University of Helsinki
Department of Computer Science
Series of Publications C, No. C–1996–2

Attribute-Oriented Induction and Conceptual Clustering

Oskari Heinonen and Heikki Mannila

Helsinki, February 1996 (Revised October 1996)

Department of Computer Science
P. O. Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki, FINLAND

Attribute-Oriented Induction and Conceptual Clustering

Oskari Heinonen[†] and Heikki Mannila[‡]

University of Helsinki, Department of Computer Science
P. O. Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki, FINLAND
E-mail: {Oskari.Heinonen, Heikki.Mannila}@cs.Helsinki.FI
Report C–1996–2, February 1996 (Revised October 1996), 7 pages

Abstract

Attribute-oriented induction is a method for knowledge discovery in databases that has recently been described and widely applied by Han et al. In this note we point out the close connection between attribute-oriented induction and conceptual clustering as presented by Pitt & Reinke, and describe how a conceptual clustering algorithm can be used to perform attribute-oriented induction.

Computing Reviews Categories and Subject Descriptors:

H.3.3 [Information Systems]: Information Storage and Retrieval – Information Search and Retrieval
I.2.6 [Computing Methodologies]: Artificial Intelligence – Learning
I.5.3 [Computing Methodologies]: Pattern Recognition – Clustering

General Terms:
Databases, machine learning, artificial intelligence.

Additional Key Words and Phrases:
Data mining, knowledge discovery, attribute-oriented induction, conceptual clustering.

[†]Work supported by the Finnish Technology Development Centre (TEKES) and the Ministry of Education.

[‡]This work was partially done while the author was visiting the Max Planck Institut für Informatik in Saarbrücken, Germany. Work supported by the Academy of Finland and the Alexander von Humboldt Stiftung.

1 Introduction

In a series of papers [4, 5] Jiawei Han and coauthors have introduced a knowledge discovery method called *attribute-oriented induction*, AOI for short. This method seems to be applicable in diverse situations, and it often produces good results. Han et al. have shown how AOI can be incorporated in a data classification process [3]. In the process, classification is performed in the event space that has been reduced by first applying AOI to the initial data relation.

In this note we point out that AOI itself can be viewed as a conceptual clustering method, as presented by Pitt & Reinke [6], with a specialized concept description language. Our conclusion is that the success of AOI is due partly to the general usefulness of conceptual clustering, and partly to the successful choice of the description language.

The paper is organized as follows. In Section 2 we briefly present the idea of AOI, and in Section 3 we define the conceptual clustering problem. The connection between the two methods is discussed in Section 4, while Section 5 is a short conclusion.

2 Attribute-oriented induction

We give here a simplified description of the attribute-oriented induction method; for details, see [4, 5]. Given a relation r over attributes $\{A_1, \ldots, A_n\}$, and for all i a generalization hierarchy H_i for the values of A_i, repeatedly perform the so called *attribute-oriented generalization step*, and combine rows that thus become equal. The attribute-oriented generalization step looks at the values of some attribute A_i of the relation r, and changes each value to the next level in the hierarchy H_i. As an algorithm, the process looks as follows.

```
begin
    while number of rows is higher than desirable do begin
        choose some attribute A;
        { perform generalization step for attribute A: }
        if more generalized levels exist in A's generalization hierarchy
            then replace the values of A with values of the next level;
        combine all rows of r that are equal;
    end;
end.
```

Table 1: Initial relation.

Color	*Diameter*
red	16
red	2
orange	7
yellow	34
green	25
blue	28
violet	8

Table 2: Generalized relation.

Color	*Diameter*
reddish	small
yellowish	large
bluish	large
bluish	small

Example 1 Consider a relation *Ball* over the schema *Ball*(*Color*, *Diameter*) shown in Table 1, and the generalization hierarchies for the attributes:

{red, orange} $\subset$ reddish
{yellow, green} $\subset$ yellowish
{blue, indigo, violet} $\subset$ bluish
{reddish, yellowish, bluish} $\subset$ ANY(*Color*)
{1..20} $\subset$ small
{21..40} $\subset$ large
{small, large} $\subset$ ANY(*Diameter*)

If the *Diameter* attribute is the first generalized, then the first two rows of the initial relation are immediately merged, as they are both changed to (red, small). If *Color* is the first attribute chosen, no immediate aggregation occurs.

By performing the generalization step once for both attributes, in either order, we get the generalized relation in Table 2. The relation could be further generalized by employing "ANY" values.

3 Conceptual clustering

In this section we present the conceptual clustering problem, following the exposition of [6] (see also [2]). Let D denote the class of all objects of interest. A description language $\mathcal{L} = (S, f)$ is a set S of expressions together with a mapping $f : S \to \mathcal{P}(D)$, where $\mathcal{P}(D)$ denotes the powerset of D. Thus for every description $\varphi \in S$, the function f associates the set of objects of D belonging to the class specified by φ.

The conceptual clustering problem is as follows. Given a subset $E \subseteq D$ of examples, find a set $\{\varphi_1, \ldots, \varphi_k\}$ of descriptions such that

(1) $\bigcup_{i=1}^{k} f(\varphi_i) = E$

(2) $f(\varphi_i) \cap f(\varphi_j) = \emptyset$ for all $1 \le i < j \le k$, and

(3) φ_i and φ_j are distant, far from each other, for each $1 \le i < j \le k$.

The definition of distance in (3) depends naturally on the description language $\mathcal{L}$ used.

The hierarchical agglomerative algorithm, also called the central agglomerative procedure [1], for conceptual clustering can now be formulated as follows.

1. Given $E = \{e_1, \ldots, e_n\}$, let φ_i be a concept such that $f(\varphi_i) \cap E = \{e_i\}$.
2. Let $\Phi_0 = \{\varphi_1, \ldots, \varphi_n\}$.
3. While Φ_h contains more than one cluster do:
   - Let $\varphi_i, \varphi_j \in \Phi_h$ be the clusters such that the distance between φ_i and φ_j is the smallest between any clusters within Φ_h.
   - Let new clustering $\Phi_{h+1} = \Phi_h \setminus \{\varphi_i, \varphi_j\} \cup \{\varphi'\}$. That is, replace φ_i and φ_j by φ', where φ' is the description such that $f(\varphi_i) \cup f(\varphi_j) \subseteq f(\varphi')$ and $f(\varphi')$ is the smallest set having this property.
4. Output the best of the clusterings $\Phi_0, \ldots, \Phi_{n-1}$, where "best" is selected according to a goodness criterion.

The *goodness* G of the clusterings is defined as a function of intracluster *tightness*, similarity within clusters, and intercluster *distance*, unlikeness between clusters. An exact formulation of various possibilities of defining goodness is given in [6]. The surprising result of [6] is that the above algorithm produces, under fairly general conditions, an optimal clustering, i.e., one that has the best possible value of goodness. The algorithm runs in polynomial time, assuming that the basic operations on descriptions can be done in polynomial time.

4 Attribute-oriented induction as conceptual clustering

Now we can show how AOI can be described as conceptual clustering. An instance of AOI consists of a relation r over schema $R(A_1, \ldots, A_n)$, where attribute A_i has domain D_i and generalization hierarchy H_i. Let G_i be the set of values occurring in H_i. Denote $E_i = D_i \cup G_i$, and let $D = E_1 \times E_2 \times \cdots \times E_n$. Then the generalized relations that AOI produces are subsets of D.

Next we define the corresponding conceptual clustering problem. The description language $\mathcal{L}$ consists of the expressions in D (the *generalized rows*). A row $\varphi = (c_1, \ldots, c_n) \in D$ denotes the set $f(\varphi)$ of objects of D, where $f(\varphi) = \{(a_1, \ldots, a_n) \mid a_i \leq_i c_i \text{ for all } i = 1, \ldots, n\}$, and $\leq_i$ denotes the specific-to-general relation of hierarchy H_i. That is, the generalized rows are the concepts, and the interpretation is provided by the generalization hierarchies.

The distance between two descriptions $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ can be defined as $\sum_{i=1}^{n} \delta_i$, where δ_i is the number of steps (the greater number in case of two different numbers) in the generalization hierarchy to the lowest, i.e. most specific, common ancestor of a_i and b_i. The tightness of a cluster can be defined based on the distance measure, for example as the distance between $(a_1, \ldots, a_n)$ and ("ANY", ..., "ANY"). This suggests that goodness, as presented by Pitt & Reinke [6], can be used as the overall quality function in the AOI process.

Now the instance of the conceptual clustering problem corresponding to the instance of AOI on relation r is defined as follows. It is easy to see that the operations which the basic AOI algorithm performs correspond exactly to the operations in the agglomerative clustering algorithm of Section 3:

- Choosing an attribute for generalization compares closely to selecting the closest clusters φ_i, φ_j. The choice of the attribute could actually be executed by selecting the closest rows, and then choosing the attribute, or several attributes, that need to be generalized in order to equalize the rows.

- Generalizing an attribute conforms to forming the description φ' for a new cluster, except that in basic AOI, performing the generalization step for an attribute causes an update of all the rows at once, i.e. several new descriptions are created.

- Combining equal rows is equivalent to forming a new clustering by adding the description φ' and subtracting descriptions φ_i and φ_j. The composition of several clusters can be affected at once in AOI.

As an algorithm, the process of conceptual clustering, incorporating the characteristics of the AOI process, looks as follows.

begin
    $h := 0$;
    $\Phi_h :=$ initial state of r;
    compute the goodness of Φ_h;
    while number of rows > 1 { *or "higher than desirable"* } **do begin**
        select the closest rows φ_i and φ_j;
        while φ_i and φ_j differ on some attribute A **do begin**
            replace the values of A with values of the next level;
        end;
        combine all rows of r that are equal;
        increase h by the reduction in the number of rows caused by
        the combination;
        $\Phi_h :=$ current state of r;
        compute the goodness of Φ_h;
    end;
    output the best of the clusterings $\Phi_0, \ldots, \Phi_h$ according to goodness;
end.

Example 2 shows the AOI process of Example 1 as conceptual clustering.

Example 2 The initial data relation in Table 1 corresponds to clustering Φ_0 in Figure 1. The subsequent clusterings are formed, when the generalization order of the attributes is *Diameter*, *Color*, *Diameter*, and finally *Color*. The process jumps over clusterings Φ_2 and Φ_5. Clustering Φ_3 corresponds to the generalized relation in Table 2.

5 Concluding remarks

We have shown that the successful data mining method of attribute-oriented induction can be viewed as conceptual clustering as described in, e.g., [6]. In retrospect this correspondence is not particularly surprising: both AOI

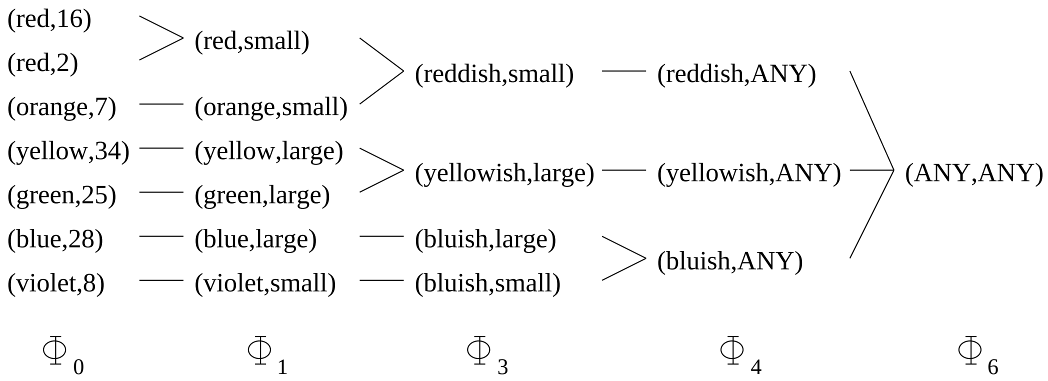

Figure 1: AOI in *Ball* relation as agglomerative hierarchical clustering.

and conceptual clustering are unsupervised methods producing disjunctive descriptions of their input.

It seems to us that this connection could be useful in both directions. One could, for example, study modifications of AOI where only two rows are modified at any given step, as in the basic conceptual clustering method. Furthermore, the quality measure(s) presented by Pitt & Reinke could give AOI a method for selecting the output generalized relation. Another possibility is to look at alternative implementations for AOI in terms of nonhierarchical clustering methods (see [1]). Conversely, conceptual clustering methods could benefit from looking at the successful choice of representation language in AOI.

References

[1] M. R. Anderberg. *Cluster Analysis for Applications*. Number 19 in Probability and Mathematical Statistics. Academic Press, New York, USA, 1973.

[2] D. Fisher and P. Langley. Approaches to conceptual clustering. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 691–697, Los Angeles, CA, USA, August 1985.

[3] J. Han, Y. Cai, and N. Cercone. Concept-based data classification in relational databases. In *Workshop Notes of 1991 AAAI Workshop on Knowledge Discovery in Databases (KDD'91)*, pages 77–94, Anaheim, CA, USA, July 1991.

[4] J. Han, Y. Cai, and N. Cercone. Knowledge discovery in databases: an attribute-oriented approach. In *Proceedings of the 18th VLDB Conference*, pages 547–559, Vancouver, British Columbia, Canada, 1992.

[5] J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Engineering*, 5(1):29–40, February 1993.

[6] L. Pitt and R. E. Reinke. Criteria for polynomial-time (conceptual) clustering. *Machine Learning*, 2(4):371–396, 1988.