Expert Systems with Applications 38 (2011) 4889–4900

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Autonomous rule induction from data with tolerances in customer relationship management

Tzu-Liang (Bill) Tseng [a], Chun-Che Huang [b,*], Yu-Neng Fan [c]

[a] Department of Mechanical and Industrial Engineering, The University of Texas at E1 Paso, 500 West University Avenue, El Paso, TX 79968, USA
[b] Department of Information Management, National Chi Nan University, No. 1, University Road, Puli, Nantou 545, Taiwan, ROC
[c] Department of Information Management, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan, ROC

ARTICLE INFO

Keywords:
Rough set theory
Rule induction
Data with tolerance
Feature selection
Customer relationship management

ABSTRACT

In this paper, application of the rough set theory (RST) to feature selection in customer relationship management (CRM) is introduced. Compared to other methods, the RST approach has the advantage of combining both qualitative and quantitative information in the decision analysis, which is extremely important for CRM. Automated decision support for CRM has been proposed in recent years. However, little work has been devoted to the development of computer-based systems to support CRM in rule induction. This paper presents a novel rough set based algorithm for automated decision support for CRM. Particularly, the approach is capable to handle real numbers instead of integer numbers through introduction of converted numbers involving tolerances. Being unique and useful in solving CRM problems, an alternative rule extraction algorithm (AREA) is presented for discovering preference-based rules according to the reducts which contain the maximum of strength index (SI) in the same case, where the data with tolerance. The empirical data set associated with CRM has proven the validity and reliability of these approaches. This research thus contributes to developing and validating a useful approach to automated decision support for CRM. This paper forms the basis for solving many other similar problems that occur in the service industry.

1. Introduction

Customer relationship management (CRM) is defined as the managing of customer relationships on an organizational level through understanding, anticipating and managing of customer needs, based on knowledge gained of the customer, to increase organizational effectiveness and efficiency and thereby increasing profitability (Brown, 2000; Handen, 2000; Nicolett, Andren, & Gilbert, 2000). Companies now recognize that CRM can contribute to a value-creation strategy because of the advantages associated with being a trusted participant in the network or set of strategic alliances that are maintained in a CRM relationship (Morgan & Hunt, 1994; Morris, Brunyee, & Page, 1998). As a result, the use of CRM strategies and tactics now serve as one of the major driving forces behind many companies' efforts to create superior value for their customers and generate a long-term revenue stream for themselves (Kothandaraman & Wilson, 2000; Ulaga & Chacour, 2000). CRM is implemented as a combination of the managerial and marketing issues with detailed collection and analysis of customer related data. Moreover, the importance of the understanding of customer behavioral moves to develop a better relationship in response to their need brings forward customer loyalty management as a critical issue for any business acting in highly competitive markets (Coner & Gungor, 2002). With a great number of customers, how do we identify their interests? The answer can be practiced through understanding of customer preference. Through customers' purchasing history, the product relevance, such as brand, material, size, color, appearance, price, quality can be studied to understand customers' preference toward particular product features (Weng & Liu, 2004). For instance, which customer-oriented features in the video game market (e.g., shape concern, function concern and quality concern, etc.) are critical and can be used to segment consumers? Obviously, feature selection is a core and effective tool to explore more about the critical customer-oriented features.

In CRM, information related to the customer-preferred features are collected by using research, interviews, meetings, questionnaires, sampling, and other techniques. This type of data is often discrete and frequently in "qualitative" format (e.g., salary level, preference level, etc.). To analyze these qualitative data for extracting the useful information to assist promoting sales is critical. Numerous approaches are applied to feature selection like genetic algorithms (Handels, Ross, Kreusch, Wolff, & Pöppl, 1999), neural network (NN) (Arana, Martí-Bonmatí, Bautista, & Paredes, 1998), knowledge engineering (Helma, 2004), Bayesian network

* Corresponding author.
E-mail address: cchuang@ncnu.edu.tw (C.-C. Huang).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.09.098

classifiers (Lee & Lee, 2006), K-nearest neighborhood (KNN) (Uncu & Türksen, 2007). However, these approaches are not imposed on processing qualitative information. They are not suitable in feature selection of CRM because the aforementioned methodologies are with population based approaches which may require several statistical assumptions and they have limitation to handle qualitative type of data in CRM. An individual object model based approach that provides a very good tool for analyzing data is preferred. One of promising approaches to deal with qualitative information and provide an individual object model based approach is the rough set approach (Kusiak, 2001).

The rough set approach seems to be of fundamental importance to AI, especially in the areas of machine learning, knowledge acquisition, decision analysis, inductive reasoning. It seems of particular importance to decision support systems and data mining. The main advantage of rough set theory is that it does not need any preliminary or additional information about data like probability in statistics, grade of membership, or the value of possibility in fuzzy set theory (Pawlak, 1998). The rough set approach is suitable for processing qualitative information that is difficult to analyze by standard statistical techniques (Srikant, Vu, & Agrawal, 1997). It integrates learning-from-example techniques, extracts rules from a data set of interest, and discovers data regularities. The use of rough set to feature selection has been proposed in literature (Swiniarski et al., 1995; Swiniarski & Nguyen, 1996). The main idea is based on calculation of a core for discrete attribute data set, containing strongly relevant features, and reducts, containing a core plus additional weakly relevant features. Numerous rough set based feature selection methodologies can be found in literature but few of then apply in the domain of CRM, where the qualitative data forms as primary information in it. In CRM, the feature selection approach attempts to eliminate as many features as possible in the problem domain, and still carry out useful and meaningful outcomes with acceptable accuracy. Having a minimal number of features often leads to establishing simple models that can be more easily interpreted. Li, Hong, and Nahavandi (2003) developed a client classifying algorithm based on rough set theory which reduced the attributes and worked out the reduced decision form to the rules of rough decision. In order to offer the right services to right persons at the right time through the right channel automatically, Chiang and Lin (2000) used rough sets to analyze to user profiles and transactions for CRM.

Tseng and Huang (2006) proposed an approach to elicit a minimum number of features from n-dimensional feature space to derive inductive rules. The approach named rule identification algorithm (RIA) which develops for discovering preference-based rules according to the reducts which contain the maximum strength. Nevertheless, alternative rules which can be defined as the rules that hold identical preference as the original decision rule and could be more attractive than the original one by a decision-maker. It is not considered in previous literature since the desired reduct rules are not unique. To generate the completed rules in the provided data set, re-running the rule extraction algorithm is time-consumption after certain desired rules are selected. Therefore, an effective alternative rule extraction algorithm is required.

Moreover, numerous rough set based algorithms have been developed for dealing with integer numbers (Lin & Cercone, 1997; Ziarko, 1994). Fairly few are designed to handle real numbers with given tolerance. In general, if the real numbers have been considered, certain issues occur as the real numbers are involved in. One critical issue is related to the precision of real numbers since any real numbers usually contain some errors caused by measurement. Two factors are usually relevant to the generation of converted numbers from real numbers. The first is the tolerance associated with each attribute while the second is the threshold value which determines if the two adjacent numbers are treated equally. Basically, considering the data set with tolerance is another challenge in the RS theory.

The heuristic approach of converted numbers which involve tolerance is introduced according to Tseng (1999). The approach assists in dealing with the problem employed by real number. It is an innovative approach in rough set field since many rough set based algorithms are developed for dealing with scaled (integer) numbers (Lin & Cercone, 1997; Ziarko, 1994) while very few are designed for real numbers. The tolerance and threshold value play an important role in this transformation procedure. Furthermore, the proposed approach also considers the alternative rules and inducts decision rules from a dynamic database, which overcomes the weakness of previous RS based approaches. Consequently, the final outcomes should be more attractive by a decision-maker comparing with the original approach. In this paper, it is assumed that the tolerance and threshold associated with each attribute are provided. A heuristic approach to determine the converted numbers from real number data set is developed and an alternative rules search methodology is also introduced. The remainder of this paper is organized as follows. Section 2 introduces the concept of modeling the tolerance containment for a number and a heuristic approach to determine the converted numbers. Section 3 presents an alternative rules search methodology, while a process of the synthesis of tolerance containment and inductive rules is illustrated in Section 4. The case study in Section 5 represents how the proposed methodologies can assist CRM. Concluding remarks are provided in Section 6.

2. Modeling data with tolerances

Tolerances are specifications for the performance characteristics of objects (Chase & Parkinson, 1991). Traditionally, tolerances have been stated in the form of a lower specification limit (LSL) and an upper specification limit (USL). Taguchi (1987) has championed a form of the tolerance statement that emphasizes the target and the importance of achieving the exact target. His form is $\tau \pm \delta$, where δ denotes the maximum allowable deviation from the target τ. Tolerance is the integration of a measurement-driven specification and the capabilities of the system, product, and/or process. The capability of the system is measured by both the mean and the variability of its important performance measures. These measures, in turn, are directly dependent on the means and variabilities of the subsystems, components that constitute the product or system. A careful analysis and assignment of tolerances for objects in the problem domain can significantly improve the quality (precision) of solutions to cope with the errors of measurement (Kusiak & Feng, 1995; Kusiak, 1993). A physician usually selects tolerances based on his/her experiences and the functionality of an object. Then a safety factor (i.e., tightening of tolerances by some factor) is used to compensate for the lack of domain knowledge. Assuming a tolerance and a threshold value are provided for each attribute in the database, the comparison of numbers is discussed next.

2.1. Interval analysis

In order to compare two adjacent numbers, the following cases are considered:

Case (a) two adjacent numbers are identical;

$$\text{overlap} = 2 \times \text{Tol} \tag{1}$$

Case (b) two adjacent numbers are superimposed;

Type (I) diff < Tol

$$\text{overlap} = 2 \times \text{Tol} - \text{diff} \tag{2}$$