On Modal Decision Logics

Tuan-Fang Fan
Department of Information Engineering
National Penghu Institute of Technology
Penghu, Taiwan
dffan@npit.edu.tw

Churn-Jung Liau
Institute of Information Science
Academia Sinica
Taipei, Taiwan
liaucj@iis.sinica.edu.tw

Y.Y. Yao
Department of Computer Science
University of Regina, Regina
Saskatchewan, Canada S4S 0A2
yyao@cs.uregina.ca

Abstract

Some modal decision logic languages are proposed for knowledge representation in data mining through the notions of models and satis£ability. The models are collections of data tables consisting of a £nite set of objects described by a £nite set of attributes. Some relationships may exist between data tables in a collection and the modalities of our languages are interpreted with respect to these relations in a Kripkean style semantics.

1 Introduction

Theory of knowledge has been a commonly important topic of many academic branches such as philosophy, psychology, economics, and arti£cial intelligence, whereas the storage and retrieval of data is the main concern of information science. In the modern experimental science, knowledge is usually acquired from observed data. The data can provide the causal-effect or associational relationship between attributes of the observed objects. However, when the amount of data is large, it becomes a dif£cult task to analyze the data and extract knowledge from them. With the aid of computers, the large amount of data stored in relational data tables can be transformed into symbolic knowledge automatically. Thus intelligent data analysis has received much attention recently.

While the data mining researches mainly concentrate on the design of ef£cient algorithms for extracting knowledge from data, to close the semantic gap between structured data and human-comprehensible concepts has been a lasting challenge for the research community[6]. This is called the interpretability problem of intelligent data analysis in [6]. Since the discovered knowledge is useful for a human user only when he can understand its meaning, the knowledge representation formalism will play an important role in the utilization of the induced rules. A good representation formalism should have clear semantics so that a rule can be effectively validated with respect to the given data tables. In this regard, logic is one of the best choices. As indicated by Zadeh[14], human usually compute with words instead of numbers, so if we can incorporate linguistically meaningful terms into the representation formalism, then the induced rules may be more useful to the human decision-makers.

The rough set theory proposed by Pawlak provides an effective tool for extracting knowledge from data tables[10]. To represent and reason about the extracted knowledge, a decision logic (DL) is also proposed in [10]. The semantics of the logic is de£ned in a Tarskian style through the notions of models and satisfaction. While DL can be considered as an instance of classical logic in the context of data tables, different generalizations of DL corresponding to some non-classical logics are also desirable from the knowledge representation viewpoint. For example, to deal with uncertain or incomplete information, some generalized decision logics have been proposed in [3, 8, 9, 11, 12].

These generalized decision logics, however, concentrate mostly on the representation of knowledge from a single data table. Though in principle, all data can be put into a single table, it is sometimes more natural to represent them by a collection of data tables. For example, in an enterprize database, the business transaction records may be stored as a collection of data tables indexed by dates. To extract knowledge from such structured data tables, we need richer representation languages than the decision logic. Among the traditional logical tools, modal logic would be one of the most appropriate candidates that can meet the need since it is a logic for reasoning about relations in a broad sense[1], whereas the knowledge extracted from multiple data tables usually concerns with the relationship of objects across different tables.

In this paper, we present a formulation of the modal decision logics based on multiple data tables. In the next section, we £rst review the decision logic proposed by Pawlak. Then, a general modal decision logic(MDL) is presented in section 3, which is followed by three case studies. They are respectively the uncertain, epistemic, and temporal decision logic. In each case, the syntax and semantics of the logics are presented and some illustrative examples are given. Finally, the summery is given a concluding section and some further research directions are also pointed out.

2 Review of Decision Logic

A data table(DT) is usually used as a regular approach for storage of data. A formal de£nition of data table is given in [10].

De£nition 1 *A data table*[1] *is a pair* $T = (U, A, \{a_T \mid a \in A\})$ *such that*

- U *is a nonempty £nite set, called the universe,*
- A *is a nonempty £nite set of primitive attributes, and*
- *for each* $a \in A$, $a_T : U \to V_a$ *is a total function, where* V_a *is the domain of values for* a. *Usually, we will simply write* a *instead of* a_T *for the functions.*

Given a data table T, we will denote its universe U and attribute A by $Uni(T)$ and $Att(T)$ respectively.

In [10], a decision logic(DL) is proposed for the representation of the knowledge discovered from data tables. The logic is called decision logic because it is particularly useful in a special kind of data tables, called *decision table.* A decision table is a data table $T = (U, C \cup D)$, where $Att(T)$ can be partitioned into two sets C and D, called condition attributes and decision attributes respectively. By data analysis, decision rules relating the condition and the decision attributes can be derived from the table. A rule is then represented as an implication between formulas of the logic. However, for a general data table, the acronym DL can also denote *data logic.*

The basic alphabet of a DL consists of a £nite set of attribute symbols $\mathcal{A}$ and for $a \in \mathcal{A}$, a £nite set of value symbols $\mathcal{V}_a$. The syntax of DL is then de£ned as follows.

[1]Also called knowledge representation system, information system, or attribute-value system

De£nition 2

1. *An atomic formula of DL is a descriptor* (a, v), *where* $a \in \mathcal{A}$ *and* $v \in \mathcal{V}_a$.
2. *The well-formed formulas (wff) of DL is the smallest set containing the atomic formulas and closed under the Boolean connectives* $\neg$, $\wedge$, *and* $\vee$.

A data table $T = (U, A)$ is for a given DL if there is a bijection $f : \mathcal{A} \to A$ such that for every $a \in \mathcal{A}$, $V_{f(a)} = \mathcal{V}_a$. Thus, by somewhat abusing the notation, we will usually denote an atomic formula as (a, v), where $a \in A$ and $v \in V_a$, if the data tables are clear from the context. Intuitively, each element in the universe of a data table corresponds to a data record and an atomic formula, which is in fact an attribute-value pair, describes the value of some attribute in a data record. Thus the atomic formulas (and so the wffs) can be veri£ed or falsi£ed in each data record. This gives rise to a satisfaction relation between the universe and the set of wffs.

De£nition 3 *Given a DL and a DT* $T = (U, A)$ *for it, the satisfaction relation* $\models$ *between* $x \in U$ *and wffs of DL is de£ned inductively as follows.*

1. $(T, x) \models (a, v)$ *iff* $a(x) = v$
2. $(T, x) \models \neg\varphi$ *iff* $(T, x) \not\models \varphi$
3. $(T, x) \models \varphi \wedge \psi$ *iff* $(T, x) \models \varphi$ *and* $(T, x) \models \psi$
4. $(T, x) \models \varphi \vee \psi$ *iff* $(T, x) \models \varphi$ *or* $(T, x) \models \psi$

If φ *is a DL wff, the set* $m_T(\varphi)$ *de£ned by:*

$$m_T(\varphi) = \{x \in U \mid (T, x) \models \varphi\}, \tag{1}$$

is called the meaning of the formula φ *in* T. *If* T *is understood, we simply write* $m(\varphi)$.

A formula φ is said to be valid in a data table T, written $T \models \varphi$ or $\models \varphi$ for short when T is clear from the context, if and only if $m(\varphi) = U$. That is, ϕ is satis£ed by all individuals in the universe.

A DL wff talks about the properties of individuals in the universe, so it is satis£ed by some individuals but falsi£ed by the others. However, the mined knowledge is usually regarding the aggregated or statistical information of all individuals. Obviously, the wffs valid in a data table represent a kind of knowledge that can be induced from the table since they hold for all individuals. However, not all kinds of useful information are in the form of valid wffs. Sometimes, even probabilistic rules are very useful from the viewpoint of knowledge discovery. To quantify the usefulness of the mined rules, some measures have been proposed in [13, 11].

In contrast with DL, where extra meta-level measures must be attached to the wffs, these measures can also be internalized to the language by the so-called generalized quanti£ers[4, 7]. This is the approach adopted by the monadic observational predicate calculus(MOPC) of [5]. A wff in DL corresponds to the open formula of MOPC, however, there is not yet the counterpart for the closed formulas of MOPC in DL. To de£ne the corresponding extension in DL, let us call the above-de£ned DL wffs individual formulas and £x a set of unary and binary quanti£ers in advance, then the aggregate formulas for a data table T are de£ned by the following formation rules:

1. if φ is an individual formula and q is an unary quanti£er, then $(q)\varphi$ is an aggregate formula,
2. if φ and ψ are individual formulas and q is a binary quanti£er, then $(q)(\varphi, \psi)$ is an aggregate formula,
3. if φ and ψ are aggregate formulas, so are $\neg\varphi$, $\varphi \wedge \psi$, and $\varphi \vee \psi$

Each quanti£er q is interpreted according to its truth function Tr_q according to [5]. For each unary quanti£er q, $Tr_q : N^2 \to \{0, 1\}$ is a 2-place function from natural number to $\{0, 1\}$ and for the binary one, $Tr_q : N^4 \to \{0, 1\}$ is a four-place function. Then the satisfaction of an aggregate formula with respect to a data table T is de£ned as follows:

1. $T \models (q)\varphi$ iff $Tr_q(|m(\varphi)|, |m(\neg\varphi)|) = 1$,
2. $T \models (q)(\varphi, \psi)$ iff $Tr_q(|m(\varphi \wedge \psi)|, |m(\varphi \wedge \neg\psi)|, |m(\neg\varphi \wedge \psi)|, |m(\neg\varphi \wedge \neg\psi)|) = 1$,
3. $T \models \neg\varphi$, $T \models \varphi \wedge \psi$, and $T \models \varphi \vee \psi$ are de£ned inductively as in the case of individual formulas.

Note that the classical quanti£ers $\forall$ and $\exists$ are de£ned with truth functions $Tr_\forall(n_1, n_2) = 1$ iff $n_2 = 0$ and $Tr_\exists(n_1, n_2) = 1$ iff $n_1 > 0$.

3 General Modal Decision Logic

Just like the models of DL are data tables, those for modal decision logic (MDL) will be structured sets of data tables.

De£nition 4 *Let I and J be two £xed sets of indices, then a structured set of data tables (SSDT) is a pair $\mathcal{S} = (\{T_i \mid i \in I\}, \{R_j \mid j \in J\})$, where each T_i is a data table and each R_j is a binary relation over $\{T_i \mid i \in I\}$.*

In this paper, we will consider only the SSDT $\mathcal{S} = (\{T_i \mid i \in I\}, \{R_j \mid j \in J\})$ satisfying the following assumptions:

- £xed attribute assumption: for $i, j \in I$, $Att(T_i) = Att(T_j)$, namely, we assume the data tables in an SSDT are homogeneous. This assumption is necessary since the atomic formula of our logic language will depend on the set of attributes, so to make our language interpretable in all data tables, they must have the same set of attributes.
- constant domain assumption: for $i, j \in I$, $Uni(T_i) = Uni(T_j)$. In other words, we assume the set of individuals keeps unchanged between different data tables. This assumption is not essential, however, it will simplify the semantics of our logic.
- £nite table assumption: I is £nite. This is a practical assumption since we will consider only £nite amount of data in the knowledge discovery process.

The syntax of MDL is an extension of DL with the following rule:

- if φ is an individual (resp. aggregate) formula, so are $[j]\varphi$ and $\langle j\rangle\varphi$ for any $j \in J$.

Given an SSDT $\mathcal{S} = (\{T_i \mid i \in I\}, \{R_j \mid j \in J\})$, the satisfaction of individual formulas are de£ned by

1. $(T_i, x) \models_\mathcal{S} [j]\varphi$ iff for all T such that $(T_i, T) \in R_j$, $(T, x) \models_\mathcal{S} \varphi$
2. $(T_i, x) \models_\mathcal{S} \langle j\rangle\varphi$ iff there exists T such that $(T_i, T) \in R_j$ and $(T, x) \models_\mathcal{S} \varphi$
3. the satisfaction of classical formulas is de£ned as in the case of DL.

The satisfaction of aggregate formulas can be analogously de£ned and is denoted by $T \models_\mathcal{S} \varphi$.

4 Case Studies

In MDL, there is a set of modal operators $[j]$ which are interpreted semantically by the binary relations R_j over the data tables of an SSDT, however, it remains unspeci£ed how the binary relations are constructed. In the following sections, we will study some cases in which the binary relations between data tables arise naturally from the application problems.

4.1 Uncertain decision logic

The £rst problem we trying to deal with is regarding uncertain data tables.

De£nition 5 *An uncertain data table is a pair $T = (U, A, \{a_T \mid a \in A\})$ such that*

- *U is a nonempty £nite set, called the universe,*
- *A is a nonempty £nite set of primitive attributes, and*

- *for each $a \in A$, $a_T : U \to 2^V_a - \{\emptyset\}$ is a total function, where V_a is the domain of values for a.*

For each $x \in U$, $a_T(x)$ denotes the set of possible values for its attribute a. Since $a_T(x)$ may contain more than one values, this means that we do not have the exact knowledge about what the value is. In particular, if $a_T(x) = V_a$, then we have null information for the particular x on its attribute a. Given an uncertain data table $T = (U, A, \{a_T \mid a \in A\})$, a *possible realization* of T is a data table $T' = (U, A, \{a_{T'} \mid a \in A\})$ such that for any $x \in U$ and $a \in A$, $a_{T'}(x) \in a_T(x)$. Let $\Xi(T)$ denote the set of all possible realization of T, then the SSDT for T is de£ned as

$$\mathcal{S} = (\Xi(T), R_u)$$

where R_u is the universal relation, i.e., for each T_i and $T_j \in \Xi(T)$, $(T_i, T_j) \in R_u$.

Thus the language of uncertain modal logic(UDL) contains only two modalities $[u]$ and $\langle u \rangle$ and we will denote them by the ordinary alethic modalities $\Box$ and $\Diamond$ respectively.

4.2 Epistemic decision logic

For a data table $T = (U, A, \{a_T \mid a \in A\})$ and a subset of attributes $B \subseteq A$, let us de£ne the information function Inf^T_B as a mapping from U to the vectors of attribute values by

$$Inf^T_B(x) = (a_T(x))_{a \in B}.$$

Then an equivalence relation $\equiv^T_B$ over U can be formulated with respect to T and B by $x \equiv^T_B y$ iff $Inf^T_B(x) = Inf^T_B(y)$. For each subset E of U, let $Inf^T_B(E)$ denote the multiset $\{Inf^T_B(x) \mid x \in E\}$. Then two data tables $T_1 = (U, A, \{a_{T_1} \mid a \in A\})$ and $T_2 = (U, A, \{a_{T_2} \mid a \in A\})$ are said to be epistemically indistinguishable with respect to $B \subseteq A$ iff $\equiv^{T_1}_B$ is the same as $\equiv^{T_2}_B$ and for each equivalence class E of $\equiv^{T_1}_B$, $Inf^{T_1}_{A-B}(E) = Inf^{T_2}_{A-B}(E)$.

Now, given a data table T, we can de£ne its epistemic SSDT as

$$\mathcal{S} = (\mathcal{T}, \{R_B \mid B \subseteq A\})$$

where $\mathcal{T}$ is the set of all data tables which are epistemically indistinguishable with T with respect to some $B \subseteq A$ and for each R_B and any two tables $T_i, T_j \in \mathcal{T}$, $(T_i, T_j) \in R_B$ iff T_i and T_j are epistemically indistinguishable with respect to B. Following the notations of epistemic logic[2], the modalities $[B]$ will be denoted by K_B.

4.3 Temporal decision logic

De£nition 6 *A (linear-time) temporal SSDT is of the form*

$$\mathcal{S} = (\{T_i \mid 0 \leq i \leq n-1\}, \{R_+, R_<\})$$

where

- *each T_i is a data table,*
- *$(T_i, T_j) \in R_+$ iff $j = i + 1$, and*
- *$(T_i, T_j) \in R_<$ iff $i < j$.*

The modalities $[+]$ and $[<]$ corresponds to the "next" and "future" operators in ordinary temporal logic and will be denoted by $\bigcirc$ and $\Box$ respectively.

5 Conclusion

References

[1] P. Blackburn. "Representation, reasoning, and relational structures: a hybrid logic manifesto". *Logic Journal of IGPL*, 8(3):339–365, 2000.

[2] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. MIT Press, 1996.

[3] T.F. Fan, W.C. Hu, and C.J. Liau. "Decision logics for knowledge representation in data mining". In *Proceedings of the 25th Annual International Computer Software and Applications Conference(COMPSAC)*, pages 626–631. IEEE Press, 2001.

[4] P. Gärdenfors, editor. *Generalized Quanti£ers*. Dordrecht: Reidel Publishing Company, 1987.

[5] P. Hájek. "Logics for data mining (GUHA rediviva)". *Neural Network World*, 10:301–311, 2000.

[6] R. Kruse, C. Borgelt, and D. Nauck. "Fuzzy data analysis: challenges and perspectives". In *Proceedings of the 8th IEEE International Conference on Fuzzy Systems*, pages 1211–1216, San Francisco, CA, 1999. IEEE.

[7] M. Krynicki, M. Mostowski, and L.W. Szczerba, editors. *Quanti£ers: Logics, Models and Computation*. Kluwer Academic Publishers, 1995.

[8] C.J. Liau and D.R. Liu. "A logical approach to fuzzy data analysis". In J.M. Zytkow and J. Rauch, editors, *Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, LNAI 1704, pages 412–417. Springer-Verlag, 1999.

[9] C.J. Liau and D.R. Liu. "A Possibilistic decision logic with applications". *Fundamenta Informaticae*, 46(3), 2001.

[10] Z. Pawlak. *Rough Sets–Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, 1991.

[11] Y.Y. Yao and C.J. Liau. "A generalized decision logic language for granular computing". In *Proceedings of the 11th IEEE International Conference on Fuzzy Systems*, page to appear. IEEE Press, 2002.

[12] Y.Y. Yao and Q. Liu. "A generalized decision logic in interval-set-valued information tables". In N. Zhong, A. Skowron, and S. Ohsuga, editors, *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*, LNAI 1711, pages 285–293. Springer-Verlag, 1999.

[13] Y.Y. Yao and N. Zhong. "An analysis of quantitative measures associated with rules". In *Proceedings of the 2nd Paci£c-Asia Conference on Knowledge Discovery and Data Mining*, pages 479–488. IEEE Press, 1999.

[14] L.A. Zadeh. "Fuzzy logic = computing with words". *IEEE Transactions on Fuzzy Systems*, 4:103–111, 1996.