Belief Classification Approach based on Dynamic Core for Web Mining database

Salsabil Trabelsi[1], Zied Elouedi[1], and Pawan Lingras[2]

[1] Larodec, Institut Superier de Gestion de Tunis, Tunisia
[2] Saint Mary's University Halifax, Canada

Abstract. In this paper, we apply a classification system based on the hybridization of rough set methodology and generalization distribution table to learn decision rules from uncertain data consisting of web usage. The uncertainty appears only in decision attributes and is handled by the belief function theory. Due to the uncertainty existing in the chosen web usage mining database, the feature selection step used to construct our classifier will be based on dynamic core approach which allows getting better performance in uncertain and large database.

1 Introduction

The Rough Sets (RS) [4] constitutes a sound basis for data mining. It is a successful technique of classification applied in several real-world applications which performs feature selection before generating rules. Many researchers have extended rough sets and its applications to accommodate uncertainty and incomplete data [1, 2, 6]. In this paper, we apply new classification approach denoted by Belief Rough Set Classifier based on Generalization Distribution Table (BRSC-GDT) [8] to discover hidden patterns from uncertain web usage database. The uncertainty appears only in decision attributes and is handled by the belief function theory as understood in the transferable belief model (TBM). The web usage mining dataset used in this paper was obtained from web access logs of the introductory computing science course at Saint Mary's University. The visits were clustered based on study patterns. Instead of using crisp assignment of a visit to one of the three clusters (Studious, Crammers, Workers), the study associated a basic belief assignment (bba). The resulting uncertain clustering was characterized using belief functions. To better handle the noise and the uncertain issues existing in the chosen web usage mining database, the feature selection step used to construct our classifier will be based on dynamic core which consists of the more stable attributes which cannot be reduced from the decision table and the more frequently appearing core in sub-decision tables created by random samples of a given decision table. Decision rules induced by means of dynamic core are more appropriate to classify new objects.

2 Rough set theory

In this section, we recall some basic notions related to RS [4]. Due to space limitations, a familiarity with rough set theory is assumed. A decision table (DT) is defined as A= (U, C, {d}), where $U = \{o_1, o_2, ..., o_n\}$ is a nonempty finite set of n objects called *the universe*, $C = \{c_1, c_2, ..., c_k\}$ is a finite set of k *condition* attributes and $d \notin C$ is a distinguished attribute called *decision*. The value set of d, called $\Theta = \{d_1, d_2, ..., d_s\}$. The rough sets adopt the concept of indiscernibility relation to partition the object set U into disjoint subsets, denoted by U/B or IND_B. The partition that includes o_j is denoted $[o_j]_B$. Let $B \subseteq C$ and $X \subseteq U$. We can approximate X by constructing the $B - lower$ and $B - upper\ approximations$ of X, denoted $\underline{B}$(X) and $\bar{B}$(X), respectively. A reduct is a minimal subset of attributes from C that preserves the lower and upper approximations and the ability to perform classifications. The core is the most important subset of attributes, it is included in every reduct.

3 Belief function theory

In this section, we briefly review the main concepts underlying the belief function theory [5] as interpreted in the $Transferable\ Belief\ Model$ (TBM) [7]. The latter is a useful model to represent quantified belief functions.

3.1 Basic concepts

Let Θ be a finite set of elementary events to a given problem, called the frame of discernment. All the subsets of Θ belong to the power set of Θ, denoted by 2^{Θ}.

The impact of a piece of evidence on the different subsets of the frame of discernment Θ is represented by a basic belief assignment (bba).

The bba is a function $m : 2^{\Theta} \rightarrow [0, 1]$ such that:

$$\sum_{E \subseteq \Theta} m(E) = 1 \tag{1}$$

The value *m(E)*, named a basic belief mass (bbm), represents the portion of belief committed exactly to the event E.

Associated with m is the belief function, denoted *bel*, corresponding to a specific bba m. The belief function *bel* assigns to every subset E of Θ the sum of masses of belief committed to every subset of E by m [5]. Contrary to the bba which expresses only the part of belief that one commits to E without being also committed to $\bar{E}$.

The belief function *bel* is defined for $E \subseteq \Theta$, $E \neq \emptyset$ as:

$$bel(E) = \sum_{\emptyset \neq F \subseteq E} m(F) \quad (2)$$

4 BRSC-GDT based on dynamic core

Our classification approach called Belief Rough Set Classifier based on Generalization Distribution Table (BRSC-GDT) [8] can generate a set of rules from uncertain databases with the minimal description length, having large strength and covering all instances. This method is based on the hybrid system GDT-RS [9]. The latter is a combinaison of the Generalization Distrution Table (GDT) and the rough sets (RS). The main steps to construct BRSC-GDT method from uncertain decision table under the belief function framework based on dynamic core approach for feature selection are as follows:
Step 1. Creation of the GDT: This step can be omitted because the prior distribution of a generalization can be calculated [9].
Step 2. Feature selection based on dynamic Core: We remove the superfluous condition attributes that are not in reducts. Nevertheless, computing reducts from uncertain and noisy data leads to results which are unstable and sensitive to the sample data. Therefore it is important to search stable reduct containing dynamic core (if it is also a core of all subtables from a given family F).
Step 3. Definition of the compound object: Considering the indiscernibily classes with respect to the condition attribute subset B as one object, called the compound object.
Step 4. Elimination of the contradictory compound objects: Removing the compound objects where the decision of their composing objects are not very similar.
Step 5. Minimal description length of decision rule: Determining the sets of all reduct values of the compound object.
Step 6. Selection of the best rules: Constructing the decision rules obtained from the local reducts for each coumpond object. Then, we select only the best rule having the best strength.

5 Experimental results

In this section, we will report the experimental results by applying our classification technique on web usage mining dataset. Table 1 summarizes the results relative to the three evaluation criteria : time requirement of learning, size of models and Perecent of the Correct Classicfication (PCC).

6 Conclusion

In this paper, belief classification approach based on rough sets named BRSC-GDT have been applied to generate a classification model from uncertain data

Table 1. The experimental results

Approaches	PCC (%)		Size	Time
	certain case	uncertain case		requirement
BRSC-GDT	85.27	89.63	37	127

consisting of web usage. The uncertainty appears only in decision attributes and is handled by the TBM. The feature selection step used to construct the BRSC-GDT is based on the calculation of dynamic core to extract more relevant and stable features for the classification process. We find interesting results that may encourage users or experts in the web domain to use our classifiers.

References

1. Grzymala-Busse, J.W.: Rough set strategies to data with missing attribute values. In: Workshop Notes, Foundations and New Directions of Data Mining, the 3-rd International Conference on Data Mining, Melbourne, Florida, 56-63, 2003.
2. Grzymala-Busse, J.W. and Siddhaye, S.: Rough Set Approaches to Rule Induction from Incomplete Data. Proceedings of the IPMU'2004, Perugia, Itally, July 4-9, Vol 2, 2004 (923-930).
3. Lingras, P. and West, C.: Interval Set Clustering of Web Users with Rough K-means, Journal of Intelligent Information Systems, Vol 23(1), 2004(5-16).
4. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Dordrecht: Kluwer Academic Publishing, 1991.
5. Shafer, G.: A mathematical theory of evidence. Princeton: Princeton University Press, 1976.
6. Skowron, A. and Grzymala-Busse, J.W.: From rough set theory to evidence theory, In Advances in the Dempster-Shafer Theory of Evidence, New York, (1994) 193-236.
7. Smets, P. and Kennes, R.: The transferable belief model. Artificial Intelligence, Vol 66 (2), (1994) 191-234.
8. Trabelsi, S., Elouedi, Z. and Lingras, P.: Rule discovery process based on rough sets under the belief function framework, IPMU 2010, LNAI 6178, (2010) 726-736.
9. Zhong, N., Dong, J.Z. and Ohsuga, S.: Data mining: A probabilistic rough set approach, In Rough Sets in Knowledge Discovery, Vol 2, (1998) 127-146.