "Wavelets and fuzzy relational classifiers: A novel spectroscopy analysis system for pediatric metabolic brain diseases," Fuzzy Sets and Systems, vol. 161, pp. 75-95, 2009.
[14] Y. Zhengmao, "Artificial-intelligence approach for biomedical sample characterization using Raman spectroscopy," Automation Science and Engineering, IEEE Transactions on, vol. 2, pp. 67-73, 2005.
[15] R. Perez-Pueyo, M. J. Soneira, and S. Ruiz-Moreno, "A fuzzy logic system for band detection in Raman spectroscopy," Journal of Raman Spectroscopy, vol. 35, pp. 808-812, 2004.
[16] S. G. Kong, Y.-R. Chen, I. Kim, and M. S. Kim, "Analysis of Hyperspectral Fluorescence Images for Poultry Skin Tumor Inspection," Appl. Opt., vol. 43, pp. 824-833, 2004.
[17] "Bio Max Website," 2012(www.biomax.us)
[18] A. Hyvärinen, "Survey on Independent Component Analysis," Neural Computing Surveys, vol. 2, pp. 94-128, 1999.