The CLUSTER3 system for goal-oriented conceptual clustering: method and preliminary results

W. D. Seeman[1] & R. S. Michalski[1, 2]
[1]*Machine Learning and Inference Laboratory, George Mason University, USA*
[2]*Institute of Computer Science, Polish Academy of Sciences, Poland*

Abstract

A conceptual clustering program CLUSTER3 is described that, given a set of objects represented by attribute-value tuples, groups them into clusters described by generalized conjunctive descriptions in attributional calculus. The descriptions are optimized according to a user-designed multi-criterion clustering quality measure. The clustering process in CLUSTER3 depends on a viewpoint underlying the clustering goal, and employs the view-relevant attribute subsetting method (VAS) that selects for clustering only attributes relevant to this viewpoint. The program is illustrated by a simple designed problem and by its application to clustering of US Congressional voting records. The ongoing research concerns application of CLUSTER3 to large and complex datasets such as collections of web pages.
Keywords: conceptual clustering, unsupervised learning, goal-oriented clustering, pattern recognition, attributional calculus, view-relevant attribute subsetting.

1 Introduction

Clustering analysis is a fundamental way of gleaning knowledge from data when little is known about the organization of the observations recorded in the data. Most clustering methods group objects into clusters solely on the basis of the relationship of the observations to each other, and output results as collections or hierarchies of clustered observations. They typically take descriptions of objects in the form of attribute-value vectors and group the descriptions on the basis of

WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060091

a predefined measure of similarity. Similar objects are grouped into the same clusters and dissimilar objects into different clusters. The so-generated clusters or hierarchies are not given any descriptions or explanations.

Research presented here concerns another approach to clustering, called conceptual clustering, originally introduced by Michalski and Stepp [1], that groups objects into clusters that represent meaningful concepts. The result of such clustering is a hierarchy of clusters together with their descriptions. The cluster descriptions are in the form of conjunctions in attributional calculus Michalski [2], a logic and representation system that combines features of propositional, predicate and multiple valued logic to facilitate machine learning.

The CLUSTER3 program is the newest and most advanced implementation of conceptual clustering. The general algorithm implemented in CLUSTER3 is described in Section 2. Presented in Section 3 are novel aspects of CLUSTER3 including the view-relevant attribute subsetting (VAS) method, which selects for clustering only those attributes relevant to a predefined viewpoint, and new criteria used for cluster evaluation, combined into a multi-criterion measure of quality using the Lexicographic Evaluation Functional (LEF) [1]. Section 4 presents results of initial testing of the program on a few designed and real-world datasets. Conclusions and future directions are in Section 5.

2 The CLUSTER3 algorithm

2.1 Clustering representation

Standard clustering methods split observations into a set of clusters without providing any generalized descriptions of the clusters. Results of such methods are often lists of points and their corresponding clusters. Therefore, introducing a new observation to an existing set of clusters usually requires performing an additional cluster analysis on the entire dataset.

CLUSTER3 presents clustering results as both a hierarchy of clustered observations and as generalized descriptions of the resulting clusters represented in attributional calculus. Attributional calculus is a language with high descriptive power whose descriptions are comparable to the structure and usage of a subset of natural language. A clustering description is a set of statements in attributional calculus, one statement per cluster. Each statement is a conjunction of attributional conditions that in the simplest form are relations between a single attribute and a subset of values of the attribute domain. An example using 3 binary variables X_1, X_2, X_3 is *[*X_1*=1]&[*X_2*=0]&[*X_3*=0]*.

2.2 General algorithm

The CLUSTER3 algorithm, presented in Figure 1, uses an iterative search to generate logically disjoint cluster descriptions that maximally adhere to the clustering quality measure (LEF; see Section 3.3). The initial preliminary cluster descriptions often contain large areas of intersection. The Nondisjoint Into Disjoint

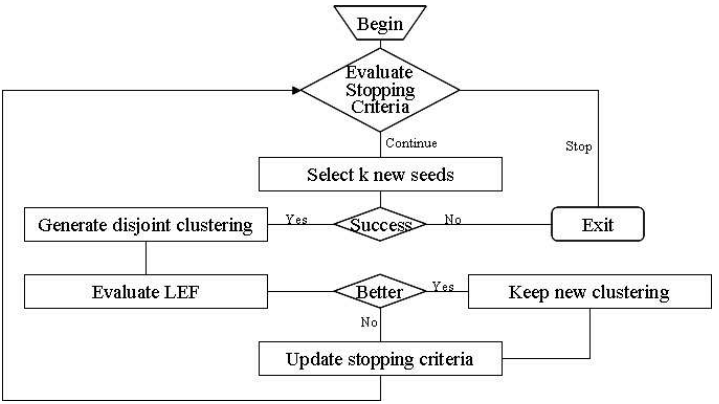


Figure 1: General algorithm for CLUSTER3.

(NID) process removes observations from the areas of intersection and places them in a multiple-covered observations list. This process then contracts the cluster descriptions and uses an agglomerative technique to add observations from this list to the clusters that best fit them. The resulting intermediary clustering description that maximizes the quality measure is selected for use in the next iteration.

Each iteration of the algorithm creates a seed set $S = \{O_1, \ldots, O_k\}$ by selecting k observations to use as seeds for generating clusters, one per cluster. The STAR methodology Michalski and Stepp [1] is applied to each seed to create a set of maximally general complexes used as the basis for cluster descriptions. A maximally general complex is the negated disjunction of all values for an attribute for all seeds except the selected seed. An attribute is excluded if its value in the selected seed appears in at least one other seed. For the set of attributes $A = \{A_1, \ldots, A_n\}$, the value of an attribute A_n for seed O_k is signified by $A_n\{O_k\}$, and the disjunction of values for A_n for all seeds except O_k is $A_n\{O_{\neg k}\}$. The negation of this disjunction is the maximally general complex which covers O_k for attribute A_n but for which no other seed is covered. A partial star is the set of all such complexes for O_k. An element of the set of partial stars $P = \{P_1, \ldots, P_k\}$ corresponding to S is defined as follows.

$$P_k = \{\neg A_1\{O_{\neg k}\}, \neg A_2\{O_{\neg k}\}, \ldots, \neg A_n\{O_{\neg k}\}\} = \{P_{k1}, \ldots, P_{kn}\}. \quad (1)$$

Elements of eqn (1) are ranked based on sparseness (see Section 3.3). The result is an ordered list of partial stars where $^{\{n\}}$ indicates the rank.

$$P_{k\{r\}} = (P_k^{\{1\}}, P_k^{\{2\}}, \ldots, P_k^{\{n\}}). \quad (2)$$

The set of potential clustering descriptions is generated by taking the cross-product of all ranked partial stars.

$$\begin{aligned}\Phi &= \{P_{1\{r\}} \times P_{2\{r\}} \times \cdots \times P_{k\{r\}}\}. \\ &= \{(P_1^{\{1\}}, \ldots, P_k^{\{1\}}), (P_1^{\{1\}}, \ldots, P_k^{\{2\}}), \ldots, (P_1^{\{n\}}, \ldots, P_k^{\{n\}})\}. \\ &= \{\Phi_1, \Phi_2, \ldots, \Phi_p\}.\end{aligned} \tag{3}$$

The elements of eqn (3) are ordered by the sums of the ranks of the partial star elements of each potential clustering description. The list is traversed in rank order and the NID procedure applied at each step. The traversal stops when the evaluation of the final disjoint clustering fails to improve after a specified number of iterations.

2.3 Advanced features

The CLUSTER3 program provides three features that allow the system to dynamically optimize and present clustering results. First, the general algorithm can be applied recursively to each clustering result to generate a clustering hierarchy. This divisive method generates clustering descriptions at each step providing recursive levels of generalizations of the observations. Second, the exceptionless NID and active overlapping search processes allow the system to generate intersecting cluster descriptions. The former retains observations which cause intersection between cluster descriptions. The latter searches all permutations of adding an observation to one or more clusters. Lastly, the program accepts a range of values for use in determining an optimal number of clusters and independently performs a clustering analysis for each value in the range. The returned final clustering description is that which maximizes the clustering quality measure.

3 Goal-oriented features

3.1 Goal-orientation

Clustering data is sometimes described as discovering the *natural classes* in the data as in Cheeseman and Stutz [3], a perspective that assumes there exists only one set of natural classes to be discovered. In contrast, CLUSTER3 presumes the possibility of clustering observations in many different ways, depending on the underlying goals of the problem to be solved. For example, a doctor may cluster prescription drugs differently depending on whether the goal is recovery time or the number and severity of drug-induced side effects.

Two goal-oriented features are implemented in CLUSTER3. The view-relevant attribute subsetting (VAS) method is an attribute selection method that selects for clustering only those attributes relevant to a particular viewpoint. The Lexicographic Evaluation Functional (LEF) is used to evaluate the quality of clustering descriptions at each stage of the process.

Table 1: Viewpoint meta-attribute subset membership for attributes describing hockey players.

Attributes	Physical	Intellectual	Leadership
Skating Speed	$M_{Physical}$	$M^C_{Intellectual}$	$M^C_{Leadership}$
IQ	$M^C_{Physical}$	$M_{Intellectual}$	$M^C_{Leadership}$
Game Knowledge	$M^C_{Physical}$	$M_{Intellectual}$	$M_{Leadership}$
# Years Experience	$M^C_{Physical}$	$M_{Intellectual}$	$M_{Leadership}$

3.2 View-relevant attribute subsetting

A clustering analysis of datasets is often performed when there is limited knowledge about the underlying organization of the data. A user has, however, usually sufficient knowledge about the attributes that characterize the individual observations in the dataset to indicate their relevance for the intended clustering. Attribute selection (a.k.a. feature selection) is a commonly used method to reduce the dimensionality of a dataset, using search and evaluation techniques to reduce the attribute space to those attributes which are most relevant as is discussed in Vafaie and DeJong [4]. We expand this concept by considering subsets of attributes that naturally group together based on their correlation to a particular viewpoint, a process referred to as view-relevant attribute subsetting (VAS).

We define a *viewpoint meta-attribute* M_v for viewpoint v as a higher level attribute used to describe a viewpoint defining a subset of attributes relevant to this viewpoint. Each viewpoint meta-attribute M_v creates two subsets - the set of attributes from A directly relevant to viewpoint v and its complement.

$$M_v = \{A_1, A_2, \ldots, A_n\}. \tag{4}$$

$$M_v^C = A \setminus M_v. \tag{5}$$

Additionally we define a *viewpoint hyper-space* $HS(A_n)$ for each attribute A_n as the set union of meta-attribute subsets M_v and their complements M_v^C to which it is a member.

$$HS(A_n) = \{M_v | A_n \in M_v\} \cup \{M_v^C | A_n \in M_v^C\}. \tag{6}$$

To illustrate these concepts, consider the attributes used to categorize hockey players and the viewpoints (physical, intellectual, leadership) used to categorize the players' abilities presented in Table 1. The following definitions are drawn from the data in this table.

$$M_{Leadership} = \{GameKnowledge, \#YearsExperience\}. \tag{7}$$

$$M^C_{Leadership} = A \setminus M_{Leadership} = \{SkatingSpeed, IQ\}. \tag{8}$$

$$HS(SkatingSpeed) = \{M_{Physical}, M^C_{Intellectual}, M^C_{Leadership}\}. \tag{9}$$

In this example, *Game Knowledge* and *# Years Experience* attributes are relevant to the viewpoint of *Leadership*. Eqn (7) defines the viewpoint meta-attribute $M_{Leadership}$ as this relevant set of attributes. Its complement, $M^C_{Leadership}$ defined in eqn (8), is the set of attributes not relevant to the viewpoint *Leadership*. Finally, eqn (9) presents the viewpoint hyper-space for the *Skating Speed* attribute. Since *Skating Speed* is only relevant to the viewpoint *Physical*, its hyper-space consists of the meta-attribute for this viewpoint and the meta-attribute complements for the other viewpoints.

Viewpoint meta-attributes can be combined using set operators to produce many possible view-relevant attribute subsets. For example, given the goal of grouping hockey players based on their combined leadership and physical ability, the resulting set of attributes is $M_{physical} \cup M_{leadership}$; written explicitly in terms of attributes as $\{SkatingSpeed, GameKnowledge, \#YearsExperience\}$.

3.3 Lexicographic evaluation functional

A critical component of all clustering systems is the measure used to evaluate the quality of a clustering (a set of clusters). Partitioning methods traditionally calculate proximity to cluster centers using Euclidean distance as in Kanungo *et al*. [5]. Density methods, as discussed in Han and Kamber [6], evaluate the concentration of points in an ϵ-neighborhood to build clusters of arbitrary shape. For categorical data, Cheeseman and Stutz [3] use Bayesian statistics and Kim *et al*. [7] use modal values to determine optimum clustering results. CLUSTER3 defines various evaluation criteria to optimize certain facets of the resulting clustering description. These criteria are combined into a single measure using the Lexicographic Evaluation Functional (LEF).

Each criterion used in the evaluation is presented as a pair of values consisting of a measurement and tolerance τ. A clustering description is considered sub-optimal if the evaluation of any particular criterion is not within τ percent of the best evaluation of that criterion thus far. All criteria are defined such that the optimality of the measure improves as the value of the measure decreases.

k	Number of cluster descriptions in final clustering
C	Set of cluster descriptions
C_k	The k^{th} cluster description from set C
A	Set of attributes in domain space
$A(C_k)$	Set of explicitly specified attributes in cluster description C_k
O	Set of observations
$O(C_k)$	Set of observations covered by cluster description C_k
$V_{A_n}(O(C_k))$	Set of values for attribute A_n appearing in observations covered by cluster description C_k
$Area(C_k)$	The area of the event space covered by cluster description C_k

Figure 2: Shorthand notations used in LEF criterion formulae.

Table 2: Commonly used evaluation criteria available in CLUSTER3.

Criterion	Evaluation
Sparseness	$\sum_{C_i \in C} (Area(C_i) - O(C_i))$
Disjointness	$-\frac{1}{2} \sum_{C_i \in C} \sum_{C_j \in C} \sum_{A_n \in A} \begin{cases} 1 & if V_{A_n}(O(C_i)) \cap V_{A_n}(O(C_j)) = \emptyset \\ 0 & otherwise \end{cases}$
Commonality	$-\sum_{C_i \in C} \sum_{A_n \in A} \begin{cases} 1 & if A_n \in A(C_i) \\ 0 & otherwise \end{cases}$
Balance	$(\sum_{C_i \in C} \left\| O - \frac{O(C_i)}{k} \right\|)/k$
Relative Balance	$(\sum_{C_i \in C} \left\| \frac{Area(C_i)}{\sum_{C_j \in C} Area(C_j)} - \frac{O(C_i)}{O} \right\|)/k$

The notations specified in Figure 2 are used in the criteria formulae defined in Table 2. These criteria are a subset of all possible criteria available in the CLUSTER3 program.

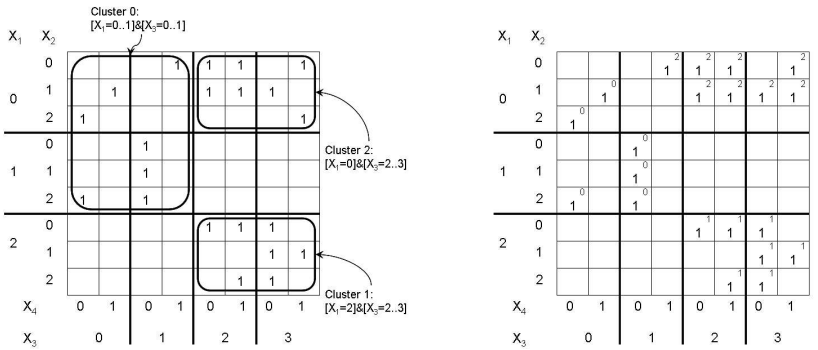
4 Experiments

4.1 Comparative testing of CLUSTER3 on a designed problem

A simple designed problem is used to illustrate the performance of CLUSTER3 in comparison to a conventional, similarity-based clustering program. The objects in the dataset to be clustered are described by four attributes with domains as follows: X_1=$\{0, 1, 2\}$;X_2=$\{0, 1, 2\}$; X_3=$\{0, 1, 2, 3\}$;X_4=$\{0, 1\}$. The dataset consists of 21 object descriptions (tuples) distributed in the space spanned over these attributes.

The dataset was tested against the implementation of Lloyd's algorithm in the KMlocal clustering application by Kanungo *et al* [5]. KMlocal was selected due to its efficient and modern implementation of the k-means algorithm. Lloyd's algorithm assigns observations to clusters using the minimum Euclidean distance between the observation and the cluster centroids. The KMlocal application was run with default parameters and 1000 stages. The CLUSTER3 program was run with default parameters and evaluation criteria of *balance* and *commonality*, with τ=10% for both criteria. The number of clusters was set to 3 for both applications.

The three simple disjoint cluster descriptions produced by CLUSTER3 are shown in the General Logic Diagram (GLD) Sniezynski *et al* [8] in Figure 3(a). A GLD is a visualization method for representing multi-attribute domain spaces in a planar grid. Each cell in the grid corresponds to a possible observation from the domain space; observed data elements are indicated by a cell value of "1". Rounded rectangles represent the resulting cluster descriptions. A similar GLD is presented in Figure 3(b) for comparison of the results of the KMlocal application. Rounded rectangles are excluded since the results are not generalized descriptions. Instead, the cluster number is indicated in the upper right-hand corner of the cell.



(a) CLUSTER3 results (b) KMlocal results

Figure 3: GLD representations of resulting clusters for designed problem.

The clusters produced by KMlocal and CLUSTER3 are similar with the exception of the observation (X_1,X_2,X_3,X_4) = (0,0,1,1). The main difference is that in addition to clusters, CLUSTER3 produced their description in the form of attributional conjunctions: $Cluster0 : [X_1 = 0..1]\&[X_3 = 0..1]; Cluster1 : [X_1 = 2]\&[X_3 = 2..3]; Cluster2 : [X_1 = 0]\&[X_3 = 2..3]$.

The descriptions produced by CLUSTER3 are general, in the sense that they not only include (cover) the observations in the dataset but also unobserved instances. This way a new observation (instance) can be easily classified by determining the description it matches. For example, a potential new observation (0,2,1,1) would be classified to Cluster 0 because it matches (satisfies) the description of that cluster. In contrast, in the case of KMlocal, it is not obvious whether this observation belongs to Cluster 0 or Cluster 2. To decide this, KMlocal would need to re-cluster the entire dataset with the new observation.

4.2 Performance of CLUSTER3 on a real-world problem

CLUSTER3 was applied to a larger real-world dataset consisting of the Senate voting record of the first session of the 108th United States Congress (Congressional Voting dataset), built from the 459 votes cast by the 100 members of the United States Senate. Each of the votes represents one attribute in the dataset. The possible values are Yea (vote for), Nay (vote against), Present (abstention), and ? (no vote cast). Additionally, viewpoint meta-attributes were defined based on the type of vote being cast (e.g. Bill Passage, Amendment, etc.).

Several experiments, shown in Table 3, were performed to determine what pivotal votes drive the congressional voting record. Pivotal votes are those that, when taken together, distinguish between resulting cluster descriptions. The first experiment attempted to discern the pivotal vote(s) resulting from clustering all votes. The result was a description whose pivotal vote, a motion to adjourn, is primarily irrelevant. The VAS and LEF methods were subsequently applied

Table 3: Results of adjusting VAS and LEF on Congressional Voting dataset.

Exp	Viewpoint Meta-Attributes	LEF Criteria (τ)	Pivotal vote(s) / Vote Topic
1	$M_{AllVotes}$	Sparseness (10%)	Vote 1: Motion to Adjourn
2	$M_{BillPassage}$	Sparseness (15%)	Vote 179: Tax Relief
3	$M_{BillPassage}$	Balance (10%) Relative Balance (15%)	Vote 51: Abortion
4	$M_{BillPassage}$	Disjointness (15%)	Vote 51: Abortion

to discern more relevant pivotal votes. The viewpoint meta-attributes selected correspond to legislative votes rather than administrative ones. The evaluation criteria used for the experiments evaluate sparseness, dissimilarity, and balance of the votes.

The experiments show that modifying VAS and LEF parameters provide insight into relevant pivotal votes. Experiments 2 and 3 only include legislative votes ($M_{BillPassage}$) but use different evaluation criteria, resulting in descriptions containing two different pivotal votes. Conversely, experiments 3 and 4 produce the same clustering descriptions using different criteria measures. Experiment 2 demonstrates that selecting clusters based on a Tax Relief legislation vote provides the clusters with the greatest density. Experiment 3 demonstrates that selecting clusters based on an Abortion legislation vote produces clusters that have the greatest voting record balance. The same result is achieved in experiment 4 when clustering is performed to minimize the number of values shared across all attributes. This leads to the possible conclusion that clustering Senators based on their Abortion votes results in clusters that have high voting dissimilarities between the clusters while still maintaining well-balanced clusters.

5 Conclusions

The CLUSTER3 program generates generalized descriptions of clusters representing meaningful concepts to facilitate understanding the organization of the observations comprising a dataset. Resulting concepts are output as conjunctive statements in attributional calculus produced by an iterative search and clustering quality evaluation process. A number of novel features have been described, including VAS and LEF, which facilitate clustering based on specification of the goals desired by the user and recommended criteria for judging the clustering quality.

The application of CLUSTER3 to a designed problem and comparison with the KMlocal similarity-based clustering method demonstrates its ability to discover simple concepts in the presence of irrelevant attributes. The capability of the VAS procedure for discovering meaningful clusters and clustering descriptions from the Congressional Voting dataset underscores its potential use in other areas. Some areas for future research consideration include remote sensing imagery, web pages, automobiles, and resumes.

Acknowledgements

The authors thank Kenneth Kaufman, Jarek Pietrzykowski, and Janusz Wojtusiak for their useful comments on earlier versions of this paper.

Much of the fundamental theory presented in this paper was developed through previous implementations of conceptual clustering [1, 9, 10]. These previous applications proved invaluable in the development of the current program.

The presented research was conducted in the Machine Learning and Inference Laboratory of George Mason University, whose research activities are supported in part by the National Science Foundation Grants No. IIS 9906858 and IIS 0097476, and in part by the UMBC/LUCITE #32 grant. The findings and opinions expressed here are those of the authors, and do not necessarily reflect those of the above sponsoring organizations.

References

[1] Michalski, R.S. & Stepp, R.E., Learning from observation: Conceptual clustering. Machine Learning An Artificial Intelligence Approach, eds. R.S. Michalski, J.G. Carbonell & T.M. Mitchell, Morgan Kaufman, chapter 11, pp. 331–364, 1983.

[2] Michalski, R.S., Attributional calculus: A logic and representation language for natural induction. Technical Report MLI 04-2, Reports of the Machine Learning and Inference Laboratory, George Mason University, Fairfax, VA, 2004.

[3] Cheeseman, P. & Stutz, J., Bayesian classification (autoclass): Theory and results. Advances in Knowledge Discovery and Data Mining, AAAI Press / MIT Press, chapter 6, pp. 153–180, 1996.

[4] Vafaie, H. & DeJong, K., Robust feature selection algorithms. Proceedings of the 1993 International Conference on Tools with AI., Boston, MA, 1993.

[5] Kanungo, T., Mount, D.M., Netanyahu, N., Piatko, C., Silverman, R. & Wu, A.Y., A local search approximation algorithm for k-means clustering. Proceedings of the 18th ACM symposium on computational geometry, 2002.

[6] Han, J. & Kamber, M., Data mining concepts and techniques. Morgan Kaufman, 2001.

[7] Kim, D., Lee, K.Y., Lee, D. & Lee, K.H., A k-populations algorithm for clustering categorical data. Pattern Recognition, **38**, pp. 1131–1134, 2005.

[8] Sniezynski, B., Szymacha, R. & Michalski, R.S., Knowledge visualization using optimized general logic diagrams. Proceedings of the Intelligent Information Processing and Web Mining Conference, IIPWM05, Gdansk, Poland, 2005.

[9] Stepp, R.E., Conjunctive conceptual clustering: a methodology and experimentation. Ph.D. thesis, University of Illinois, Urbana, 1984.

[10] Fischthal, S., A description and users guide for cluster2c++ a program for conjunctive conceptual clustering. Technical Report MLI 97-10, Reports of the Machine Learning and Inference Laboratory, George Mason University, Fairfax, VA, 1997.