Presented at the Workshop on
Machine Learning, Carnegie-Mellon
University, Pittsburgh, PA,
July 16-18, 1980

80-8

80-8

Learning from Observation: Experiments in Conceptual Clustering

Robert Stepp
University of Illinois
Urbana Illinois 61801

Introduction

Learning from Observation is a type of machine learning in which the input data are unclassified events, and the objective is to determine an underlying structure of the events by partitioning them into clusters which are defined by a set of concepts selected from a universe of concepts. An approach is presented (conceptual clustering) which partitions the data into a number of conceptual components described by conjunctive statements (VL_1 complexes) which together fit the collection of events best according to a specific criterion. The set of conceptual descriptions of each cluster forms a taxonomic description of the collection of events.

A conceptual clustering algorithm created by Michalski [1] has been implemented in an inductive program called CLUSTER/paf. This program accepts events (or object representations) consisting of a number of multivalued feature variables. Nominal and linearly ordered variable scales are handled with specific generalization techniques applied to each. Given a set of events, the program creates a number of clusters each described by a conjunctive statement (called a complex) in the VL_1 (Variable-valued Logic system 1) language. (VL_1 expressions and the various generalization mechanisms which operate on them are described in [2].)

A complex denotes a portion of the feature space and is arranged to cover one or more of the observed events while not overlaping any other complex. Because of generalization, complexes often cover a portion of the feature space including some <u>unobserved</u> events along with the observed events. The number of such unobserved events, called the <u>sparseness</u> of the complex, indicates the degree of overgeneralization of the complex with respect to the observed events it covers. Total sparseness (summed over all complexes) is one good indicator (see [1] for others) of the overall fit of the complexes to the observations, and minimizing total sparseness can be used (as it is in the following experiments) as a criterion of optimality to guide the heuristic search for the best (locally optimum) clustering.

A flow diagram of the algorithm used by CLUSTER/paf is given in Figure 1. On each iteration, a number of clusters (given as k) are formed which partition the event set E. The termination criterion is adjustable and is normally the lack of further improvement in the degree of fit of the complexes to the events in a number of "look ahead" iterations. For the experiments which follow, the stopping criterion was the absence of further diminishment in total sparseness.

The central step in each iteration uses the seed events to help form the required number of non-overlaping VL_1 complexes, one for each cluster. The program creates a VL_1 complex covering each seed event and as many surrounding events as it can and yet achieve the best fit (minimal sparseness). This is a combinatorially explosive task which necessarily incorporates a best-first search of the alternative choices of complexes.

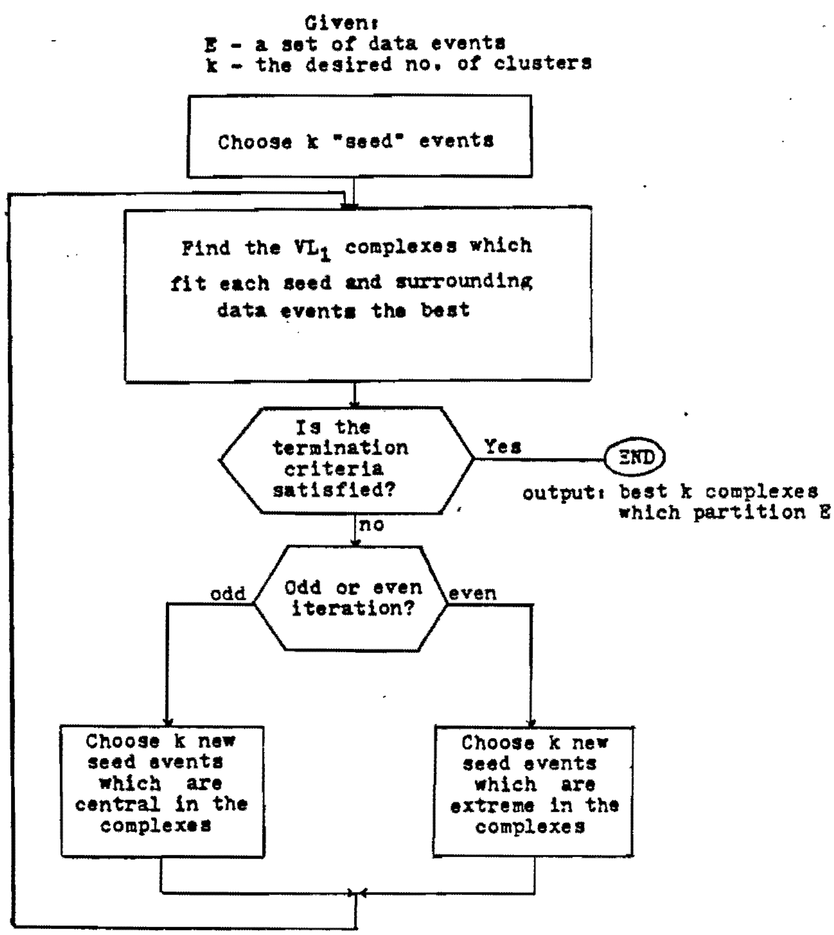

Figure 1
Flow diagram of CLUSTER/paf

The CLUSTER/paf program implements a sequential algorithm for exploring the search tree which does not guarantee that the complexes will not overlap. When overlaping complexes do occur, a mechanism is provided to try to "fix" them by making minor adjustments. Such a fixup is not always possible and in cases of irreparable overlaping complexes, the events which could not be covered without causing an overlap are left uncovered as "noise" in the collection.

After producing a new partitioning, new seeds are selected as representatives of the new clusters. Usually a representative event is one which exhibits the central tendencies of a group, however in CLUSTER/paf a principle of adversity is incorporated with the representative event being an extremal (border) event on even iterations and a central event on odd iterations. This design is provided to speed up the convergence to the optimal choice of seed events. Experience with the program shows that this goal may have been acheived as it has been observed that this technique sometimes may waste one iteration (i.e. one using extremal events) but at other times can save several iterations by using the "surprise" effect of the quite different extremal seeds to skip to a more interesting set of clusters. (A thorough study of this behavior remains incomplete.)

The general operation of the conceptual clustering algorithm, i.e. determining representative (seed) events and then the best partitioning (set of complexes) with respect to those representative events, is an adaptation of Dynamic Clustering which is thoroughly discussed in [5].

Experiment 1: Soybean Disease Clustering

This experiment was designed to test the effectiveness of conceptual clustering in revealing useful partitionings of the observed events, and to gauge the effect of non-optimal choices for k (the number of clusters). The input data consisted of 47 events representing cases of diseased soybean plants via 34 multivalued variables. Because the cases were known to describe plants with one of four diseases (a result of a previous study [3]) the number of clusters was set to 4. The CLUSTER/paf program obtained an optimal partitioning after 4 iterations and it was found that the clusters in that partition each contained plants with the same disease. In addition the VL_1 statements for each cluster provided a description which revealed the differences in the symptoms of

Feature Variable	Diaporthe Stem Canker cluster 1	Charcoal Rot cluster 2	Rhizoctonia Root Rot cluster 3	Phytophthora Rot cluster 4
PLANT STAND:	Normal	Normal		Less Than Normal
PRECIPITATION:	Above Normal	Below Normal	Above Normal	Normal or Above
TEMPERATURE:	Normal	Normal or Above	Below Normal	Normal or Below
DAMAGED AREA:	Scat. or Low	Wh. Fields,Upland	Low Areas	Wh.Fields, Low Areas
STEM CANKERS:	Above 2nd node	Absent	Below soil line	Below or sl. above soil
CANKER LESION COLOR:	Brown	Tan	Brown	D. brown or Black
FRUITING BODIES ON STEM:	Present	Absent	Absent	Absent
EXTERNAL DECAY OF STEM:	Firm and Dry	Absent	Firm and Dry	Abs. or Firm and Dry
INTERNAL DISCOLORATION OF STEM:	None	Black	None	None
SCLEROTIA:	Absent	Present	Absent	Absent
FRUIT PODS:	Normal	Normal	Few or None	
ROOTS:	Normal	Normal	Normal or Rotted	Rotted

Figure 2
Distinguishing Features of the Clusters of Diseased Plants Produced by CLUSTER/paf

[TIME=JULY..OCT] [PLANT STAND=NORMAL]
[PRECIPITATION=ABOVE NORMAL]
[TEMPERATURE=NORMAL]
[CROP HISTORY=SAME FOR SEVERAL YEARS]
[DAMAGED AREA=SCATTERED or LOW AREAS]
[SEVERITY=POTENTIALLY SEVERE or SEVERE]
[SEED TREATMENT=NONE or FUNGICIDE]
[PLANT HEIGHT=ABNORMAL] [LEAVES=ABNORMAL]
[LEAF SPOTS=ABSENT] [SHOTHOLING=ABSENT]
[LEAF MALFORMATION=ABSENT] [LEAF MILDEW=ABSENT]
[STEM=ABNORMAL] [MYCELIUM=ABSENT]
[STEM CANKERS=ABOVE 2nd NODE]
[LESION COLOR=BROWN or NA]
[FRUITING BODIES=PRESENT] [ROOTS=NORMAL]
[EXTERNAL DECAY=FIRM AND DRY] [SCLEROTIA=ABSENT]
[INTERNAL DISCOLORATION OF STEM=NONE]
[FRUIT PODS=NORMAL] [SEED=NORMAL]

Figure 3a

Description given by CLUSTER/paf for the cluster containing 10 cases of Diaporthe Stem Canker.

[TIME=AUG..SEPT]
[PRECIPITATION=NORMAL or ABOVE]
[TEMPERATURE=NORMAL or ABOVE]
[CROP HISTORY=SAME FOR SEVERAL YEARS]
[STEM CANKERS=ABOVE 2nd NODE]
[LESION COLOR=BROWN]
[FRUITING BODIES=PRESENT]
[FRUIT PODS=NORMAL]

Figure 3b

Expert description of the Diaporthe Stem Canker disease given by a plant pathologist.

each disease (see Figure 2). Figure 3 shows the complete VL_1 statement for one cluster along with a descriptive statement (also coded in VL_1) given by an expert plant pathologist. The two statements agree almost completely, though the machine-generated statement contains many variables which the expert did not mention (indicating that they may be less relevant) and the TIME (of occurrance) range indicated by the expert is a subset of the TIME range actually observed. The potential utility of the machine produced statement seems to be excellent.

The CLUSTER/paf program was run additionally with k=2 and k=6 to observe the effects of too few and too many clusters with respect to the real world structure of the data. (This is crucial since normally a researcher has only a guess as to the number of clusters which would be most relevant).

In the case of the soybean diseases, the 2-partitions and 6-partitions were well behaved. In the former case, each cluster contained those cases involving two of the four diseases (i.e. cases of the same disease remained bound together) while in the second case each cluster contained cases of only a single disease with two additional clusters formed by subdividing two disease categories. This ability to recognize strong patterns in the data, even without a perfect choice for k, is an important property.

The same soybean data was also fed to a standard numerical taxonomy program. The dendrogram obtained using unweighted simple matching scores showed the same partitioning into single-disease taxonomical units, however, being a non-conceptual process, the description of each cluster was not produced directly and was not the basis for the resulting taxonomy.

Experiment 2: Taxonomy of Spanish Folksongs

This experiment illustrates the use of hierarchical conceptual clustering to create a multilevel taxonomy of a set of objects. One hundred old Spanish folksongs each described by 22 binary and multilevel variables (musicological features of the songs) were partitioned by CLUSTER/paf into two clusters, forming the most general part of a dendrogram. Then each cluster was separately partitioned into two clusters, forming more specific parts of the dendrogram. This process was repeated to complete the dendrogram where at each level VL_1 statements provide a description of each grouping and are a source of the differences, in terms of the given musicological features, denoting each division of the dendrogram. This is similar to a decision tree whose leaves are individual observed events and clearly presents a structure for describing the collection. Figure 4 shows the top five layers in the complete dendrogram obtained for the Spanish songs. The most fundamental concept (monophonic vs polyphonic) splits the songs into two main categories. Note that the monophonic songs are then split according to the degree of rubato (the full range of rubato, 0-5, was broken into two intervals by CLUSTER/paf at the point which yielded the best fitting complexes) while the polyphonic songs are split according to the degree of embellishment.

In a study of these folksongs by Poveda [4], traditional numerical taxonomy methods were employed which generated a partition but without a description of each grouping. The complexes produced by CLUSTER/paf have proved to be useful to him.

Conclusion

Conceptual clustering, as implemented in the experimental program CLUSTER/paf, appears to be a highly useful tool for learning from observation. This is because it produces a comprehendable concept in the form of a VL_1 complex describing each cluster in the partition, and reveals an underlying structure for the observed events. Further analysis of the behavior of the algorithm and the extension of the types of concepts over which it operates are ongoing research activities.

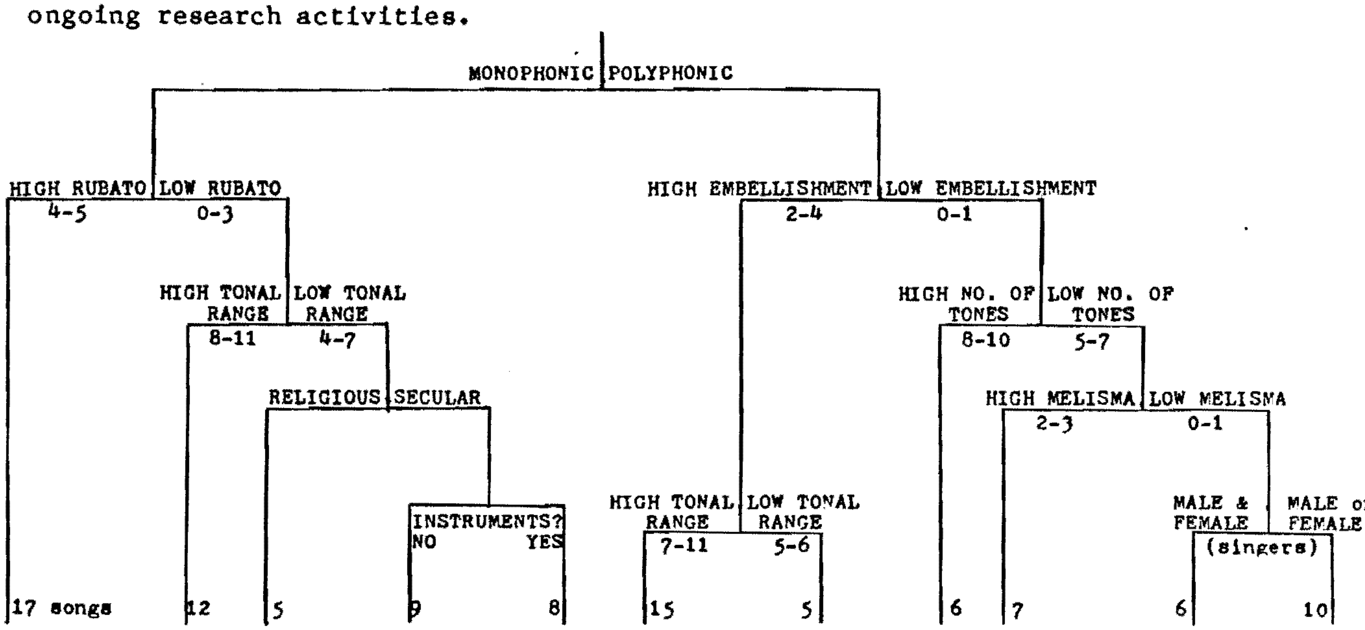

Figure 4

Taxonomy of 100 Spanish Folksongs

Acknowledgements

The author acknowledges the ideas and support provided by R. S. Michalski which contributed to the implementation of CLUSTER/paf. This work was supported by a grant from the National Science Foundation, No. MCS-76-22940.

References

[1] Michalski, R. S., KNOWLEDGE ADQUISITION THROUGH CONCEPTUAL CLUSTERING: A Theoretical Framework and an Algorithm for Partitioning Data into Conjunctive Concepts, A Special Issue on Knowledge Acquisition And Induction, Policy Analysis and Information Systems, No. 3, 1980.

[2] Michalski, R. S., Variable-Valued Logic and Its Aplications to Pattern Recognition and Machine Learning, Chapter in the monograph: Multiple-Valued Logic and Computer Science, edt. David Rine, North-Holland Publishers, 1975.

[3] Michalski, R. S. and Chilausky, R. L., Learning By Being Told and Learning From Examples: an experimental comparison of two methods of knowledge acquisition in the context of developing and expert system for soybean disease diagnosis, A Special Issue on Knowledge Acquisition and Induction, Policy Analysis and Information Systems, No. 2, 1980.

[4] Poveda, P., Classification of Folksongs According to the Principles of Numerical Taxonomy, unpublished report to become a thesis in the Department of Music, University of Illinois.

[5] Diday, E. and Simon, J. C., Clustering Analysis, Chapter in Communication and Cybernetics 10, Ed. K. S. Fu, Springer-Verlag, Berlin, Heidelberg, New York, 1976.