Lower Bounds on Cardinality of Reducts for Decision Tables from Closed Classes

Azimkhon Ostonov and Mikhail Moshkov
Computer, Electrical and Mathematical Sciences & Engineering Division
and Computational Bioscience Research Center
King Abdullah University of Science and Technology (KAUST)
Thuwal 23955-6900, Saudi Arabia
{azimkhon.ostonov,mikhail.moshkov}@kaust.edu.sa

Abstract

In this paper, we consider classes of decision tables closed under removal of attributes (columns) and changing of decisions attached to rows. For decision tables from closed classes, we study lower bounds on the minimum cardinality of reducts, which are minimal sets of attributes that allow us to recognize, for a given row, the decision attached to it. We assume that the number of rows in decision tables from the closed class is not bounded from above by a constant. We divide the set of such closed classes into two families. In one family, only standard lower bounds $\Omega(\log \operatorname{cl}(T))$ on the minimum cardinality of reducts for decision tables hold, where $\operatorname{cl}(T)$ is the number of decision classes in the table T. In another family, these bounds can be essentially tightened up to $\Omega(\operatorname{cl}(T)^{1/q})$ for some natural q.

Keywords: decision table, closed class, reduct.

1 Introduction

Decision tables are a well-known way of presenting the information needed to make decisions. These tables are used, in particular, in data analysis, including classification problems, in modeling and studying problems related to combinatorial optimization, fault diagnosis, computational geometry, etc. [1, 2, 3, 4, 5, 6, 7, 9, 10]. Note that finite information systems with a selected decision attribute, data sets with a selected class attribute, and partially defined Boolean functions studied in various fields of data analysis as representations of decision problems can naturally be interpreted as decision tables.

In this paper, we consider classes of decision tables closed under removal of attributes (columns) and changing of decisions attached to rows. The most natural examples of such classes are closed classes of decision tables derived from information systems: the set of decision tables corresponding to problems over an information system forms a closed class of decision tables. However, the family of all closed classes of decision tables is essentially wider than the family of closed classes derived from information systems. In particular, the union

of classes derived from two information systems is a closed class, but generally, there is no an information system for which this union is the closed class derived from it.

For decision tables from closed classes, we study lower bounds on the minimum cardinality of (decision) reducts, which are minimal sets of attributes that allow us to recognize, for a given row of the table, the decision attached to it. Reducts are one of the main notions of rough set theory in which they are used to choose appropriate features, to solve classification problems, and to compress the knowledge [2, 7, 8, 11, 12]. The bounds on the minimum cardinality of reducts are of significant interest for rough set theory.

In this paper, we assume that the number of rows in decision tables from the closed class is not bounded from above by a constant. We divide the set of such closed classes into two families. In one family, only standard lower bounds $\Omega(\log \operatorname{cl}(T))$ on the minimum cardinality of reducts for decision tables hold, where $\operatorname{cl}(T)$ is the number of decision classes in the table T. In another family, these bounds can be essentially tightened up to $\Omega(\operatorname{cl}(T)^{1/q})$ for some natural q. The obtained results can be useful for the specialists in data analysis.

The present paper consists of six sections. Sections 2 and 3 contain main definitions and some results related to the decision tables and to the closed classes of decision tables. In Sect. 4, we discuss lower bounds on the cardinality of reducts and, in Sect. 5 – examples related to the closed classes of decision tables derived from information systems. Section 6 contains short conclusions.

2 Decision Tables

Let B be a nonempty finite set with k elements, $k \geq 2$. A B-decision table T is a rectangular table with n columns labeled with attributes (really names of attributes) in which rows are pairwise different tuples from B^n that are labeled with nonnegative integers (decisions). Denote by Rows(T) the set of rows of the table T, N(T) the number of rows in T, and cl(T) the number of different decisions attached to rows of T (the number of decision classes in the table T). The number n will be called the dimension of the table T and will be denoted dim T.

A test for the table T is a set of attributes (columns) of the table T such that any two rows of the table T labeled with different decisions are different in at least one of the considered columns. A reduct for the table T is a test for T each proper subset of which is not a test. We denote by R(T) the minimum cardinality of a reduct for the table T. If $\operatorname{cl}(T) < 2$, then R(T) = 0.

Denote by [T] the set of decision tables that can be obtained from T in the following way: we can remove from T an arbitrary number of attributes (columns), keep only one row from each group of equal rows in the obtained table, and change in an arbitrary way decisions attached to the remaining rows.

A decision table T with n columns will be called quasicomplete if there exist two-elements subsets B_1, \ldots, B_n of the set B such that

$$B_1 \times \cdots \times B_n \subseteq \text{Rows}(T)$$
.

We denote by I(T) the maximum dimension of a quasicomplete table from [T]. The next statement follows immediately from Theorem 4.6 [5].

Lemma 1. For any B-decision table T with $cl(T) \geq 2$,

$$N(T) \le (k^2 \dim T)^{I(T)}.$$

3 Closed Classes of Decision Tables

Let C be a set of B-decision tables. This set will be called a closed class of decision tables if $C = \bigcup_{T \in C} [T]$. The closed class C will be called nondegenerate if the number of rows in tables from C is not bounded from above by a constant.

We now define a parameter I(C) of a nondegenerate closed class C of decision tables. If the parameter I is bounded from above by a constant on tables from the class C, then $I(C) = \max\{I(T) : T \in C\}$. Otherwise, $I(C) = +\infty$.

Let us consider the behavior of the function

$$N_C(n) = \max\{N(T) : T \in C, \dim T \le n\}$$

defined on the set of natural numbers. This function characterizes the growth in the worst case of the number of rows in decision tables from the class C with the growth of their dimension.

Lemma 2. Let C be a nondegenerate closed class of B-decision tables.

- (a) If $I(C) < +\infty$, then $N_C(n) \le (k^2 n)^{I(C)}$ for any natural n.
- (b) If $I(C) = +\infty$, then $2^n \le N_C(n) \le k^n$ for any natural n.

Proof. (a) Let $I(C) < +\infty$. Using Lemma 1, we obtain that $N_C(n) \leq (k^2 n)^{I(C)}$ for any natural n.

(b) Let $I(C) = +\infty$ and n be a natural number. The inequality $N_C(n) \leq k^n$ is obvious. Since $I(C) = +\infty$, there exists a quasicomplete table $T_n \in C$ with dim $T_n = n$. It is clear that $N(T_n) \geq 2^n$. Therefore $2^n \leq N_C(n)$.

4 Bounds on Cardinality of Reducts

First, we prove an auxiliary statement.

Lemma 3. Let C be a nondegenerate closed class of B-decision tables and T be a decision table from C with $cl(T) \geq 2$. Then

$$N_C(R(T)) \ge \operatorname{cl}(T).$$

Proof. Let R(T) = m and $\{f_1, \ldots, f_m\}$ be a reduct with the minimum cardinality for the table T. We denote by T' a table from [T], which is obtained from T by the removal of all attributes with the exception of f_1, \ldots, f_m . Then the number of rows in the table T' should be at least the number of decision classes in T, i.e., $N(T') \geq \operatorname{cl}(T)$. It is clear that $N(T') \leq N_C(m)$. Therefore $N_C(m) \geq \operatorname{cl}(T)$.

Theorem 1. Let C be a nondegenerate closed class of B-decision tables.

- (a) If $I(C) < +\infty$, then $R(T) \ge \operatorname{cl}(T)^{1/I(C)}/k^2$ for any table $T \in C$ with $\operatorname{cl}(T) \ge 2$.
- (b) If $I(C) = +\infty$, then $R(T) \ge \log_k \operatorname{cl}(T)$ for any table $T \in C$ with $\operatorname{cl}(T) \ge 2$.
- (c) If $I(C) = +\infty$, then the inequality $R(T) \ge \log_2 \operatorname{cl}(T) + 1$ does not hold for infinitely many tables T from the class C for which both the dimension and the number of decision classes are not bounded from above by constants.
- Proof. (a) Let $I(C) < +\infty$, $T \in C$, $\operatorname{cl}(T) \geq 2$, and R(T) = m. From Lemma 2 it follows that $N_C(m) \leq (k^2m)^{I(C)}$. By Lemma 3, $N_C(m) \geq \operatorname{cl}(T)$. Therefore $(k^2m)^{I(C)} \geq \operatorname{cl}(T)$ and $m \geq \operatorname{cl}(T)^{1/I(C)}/k^2$.
- (b) Let $I(C) = +\infty$, $T \in C$, $\operatorname{cl}(T) \geq 2$, and R(T) = m. From Lemma 2 it follows that $N_C(m) \leq k^m$. By Lemma 3, $N_C(m) \geq \operatorname{cl}(T)$. Therefore $k^m \geq \operatorname{cl}(T)$ and $m \geq \log_k \operatorname{cl}(T)$.
- (c) Let n be a natural number. Since $I(C) = +\infty$, there exists a quasicomplete decision table T_n from C with dim $T_n = n$ and $\operatorname{cl}(T_n) \geq 2^n$. Let us assume that $R(T_n) \geq \log_2 \operatorname{cl}(T) + 1$. Then $R(T_n) \geq \log_2 2^n + 1 = n + 1$. It is obvious, that $n \geq R(T_n)$. Thus, the inequality $R(T_n) \geq \log_2 \operatorname{cl}(T_n) + 1$ does not hold.

The statement (c) shows that the bound from the statement (b) cannot be improved essentially.

5 Closed Classes of Decision Tables Derived from Information Systems

The most natural examples of closed classes of decision tables are classes derived from infinite information systems. An infinite information system is a triple U = (A, F, B), where A is an infinite set of objects called universe, B is a finite set with k elements, $k \geq 2$, and F be an infinite set of functions from A to B called attributes. A problem over U is specified by a finite number of attributes $f_1, \ldots, f_n \in F$ that divide the universe A into nonempty domains in each of which values of attributes f_1, \ldots, f_n are fixed. Each domain is labeled with a decision. For a given object $a \in A$, it is required to recognize the decision attached to the domain to which the object a belongs. A decision table corresponds to this problem in the following way: the table contains n columns labeled with attributes f_1, \ldots, f_n , rows of this table correspond to domains and are labeled with decisions attached to the domains.

We denote by Tab(U) the set of decision tables corresponding to all problems over the information system U. One can show that Tab(U) is a nondegenerate closed class of decision tables. We will say that this class is derived from the information system U.

A subset $\{f_1, \ldots, f_p\}$ of the set F is called independent if there exist two-element subsets B_1, \ldots, B_p of the set B such that, for any tuple $(b_1, \ldots, b_p) \in B_1 \times \cdots \times B_p$, the equations system

$$\{f_1(x) = b_1, \dots, f_p(x) = b_p\}$$

has a solution from A. If, for any natural p, the set F contains an independent subset, which cardinality is equal to p, then $I(\text{Tab}(U)) = +\infty$. Otherwise, I(Tab(U)) is the maximum cardinality of an independent subset of the set F.

We now consider examples of infinite information systems from the book [6].

Example 1. Let P be the Euclidean plane and l be a straight line in P. This line divides the plane into two open half-planes h_1 and h_2 , and the line l. We correspond an attribute to the line l. This attribute takes the value 0 on points from h_1 and the value 1 on points from h_2 and l. We denote by F_P the set of attributes corresponding to all lines in P and consider the information system $U_p = (P, F_P, \{0, 1\})$. There are two lines that divide the plane P into four domains, but there are no three lines that divide P into eight domains. Therefore $I(\text{Tab}(U_P)) = 2$ and, for any table $T \in \text{Tab}(U_P)$ with $\text{cl}(T) \geq 2$, $R(T) \geq \text{cl}(T)^{1/2}/4$. This lower bound is essentially tighter than the standard bound $R(T) \geq \log_2 \text{cl}(T)$.

Example 2. Let m and t be natural numbers. We denote by $\operatorname{Pol}(m)$ the set of polynomials with integer coefficients that depend on variables x_1, \ldots, x_m . We denote by $\operatorname{Pol}(m, t)$ the set of polynomials from $\operatorname{Pol}(m)$, which degree is at most t. We define information systems U(m) and U(m,t) in the following way: $U(m) = (\mathbb{R}^m, F(m), E)$ and $U(m,t) = (\mathbb{R}^m, F(m,t), E)$, where \mathbb{R} is the set of real numbers, $E = \{-1,0,+1\}$, $F(m) = \{\operatorname{sign}(p) : p \in \operatorname{Pol}(m)\}$, $F(m,t) = \{\operatorname{sign}(p) : p \in \operatorname{Pol}(m,t)\}$, and $\operatorname{sign}(x) = -1$ if x < 0, $\operatorname{sign}(x) = 0$ if x = 0, and $\operatorname{sign}(x) = +1$ if x > 0. One can show that $I(\operatorname{Tab}(U(m))) = +\infty$ and $I(\operatorname{Tab}(U(m,t))) < +\infty$. Therefore, for any natural m and any table T from $\operatorname{Tab}(U(m))$ with $\operatorname{cl}(T) \geq 2$, $R(T) \geq \log_3 \operatorname{cl}(T)$ and this bound cannot be tightened essentially. For any natural m and t and any table T from $\operatorname{Tab}(U(m,t))$ with $\operatorname{cl}(T) \geq 2$, $R(T) \geq \operatorname{cl}(T)^{1/q}/9$ for some natural q.

6 Conclusions

In this paper, we divided the set of nondegenerate closed classes of decision tables into two families. For closed classes from one family, only standard lower bounds $\Omega(\log \operatorname{cl}(T))$ on the minimum cardinality of reducts for decision tables hold, where $\operatorname{cl}(T)$ is the number of decision classes in the table T. For closed classes from another family, these bounds can be essentially tightened up to $\Omega(\operatorname{cl}(T)^{1/q})$ for some natural q.

Acknowledgements

Research reported in this publication was supported by King Abdullah University of Science and Technology (KAUST).

References

- [1] Boros, E., Hammer, P.L., Ibaraki, T., Kogan, A.: Logical analysis of numerical data. Math. Program. **79**, 163–190 (1997)
- [2] Chikalov, I., Lozin, V.V., Lozina, I., Moshkov, M., Nguyen, H.S., Skowron, A., Zielosko, B.: Three Approaches to Data Analysis Test Theory, Rough Sets and Logical Analysis of Data, *Intelligent Systems Reference Library*, vol. 41. Springer (2013)
- [3] Fürnkranz, J., Gamberger, D., Lavrac, N.: Foundations of Rule Learning. Cognitive Technologies. Springer (2012)

- [4] Humby, E.: Programs from Decision Tables, *Computer Monographs*, vol. 19. Macdonald, London and American Elsevier, New York (1973)
- [5] Moshkov, M.: Time complexity of decision trees. In: J.F. Peters, A. Skowron (eds.) Trans. Rough Sets III, Lecture Notes in Computer Science, vol. 3400, pp. 244–459. Springer (2005)
- [6] Moshkov, M., Zielosko, B.: Combinatorial Machine Learning A Rough Set Approach, Studies in Computational Intelligence, vol. 360. Springer (2011)
- [7] Pawlak, Z.: Rough Sets Theoretical Aspects of Reasoning about Data, *Theory and Decision Library: Series D*, vol. 9. Kluwer (1991)
- [8] Pawlak, Z., Skowron, A.: Rudiments of rough sets. Inf. Sci. 177(1), 3–27 (2007)
- [9] Pollack, S.L., Hicks, H.T., Harrison, W.J.: Decision Tables: Theory and Practice. John Wiley & Sons (1971)
- [10] Rokach, L., Maimon, O.: Data Mining with Decision Trees Theory and Applications, Series in Machine Perception and Artificial Intelligence, vol. 69. World Scientific (2007)
- [11] Slezak, D.: Approximate entropy reducts. Fundam. Informaticae **53**(3-4), 365–390 (2002)
- [12] Stawicki, S., Slezak, D., Janusz, A., Widz, S.: Decision bireducts and decision reducts a comparison. Int. J. Approx. Reason. 84, 75–109 (2017)