

# A Proposed Method of Conceptual Clustering for Structured and Decomposable Objects

Douglas Fisher

Department of Information and Computer Science  
University of California, Irvine, CA. 92717

## Abstract

A method of clustering structured, decomposable objects in the presence of background knowledge is proposed, with the ability to construct its own 'bias' (generalization hierarchies), thus allowing it to evolve from search-intensive (knowledge limited) behavior to a knowledge intensive (search limited) behavior.

## I. INTRODUCTION

Conceptual clustering is a process abstraction originally defined by Michalski [3] as an extension of numerical taxonomy, a class of methods developed by social and natural scientists for creating classification schemes over object sets. Conceptual clustering methods discover concept (intensional) descriptions for object classes (ie. clusters) participating in a classification scheme and use these concept descriptions to evaluate the *quality* of object classes, whereas methods of numerical taxonomy yield only extensional object class representations. Two problems must be addressed in conceptual clustering.

- The *aggregation* problem involves determining useful subsets of an initial object set. Thus it consists of identifying a set of object classes, each defined as an extensionally enumerated set of objects.
- The *characterization* problem involves determining a useful characterization (concept) for some (extensionally defined) object class, or for each of multiple object classes. This is simply the problem of *learning from examples*.

Current conceptual clustering methods exploit well-understood methods of learning from examples, by making such a process subordinate to a higher-level aggregation process. This interaction can be viewed as a two-tiered search, which can be used to frame a number of conceptual clustering algorithms [1].

## II. LIMITATIONS OF EXISTING CONCEPTUAL CLUSTERING SYSTEMS

Current conceptual clustering systems are limited in a number of important respects. Foremost among these is the limited object and concept languages utilized. Present systems allow objects to be represented in terms of attribute - value pairs. This language can be extended in two ways.

*Proceedings of the Third International Machine Learning Workshop, June 24-26, 1985. Skytop, PA pp. 38-40.*

- *Structured* object representations can be allowed, where relations between attribute values of an object can be explicitly represented.
- *Decomposable* object representations can be allowed, where attribute values of an object are themselves objects which may be further decomposed.

A second means by which current conceptual clustering systems could be extended is to endow them with the ability to utilize *background knowledge* to augment object descriptions. Background knowledge as used by Vere [4] is a body of relations defined over the attribute values of objects. For example, to identify a concept, 'straight' in poker, it is not sufficient to simply know the cards of individual hands (5 of spades, 4 of hearts, 3 of spades, ...), but background knowledge (5 is just greater than 4, 4 is just greater than 3, ...) is also required.

### III. CLUSTERING STRUCTURED AND DECOMPOSABLE OBJECTS

A method of clustering structured and decomposable objects in the presence of background knowledge is currently being investigated. The method, originally inspired by Vere's THOTH system [4], constructs a hierarchical classification scheme in a bottom-up manner. Given a set of objects, aggregation of objects is accomplished by repeatedly 'fusing' individuals to form higher-level classes. Characterization of these classes is dependent on aggregation and characterization over lower-level components of class members. Thus, the clustering task is recursively defined. Aggregation of higher-level objects constrains aggregation at lower levels. Consider this example: Assume the algorithm observes a number of (teacher generated) solution paths (a sequence of operator applications) for solving linear equations of 1 variable (ie.  $ax + b = c$ , where  $1x + 0 = d$  is a 'solution'). The operators (represented as *relational productions* [4]) used by the teacher (but not known by the clustering algorithm) correspond to 'subtract' (ie.  $ax + b = c \Rightarrow ax + 0 = \langle c - b \rangle$ , where  $b > 0$ ), 'add' ( $ax + b = c \Rightarrow ax + 0 = \langle c + b \rangle$ , where  $b < 0$ ), and 'divide' (ie.  $ax + 0 = c \Rightarrow 1x + 0 = \langle c/a \rangle$ ) The algorithm is given only a set of relational production instances (eg.  $3x - 2 = 1 \Rightarrow 3x + 0 = 3$ ), each considered an 'object'. Each production has two components - a 'before' and 'after' state. Each state corresponds to a linear equation with (integer) components a, b, and c. Intuitively, the algorithm is expected (through clustering) to identify production classes, each of which corresponds to an operator class for transforming linear equations. In aggregating production instances, classes corresponding to 'subtract', 'add', and 'divide' operator instances will be generated. By recursively clustering over the components of the members of each of these classes, hierarchical classifications over the bottom-level integer components are generated. These hierarchies (containing nodes corresponding to integer classes such as  $\langle \text{positive} \rangle$ ,  $\langle \text{nonpositive} \rangle$ , and  $\langle \text{odd} \rangle$ ) in turn can be used to guide the characterization process of higher-level objects.

As clustering proceeds, we would expect that the structuring of low-level objects would serve to increasingly constrain the search for higher-level object class characterizations. Thus, one hope of the proposed work is that a clustering system can be realized which evolves from search-intensive behavior to increasingly search-limited behavior.

#### IV. VALIDATING THE PROPOSED ALGORITHM

We wish to validate the algorithm in a variety of domains. Methods of conceptual clustering have thus far been tested on such domains as micro-computers, Spanish folk songs, and animals. However, the behavior of and classification scheme resulting from a conceptual clustering process are open to many varied interpretations. For example, GLAUBER, a system for discovering qualitative empirical laws in the domain of chemistry, by Langley, et.al [2] has been framed as a conceptual clustering system [1]. In creating a classification over a set of molecular structures, concepts derived for molecule groups can be viewed as scientific laws relating these groups. A similar view can be taken of another discovery system, BACON [2]. Of interest also are problem-solving tasks in the domains of algebra and integral calculus.

#### V. CONCLUDING REMARKS

A conceptual clustering system for structured and decomposable objects in the presence of background knowledge has been proposed. The algorithm is to be tested in a number of domains, including integral calculus, algebra, as well as interpreting the algorithm's behavior as scientific discovery. Investigations into the algorithm's ability to build its own 'bias' should demonstrate that clustering becomes increasingly directed (ie. less search intensive) over the duration of a given clustering task. The view throughout this work is that conceptual clustering, despite its original motivation as a data analysis tool, may be viewed as concept formation with behavior open to widely varying interpretations.

#### References

- [1] Fisher, D. and Langley, P. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (forthcoming), 1985.
- [2] Langley, P., Zytkow, J., Simon, H., and Bradshaw, G. In: *Machine Learning, Volume 2* (forthcoming), 1985.
- [3] Michalski, R. *International Journal of Policy Analysis and Information Systems* 4 3, 1980, 219-243.
- [4] Vere, S. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, 1977, 349-355.