Discovering Classification Knowledge in Databases Using Rough Sets

Ning Shan, Wojciech Ziarko, Howard J. Hamilton and Nick Cercone

Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
E-Mail: {ning,ziarko,hamilton,nick}@cs.uregina.ca

Abstract

The paper presents an approach to data mining involving search for complete, or nearly complete, domain classifications in terms of attribute values. Our objective is to find classifications based on interacting attributes that provide a good characterization of the concept of interest by maximizing predefined quality criteria. The paper introduces the notion of the classification complexity and several other measures to evaluate quality.

Introduction

Knowledge discovery in databases (KDD) or data mining has recently attracted great interest and research activity (Fayyad et al. 1996; Piatetsky-Shapiro & Frawley 1991; Ziarko 1994). One goal is the discovery of relationships between data values. The standard approach to this problem, as discussed by many authors, is to attempt to induce classification rules from data using inductive algorithms (Fayyad et al. 1996; Piatetsky-Shapiro & Frawley 1991; Shan et al. 1995a, 1995b; Ziarko 1994; Ziarko & Shan 1994). A *classification* is a partition of the objects in the domain into equivalence classes based upon the condition attributes. The rules, which have the format "if *conditions* then *decision*," normally properly reflect relationships occurring in the database. Unfortunately, the validity of rules in the domain from which the database was collected is typically questionable.

We observe that (a) databases are often highly incomplete, that is, in the selected attribute-value representation many possible tuples are not present in the database; and (b) in particular, insufficient tuples exist in the database to reliably estimate the probability distributions. As long as we lack sufficient evidence about the completeness of the classification and the reliability of the associated probability estimates, rule extraction from the data and the application of these rules to the original domain is of limited utility. In our approach, rule induction, although important, is only applied after ensuring the completeness of the classification and the reliability of the probability estimates.

We perform three steps prior to rule induction. First, we generalize the condition attributes as necessary to increase the credibility of the classification. Then we identify clusters of interacting attributes, *i.e.*, those with direct or indirect dependencies. Finally, we search for credible classifications of the database tuples based on these clusters. A classification is *credible* if it is complete or almost complete with respect to the domain from which the database was collected. We use evaluation parameters, such as classification complexity, to constrain the search for credible classifications. Classifications which result in the good approximation of the concept of interest, in the rough sets sense (Katzberg & Ziarko 1994; Pawlak 1991), are subsequently selected to obtain the classification rules.

Classifications on Rough Sets

A relational database can be viewed as an information system (Pawlak 1991). Formally, an *information system* S is a quadruple $\langle U, A, V, f\rangle$, where U is a nonempty set of objects called *universe*; A is a finite set of attributes consisting of *condition attributes* C and *decision attributes* D such that $A = C \cup D$ and $C \cap D = \emptyset$; $V = \bigcup_{p \in A} V_p$ is a nonempty finite set of values of attributes A and V_p is the *domain* of the attribute p (the set of values of attribute p); $f : U \times A \to V$ is an *information function* which assigns particular values from the domain of attributes A to objects such that $f(x_i, p) \in V_p$ for all $x_i \in U$ and $p \in A$.

Any subset of condition attributes defines a classification of the universe of objects U as follows. Let B be a nonempty subset of C, and let x_i, x_j be members of U. The *projection* of the function f onto attributes belonging to the subset B will be denoted as f_B. A binary relation $R(B)$, called an *indiscernibility relation*, is first defined as follows:

$$R(B) \quad = \quad \{(x_i, x_j) \in U^2 : f_B(x_i) = f_B(x_j)\} \quad (1)$$

We say that x_i and x_j are indiscernible by a set of attributes B in S iff $f(x_i, p) = f(x_j, p)$ for every $p \in B$. $R(B)$ is an equivalence relation on U for every $B \subset C$ which classifies the objects in U into a finite, preferably small, number of *equivalence classes.* The set of equivalence classes is called the *classification* $R^*(B)$. The pair $\langle U, R(B) \rangle$ is called an *approximation space* (Pawlak 1991).

The above model cannot, however, be directly applied to most KDD problems. A database represents only a subset (a sample) U' of the universe U about which we are trying to discover something. Depending on the selection of the information function f, the subset of the attributes B, the size and the distribution of objects in the sample U', we may or may not have all values of the information function f_B in our database. If all values are present then our knowledge about the classification is *complete* (despite not having all domain objects in the database); otherwise our knowledge about the classification is *incomplete.* To properly reason about the relationships occurring in U, the classification must be complete; otherwise, false conclusions may be drawn.

Reduction of Classification Complexity

In KDD-related problems, the universe U is finite and it is highly desirable for it to be small. Only finite classifications are "*learnable*," *i.e.*, we can potentially acquire complete knowledge about such classifications. Unfortunately, most finite classifications are not learnable due to the excessively large number of possible equivalence classes. Only a small fraction of all possible classifications expressible in terms of the indiscernibility relation are learnable.

Classification Complexity

To evaluate the computational tractability of the finite classification learning problem, we introduce the notion of *classification complexity,* defined as the number of equivalence classes in the classification. In practice, this number is usually not known in advance. Instead, a crude upper bound on the classification complexity for a subset of attributes $B \subseteq C$, can be computed "a priori" by (2)

$$TC(B, V) = \prod_{p \in B} card(V_p) \qquad (2)$$

The quantity $TC(B, V)$ is called the *theoretical complexity* of the set of attributes B given the set of values V of the attributes B. If the number of attributes and the size of the domain V_p for each attribute is large, then $TC(B, V)$ grows exponentially large. It is very difficult to find a credible classification based on a large number of attributes unless the attributes are strongly dependent (*e.g.*, functionally dependent) on each other (limiting the number of equivalence classes).

Complexity Reduction

Complexity reduction increases the credibility of the classification by generalizing condition attributes. The information generalization procedure (Shan et al. 1995a, 1995b) applies *attribute-oriented concept tree ascension* (Cai, Cercone & Han 1991) to reduce the complexity of an information system. It generalizes a condition attribute to a certain level based on the attribute's *concept tree*, which is provided by knowledge engineers or domain experts. Trivially, the values for any attribute can be represented as a *one-level* concept tree where the root is the most general value "ANY" and the leafs are the distinct values of the attribute.

The following algorithm, adapted from (Shan et al. 1995a, 1995b), extracts a generalized information system. Two thresholds (the *attribute threshold* and the *theoretical complexity threshold*) constrain the generalization process. Condition attributes are generalized by ascending their concept trees until the number of values for each attribute is less than or equal to the user-specified attribute threshold for that attribute and the theoretical complexity of all generalized attributes is less than or equal to the user-specified theoretical complexity threshold. For each iteration, one attribute is selected for generalization (this selection can be made in many ways (Barber & Hamilton 1996). Lower level concepts of this attribute are replaced by the concepts of the next higher level. The number of possible values at a higher level of an attribute is always smaller than at a lower level, so the theoretical complexity is reduced.

```
Algorithm EGIS*:
Input: (i) The original information system $S$
           with a set of condition attributes $C_i$ ( $1 \leq i \leq n$);
       (ii) a set $H$ of concept trees, where each $H_i \in H$
           is a concept hierarchy for the attribute $C_i$.
       (iii) $t_i$ is a threshold for attribute $C_i$, and $d_i$ is
           the number of distinct values of attribute $C_i$;
       (iv) $TC$ defined by user is a theoretical
           complexity threshold.
Output: The generalized information system $S'$
$S' \leftarrow S$
$TC_1 = \prod_{i=1}^{n} d_i$
while $TC_1 > TC$ and $\exists d_i > t_i$ do
    Select an attribute $C_i \in C$ such that $\frac{d_i}{t_i}$ is maximal
    Ascend tree $H_i$ one level and make appropriate
    substitutions in $S'$
    Remove duplicates from $S'$
    Recalculate $d_i$
    Recalculate $TC_1 = \prod_{i=1}^{n} d_i$
endwhile
```

The side effect of such a transformation is an imprecise concept representation in terms of the rough *lower bound* and *upper bound* (Pawlak 1991).

Quality of Classification

Each combination of values of the decision attribute is a concept. Our main goal is to identify a credible classification for each such concept $F \in R(D)$, based on some interacting attributes B. To evaluate the quality of the classification $R^*(B)$ with respect to the concept F, we use the following criterion (Ziarko & Shan 1995):

$$Q_B(F) = \beta \sum_{E \in R^*(B)} P(E) \times |P(F|E) - P(F)|, \qquad (3)$$

and $\beta = \frac{1}{2P(F)(1-P(F))}$.

Criterion (3) represents the average gain in the quality of information, reflected by $P(F|E)$, used to make the classificatory decision F versus $\neg F$. In the absence of the classification $R^*(B)$, the only available information for this kind of the decision is the occurrence probability $P(F)$. The quantity β is a normalization factor to ensure that Q_B is always within the range $[0, 1]$, with 1 corresponding to the exact characterization of the concept (that is, when for every equivalence class E, $P(F|E)$ is either 0 or 1) and 0 corresponding to the situation where the distribution of F within every equivalence class E is the same as in the universe U.

Identifying Interacting Attributes

Local Discovery of Interacting Attributes

The local discovery of interacting attributes has been reported in (Ziarko & Shan 1995). All condition attributes are grouped into disjoint clusters without considering the decision attribute(s). Each *cluster* contains attributes which are directly or indirectly dependent upon each other.

DEP is a generalization of the concept quality measure Q_B (Ziarko & Shan 1995). $DEP(X, Y)$ measures degree of dependency between two groups of attributes X and Y:

$$DEP(X, Y) = \sum_{E \in R^*(Y)} P(E) Q_X(E). \qquad (4)$$

The degree of dependency (4) is the average gain in the quality of the characterization of equivalence classes of $R(Y)$ in the approximation space $\langle U, R(Y) \rangle$. Here, we are not concerned about the direction of the dependency. The crucial question is whether $DEP(X, Y)$ or $DEP(Y, X)$ is greater than or equal to a *dependency threshold*, τ, which is the minimum recognizable dependency level. The generalization procedure for attribute clusters is as follows:

```
Local ClusterGen Algorithm:
Input : $C$ is a set of condition attributes and
        $\tau$ is a dependency threshold.
Output : $AttriCluster$ is a set of attribute clusters.
$AttriCluster \leftarrow \emptyset$
while $C \neq \emptyset$ do
    $ACluster \leftarrow a \in C$
    $C \leftarrow C - \{a\}$
    forall attribute $x \in ACluster$ do
        forall attribute $y \in C$ do
            $MaxDep \leftarrow Max(DEP(\{x\}, \{y\}),$
                    $DEP(\{y\}, \{x\}))$
            if $MaxDep \geq \tau$ then
                $ACluster \leftarrow ACluster \cup y$
                $C \leftarrow C - \{y\}$
            endif
        endfor
    endfor
    $AttriCluster \leftarrow AttriCluster \cup ACluster$
endwhile
```

Global Discovery of Interacting Attributes

The global discovery of interacting attributes is a novel contribution of this paper. A subset of condition attributes is selected based on their relevance to the decision attribute(s).

```
Global ClusterGen Algorithm:
Input : $C$ is a set of condition attributes
        $D$ is a set of decision attributes, and
        $\tau$ is a dependency threshold.
Output : $AttriCluster$ is a set of attribute's clusters.
$AttriCluster \leftarrow a \in C$
$C \leftarrow C - \{a\}$
$Dep \leftarrow DEP(AttriCluster, D)$
while $C \neq \emptyset$ and $Dep < \tau$ do
  forall attribute $\overline{a} \in AttriCluster$ do
      $C' \leftarrow AttriCluster \cup \{a\}$
      $Dep_a \leftarrow DEP(C', D)$
  endfor
  Find the attribute $x$ that has the maximum value of $Dep_a$
  $AttriCluster \leftarrow AttriCluster \cup \{x\}$
  $C \leftarrow C - \{x\}$
  $Dep \leftarrow DEP(AttriCluster, D)$
endwhile
```

Search for Domain Classifications

Finally, we search for acceptable classifications (*i.e.*, any subset B of a cluster that satisfies the following criteria: (1) the cardinality of B is at most $MaxSize$; (2) the theoretical complexity of B is at most TC; (3) the size of equivalence classes in $R^*(B)$ is at least $MinESize$; and (4) the credibility of the classification is at least $MinCred$.

Algorithm **ClassificationSearch** considers all subsets whose size is at most $MaxSize$. The algorithms starts with an empty subset of attributes and incrementally expands the generated subsets as long as they meet the specified criteria.

```
ClassificationSearch Algorithm:
Input : Cluster is a cluster of attributes generated by
        the ClusterGen algorithm, MaxSize is an
        attribute number threshold, TC is a theoretical
        complexity threshold, MinESize is the
        equivalence class threshold, and MinCred is the
        minimum credibility for the classification.
Output : ClassificationSet is a set of attribute subsets
        of Cluster.
$ClassificationSet \leftarrow \emptyset$
$SUB \leftarrow \{\emptyset\}$
while $SUB \neq \emptyset$ do
    forall subset $X \in SUB$ do
        $MaxSize_X \leftarrow$ the number of attributes in $X$
        $TC_X \leftarrow$ the theoretical complexity of $X$
        $MinESize_X \leftarrow Min_{E_i \in R^*(X)}(|E_i|)$
        $ECount_X \leftarrow$ the number of equivalence classes in $X$
        if $(MaxSize_X > MaxSize)$ or $(TC_X > TC)$ or
            $(MinESize_X < MinESize)$ or
            $(ECount_X/TC_X < MinCred)$ then
            Remove $X$ from $SUB$
        endif
    endfor
    if $SUB \neq \emptyset$ then
        $ClassificationSet \leftarrow ClassificationSet \cup SUB$
        if $Cluster \notin SUB$ then
            $SUB' \leftarrow \emptyset$
            forall subset $X \in SUB$ do
                $Cluster_1 \leftarrow$ all attributes in $Cluster$ which
                    have higher numbered than attributes in $X$
                forall attribute $x \in Cluster_1$ do
                    $SUB' \leftarrow SUB' \cup (X \cup \{x\})$
                endfor
            endfor
            $SUB \leftarrow SUB'$
        endif
    endif
endwhile
```

This algorithm appears exponential in the size of subset B, but it can be reworked to be $2^{MaxSize}$, which is a constant. Large subsets of attributes correspond to prohibitively complex and unlearnable classifications, so the given method is quite feasible for relatively simple classifications which are of the most interest.

Summary and Conclusions

We described an approach to database mining based upon searching for domain classifications. The goal of the search is to find a classification or classifications which jointly provide a good, in the rough sets sense, approximation of the concept of interest. We have introduced measures to evaluate the classifications, such as classification complexity measures, quality of classification and connectivity measures to evaluate the degree of connection between classifications, attributes or the concept of interest. We also presented algorithms for complexity reduction, local and global discovery of interacting attributes, and classification search. After classification search, further steps in the rough sets approach to knowledge discovery involve classification analysis and simplification, rule induction and prediction, if required by an application. These aspects have been omitted here as they are presented in detail in other publications (Shan et al. 1995a, 1995b; Ziarko & Shan 1994).

References

Barber, B. and Hamilton, H.J. 1996. Attribute Selection Strategies for Attribute-Oriented Generalization. In *Proceedings of the Canadian AI Conference (AI'96)*, Montreal, Canada.

Cai, Y., Cercone, N., and Han, J. 1991. Attribute-Oriented Induction in Relational Databases, In *Knowledge Discovery in Database*, pp. 213-228.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (eds.) 1996. *Advances in Knowledge Discovery and Data Mining*, AAAI Press/The MIT Press.

Katzberg, J. and Ziarko, W. 1994. Variable Precision Rough Sets with Asymmetric Bounds. In *Rough Sets, Fuzzy Sets and Knowledge Discovery*, pp. 167-177.

Pawlak, Z. 1991. *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer.

Piatetsky-Shapiro, G. and Frawley, W. J. (eds.) 1991. *Knowledge Discovery in Databases*, AAAI/MIT Press.

Shan, N., Hamilton, H.J., and Cercone, N. 1995a. GRG: Knowledge Discovery Using Information Generalization, Information Reduction, and Rule Generation. *International Journal of Artificial Intelligence Tools*, in press.

Shan, N., Ziarko, W., Hamilton, H.J., and Cercone, N. 1995b. Using Rough Sets as Tools for Knowledge Discovery. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD95)*, Montreal, Canada, pp. 263-268.

Ziarko, W. (ed.) 1994. *Rough Sets, Fuzzy Sets and Knowledge Discovery*, Springer-Verlag.

Ziarko, W. and Shan, N. 1994. KDD-R: A Comprehensive System for Knowledge Discovery in Databases Using Rough Sets. In *Proceedings of the International Workshop on Rough Sets and Soft Computing (RSSC'94)*, pp. 164-173.

Ziarko, W. and Shan, N. 1995. On Discovery of Attribute Interactions and Domain Classifications, In Lin. T.Y (ed.), Special Issue in *Journal of Intelligent Automation and Soft Computing*, in press.