In Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97) Workshop on "AI and Knowledge Management", AAAI Workshop Technical Report WS-97-09, pp. 15-24. July 27-28, Providence, Rhode Island, AAAI Press, Menlo Park, CA. 1997.

A Multistrategy Learning Approach to Flexible Knowledge Organization and Discovery

Seok Won Lee, Scott Fischthal[1] and Janusz Wnek[2]

Machine Learning and Inference Laboratory, M.S. 4A5
George Mason University
4400 University Drive
Fairfax, VA, 22030-4444
{swlee, sfischt, jwnek}@aic.gmu.edu

Abstract

Properly organizing knowledge so that it can be managed often requires the acquisition of patterns and relations from large, distributed, heterogeneous databases. The employment of an intelligent and automated *KDD* (Knowledge Discovery in Databases) *process* plays a major role in solving this problem. An integration of diverse learning strategies that cooperatively performs a variety of techniques and develops high quality knowledge can be a productive methodology for addressing this process. *AqBC* is a multistrategy knowledge discovery approach that combines *supervised inductive learning* and *unsupervised Bayesian classification*. This study investigates reorganizing the knowledge with the aid of an unsupervised Bayesian classification system, *AutoClass*. AutoClass discovers interesting taxonomies from databases. These taxonomies can also be represented as new attributes via constructive induction. The robust inductive learning system, *AQ15c*, can then learn useful concept relationships between knowledge objects. AqBC applied to two different sample problems yields not only simple but also meaningful knowledge due to the systems that implement its parent approaches.

Introduction

As a way of exploiting an organization's existing knowledge assets, it is essential to develop a wide variety of knowledge management (KM) solutions and strategies. In addition to the people and process aspects of knowledge management solutions are the technical aspects. An explosively growing subset of an organization's knowledge assets is often a collection of large scale databases that far exceed our ability to analyze them using standard tools. Required as part of the technical aspect of knowledge management solutions is a new approach for intelligent and automated knowledge discovery (Fayyad et al. 1996) and reorganization. By discovering and reorganizing this knowledge, it becomes easier to apply it to an organization's decision making.

We present AqBC (Lee 1996), a multistrategy knowledge discovery approach to concept learning. AqBC extracts and organizes new knowledge, determines meaningful descriptions and applies the newly acquired knowledge in supervised learning. These descriptions and knowledge grow out of patterns identified by AqBC. A clustering method using unsupervised Bayesian classification, generates the newly organized knowledge, while a supervised inductive rule learning system generalizes the descriptions and expresses them in variable valued logic. These new concepts expand the knowledge representation space for the supervised inductive learning system.

The system employs constructive induction to create and enhance the knowledge representation space with the aid of the unsupervised Bayesian classifier, AutoClass (Cheeseman et al. 1996). AutoClass provides a maximum posterior probability grouping objects into classes. The constructed classes define abstract concepts, with descriptions learned from class members using the inductive learning system, Aq15c (Wnek et al. 1995). The abstract concept descriptions are then used to improve and expand the original representation space. This expanded representation space serves as a final setting for supervised concept learning of any attribute from the original examples by employing Aq15c.

The multiple stage concept learning has the following properties:

- The task of inferring a set of classes and class descriptions that best fit and explain a given data set is placed on a firm theoretical foundation using Bayesian statistics.

[1] Also with Lockheed Martin Federal Systems, Gaithersburg, MD.

[2] Also with Science Applications International Corp., Tysons Corner, VA.

- The abstract concept descriptions learned in the first stage can illustrate and associate the corresponding concept descriptions learned in the second stage, which generates a set of simple descriptive rules. This way, the hierarchical hypotheses structures discovered from the nested classifications provide valuable information that cannot be obtained from either system alone.

The first MONK problem, MONK1, (Thrun et al. 1991) and US census data (obtained from the US Census Bureau's web page) have been used for experimentation. Other statistical classification methods (K-means centroid & Ward hierarchical clustering) were also applied to this data to compare and analyze the results. The diagrammatic visualization system, DIAV (Wnek 1995) graphically interprets the knowledge representation spaces and shows the changes in the representation space caused by constructive induction. In this paper, we show that the newly created knowledge facilitates classification and, in turn, problem solving that employs classification or pattern recognition in large databases.

The Significance of Improving the Knowledge Representation Space

Constructive Induction (CI) is a concept proposed in the field of inductive concept learning (Michalski 1978) to solve learning problems in which the original representation space is inadequate for the problem at hand and needs to be improved in order to correctly formulate the knowledge to be learned. In other words, constructive induction hypothesizes new knowledge using a search process. In our study, we search for the best representation space transformation by applying the unsupervised Bayesian classifier, AutoClass.

CI is based on the idea that the quality of the knowledge representation space is the most significant factor in concept learning. If the representation space is of high quality, i.e. the chosen attributes/descriptive terms are highly relevant to the problem at hand, learning will be relatively easy and will likely produce easily understood hypotheses with high predictive accuracy. If the quality of the representation space is low, (i.e. the attributes are less relevant to the problem) learning will be complex and no method may be able to produce good hypotheses. CI searches for patterns in data, learned hypotheses, and knowledge from experts, using them to create a new knowledge representation space (Wnek & Michalski 1994).

The illustration of this problem is given by (Arciszewski et al.1995). "Let us suppose that the problem is to construct a description that separates points marked by "+" from points marked by "-" (Figure 1A). In this case, the problem is easy because "+" points can be separated from "-" points by a straight line or a rectangular border.

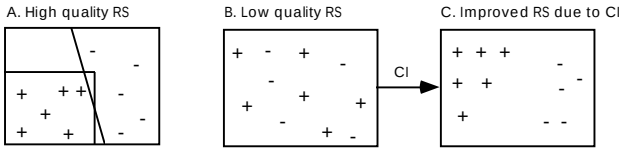


Figure 1. High vs. Low quality representation spaces for concept learning

Let us suppose now that "+"s and "-"s are distributed as in Figure 1B. In this case, "+"s and "-"s are highly intermixed, which may indicate that the representation space is inadequate for the problem at hand. A traditional approach is to draw complex boundaries that will separate these two groups. The CI approach searches for a better representation space, such as shown in Figure 1C, in which the two groups are well separated.

Conducting constructive induction thus requires mechanisms for generating new, more problem-relevant dimensions of the knowledge representation space (attributes or descriptive terms) as well as for modifying or removing less relevant dimensions from among those initially provided. In other words, a constructive induction system performs a problem-oriented transformation of the knowledge representation space. Once an appropriate representation space is found, a relatively simple learning method may suffice to develop a desirable knowledge structure (in this case, a description that separates the two groups of points)."

In order to find an appropriate representation space, clustering is an important way of summarizing and explaining data. A clustering system accepts a set of object descriptions (events, observations, facts) and produces a classification scheme over the observations. This system does not require an "oracle" to preclassify objects, but instead uses an evaluation function to discover classes that provide "good" conceptual descriptions. In our AqBC approach, the unsupervised Bayesian classifier, AutoClass, plays the clustering role and searches for the best model or classifications. Its evaluation function is based on Bayesian statistics. Once AutoClass finds the best model, AqBC creates a new attribute called *class* and augments the original data set with it. This new set of attributes is then passed to the second phase, which employs the supervised inductive learning system, AQ15c. The supervised learning should benefit from this new knowledge representation space for learning target concepts and produce more accurate descriptions of the target concepts.

To prove the quality of effectiveness of this approach and illustrate its properties, we applied AqBC to MONK1. The original representation space with the training examples denoted by "+" and "-" is shown in Figure 2A using DIAV.

Like most real world problems, the initial representation space is very disordered and ill-structured. In the improved representation space (Figure 2B), the examples representing the target concept are more ordered and therefore easier to learn. The positive examples are properly grouped in the middle value of the constructed attribute, while the negative ones group together in the first and last values. The descriptive rule sets learned by the AqBC approach are:

```
Positive-outhypo

1 [x7=2]

  Where
  class0 (x7=2) is:
  1 [x5=1]            (t:29, u:21)
  2 [x1=3] & [x2=3]   (t:17, u:13)
  3 [x1=2] & [x2=2]   (t:15, u:12)
  4 [x1=1] & [x2=1]   (t:9, u:8)

Negative-outhypo

1 [x7=1,3]

  Where
  class1 (x7=1) is:

  1 [x1=2..3] & [x2=1] & [x5=2..4] (t:20, u:20)
  2 [x1=2] & [x2=3] & [x5=2..4]    (t:6, u:6)

  class2 [x7=3] is:

  1 [x1=1] & [x2=2..3] & [x5=2..4] (t:31, u:31)
  2 [x1=3] & [x2=2] & [x5=2..4]    (t:5, u:5)

t is the total number of examples covered by a rule
u is the number of examples uniquely covered by the rule
```

In this case, **x7** is a new attribute that represents the clusters, *class0*, *class1* and *class2*, created by AutoClass. This new attribute augments the original attribute set and significantly changes the shape of the representation space. The resulting space can be divided by a single three-valued attribute, so the rules describing the concepts are trivial.

The combined approach successfully captures these meaningful subconcepts and thus successfully solves the problem (100% accuracy on both the test and training sets). Since most real world problems do not provide an "oracle" to guide the learning task, an unsupervised classifier is an attractive device to modify the representation space. We will discuss in later sections how AqBC addresses this issue in an application to US census data.

AqBC: A Multistrategy Approach for Constructive Induction-based Knowledge Discovery

In order to present a self-contained paper, we describe each module of this proposed approach below.

AQ15c Inductive Learning System

AQ15c (Wnek et al. 1995) is a C language reimplementation of AQ15 (Michalski et al. 1986). AQ-family of inductive learning programs implements the

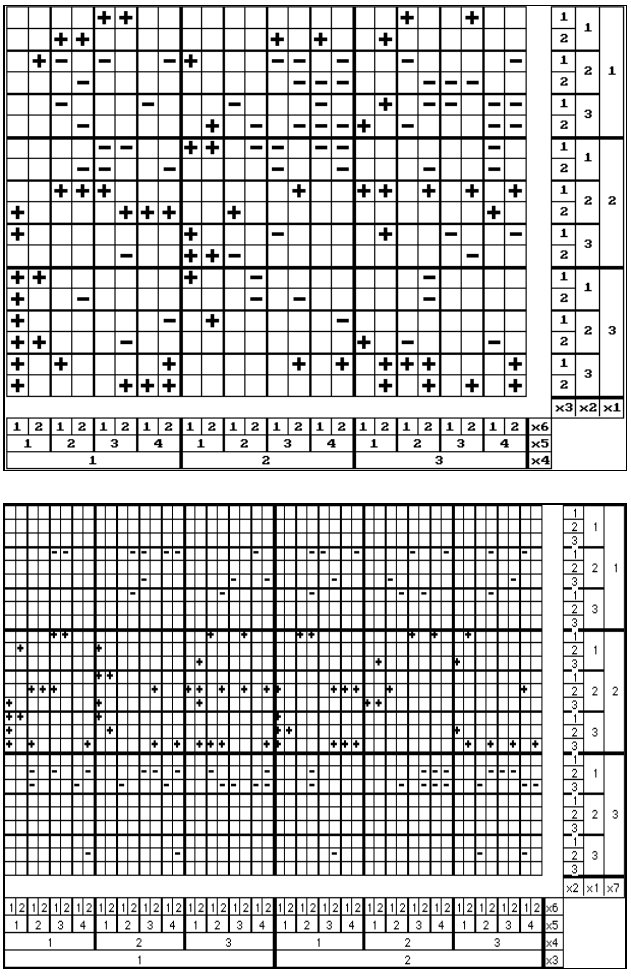


Figure 2. Diagrammatic visualization of the first MONK problem representation space: A) the initial representation space; B) the improved representation space due to the AqBC approach

STAR method of inductive learning (Michalski 1983a).It is based on the AQ algorithm for solving the general covering problem (Michalski 1969). AQ15c learns decision rules for a given set of decision classes from examples.AQ15c can generate decision rules that represent either *characteristic* or *discriminant* concept descriptions, or an intermediate form, depending on the settings of its parameters. A characteristic description consists of properties that are true for all objects in the concept. The simplest characteristic description is in the form of a single conjunctive rule. The most desirable characteristic rules have the longest conditions, listing as many common properties of objects of the given decision class as possible. A discriminant description consists of only the properties that discriminate a given concept from a fixed set of other concepts. The most desirable discriminant rules have the shortest conditions (Michalski 1983).

Concept Representation. Concept learning tasks strongly depend on the concept representation space and the representational formalism. A *concept representation space* is the set of all descriptors used in describing the concept. A *representational formalism* defines ways of constructing descriptions (i.e. a syntax) from descriptors. An example of a concept representation space is one in which descriptors are attributes with predefined sets of values. Considering attributes as dimensions spanning a multidimensional space, concept instances map to points in this space. An example of a representational formalism is predicate calculus with its set of logical operators. AQ15c is a program that learns concept descriptions from examples. It uses the Variable-Valued Logic system VL1 (Michalski 1973), a form of propositional logic, as its representational formalism that defines representation spaces in terms of attribute sets, where attributes may have multiple, discrete values. The representation space and the set of concepts for learning must be supplied by an oracle.

AQ15c Implementation. AQ15c acquires decision or classification rules from examples and counterexamples, and from previously learned decision rules. When learning rules, AQ15c uses 1) background knowledge in the form of rules (input hypotheses), 2) the definition of descriptors and their types and 3) a rule preference criterion that evaluates competing candidate hypotheses. Each training example characterizes an object, and its class label specifies the correct decision associated with that object. The generated decision rules are expressed as symbolic descriptions involving relations between objects' attribute values. The program performs a heuristic search through a space of logical expressions, until it finds a decision rule that best satisfies the preference criterion while covering all positive examples, but no negative examples.

Knowledge Representation. Each *decision class* in the training set is defined by a set of events. For each class the algorithm generates the best decision rule set (according to user-defined criteria) that is complete and consistent with respect to the input events. A *complete* rule set is one that covers all of the positive examples. A *consistent* rule does not cover any negative examples. The user may provide initial decision rules to the program. These rules are treated as initial hypotheses. Each decision rule is described by one or more conditions, all of which must be met for the rule to apply.

A *condition* is a relational statement. A *rule* is a conjunction of conditions. A *hypothesis* is a disjunction of rules that together describe a concept. The following is an example of a hypothesis consisting of two rules:

```
Flag-outhypo
 1 [color = red, white, blue] & [stripes = 13]
 & [stars = 50]
 2 [color = red, white, blue] & [stripes = 3]
```

A hypothesis is satisfied if any of its rules are satisfied, while a rule is satisfied if all of its conditions are satisfied. A condition is satisfied if the term takes one of the values in the reference. The hypothesis shown in the above example can be interpreted as follows: An object is a flag if: its color is red, white, or blue, and it has 13 stripes and 50 stars, **OR** its color is red, white, or blue, and it has 3 stripes.

AutoClass: An Unsupervised Bayesian Classifier

AutoClass is an unsupervised Bayesian classification system that looks for a maximum posterior probability classification (Cheeseman et al. 1996). The system infers classes based on Bayesian statistics, deriving a belief network via probability theory.

The idea of accumulating and applying auxiliary evidence here can be mapped into the constructive induction mechanism that employs a new attribute which summarizes the data patterns. The new attribute's degree of belief is very high because it is generated from the best model of Bayesian classification. Therefore, this new attribute can potentially reorganize and improve the knowledge representation space. The theory of how Bayesian learning is applied in AutoClass, summarized from (Hanson et al. 1991), is described below.

Let E denote a set of *evidence* and H a set of possible *hypotheses* that can conceptualize a set of combinations in E. Assume that the sets of possible evidence E and possible hypotheses H are mutually exclusive power sets.

P(*ab*|*cd*) represents a real number probability describing a *degree of belief in the conjunction of propositions a* and *b* conditioned on the assumption that the given propositions *c* and *d* are true. Let $\pi(H|E)$ denote a *posterior probability* describing a belief in H after observing evidence E. Let $L(E|H)$ denote a *likelihood* containing a theory of how likely it would be to see each possible evidence combination E in each possible set of hypotheses H. Beliefs are non-negative ($0 \le P(a|b) \le 1$) and normalized ($\sum_H \pi(H) = 1$ and $\sum_E L(E|H) = 1$). The combination of likelihood and priors produces a *joint probability* $J(EH) \equiv L(E|H)\pi(H)$ of both *E* and *H*. Bayes's rule shows how beliefs should change when evidence is obtained by normalizing the joint probability.

$$\pi(H|E) = \frac{J(EH)}{\sum_H J(EH)} = \frac{L(E|H)\pi(H)}{\sum_H L(E|H)\pi(H)}$$

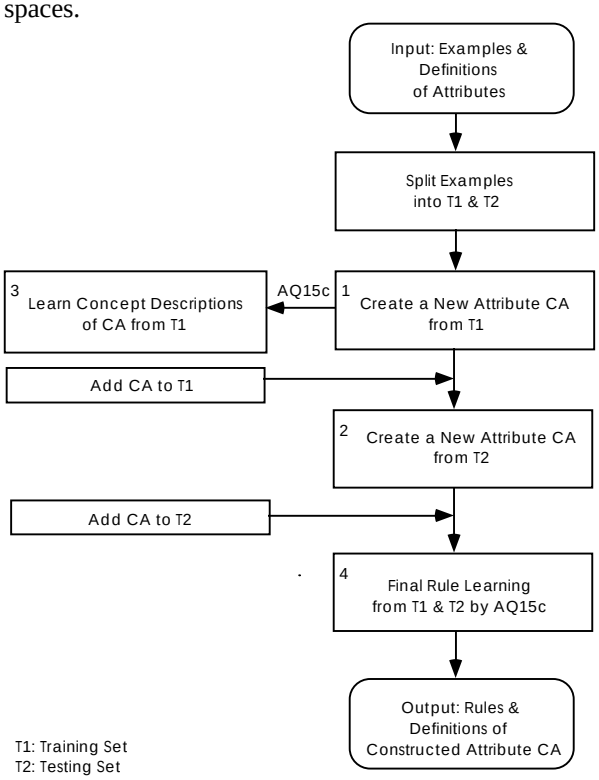
A possible set of hypotheses *H* with an associated likelihood function in any given situation indicates what evidence we expect to observe when the chosen set of hypotheses is true. A set of prior probabilities corresponding to this set of hypotheses should be obtained with a set of evidence. Bayes's rule then specifies the appropriate posterior beliefs about the hypothesis. These posterior probabilities and a *utility over hypotheses*, *U(H)*, which describes preferences for each individual hypothesis, can thus be combined to obtain the maximum expected utility:

$$EU(A) = \sum_H U(H)\pi(H|EA)$$

AutoClass discovers various kinds of knowledge including classifications and causal mechanisms between cases. In this work, AqBC discovers the classifications that can be used to improve the knowledge representation space.

The AqBC Multistrategy Learning Methodology

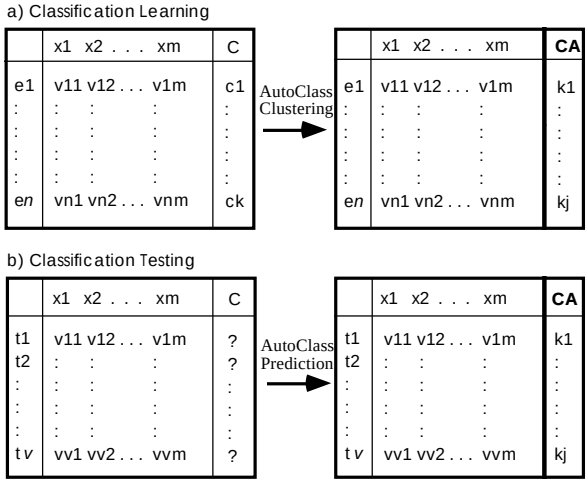
AqBC is a multistrategy knowledge discovery approach that combines unsupervised Bayesian classification with supervised inductive learning. Figure 3 shows the general structure of the AqBC approach. AqBC can be applied to two different goals. First, AutoClass provides classifications that help the user generate "expert knowledge" for a potential target concept when a data set without predefined classifications is given to the supervised inductive learning system. The inductive learning system AQ15c learns the concept descriptions of these classifications (step3 in Fig. 3). Second, when the data set is already divided into classes, AqBC repeatedly searches for the best classifications using AutoClass, then uses the best one for creating and modifying more suitable and meaningful knowledge representation spaces.



T1: Training Set
T2: Testing Set

CA: New attribute with values that are classes generated by AutoClass
In step 1, concept attribute (C) is included in T1. In step 2, C is not included in T2.

Figure 3. General structure of AqBC approach



C: Concept Attribute
CA: New attribute with values that are classes generated by AutoClass

Figure 4. Creating a new attribute for a)training and b)testing data set by use of AutoClass, where,

vnm : a value of attribute xm of nth example

$c1 \le C \le ck$: C is a Target Concept Attribute.
k is the # of target concepts.

$k1 \le CA \le kj$: CA is a classification label generated by AutoClass
j is the # of classifications.

Table 1. Definitions of attributes

1. % Asian Residents
2. % Black
3. Population Density
4. % Elderly Residents
5. % Foreign Residents
6. % Residents speaking foreign language
7. Population Growth
8. Land Area
9. Population
10. Infant Mortality
11. Crime Rate
12. % Single Dwellers
13. % Single Parents
14. % Hispanic Residents
15. July Temperature (median)
16. Precipitation (annual, inches)
17. % Jobs held are in manufacturing
18. Change in labor force from 1980
19. % Females in workplace
20. % Commuters using public transit
21. Unemployment %
22. Rent
23. % Dwellings that are condominiums
24. Home value
25. % Population who rent
26. Poverty rate
27. % Housing units built before 1939
28. Income
29. % Residents on public assistance
30. % of children in elementary or secondary school
31. % Adults over 55
32. % Holding bachelors degrees
33. Has major league baseball team
34. Has NFL football team
35. Has NBA basketball team
36. Is a state capital

Figure 4 shows the constructive induction process in which a new attribute is created for training and testing. In the first phase, a new attribute, in which the values are class labels generated by AutoClass, is added to the data table under the name of *CA* (Step 1 & 2 in Fig. 3). Note that one of the attributes provided to AutoClass is the original target concept *C* (Fig. 4-a). However, *C* is not included when we are dealing with the testing data set (Fig. 4-b) since the system is not supposed to know the given class.

The modified data table generated by this phase, which now includes *CA*, is used by AQ15c to learn the new concept description for *C*. In other words, the concept description of the original classification is learned from the modified knowledge representation space. The only change is the augmentation of the table with the new attribute. It is at this time that a separate AQ15c run generates the concept descriptions for the classes represented by *CA*.

In the second phase, the testing data set, an unseen data set with the original concept attribute *C* containing the value "?" (don't care) for all examples, is input to the AutoClass classifier trained in the first phase (see Figure 4b). By not providing the original classifications, the original properties of the testing data set can be preserved and AutoClass's ability to correctly classify previously unseen examples can be verified. The technique of constructing classification labels is based on previous constructive induction methodology (Wnek & Michalski 1994).

The improvement of the knowledge representation space is already demonstrated with the MONK1 problem in the previous section. Such an approach can potentially allow us to address large, complicated data sets not recognized well by a supervised inductive learner alone, by having an unsupervised Bayesian classifier as an oracle. Then the robust supervised inductive learner learns the target concept descriptions, providing clear, understandable rules describing these classes. Thus, the system takes advantage of the strengths of its component subsystems.

An Experimental Application to the US Census Data

US Census Data

The above demographics database (Table 1) is adapted from US census data on all US cities with a 1990 population of over 200,000 residents. 77 records containing the above 36 attributes were used for the experiments.

The sizes of the domains of these attributes vary widely. Some are binary, such as the last four attributes; others have as many as 70 different linear values. Most of the attributes have 10-20 values. The attributes also vary in their distribution; some of the attributes have outliers (for example, the population attribute has a long thin tail on which New York, Chicago and Los Angeles reside). Some of the distributions are normal, while others are uniform. The dataset thus introduces a wide variety of problems for any system that must deal with it, particularly in view of the small number of examples. Given such a dataset, an important problem is how to organize it in such a way so that useful knowledge can be extracted from it.

Learning Concept Descriptions from AutoClass classifications

AQ15c learns the concept description of the classifications obtained from the AutoClass (attribute *CA*), as described earlier. The following rules describe the two values of the *CA* attribute, representing the two concepts discovered by AutoClass. Note that, for all rules presented below, the values have been mapped back to the original data from the normalized values presented to the system.

```
class0-outhypo

1   [black = 4..76%] & [elderly = 4..18%] &
    [pop_growth < 32%] & [population > 385,000]
    (t:41, u:26)
2   [manufacturing_jobs = 7..27%] & [rent <
    $475/mo] & [poverty = 9..17%]   (t:20, u:5)
```

class0 appears to represent cities with moderate to large populations and without explosive growth OR with low average monthly rent and moderate poverty rates. When we look at the actual cities that fall into this class, we see most are cities that are very large, with stable or declining populations – particularly those of the Eastern seaboard, the old South and the Rust Belt. Of the 32 largest cities in the US, 30 fall into this class. The only exceptions are the high tech, rapidly growing San Jose, and El Paso, which is also growing extremely rapidly, due primarily to Hispanic immigration.

```
class1-outhypo

1   [foreign_speak = 4..70%] &
    [land_area < 242 sq miles] &
    [renters = 0..21%] & [old_housing = 0..19%]
    & [income = $22K..46K]       (t:22, u:19)
2   [population = 326K..385K] &
    [infant_mortality = 6..18] &
    [renters = 6..40%] & [state_capital = NO]
    (t:11, u:8)
3   [black = 43%] & [foreign < 19%]   (t:1, u:1)
```

class1 appears to represent small cities with relatively low rental rates or cities with moderately small populations that aren't state capitals. Most of these cities are either in the Sun Belt or are small cities outside of the old South and Rust Belt.

A weakness of this second class is that it has a conjunction that covers only one example. It has been found that rule truncation (in which less important selectors are removed) can often produce more general and effective results (Zhang & Michalski 1989), often by removing conjunctions that essentially cover exceptions (such as the one city covered by the third conjunction in this class). This class is a strong candidate for the application of such a method.

While AutoClass itself ranked the two cluster solution higher, for comparison we also analyzed the best three cluster solution (not shown here) discovered by AutoClass. We observed that this lower rated solution was, in fact, less attractive than the two cluster solution. For example, in the three cluster solution, one cluster has conjunctions which appear to contradict each other, with one conjunction indicating cities with low foreign speaking populations and another indicating cities with very high Hispanic populations. For this reason, we chose to use the two cluster solution shown above.

Using AqBC for Knowledge Discovery

AQ15c learns the abstract concept descriptions of the given classifications from AutoClass in the first phase. Now we augment the original knowledge representation space with the "class" label attribute, allowing AQ to learn new knowledge which was difficult to extract from the original representation space. We can now choose our target concepts from among any of the system attributes. In this case, we choose two concepts based on the *population* attribute, where these concepts, *large cities* and *moderately sized cities*, are represented by two classes. The first class is population over 385,000 and the second class is population from 200,000 to 385,000.

The following experiments use the constructed attribute from the two class clustering.

```
population_0-outhypo

1   [black=7..75%] & [foreign = 2..59%] &
    [home_value = 2..32] & [class=0](t:34, u:24)
2   [pop_density > 2250] & [single = 24..31%] &
    [july_temperature = 74..93F] &
    [precipitation < 56] &
    [manufacturing_jobs < 24%] &
    [change_in_labor < 43%] & [renters < 56%]
    (t:14, u:9)
3   [foreign = 4..27%] & [foreign_speak > 1%] &
    [pop_growth < 32%] & [july_temperature =
    0..24F] & [precipitation = 4..43] &
    [change_in_labor > -7%] & [renters < 56%]
    (t:11, u:4)
4   [single_parent = 17%]        (t:1, u:1)

population_1-outhypo

1   [hispanic > 66%] & [manufacturing_jobs =
    6..26%] & [class=1]          (t:21, u:15)
2   [pop_density < 6750] & [land_area < 129] &
    [manufacturing_jobs < 22%] & [unemployment ≠
    very high] & [renters = 41..61%] &
    [holds_bachelors < 28%] & [has_nba_team =
    NO]                          (t:9, u:5)
3   [single = 20..23%]           (t:5, u:1)
```

The constructed attribute separates large cities from small ones almost perfectly. As stated above, 30 of the 32 largest cities (and 34 of the 38 largest cities that fit into population class 0) are labeled 0 for the constructed attribute "class". This greatly simplifies the rule learner's

task for discovering simple rules representing the target concept "population". As with the constructed attribute, rule truncation would further simplify the resulting rules; the last conjunction of both rules has little generalization value and would likely be removed.

The Census database has a much larger attribute set than the MONK1 problem, and its attributes also have a wider range of values. Because of this, the Census database provides a somewhat more realistic case study. More than half the attributes of the database are dropped altogether in the final rules, while still capturing the target concept. In addition, the constructed attribute plays a major role in discriminating the two classes; the **class** attribute is part of the most powerful conjunction in each class description. Thus, constructive induction turns out to be important to understanding the target concept.

Comparison with the Statistical Approaches

Statistical techniques in general pose important usability issues, in addition to the problem of comprehending the results. Resulting clusters are often difficult to interpret, due to the lack of a symbolic description. Human intervention and context dependency are both often required to generate good results. Furthermore, standard statistical techniques rely on distance measures, which are not always appropriate for all problems. The descriptive rules generated by AQ15c make the clusters generated by the Bayesian classification substantially more understandable.

For comparison, we applied standard statistical clustering methods to the data in place of AutoClass as the unsupervised classification engine. The two approaches compared are a simple K-means centroid method and a Ward hierarchical clustering (Sharma 1996). The Ward clustering, shown in the dendrogram of Figure 5, indicates an optimal number of clusters as three. The class attribute constructed by K-means has three classes without any clear meaning. The clusterings produced by the two methods are strikingly similar, but not especially meaningful. More importantly, the constructed attribute had very little effect on the rules learned by AQ15c (see Figure 6). In the K-means clusterings, the new attribute, **class**, is only used in the second conjunction of the second population class (population_1). In the Ward clustering (not shown here, but with similarly unclear rules), the **class** attribute is not used at all in the final target concept descriptions.

Another important point to consider is that, even when these statistical approaches are used to cluster and reorganize knowledge bases and databases, our approach can use AQ15c to generate descriptive rules exposing the nature of the generated classes. While the rules shown in Figure 6 do not seem to imply an interesting structure created by the statistical clustering, they do provide us with a great deal more insight about these clusters than would simply looking at the class member lists alone. The use of a conceptual clustering tool, such as Cluster/2 (Michalski & Stepp 1983b) could also prove helpful as the unsupervised classification system, since it generates descriptive rules as part of its clustering algorithm.

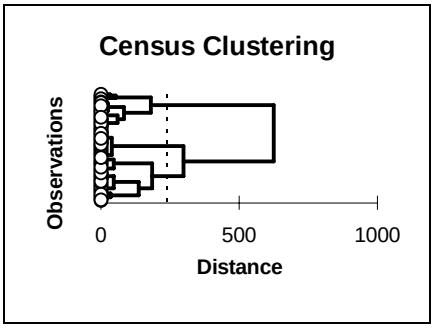


Figure 5. Ward clustering of US census data

```
class1-outhypo
1 [single_parent < 36%] & [hispanic < 29%] &
  [unemployment < very high] & [rent < $600] &
  [old_housing = 0..51%]      (t:36, u:36)

class2-outhypo
1 [single_parent > 30%] & [hispanic < 36%] &
  [use_public_transit > 3%]   (t:24, u:24)

class3-outhypo
1 [black = 0..26%] & [foreign_speak = 22..70%] &
  [old_housing < 57%]         (t:17, u:17)


population_0-outhypo

1 [elderly < 15%] & [hispanic < 70%] &
  [change_in_labor = -6..35] &
  [use_public_transit = 3..34%] & [rent > 325] &
  [renters < 72%] & [old_housing = 3..58%]
  (t:30, u:9)
2 [foreign > 2%] & [single > 23%] &
  [precipitation = 4..39] & [manufacturing_jobs
  < 16%] & [public_asst > 4%]       (t:29, u:9)
3 [land_area = 91..774] & [precipitation =
  27..48] & [holds_bachelors = 13..27%]
  (t:18, u:4)
4 [single_parent = 39..40%]         (t:6, u:1)

population_1-outhypo

1 [elderly < 18%] & [land_area < 129] & [single
  < 40%] & [single_parent < 44%] &
  [precipitation = 6..60] &
  [renters = 42..58%] & [in_school =
  moderate..high] & [has_nba_team = NO]
  (t:16, u:14)
2 [foreign < 60%] & [manufacturing_jobs < 22%] &
  [change_in_labor > 0] & [condo > 1%] &
  [poverty > 6%] & [has_nfl_team = NO] &
  [has_nba_team = NO] & [state_capital = NO] &
  [class=1..2]           (t:11, u:9)
3 [pop_density < 3750] & [old_housing = 12..18%]
  (t:4, u:2)
```

Figure 6. K-means clusterings

Conclusions

This paper presents a multistrategy approach for flexible knowledge reorganization and discovery that integrates supervised AQ inductive rule learning with unsupervised Bayesian classification for creating a meaningful knowledge representation space and discovering high quality knowledge. The patterns acquired by the Bayesian classifier make little sense until AQ provides an understandable descriptive representation that encompasses new knowledge. In other words, the data and information presented to AqBC becomes knowledge when the patterns acquired by the system are realized into understandable concepts. This new knowledge can then be applied in organizational decision making.

The AqBC approach uses the methodology of constructive induction, which modifies and improves the representation space with the aid of classifications obtained from the unsupervised Bayesian classifier, AutoClass. This methodology was applied to the MONK1 problem and a US census dataset. In the experiments, AqBC constructed new attributes used in relatively simple, meaningful rules describing the target concepts. The problems inherent in the nonsymbolic nature of both the Bayesian and the distance based statistical techniques are often mitigated by applying AQ to learn descriptive representations for both the resulting clusters and the target concepts. This multistrategy approach produces important new knowledge organization as it creates human-understandable class descriptions and new attributes that simplify those class descriptions.

A key result of this research is the generation of a simple, descriptive taxonomy for a database that can be used for sensibly partitioning into smaller databases. Another potential experiment would be to use the taxonomy produced via the AqBC method to perform supervised classification with many different decision attributes from the current set rather than just the population attribute. Other future research will focus on developing new strategies combining various statistical (Sharma 1996) and conceptual (Michalski & Stepp 1983b,c) classification methods for constructing better knowledge representation spaces from large data sets.

Not only Bayesian classifiers benefit from the use of AQ15c to learn the descriptive representations of the generated classes. In addition to the distance-based clustering methods described above, other subsymbolic systems such as SOFMs (Kohonen 1995) and k-NN methods can be employed as the unsupervised classification engine. Multiple engines may also be used to perform additional constructive induction prior to unsupervised classification. For example, AQ17-DCI (Bloedorn & Michalski 1991) and AQ17-HCI (Wnek & Michalski 1994) construct new features based on interrelationships among existing ones.

Varying the rule learner based on the application may also prove productive. If mathematical formulae are desired instead of conjunctive rules, a system such as ABACUS (Falkenhainer & Michalski 1990) could be employed in place of AQ15c. There are still many challenging real world applications to which this multistrategy approach could be applied for new knowledge discovery.

Acknowledgments

The authors thank Dr. Ryszard Michalski for his helpful comments and criticism. In addition, the authors thank the AutoClass group for making their software readily available. The first author greatly appreciates the support of a Doctoral Fellowship from the School of Information Technology and Engineering at George Mason University. This research was conducted in the Machine Learning and Inference Laboratory at George Mason University. The Laboratory's research activities are supported in part by the National Science Foundation under grants IRI-9020266 and DMI-9496192, in part by the Advanced Research Projects Agency under grants F49620-92-J-0549 and F49620-95-1-0462, administered by the Air Force Office of Scientific Research, and in part by the Office of Naval Research under grant N00014-91-J-1351.

References

Arciszewski, T., Michalski, R. S., and Wnek, J. 1995. Constructive Induction: the Key to Design Creativity. Reports of the Machine Learning and Inference Laboratory, MLI 95-10, George Mason University, Fairfax, VA.

Bloedorn, E., and Michalski, R. S. 1991. Constructive Induction from Data in AQ17-DCI: Further Experiments. Reports of the Machine Learning and Inference Laboratory, MLI 91-12, George Mason University, Fairfax, VA.

Cheeseman, P., and Stutz, J. 1996. Bayesian Classification (AutoClass): Theory and Results. In *Advances in Knowledge Discovery and Data Mining*. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R, eds.: AAAI Press, Menlo Park.

Falkenhainer, B. C., and Michalski, R. S. 1990. Integration Quantitative and Qualitative Discovery in the ABACUS System. In *Machine Learning: An Artificial*

Intelligence Approach, Vol. III. Kodratoff, Y., and Michalski, R. S., eds.: Morgan Kaufmann, San Mateo.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. eds. 1996. *Advances in Knowledge Discovery and Data Mining.*: The AAAI Press, Menlo Park.

Hanson, R., Stutz, J., and Cheeseman, P. 1991. Bayesian Classification Theory, Technical Report FIA-90-12-7-01, NASA Ames Research Center, Artificial Intelligence Branch.

Heckerman, D. 1990. Probabilistic interpretations for Mycin's certainty factors. In *Readings in Uncertain Reasoning*. 298-312. Shafer, G., and Pearl, J. eds.: Morgan Kaufmann, San Mateo.

Kohonen, T. 1995. *Self Organizing Maps.*: Springer-Verlag, Heidelberg

Lee, S. W. 1996. Multistrategy Learning: An Empirical Study with AQ + Bayesian Approach, Reports of the Machine Learning and Inference Laboratory, MLI 96-10, George Mason University, Fairfax, VA.

Michalski, R. S. 1969. On the Quasi-Minimal Solution of the General Covering Problem. In Proceedings of the International Symposium on Information Processing (FCIP 69), Vol. A3 (Switching Circuits), 125-128, Yugoslavia, Bled, October 8-11.

Michalski, R. S. 1973. AQVAL/1-Computer Implementation of a Variable-Valued Logic System VL1 and Examples of its Application to Pattern Recognition. In Proceedings of the First International Joint Conference on Pattern Recognition, Washington D. C.

Michalski, R. S. 1978. Pattern Recognition as Knowledge-Guided Computer Induction, Technical Report No. 927, Department of Computer Science, University of Illinois, Urbana-Champaign.

Michalski, R. S. 1980. Knowledge acquisition through conceptual clustering: A theoretical framework and algorithm for partitioning data into conjunctive concepts. *International Journal of Policy Analysis and Information Systems*, 4: 219-243.

Michalski, R. S. 1983a. A Theory and Methodology of Inductive Learning. *Artificial Intelligence*. Vol.20, 111-116.

Michalski, R.S., and Stepp, R. E. 1983b. Learning from Observation: Conceptual Clustering. In *Machine Learning: An Artificial Intelligence Approach*. Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. eds.: Morgan Kaufmann, San Mateo.

Michalski, R. S., and Stepp, R. E. 1983c. Automated Construction of Classifications: Conceptual Clustering Versus Numerical Taxonomy. *IEEE Transactions on PAMI* 5:4, 396-410.

Michalski, R. S., Mozetic, I., Hong, J., and Lavrac, N., 1986. The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. In Proceedings of AAAI-86, 1041-1045, Philadelphia, PA.

Michalski, R. S., and Kaufman, K. A. 1997. Data Mining and Knowledge Discovery: A Review of Issues and a Multistrategy Approach. In *Machine Learning and Data Mining: Methods and Applications.* Michalski, R. S. Bratko, I., and Kubat M. eds.: John Wiley & Sons. Forthcoming**.**

Sharma, S. 1996. *Applied Multivariate Techniques.*: John Wiley, New York.

Thrun, S. B., et al. 1991. The MONK's problems: A Performance Comparison of Different Learning Algorithms. Carnegie Mellon University.

Wnek, J., Kaufman, K., Bloedorn, E., and Michalski, R. S. 1995. Inductive Learning System AQ15c: The Method and User's Guide. Reports of the Machine Learning and Inference Laboratory, MLI 95-4, George Mason University, VA.

Wnek, J., and Michalski, R. S. 1994. Hypothesis-driven Constructive Induction in AQ17-HCI: A method and experiments. *Machine Learning,* Vol.14, No.2, 139-168.

Wnek, J. 1995. DIAV 2.0 User Manual: Specification and Guide through the Diagrammatic Visualization System, Reports of the Machine Learning and Inference Laboratory, MLI 95-4, George Mason University, Fairfax, VA.

Zhang, J., and Michalski, R. S. 1989. Rule Optimization Via SG-TRUNC Method. In Proceedings of the Fourth European Working Session on Learning.