Knowledge-Based Systems 91 (2016) 287–292

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Multi-granular mining for boundary regions in three-way decision theory

Jie Chen [a,b,c], Yan-ping Zhang [a,b,c], Shu Zhao [a,b,c,*]

[a] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, PR China
[b] Center of Information Support & Assurance Technology, Anhui University, PR China
[c] School of Computer Science and Technology, Anhui University, Hefei, Anhui Province 230601, PR China

ARTICLE INFO

Article history:
Received 25 January 2015
Revised 15 September 2015
Accepted 6 October 2015
Available online 24 October 2015

Keywords:
Boundary regions
Multi-granular three-way decision algorithm
Covering algorithm
Multiple-views of granularity
Pairs of heterogeneous points

ABSTRACT

In three-way decision theory, all samples are divided into three regions: a positive region, a negative region, and boundary regions. A lack of detailed information may make a definite decision impossible for samples in boundary regions, and hence the third non-commitment option is used. Reducing boundary regions is a new problem. In this paper, the multi-granular three-way decision (*MGTD*) algorithm is presented to reduce boundary regions. At the beginning of the multi-granular process, samples are divided using the covering algorithm, which does not need a threshold. Then pairs of heterogeneous points (*HPs*) are defined in boundary regions to obtain diversity information. This detailed information is used to define attribute subsets. Eventually, boundary regions are further investigated using multiple-views of granularity. Each view corresponds to an attribute subset. Experiments have shown that the *MGTD* algorithm is beneficial for reducing boundary regions and improving classification precision in most cases.

1. Introduction

In the conventional two-way decision model, there are only two options for a decision: positive or negative, regardless of whether information is lacking. This approach may result in wrong decisions when the information is insufficient. To address this issue, Yao proposed a three-way decision model, which extends two-way decision theory by incorporating an additional choice: the boundary decision, from Pawlak rough sets to probability rough sets (*DTRS*) [1–4]. Hence, a sample is classified into the positive, negative, or boundary region based on three-way decision theory. These classes can be interpreted as acceptance, rejection, and uncertainty. In recent years, researchers have focused on three-way decision theory, for example using new probability rough sets [5], multi-granulation rough sets [6,7], multi-granulation decision-theoretic rough sets [8], multi-granulation rough set based covering [9], neighborhood-based multi-granulation rough sets [10], and improved three-way decision models [11–18]. Three-way decision theory has been widely used in many applications such as spam filtering [19–21], text classification [22], medical diagnosis [23,24], management theory [25–27], social judgment theory [28], paper review [29], risk preferences for decision-making [30], oil exploration decisions [31], and automatic clustering [32–34].

The main superiority of three-way decision theory compared to two-way decision theory is the utility of the boundary decision [1]. In three-way decision theory, the boundary decision is regarded as a feasible decision choice when the available information for decision-making is too limited to make a proper decision. This is similar to human decision-making strategy in practical decision problems. However, it is also a disadvantage of three-way decision theory that the boundary regions need further investigation. Reducing boundary regions poses a new problem [35].

Samples in boundary regions with a non-commitment decision may be further investigated. Some researchers have focused on the processing of boundary regions. Li and Zhou used the idea of a tri-training algorithm [36] and proposed two tri-training algorithms, *TW* and *TR*, to reduce boundary regions [35]. Yao proposed sequential three-way decisions to make a definite acceptance or rejection decision for uncertain samples [37]. These methods investigated boundary regions using original information. However, samples were placed into the boundary region because the original information was not sufficient to decide. Therefore, more detailed information is needed to make further decisions. In an effort to solve this problem, the authors have proposed a method to obtain diversity information from samples in boundary regions based on the constructive covering algorithm(*CCA*).

* Corresponding author at Anhui University School of Computer Science and Technology No. 111, Jiulong Road Hefei, Anhui Province 230601 China.
Tel.: +86 13856002964.
E-mail addresses: chenjie200398@163.com (J. Chen), zhaoshuzs2002@hotmail.com (S. Zhao).
URL: http://ailab.ahu.edu.cn (S. Zhao)

http://dx.doi.org/10.1016/j.knosys.2015.10.020
0950-7051/

The *CCA* is a constructive supervised learning algorithm that maps all samples in the data set to an $n+1$-dimensional sphere S^{n+1}. A three-way decision model based on the *CCA* was proposed by Zhang and Xing [38] to overcome a major challenge with three-way decision models. This challenge is the acquisition of a set of pairs of thresholds (α, β). Thresholds are calculated by minimizing the decision loss [1]. The *CCA* does not need thresholds α and β, but classifies samples using their own characteristics. A cost-sensitive three-way decision model based on the *CCA* (*CCTDM*) [39] and a robust three-way decision model based on the *CCA* (*RTDM*) [40] have been proposed by Zhang and Zou to improve performance.

Based on the *CCA*, pairs of heterogeneous points(*HPs*) are found in boundary regions to obtain diversity information. This detailed information is used to define attribute subsets. Each attribute subset corresponds to a view of granularity. Eventually, samples in boundary regions are classified using different attribute subsets. Therefore, boundary regions are mined in a multi-granular fashion.

The rest of this paper is organized as follows: in Section 2, related research in the literature is introduced. In Section 3, the multiple-views of granularity in the *CCA* are introduced in detail, and a multi-granular three-way decision (*MGTD*) algorithm for boundary regions is proposed. Section 4 presents the analysis of the experimental results, and conclusions are drawn in Section 5.

2. Related work

2.1. An overview of three-way decisions

In a three-way decision model, decision actions are denoted by $A = \{a_P, a_N, a_B\}$, representing *POS*, *NEG*, and *BND* decisions and called positive, negative, and boundary regions respectively. The positive region *POS* consists of those objects that are accepted as satisfying the conditions, and the negative region *NEG* consists of those objects that are rejected as not satisfying the conditions. *BND* decisions are a third decision choice, which means that more information must be collected before making a further precise decision.

In three-way decision theory, the data set is denoted as a decision information table $S = (U, At = C \bigcup D, \{V_a | a \in At\}, \{I_a | a \in At\})$ [41]. By introducing a pair of thresholds (α, β), $\alpha \succ \beta$, three regions can be constructed as follows:

$$POS_{(\alpha,\beta)}(v) = \{x \in U | v(x) \succeq \alpha\}$$
$$NEG_{(\alpha,\beta)}(v) = \{x \in U | v(x) \preceq \beta\}$$
$$BND_{(\alpha,\beta)}(v) = \{x \in U | \beta \prec v(x) \prec \alpha\}$$

The cost λ_{ij} forms a matrix denoted as $(\lambda_{ij})_{2\times 3}$ since $i \in \{P, B, N\}$, and $j \in \{P, N\}$. Normally, the costs of a right decision are less than those of a wrong decision, and therefore $\lambda_{PP} \le \lambda_{BP} \le \lambda_{NP}$ and $\lambda_{NN} \le \lambda_{BN} \le \lambda_{PN}$.

Based on the properties of *DTRS*, thresholds (α, β) can be determined using the cost matrix $(\lambda_{ij})_{2\times 3}$ since $i \in \{P, B, N\}$ and $j \in \{P, N\}$ can be described as follows:

$$\alpha = \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})} \quad (1)$$

$$\beta = \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})} \quad (2)$$

2.2. Three-way decision model based on the CCA

For three-way decision models, the acquisition of a set of pairs of thresholds (α, β) presents a major challenge. Thresholds are calculated by minimizing the decision loss [1]. The *CCA* is a constructive supervised learning algorithm that maps all samples in the data set to an $n+1$-dimensional sphere $S^n + 1$. Sphere neighborhoods are used to classify the samples [42]. The *CCA* can construct neural networks (*NNs*) based on sample characteristics. Three-way decision models based on the *CCA* do not need the thresholds (α, β). Details of the *CCA* are presented below.

Definition 2.1 (Cover)**.** Assume that the domain of input vectors is a bounded set X of an n-dimensional space. A transformation T can be defined as:

$$T : X \to S^{n+1}, T(x) = (x, \sqrt{\gamma(\varphi)^2 - |x|^2}), x \in X \quad (3)$$

In this way, all points in X are projected upward onto S^{n+1} by transformation T. Notably, in this situation, a neuron (ω, φ) corresponds to a characteristic function of a "sphere neighborhood" on S^{n+1} with ω as its center and $\gamma(\varphi)$ as its radius. This sphere neighborhood can cover a number of input vectors belonging to the same class t. This sphere is therefore called a cover, $c(t)$.

Given a set of training samples $X = \{x_1, x_2, \ldots, x_n\}$, $(i = 1, 2, \ldots, n)$, which is the set in n-dimensional Euclidean space. Then $A_i = \{A_i^1, A_i^2, \ldots, A_i^m\}$ is an m-dimensional characteristic attribute of the ith sample. The *CCA* finally obtains a set of covers $C = \{C_1^1, C_2^1, \ldots, C_{s_1}^1, C_1^2, C_2^2, \ldots, C_{s_2}^2, \ldots, C_1^k, C_2^k, \ldots, C_{s_k}^k\}$, where C_i^j represents the ith cover of the jth category. It is assumed that $C^j = \bigcup C_i^j$, $i = \{1, 2, \ldots, s_j\}$. C^j represents all covers of the jth category samples.

In a three-way decision model, only two categories y_1 and y_2 are assumed. Each category has at least one cover. The covers of C^{y_1} and C^{y_2} are $\{C_1^{y_1}, C_2^{y_1}, \ldots, C_r^{y_1}\}$ and $\{C_1^{y_2}, C_2^{y_2}, \ldots, C_s^{y_2}\}$, respectively, i.e., $C^{y_1} = \{C_1^{y_1}, C_2^{y_1}, \ldots, C_r^{y_1}\}$, $C^{y_2} = \{C_1^{y_2}, C_2^{y_2}, \ldots, C_s^{y_2}\}$. $POS(C^{y_1})$ is defined by the difference of unions $\bigcup C_i^{y_1} - \bigcup C_j^{y_2}$, $NEG(C^{y_1})$ by $\bigcup C_j^{y_2} - \bigcup C_i^{y_1}$, and $BND(C^{y_1})$ by the rest, where $i = \{1, 2, \ldots, s\}, j = \{1, 2, \ldots, t\}$. Likewise, $POS(C^{y_2})$ is equal to $NEG(C^{y_1})$; $NEG(C^{y_2})$ is equal to $POS(C^{y_1})$; and $BND(C^{y_1})$ is equal to $BND(C^{y_2})$.

3. Multi-granular theory based on the *CCA*

In this section, a *MGTD* model is proposed to reduce boundary regions, and its advantages over other three-way decision algorithms are demonstrated. The *MGTD* is based on the multiple-views of granularity in the *CCA* (*MVCA*). Therefore, the principle of *MVCA* will be introduced, and then *MGTD* will be described in detail.

3.1. Multiple-views of granularity in the CCA

In real-world decision making, it is possible to consider multiple-views that eventually lead to two-way decisions. With each view, more new information is acquired. This paper presents a new approach using the notion of multiple-views of granularity. The whole attribute set of samples is divided into different subsets, with each subset representing a view of granularity. According to the definition of the pairs of *HPs*, diversity information will be obtained more and more accurately. The *MVCA* is proposed. The definition of *HPs* and their attribute bias is presented below.

Definition 3.1 (Pairs of Heterogeneous Points (*HPs*))**.** Assume that there is a sample set $X = \{x_1, x_2, \ldots, x_n\}$, where n is the number of samples, and let attribute set $A = \{A_1, A_2, \ldots, A_m\}$, where m is the number of attributes. For $\forall x_i \in X, \exists x_j \in X$, such that $y(x_i) \neq y(x_j)$ satisfy:

$$HPs(i, j) = \{(x_i, x_j) | \forall k, x_k \in X, d(x_i, x_j) \le d(x_i, x_k), y(x_k) = y(x_j) \neq y(x_i)\} \quad (4)$$

where $y(x_i)$ is the class of sample x_i and $d(x_i, x_j)$ denotes the Euclidean distance between x_i and x_j. Sample x_j is the nearest sample to sample x_i. Therefore, (x_i, x_j) is called a pair of *HPs*. In Fig. 1, the nearest other sample to sample 1 is sample a, and therefore $(1, a)$ is *HPs*. Likewise, $(2, b)$, $(3, c)$, $(4, d)$ are *HPs*.