83-27

MED. INFORM. (1983), VOL. 8, NO. 3, 187–195

A logic-based approach to conceptual data base analysis

R. S. MICHALSKI, A. B. BASKIN and K. A. SPACKMAN
Department of Computer Science and School of Clinical Medicine,
University of Illinois at Urbana-Champaign,
1408 West University Avenue, Urbana, Illinois 61801, USA

(*Received December 1982*)

Keywords: Data base analysis; Inductive inference; Conceptual clustering.

1. Introduction

Data bases are constructed for two major reasons: to keep track of information (data base management); and to learn more about the phenomena which produce the data (data base analysis). Data base *storage* and *retrieval* techniques extend human memory and make possible the management of large sets of specific facts. Data base *analysis* techniques extend our ability to detect and generalize trends shown in the data. Typical data base analysis techniques include statistical tests, discriminant function analysis, and probabilistic techniques.

Most data base analysis techniques operate on numeric data and assume that arithmetic operations such as addition and multiplication are applicable to individual data elements. Also, many techniques require that a complete set of attribute values be supplied for each data object (e.g., a patient). The above restrictions are a significant disadvantage in view of the growing need to store and manipulate incomplete and imprecise data in the life sciences. In clinical medicine, the data completeness condition can only be met for small subsets of the available data and where collection is rigidly monitored.

In addition, it is important that a data base analysis system can be easily used by non-technical personnel, which implies that the method of interaction with the system must be very simple and natural. In recognition of all these needs, researchers have undertaken efforts to build systems able to analyse data using new, non-statistical approaches and applying a high-level formal language [2, 6, 7], or natural language [1], as a medium for interaction with a system.

This paper presents a logic-based data base analysis system that meets the above needs by using logical operations on non-numeric data (nominal data) and numeric operations only where such operations are most appropriate. The interaction with the system is done using a very simple and easy-to-use formal language based on the variable-valued logic calculus. Combined with the ability to handle incompletely- or imprecisely-specified data, this system provides a new and potentially widely applicable tool for the analysis of data bases.

This is a revised and expanded version of a paper (Copyright © 1982 IEEE) given at the Sixth Annual Symposium on Computer Applications in Medical Care (SCAMC-6), 30 October–2 November 1982 in Washington DC and published in the *Proceedings* thereof, pp. 792–796.

2. Data base analysis

In the biological sciences, data bases are developed to facilitate the formulation of new theories or the validation of existing ones. Clinical data bases, in particular, may be used for retrospective studies to derive new medical hypotheses, or for prospective studies to test and validate proposed hypotheses. In both cases, their ultimate purpose is to monitor and modify patient care.

Among important tasks of data base analysis one may list:

(*a*) a determination of important features and patterns in the data base (*descriptive analysis*); and
(*b*) a derivation of correct and 'most cost-effective' classification rules (*predictive analysis*).

Descriptive analysis is a process in which normative information is extracted from the data base and used to identify structure in the data. The grouping of similar patients into syndromes and the identification of the important characteristics of each syndrome are examples of such analysis.

Predictive analysis is a process in which rules for classifying data elements into known categories are derived from the data base. The development of minimal cost diagnostic rules for predicting the disease of new patients on the basis of accumulated data is an example of predictive analysis.

In general, descriptive analysis techniques (whether done by a human expert or by machine) are used to produce a classification of records in a data base before predictive analysis techniques can be applied. The data base analysis system described in this paper, called QUIN [2], includes inference operators for both descriptive and predictive data base analysis.

3. The QUery and INference (QUIN) system

The QUIN system developed at the University of Illinois is an experimental data analysis system which uses a relational data base management scheme to store/retrieve data and various inference operators to conceptually analyse the data. Operators for such analysis include:

(i) clustering the data into subsets corresponding to certain concepts;
(ii) determining the most cost-effective classification or decision rules for the data;
(iii) selecting most relevant attributes; and
(iv) selecting most representative examples of data (e.g., classical cases of diseases).

In this paper we discuss and illustrate the first two operators. The other two are described elsewhere [2, 8].

The QUIN program provides a concise and human-oriented interface to relational data base functions as well as the inference operations. The command language uses variable-valued logic [3] as a relational calculus and its syntax is similar to Codd's relational data sublanguage ALPHA [4].

Using the variable-valued logic (VL) language interactively, a data base can be built, perused, and analysed. At any step, only a portion of the data base is considered, and the analysis results provide inputs to the subsequent steps. For instance, in a data base containing records for several thousand patients, each record

considered at any one time. To facilitate this, the system allows the user to identify the most representative disease cases or most relevant attributes before dealing with the entire set of patient data.

4. Data base management with QUIN

The QUIN program uses relational table format to store and retrieve data. Apart from some normalization requirements which are not important for this paper, a relational table in QUIN consists of a rectangular array of names/numbers arranged into rows and columns. Each table is given a name. The topmost row contains labels (attribute names) for the columns. Subsequent rows contain attribute values for a single data item such as a patient record.

As an example of a relational data base, consider a collection of data about twin births. Such a data base might contain a table showing the identification number of the mother (ID), the number of pregnancies for that mother (GR), the number of previous full term pregnancies (PTP), the number of previous premature deliveries (PPP), and the number of previous abortions (PAB). The QUIN table is represented in table 1.

Another relational table (table 2) illustrating the use of non-numeric attributes might list the delivery methods of twin *A* and twin *B* (DMA and DMB respectively). In this table TBE is total breech extraction, MFD is mid forceps delivery and NSD is normal spontaneous delivery.

The QUIN system provides commands to define new tables, add rows to a given table, change rows, and delete rows. Existing tables may be concatenated together (joined) if they share a common attribute such as ID in the tables above. In the latter case, for each value of ID, the resulting table will have a row containing the original rows associated with this value. Thus, the join of the two tables above would be as in table 3.

The complete data base about twin births can be assembled by multiple joins of the individual tables using the mother's ID number to determine which rows to match. QUIN allows columns to be selected from a table by typing the table name

Table 1. Relational table for twin births.

Mother

ID	GR	PTP	PPP	PAB
6	3	1	1	0
16	3	1	0	0
165	2	0	1	0
.	.	.	.	.
.	.	.	.	.

Table 2. Relational table for delivery methods for twins *A* and *B*.

Delmethod

ID	DMA	DMB
6	MFD	TBE
16	NSD	TBE
165	NSD	TBE
.	.	.
.	.	.

Table 3. 'Join' of tables 1 and 2 above.

Mother* Delmethod

ID	GR	PTP	PPP	PAB	DMA	DMB
6	3	1	1	0	MFD	TBE
16	3	1	0	0	NSD	TBE
165	2	0	1	0	NSD	TBE
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Table 4. Result of QUIN command check.

ID	GR	PTP	PPP	PAB
16	3	1	0	0

Table 5. Improbable finding from retrieval request.

ID	GA	WTA	WTB
6	28	2214	2221

and then the names of the columns desired. For example, the table Delmethod (ID, DMA) with only columns ID and DMA contains data about the identity and the delivery method of the first twin only.

The three examples below illustrate how the logical expressions of the VL language can be used with data management operations. Complex retrievals are easily specified using the command 'get' with logical restrictions on the retrieval request and can be used for quality control.

In the twin data base described above, all data has been joined together in a single table called 'Twin'. The following QUIN command can be used to find any records in which the number of pregnancies is not equal to the number of previous full term pregnancies plus the number of previous premature deliveries plus the number of abortions plus one (for the present pregnancy):

$$\text{get Twin (ID, GR, PPP, PTP, PAB)}:[GR \neq PTP + PPP + PAB + 1]$$

which produces table 4, showing that only one inconsistent data record exists and should be checked.

An improbable finding such as an infant born at less than 28 weeks gestation with a weight of over 2000 grams can be detected with the following retrieval request:

$$\text{get Twin (ID, GA, WTA, WTB)}:[GA < 28]\&([WTA > 2000] \vee [WTB > 2000])$$

which will detect an improbable birth weight in either twin and produce a table containing one implausible data item (see table 5). In this, GA is the gestational age in weeks, WTA/WTB the weight of the respective twin in grammes.

Finally, the example below checks for an improbable delivery sequence in which the first twin was delivered by c-section and the second twin was not:

$$\text{get Twin (ID, DMA, DMB)}:[DMA = CS]\,\&\,[DMB \neq CS]$$

which produces an empty table showing that no such data records exist.

5. Data base analysis with QUIN

The operators invoked by QUIN to perform descriptive analysis and predictive analysis are called CLUSTER [5] and DIFF (realized by inductive learning program AQ11 [6]), respectively. The CLUSTER operator takes a table of data elements (rows) and attempts to partition the data elements into a specified number of groups. The operator searches for *conceptual* rather than *statistical* groupings (clusters) which can be described with logical statements. The DIFF (differentiate) operator searches for most economical predictive rules which distinguish two or more already-identified classes of data elements. Rules induced are optimized according to a generalized cost criterion.

The operators CLUSTER and DIFF can best be illustrated by a simple example. Although the data below was taken from a real data base, only a subset of the available descriptors were used. The number of patients and the complexity of the example have been kept small for purposes of illustration.

5.1. *CLUSTER operator*

In the example, the CLUSTER operator in QUIN is used for descriptive analysis of a data base containing the records of patients with craniosynostosis syndromes. The operator is used to partition the data base into subgroups in much the same way a physician might divide the patients into different syndromes. After the CLUSTER operation partitioned the data, the DIFF operator was used to determine the simplest rules for predicting membership in each of the 'syndromes' found by the CLUSTER operator. Figure 1 shows the 18 attributes which were used to characterize a patient. The attributes may take the values present(+) or absent (−). Figure 2 shows the data base of patients where each row specifies attribute values for each patient.

The data table used by CLUSTER contains no syndrome classification. Even though we know the syndrome classifications which have been assigned by physicians, this information was not provided to the CLUSTER operator. In the

Attributes

Symbolic name	*Full name*
A	craniostenosis or craniosynostosis
B	facial assymetry
C	flat forehead or low-set hairline
D	malformed or low-set ears
E	hearing impairment
F	ptosis
G	proptosis or exophthalmos
H	strabismus
I	tear duct stenosis or excessive tearing
J	cleft palate
K	high arched palate
L	midface or maxillary hypoplasia
M	spinal malformations
N	complete syndactyly of fingers
O	impaired CNS function
P	complete syndactyly of toes
Q	cutaneous syndactyly of fingers or webbing of fingers
R	cutaneous syndactyly of toes or webbing of toes

Figure 1. The 18 attributes used in the example study.

	DATA																	
Pt no	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	+	+	−	−	−	−	+	−	−	+	−	+	−	+	−	+	−	−
2	+	+	−	−	−	−	−	−	−	−	+	+	−	+	−	+	−	−
3	+	−	−	+	+	−	−	−	−	−	+	+	−	+	−	+	−	−
4	+	−	−	−	−	−	−	+	−	+	+	+	−	+	−	+	−	−
5	+	+	−	−	−	−	+	−	−	−	−	+	−	+	−	+	−	−
6	+	−	−	−	−	−	+	−	−	−	+	+	−	+	−	+	−	−
7	+	−	−	−	−	−	+	−	−	−	+	+	−	−	−	−	−	−
8	+	−	−	−	−	+	+	−	−	−	−	+	−	−	−	−	−	−
9	+	−	−	−	−	−	+	−	−	−	−	+	−	−	−	−	−	−
10	+	−	−	−	−	−	+	−	−	−	−	+	+	−	+	−	−	−
11	+	−	−	+	−	−	+	−	−	−	+	+	+	−	+	−	−	−
12	+	−	−	−	−	−	+	−	−	−	+	+	−	−	−	−	−	−
13	+	−	+	−	−	−	−	−	−	−	−	+	−	−	−	−	+	−
14	+	−	+	−	−	−	+	−	+	−	−	+	−	−	−	−	−	+

Figure 2. The 14 data records with variable names and values abbreviated.

first experiment the CLUSTER operator was applied to cluster the patient data into only two classes, even though three syndromes are present in the data (Apert, Crouzon, and Saethre-Chotzen).

The best grouping of patient records found by CLUSTER for two groups was:

Group	Rule	Cluster index
1	[A = +]&[B = −]&[E = −]&[H = −]&[J = −]&[L = +]&[N = −]&[P = −]	1016
2	[A = +]&[C = −]&[F = −]&[I = −]&[L = +]&[M = −]&[O = −]&[P = +]&[Q = −]&[R = −]	122

which can be paraphrased:

1 present: craniostenosis, maxillary hypoplasia
absent: facial asymmetry, strabismus, cleft palate, complete syndactyly of hands or feet

2 present: craniostenosis, maxillary hypoplasia, complete syndactyly of hands and feet
absent: low-set hairline, ptosis, tear-duct stenosis, spinal malformations, CNS impaired, webbing

The 'cluster index' is a measure of the 'fit' between the cluster description and the data (the smaller the index the better the fit).

From a comparison of the rules above with the data base, it can be shown that class 1 contains both the Crouzon and Saethre-Chotzen syndromes while class 2 contains only the Apert syndrome. Furthermore, the large value of the cluster index for class 1 indicates that further clustering of the data might produce a better classification. Indeed, when the CLUSTER operation was applied to cluster the data into three groups, the following results were obtained:

Group	Rule	Cluster index
1	[A = +]&[C = +]&[L = +]&[B = −]&[D = −]&[E = −]&[F = −]&[H = −]&[J = −]&[K = −]&[M = −]&[O = −]&[N = −]&[P = −]	14

2 [A = +]&[L = +]&[N = +]&[P = +]&[C = −]&[F = −]&
[I = −]&[M = −]&[O = −]&[R = −] 122

3 [A = +]&[G = +]&[L = +]&[B = −]&[C = −]&[E = −]&
[J = −]&[N = −]&[P = −]&[R−] 26

which can be paraphrased:

1 present: craniostenosis, low-set hairline, maxillary hypoplasia
absent: facial asymmetry, ears malformed/low-set, hearing impairment, ptosis, strabismus, cleft palate, high arched palate, spinal malformations, impaired CNS function,. complete syndactyly-fingers and toes,

2 present: craniosynostosis, maxillary hypoplasia, complete syndactylyl-fingers and toes
absent: low-set hairline, ptosis, tear-duct stenosis, spinal malformations, impaired CNS function, webbing

3 present: craniosynostosis, proptosis, maxillary hypoplasia
absent: facial asymmetry, low-set hairline, hearing impairment strabismus, tear duct stenosis, cleft palate, complete syndactyly fingers and toes, webbing

Notice that now the cluster indices are comparable and much lower than the index of the first cluster in the previous case's two groups.

Among the 14 patient records used, the grouping above splits the patients into exactly the same three syndromes as the human experts do. The low cluster indices suggest a good fit between the clusters and the data.

5.2. *DIFF operator*

The DIFF operator in QUIN invokes the inductive learning program AQ11 to determine the most economical rules differentiating between given collections of observations. When applied to discriminate between the three clusters obtained above, the following discrimination rules were obtained:

Group 1	(Saethre-Chotzen)	[C = +]
Group 2	(Apert)	[P = +]
Group 3	(Crouzon)	[G = +]&[N = −]&[R = −]

which can be paraphrased:

Saethre-Chotzen: low-set hairline/flat forehead;
Apert: complete syndactyly of toes;
Crouzon: proptosis and no complete syndactyly of fingers and no webbing of the toes.

Unlike the CLUSTER operator which describes the conceptually-important features for defining a class (even when they are common to all classes), the DIFF operator finds only those attributes which differentiate between the members of each class. The discrimination rules given by the program accurately differentiate the three syndromes given that one of the syndromes is present.

Because of the small number of patient records used, the discrimination rules determined by the operator may not be perfectly reliable in differentiating unusual

cases of these syndromes. In such a case, the DIFF operator can be applied repetitively using new data until the rules are sufficiently refined. In our case, the rules listed above do correctly separate all of the patient records.

6. Comparison with traditional techniques

It is useful to compare the logical data base analysis described here with other kinds of analysis that use statistics, probability and numerical taxonomy. CLUSTER has two advantages over traditional numerical taxonomic techniques: first, it is capable of handling nominal or ordinal values as well as numeric data; second, it not only generates clusters but provides descriptions of the clusters in conceptual or logical form which can be critiqued by the investigator. In addition, the criteria on which the clusters are generated (the 'similarity measure') may be based on concepts other than numeric distance, and this results in clusters that tend to match human solutions more closely [6].

The DIFF command generates rules for differentiating classes of events based on logical and numeric information. This means that, unlike most probabilistic techniques, the distinction between two classes of events can be made even with incomplete information. The rules are optimized to be simple and to include the factors considered most important by the investigator (via weighting), and they are easily understood and critiqued. Although QUIN supports primarily logical operations on individual data items, it also is capable of statistical measures (e.g., the mean, variance, chi square).

When using QUIN, it is important to realize that it is difficult to fit certain types of data into the relational formalism. Consider a data base which contains descriptions of abnormalities found in patients. Let each patient occupy a row in the relational table, and each abnormality be a column. It is readily apparent that these will be large sparse tables because of the limited number of abnormalities that occur in any one patient. Such large sparse tables are computationally expensive as inputs to the analysis procedures and generally must be collapsed using the relational table operations. Thus, the analysis process is an iterative process involving judgements on each cycle.

7. Conclusion

One objective of QUIN is to provide the medical researcher with a tool which enhances his ability to discover new syndromes (descriptive analysis) and generate rules for differentiating known syndromes (predictive analysis). It should be emphasized that the intellectual involvement of the clinician is crucial to the success of the analysis algorithms, because he must make the initial observations, record them, decide which descriptors (symptoms and signs) are to be measured and finally, which examples are to be used for analysis. He must choose a set of descriptors large enough to adequately characterize the patient data but which does not exceed the computational limits of the algorithms.

The major benefit of the proposed logic-based analysis is the potential simplification of time consuming non-intellectual activities associated with clinical investigations in which large amounts of data are to be analysed. Automated data base analysis should allow a researcher more time for productive thought and less time spent doing tedious and boring tabulations, assessments and initial hypothesis formation.

Acknowledgements

The authors acknowledge the partial support provided by the Office of Naval Research under Grant No. N000 14-82-K-0186, the National Science Foundation under Grant No. MCS 82-05166 and the National Library of Medicine under Grant No. PHS NLM 5 T15 LM07011. They thank the Center for Craniofacial Anomalies, University of Illinois at the Medical Center for providing the data for the craniosynostosis example. Gratitude goes also to Dr Lance Rodewald for helpful comments and the examples using QUIN on the data for twins.

References

1. SHAPIRO, A. R. (1980), A System for Conceptual Analysis of Medical Practices. *Fourth Annual Symposium on Computer Applications in Medical Care* (IEEE Computer Society, New York).
2. SPACKMAN, K. A. (1983), QUIN: Integration of Inferential Operations in a Relational Database. University of Illinois Department of Computer Science Master's Thesis.
3. MICHALSKI, R. S. (1974), Variable-Valued Logic: System VL_1. *1974 International Symposium on Multiple-valued Logic* (West Virginia University, Morgantown, West Virginia).
4. CODD, E. F. (1971), A Data Base Sublanguage Founded on the Relational Calculus. *Proceedings of ACM SIGFIDET Workshop on Data Description, Access and Control.*
5. STEPP, R. E. (1980), Learning from Observations: Experiments in Conceptual Clustering. *Workshop on Current Developments in Machine Learning* (Carnegie-Mellon University, Pittsburgh).
6. MICHALSKI, R. S. and LARSON, J. B. (1978), Selection of Most Representative Training Examples and Incremental Generation of VL_1 Hypotheses: the underlying methodology and the description of programs ESEL and AQ11. *University of Illinois Department of Computer Science Report No. UIUCDCS-R-78-867.* (University of Illinois, Urbana IL).
7. MICHALSKI, R. S. and STEPP, R. E. (1981), Concept-based Clustering versus Numerical Taxonomy. *University of Illinois Department of Computer Science Report No. UIUCDCS-R-81-1073* (University of Illinois, Urbana IL).
8. GRAMM, S. A. (1982), ESEL/2: A program for selecting the most representative training events for inductive learning. *Report No. ISG 82-4, Intelligent Systems Group, Department of Computer Science, University of Illinois* (University of Illinois, Urbana IL).