EXPLORATORY AND DIRECTED ANALYSIS OF MEDICAL INFORMATION
VIA DYNAMIC CLASSIFICATION TREES

CHRISTOPHER HUGHES
College of Medicine, College of Engineering
University of Illinois, Chicago, Illinois 60612

ABSTRACT

Given the proliferation of computers in the medical domain, computerized clinical patient data bases are rapidly becoming an important source of medical knowledge. However, medical data are often voluminous, incomplete and non-numeric, making analysis with traditional statistical techniques difficult at best. The purpose of this presentation is to describe a generic medical data analysis system called FRID which overcomes these restrictive characteristics. Given its incorporation of partitioning heuristics, FRID can be used to initiate broad-based exploratory analysis. Alternatively, the system's flexible design also allows a specific search for the probability of a disease given a single symptom. Results from experiments using this system will be presented as well as plans for its future use.

Introduction

Given the current proliferation of computers in medicine, clinical patient databases are being developed on a much larger scale than in the past. One of the results of this computerization of patient information is that others have been able to follow the work of de Dombal [1] and Michalski [2], namely, using statistical analysis of computerized patient data as an important source of medical knowledge. For example, Hier and Atkinson have used NIH's NINCDS Stroke Data Bank to develop a stroke assessment Rule Based Expert System [3]; Fattu and Blomberg used a database to develop an expert system to help assess the presence of anemia [4]; Weiss and Kulikowski have used a patient database to assess the consistency of rules derived from experts in the field of Rheumatology [5]; while similar work, focusing on the field of Cardiology is conducted by Uretz and Denes [6].

Given a consistent, carefully developed database, these examples illustrate the fact that statistical analysis can easily extract the routine, regimented, categorical diagnostic knowledge referred to by Pauker [7]. Furthermore, with the proper analysis tools, these databases constitute important investigative research domains, capable of providing physicians with detailed and accurate information concerning their patient population. Safran and Sobel appear to agree, given their development of an interactive statistical analysis facility within a hospital information system [8].

Yet, medical data are often voluminous, incomplete and non-numeric, making the use of traditional statistical analysis packages difficult at best. In light of these developments, it was our intention to design a generic medical data analysis system which would not suffer from the restrictions and assumptions of commonly-used statistical techniques. The result of this work was the interactive analysis system depicted in figure 1.

The FRID System, An Overview

The FRID system (Finding Rules in Data) is an analysis package consisting of an data base and four program modules which each perform a separate task. The data base is a collective store of numerous medical cases which is organized into three separate files. As new patient cases are presented to the system (in the form of case input files) the compilation module updates the data base. To perform this task the module must send the compiled patient information to the file and memory manager which has sole control of (and access to) the data base.
The main purpose of the FRID system is to perform exploratory or directed analysis of the data base. Given the analysis instructions found in the search specification packet, the analysis tree generator builds dynamically allocated classification trees which attempt to find correlations within the collective store of information. This is accomplished using a stepwise heuristic search described later. The results of the analysis are then returned to the user via the user interface.

Information in the data base can be symbolic, numeric or temporal. If the analysis tree generator is performing analysis on numeric information, it is often necessary to cluster these data so that correlations can be found. Upon request from the analysis tree generator, the clustering module will take a collection of numeric data and return ranges of density along with corresponding frequency counts.

The mode density clustering module of the FRID system takes advantage of the range of medical data without becoming too specific and impractical for other uses. Unlike most mode/density clustering algorithms which assume data clusters will be separated by gaps of zero density, the FRID clustering system notes that most numerical medical data, with few exceptions, fall within narrow ranges and therefore density clusters will be seen more as "Humps" in an uniformly dense range of analysis. The purpose of the analysis is therefore to recognize the humps, determine the valleys between them and separate them into different clusters.

Two other facts about numeric medical data also influenced the design of this clustering system. First, numeric medical data are often voluminous which indicates that the system must be relatively fast in order to be effective. Secondly, precise values for range bounds are not as important as the approximate ranges.
themselves. The combination of these two facts suggested that costly mean square
deviations and successive variance calculations could probably be avoided without sacrificing the accuracy necessary to be effective. Therefore, the system was
designed around the concept of variable sized clustering ranges; or put in terms of histogram construction -variable sized bins. By using simple calculations of midpoint and mean (one floating point division each) fast and accurate
clustering was performed without being needlessly, computationally expensive.

User Interface

The command driven user interface allows an investigator to specify data
base analysis ranging from completely heuristic exploratory searches to detailed
and specific pair-wise tabulation. By reviewing the results interactively, the user can test hypothesis and determine the most predictive characteristics of specified diseases.

Though command oriented, the user interface has extensive help facilities
which provide command completion to the experienced user and display of options
(at any point) for the novice. Within the interactive user interface there exists
two levels of commands. The first level is
known as the root and from this prompt the user can perform numeric clustering, data
base updating and data base statistic display. Alternatively an option may be
chosen to descend to the second level of commands where the user specifies all the
parameters necessary to initiate a specific analysis of the data base.

At the top command level, the List
Stats command, in its simplest form, displays the total number of occurrences
of each class (and its respective attributes) in the data base. If desired, these results can be saved in a designated file. This type of information is important for assessing the general prevalence of symptoms within the patient population.

The Clustering command causes the
FRID clustering module to perform cluster analysis on the numeric class specified by
the user. Results are presented in the
form of ranges and the number of instances
of the specified class that fall within
that range.

Data Base File Structure

As described in detail later, the
basic datum of the FRID data base is
denoted by: "Class = Attribute". Using the
context of medicine, valid examples would be: "Age < 45", "Bacteria = E. Coli", or
"Diagnosis = Metastatic Stage I Liver Cancer". The data base is designed to store
multiple instances of these classes and attributes by maintaining three distinct files. The first file contains all the names and information associated with the
individual classes and attributes. The second file, records all the instances in
which these classes and attributes were found. Finally, space within these two
files is incrementally allocated with the use of free space pointers and other
status information kept in the status file.

Data Base Addressing System

Since medical data are often extensive, the entire FRID data base could
not reasonably be expected to fit into memory at any one time, an internal
addressing scheme was needed that would allow disk storage to be used effectively
as an extension of the system's main memory. Unity of address representation
between the two storage media would allow
necessary sections of the data base to be
held in memory (for use) while unnecessary
segments could be left in secondary
storage until needed.

To meet these criteria, an internal, page-based, addressing scheme was
developed using 32-bit integers. With this scheme each file within the data base
system can be expanded to a limit of 4096 Megabytes and swapped into memory in 64K
byte segments using a least-frequently-
used paging algorithm.

Data Analysis in the Medical Domain

The purpose of the FRID system is to
discover associational relationships
between different variables, variables in
the FRID system are represented as classes
while the different values that a variable
can attain are referred to as attributes.
In a medical context, finding
relationships among different variables
classes) has always been an important
aspect of research. Common examples range
from: What patient characteristics
constitute effective predictors of
potential heart disease? to, How often is
lung cancer found in people who have long
histories of cigarette smoking? The former
case exemplifies the need for analysis
techniques which can assess the strength
of association or predictability of one
class versus another. The latter
represents a simple posterior probability
calculation of a disease given a symptom.
Typically, medical researchers will collect large amounts of data from a large number of patients. Those data are then represented as a collection of "class = attribute" pairs which are grouped as instances. In this case, an instance represents all the data on a single patient. At this point, the instances are searched to determine how many times a class, with a specific attribute value, is found in the same instance (patient) as another class - with a specific attribute value. For example, how many times is "smoking = present" found in the same instance as "lung cancer = present".

Philosophically, data structure and organization imply inherent cognitive association. Since one of the purposes of FRID is to discover associations, the structure of the data base was kept to a minimum to avoid biasing data relationships. For this reason, the grouping of class/attribute pairs into instances was the predominant structure imposed on the FRID data base.

**Exploratory Analysis**

The analysis/discovery system is based on the step-wise analysis of classification trees. This technique has been available for many years but has recently been shown to be very effective as an exploratory tool in data analysis by Breiman et al. [9]. These investigators note that most traditional statistical analysis techniques assume data sets to be: small in size, highly structured and homogeneous (i.e. relationships between variables are constant over the data space). By contrast, computerized medical information possesses none of these characteristics. Classification tree analysis is a dynamic, recursive, computationally efficient process using context insensitive partitioning heuristics to cluster data sets. Given these generic qualities, classification trees are not restricted by the assumptions of other analysis techniques.

Breiman and others have experimented with a number of different partitioning heuristics and found that the quality of the classification tree is remarkably insensitive to the specific heuristic used. One such heuristic found to be effective is based on the information measurement of entropy. This definition of entropy has been used mostly in the field of communication and information theory [10,11] but has also been effective as a partitioning criteria.

Classification programs using heuristics have been used by Breiman in a number of medical and non-medical contexts. Results from these experiments found that whether using binary data, categorical medical data, or mixed (categorical, numerical, binary) medical data, these classification techniques performed as well or better than comparable logistic regression or step-wise discriminant analysis [9, Breiman, chapter 6]. Others such as Quinlan have found classification trees using this heuristic to be equally applicable and effective in other fields as well [12]. The computational efficiency of classification trees, in combination with their non-parametric nature represented sufficient justification for their use in this context.

**Search Specification**

To facilitate communication between the user interface and the analysis system, a standardized search specification packet was developed. This packet carries all the instructional information necessary to completely specify a detailed data base analysis. The search parameters are shown below:

- **Sort Flag**
- **Search Limit**
- **Search Depth**
- **Search Roots**
- **Search Domain**
- **Correlation Class**

The correlation class is the dependent variable and the search domain is a list of all the classes from the data base that are to be searched (and possibly used) to form the classification tree. This list can range in size from: the entire data base, to a single class. The former would constitute an overly enthusiastic heuristic search while the latter could be used to calculate the posterior probability of a disease given the presence of a specific symptom. Within the search domain, the user may designate specific attributes that should be ignored during the search. In the case of numeric classes, specific ranges of values may be given to narrow the data base search.

The search limit parameter restricts the number of instances searched for each class and attribute. This is useful for limiting the analysis of classes and attributes having a large number of instances. Given the reverse chronological structure of the data base, this parameter is also helpful in assessing recent trends within a patient population.
The depth parameter allows the user to specify the depth of the analysis, or in other words, the height of the classification tree. For example, to specify an exploratory search of ten patient variables for the two most predictive, in combination, the depth would be set to two.

The roots parameter specifies how many analysis roots the system should use during analysis. Normally, the classification heuristics order the domain variables according to their ability to predict a value of the correlation class. Using the example above, if the user wanted to order the variables according to their predictive quality, used alone, the depth would be set to ten and the roots would be set to one to provide results for all ten variables. Therefore, the combination of depth and roots can focus the level of breadth and detail required by the user.

Results of System Utilization

FRID was initially used in collaboration with the Neurology Department at the University of Illinois to analyze a database of dementia cases. This database contains 100 patients and 64 different variables ranging from laboratory results (CBC, CSF protein levels, etc.) to cognitive assessments (Mini Mental Status Exam, Mattis score, etc.) to EEG results (background rhythm, slow waves, sharp waves, etc.). Detailed results are in press at the present time [13], but for the purposes of this discussion, only the results of the exploratory analysis will be presented.

One indication of system effectiveness would be the discovery of strong relationships among variables that were correlated by their definition. In fact, such relationships were found by the FRID system. For example, the purpose of the Modified Hachinski test is to assess the presence of Infarct Dementia. The fact that FRID discovered that patients with Infarct Dementia were 12 times more likely to have high Hachinski scores than low scores was considered indicative of proper system functioning. Further examples of confirmatory results include: 75% of patients with the neurological finding of Alcohol related ataxia had a diagnosis of Alcoholic dementia, and given the absence of seizures, patients with abnormal liver function tests were six times more likely to have Alcoholic dementia than those with normal liver function tests. Finally, FRID discovered that Alzheimer's patients rarely present with seizures, often have high cognitive Mattis and cognitive attention scores, display non focal findings on neurological exam, and rarely present with alcoholism. All of these findings were considered consistent with our current understanding of this disease.

REFERENCES