Reduct Calculation and Discretization of Numeric Attributes in Sparse Decision Systems

Wojciech Swieboda and Hung Son Nguyen

Institute of Mathematics, The University of Warsaw,
Banacha 2, 02-097, Warsaw Poland

Abstract. In this paper we discuss three problems in Data Mining Sparse Decision Systems: the problem of short reduct calculation, discretization of numerical attributes and rule induction. We present algorithms that provide approximate solutions to these problems and analyze the complexity of these algorithms.

1 Introduction

In the paper we discuss algorithms for Data Mining [3] Sparse Decision Tables. We first review basic notions of Information Systems, Decision Systems and Rough Set Theory [9]. We introduce a convenient representation for sparse decision tables and finally discuss algorithms for short reduct calculation, discretization and rule induction.

2 Rough Set Preliminaries

An *information system* is a pair $\mathbb{I} = (\mathbb{U}, \mathbb{A})$ where $\mathbb{U}$ denotes the *universe of objects* and $\mathbb{A}$ is the set of *attributes*. An attribute $a \in \mathbb{A}$ is a mapping $a : U \to V_a$. The co-domain V_a of attribute a is often also called the *value set* of attribute a.

A *decision system* is a pair $\mathbb{D} = (\mathbb{U}, \mathbb{A} \cup \{dec\})$ which is an information system with a distinguished attribute $dec : U \to \{1, \ldots, d\}$ called *a decision attribute*. Attributes in $\mathbb{A}$ are called *conditions* or *conditional attributes* and may be either *nominal* or *numeric* (i.e. with $V_a \subseteq \mathbb{R}$).

Throughout this paper n will denote the number of objects in a decision system and k will denote the number of conditional attributes.

3 Sparse Data Sets and Decision Systems

In many situations a convenient way to represent the data set is in terms of Entity-Attribute-Value (EAV) Model [11], which encodes observations in terms of triples. For an information system $I = (\mathbb{U}, \mathbb{A})$, the set of triples is $\{(u, a, v) : a(u) = v\}$. This representation is especially handy for information systems with numerous attributes, missing or default values. Instances with missing and default values are not included in EAV representation, which results in compression of the data set. In this paper we are only dealing with default values. Their interpretation/semantics is the same as of any other attribute. In practice we store triples corresponding to numeric attributes and to

Table 1. A typical decision system with symbolic attributes represented as a table. Attributes *Diploma, Experience, French* and *Reference* are *conditions*, whereas *Decision* is the decision attribute. All conditional attributes in this decision system are nominal

	Diploma	Experience	French	Reference	Decision
x_1	MBA	Medium	Yes	Excellent	Accept
x_2	MBA	Low	Yes	Neutral	Reject
x_3	MCE	Low	Yes	Good	Reject
x_4	MSc	High	Yes	Neutral	Accept
x_5	MSc	Medium	Yes	Neutral	Reject
x_6	MSc	High	Yes	Excellent	Accept
x_7	MBA	High	No	Good	Accept
x_8	MCE	Low	No	Excellent	Reject

Table 2. A decision system in which all conditional attributes are numeric

	a_1	a_2	a_3	Decision
x_1	0	1.3	0	F
x_2	3.3	0.9	0	F
x_3	0	1.5	0	F
x_4	0	1.2	2.5	F
x_5	0	1.3	3.6	F
x_6	3.7	2.7	2.4	T
x_7	4.1	1.0	2.8	T

symbolic attributes in two separate tables, and store decisions (which we assume are never missing) of objects in a separate vector.

Another related representation, more general then EAV model, is Subject-Predicate-Object (SPO), and is used e.g. in Resource Description Framework (RDF) Model and implemented in several Triplestore databases.

4 Problems for Sparse Decision Systems

In our paper we address the following problems for Sparse Decision Systems:

1. Finding a short reduct or a superreduct [1].
   A *reduct* is a subset of attributes $R \subseteq A$ which guarantees discernibility of objects belonging to different decision classes.
2. Discretization of numerical attributes [6].
   Discretization of a decision system is determining a set of cuts on numerical attributes so that the induced partitions (i.e. intervals between cutpoints) guarantee discernibility of objects belonging to different decision classes.
3. Generating set of rules or dynamic rules [1].

Table 3. EAV representation of decision system in table 1. The default values (omitted in this representation) for consecutive attributes are 'MBA', 'Low', 'Yes' and 'Excellent'

Entity	Attribute	Value
x_1	a_2	Medium
x_2	a_4	Neutral
x_3	a_1	MCE
x_3	a_4	Good
x_4	a_1	MSc
x_4	a_2	High
x_4	a_4	Neutral
x_5	a_1	MSc
x_5	a_2	Medium
x_5	a_4	Neutral
x_6	a_1	MSc
x_6	a_2	High
x_7	a_2	High
x_7	a_3	No
x_7	a_4	Good
x_8	a_1	MCE
x_8	a_3	No

Entity	Decision
x_1	Accept
x_2	Reject
x_3	Reject
x_4	Accept
x_5	Reject
x_6	Accept
x_7	Accept
x_8	Reject

Table 4. EAV representation of decision system in table 2. The default value (omitted in this representation) for each attribute is 0

Entity	Attribute	Value
x_1	a_2	1.3
x_2	a_1	3.3
x_2	a_2	0.9
x_3	a_2	1.5
x_4	a_2	1.2
x_4	a_3	2.5
x_5	a_2	1.3
x_5	a_3	3.6
x_6	a_1	3.7
x_6	a_2	2.7
x_6	a_3	2.4
x_7	a_1	4.1
x_7	a_2	1.0
x_7	a_3	2.8

Entity	Decision
x_1	T
x_2	T
x_3	T
x_4	T
x_5	T
x_6	T
x_7	T

References

1. Bazan, J.G., Nguyen, H.S., Nguyen, S.H., Synak, P., Wróblewski, J.: Rough set algorithms in classification problem pp. 49–88 (2000)

Table 5. A discretized version of the decision system presented in table 2.

	a_1	a_2	a_3	Decision
x_1	$(-\infty, +\infty)$	$(1.25, +\infty)$	$(-\infty, 1.2]$	F
x_2	$(-\infty, +\infty)$	$(-\infty, 1.1]$	$(-\infty, 1.2]$	F
x_3	$(-\infty, +\infty)$	$(1.25, +\infty)$	$(-\infty, 1.2]$	F
x_4	$(-\infty, +\infty)$	$(1.1, 1.25]$	$(1.2, +\infty)$	F
x_5	$(-\infty, +\infty)$	$(1.25, +\infty)$	$(1.2, +\infty)$	F
x_6	$(-\infty, +\infty)$	$(1.25, +\infty)$	$(1.2, +\infty)$	T
x_7	$(-\infty, +\infty)$	$(-\infty, 1.1]$	$(1.2, +\infty)$	T

2. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, New York, 2. edn. (2001)
3. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. MIT Press (2001), `http://mitpress.mit.edu/026208290X`
4. Hastie, T., Tibshirani, R., Friedman, J.H.: The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations. New York: Springer-Verlag (2001)
5. Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: A tutorial (1998)
6. Nguyen, H.S.: Discretization problem for rough sets methods. In: Polkowski and Skowron [10], pp. 545–552, `http://dx.doi.org/10.1007/3-540-69115-4\_75`
7. Nguyen, H.S.: Approximate boolean reasoning: Foundations and applications in data mining (2006)
8. Pawlak, Z.: Rough sets. International Journal of Information and Computer Sciences 11(5), 341–356 (1982)
9. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Springer, Formerly Kluwer Academic Publishers, Boston, Dordrecht, London (1991)
10. Polkowski, L., Skowron, A. (eds.): Rough Sets and Current Trends in Computing, First International Conference, RSCTC'98, Warsaw, Poland, June 22-26, 1998, Proceedings, Lecture Notes in Computer Science, vol. 1424. Springer (1998)
11. Stead, W.W., Hammond, W.E., Straube, M.J.: A chartless record – is it adequate? Journal of Medical Systems 7, 103–109 (1983), `http://dx.doi.org/10.1007/BF00995117`, 10.1007/BF00995117