ARTICLE IN PRESS

IJA:7978

International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Tri-partition neighborhood covering reduction for robust classification [☆]

Xiaodong Yue [a], Yufei Chen [b,*], Duoqian Miao [b], Jin Qian [c]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
[b] College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China
[c] School of Computer Engineering, Jiangsu University of Technology, Changzhou 213001, China

ARTICLE INFO

Article history:
Received 20 June 2016
Received in revised form 4 October 2016
Accepted 10 November 2016
Available online xxxx

Keywords:
Neighborhood covering reduction
Three-way decision
Rule extraction

ABSTRACT

Neighborhood Covering Reduction extracts rules for classification through formulating the covering of data space with neighborhoods. The covering of neighborhoods is constructed based on distance measure and strictly constrained to be homogeneous. However, this strategy over-focuses on individual samples and thus makes the neighborhood covering model sensitive to noise and outliers. To tackle this problem, we construct a flexible Tri-partition Neighborhood for robust classification. This novel neighborhood originates from Three-way Decision theory and is partitioned into the regions of certain neighborhood, neighborhood boundary and non-neighborhood. The neighborhood boundary consists of uncertain neighbors and is helpful to tolerate noise. Besides the neighborhood construction, we also proposed complete and partial strategies to reduce redundant neighborhoods to optimize the neighborhood covering for classification. The reduction process preserves lower and upper approximations of neighborhood covering and thereby provides a flexible way to handle uncertain samples and noise. Experiments verify the classification based on tri-partition neighborhood covering is robust and achieves precise and stable results on noisy data.

1. Introduction

In the area of soft computing, Rough Set theory has been proven to be an effective tool for uncertain data analysis [27–29]. Classic rough set model specializes in feature selection and rule extraction from the table-formed symbolic data and thus is widely applied in the tasks of structural data mining, concept learning and rule-based expert systems [7,24,31]. To handle the data of both symbolic and numerical features, rough set model was extended to neighborhood systems [15,20,21,37]. Different from the equivalence classes defined for symbols in classic rough sets, the neighborhoods in numerical data space are basic granules to constitute Neighborhood Rough Sets [35,44]. For a sample, its neighborhood consists of the neighbors surrounding it [22,25,32]. Neighborhood rough sets actually provide us a way to formulate data space on neighborhood level [20,43]. From the view of topology, it has been proven that the concepts derived from neighborhood spaces are more general than from data-level spaces [21]. This indicates that transforming original data space into a neighborhood system will facilitate the data generalization [35,37].

☆ This paper is part of the virtual special issue on tri-partition, edited by Davide Ciucci and Yiyu Yao.

* Corresponding author.
E-mail address: april337@163.com (Y. Chen).

http://dx.doi.org/10.1016/j.ijar.2016.11.010
0888-613X/© 2016 Elsevier Inc. All rights reserved.

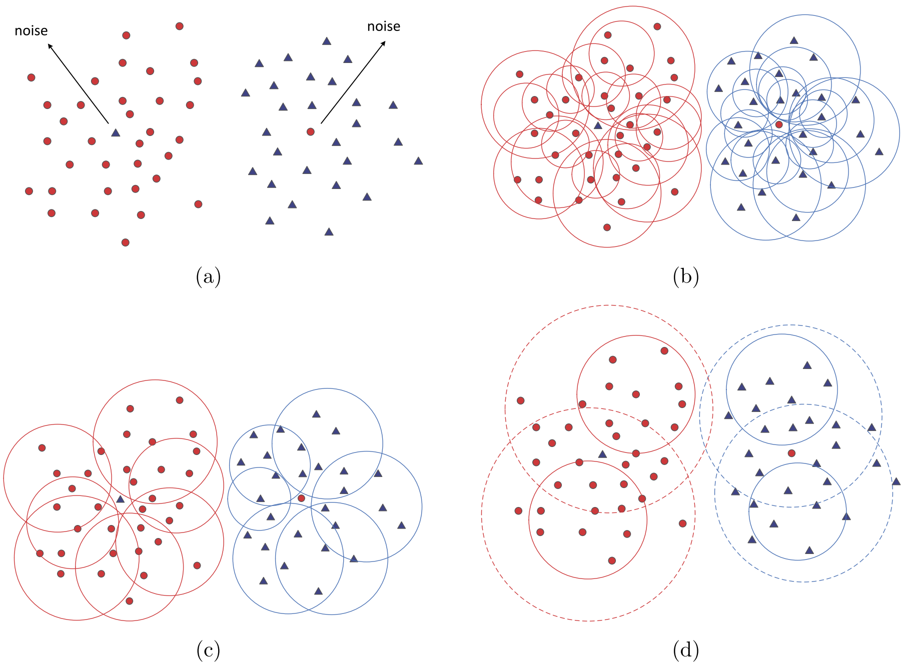

Fig. 1. Neighborhood covering reduction on noisy data, (a) original data with noise, (b) initial neighborhood covering, (c) reduced neighborhood covering, (d) flexible covering of tri-partition neighborhoods.

To formulate neighborhood systems of data space, neighborhood rough sets were proposed to form neighborhood coverings of data samples and the union of neighborhoods constitute an approximate representation of data space [16,26,33]. Based on this, the methods of Neighborhood Covering Reduction (NCR) were further proposed to approximate data space with neighborhoods for learning tasks [6,36]. Specifically, the samples are initially grouped into neighborhoods based on distance measure and then redundant neighborhoods are iteratively reduced to generate a concise covering of data space for classification [12,41] and feature selection [14,15]. In most existing neighborhood covering models, neighborhoods are constrained to be homogeneous, i.e. all the samples in a neighborhood should belong to the same class, and the boundary of neighborhood is decided by the nearest heterogeneous sample. However, this strategy of neighborhood construction over-focuses on individual samples and thus makes the model sensitive to noise and outliers. Fig. 1(a–c) shows the disturbance of noise to the formulation of neighborhood covering. Because of the embedded noisy data, the neighborhoods over-partition the data space and the distribution of samples belonging to the same class is broken up.

Aiming to handle the problem above, we expect to construct more flexible neighborhoods for robust classification. The methodology of neighborhood construction originates from the theory of Three-way Decision (3WD) [38–40]. Generally speaking, in the process of three-way decision, knowledge are extracted from the data with uncertainty through tri-partitioning data space into Positive, Negative and Boundary regions [1,10,13]. From the decision view, the three regions correspond to the cases of acceptance, rejection and non-commitment (uncertain case) respectively [5,19,23]. Similarly, in a neighborhood system, a flexible neighborhood should contain certain and uncertain neighbors to tolerate noise and thereby could be tri-partitioned into the regions of certain neighborhood, uncertain boundary and non-neighborhood. The certain neighborhood consists of the neighbors certainly belonging to the same class, the neighborhood boundary involves uncertain neighbors may belong to different classes and the non-neighborhood region denotes the samples outside the neighborhood.

As introduced above, motivated by three-way decision, we extend traditional neighborhood to Tri-partition Neighborhood, briefly denoted by T-neighborhood, to improve the robustness of neighborhood-based model for classification. T-neighborhood consists of inner and outer parts, which involve certain and uncertain neighbors respectively, see Fig. 1(d). For a sample, its inner and outer neighborhoods are decided according to the proportion of the heterogeneous samples surrounding it. As neighborhood coverings form an approximation of data space, the coverings of T-neighborhoods will lead to multilevel approximations of data space. Specifically, the union of inner neighborhoods constitutes the Lower Approximation of data space, in which the samples in neighborhoods certainly belong to a class. And the outer neighborhoods comprise the Upper Approximation of data space, in which the samples in neighborhoods belong to a class with uncertainty. With the lower and upper approximations of data space, the covering of T-neighborhoods is able to tolerate uncertain samples and thus weaken the disturbance of noise.

Besides the neighborhood construction, the reduction strategies of T-neighborhoods are also proposed to optimize neighborhood coverings for classification. The reduction objectives aim to filter out redundant neighborhoods while preserving the approximations of data space. In the reduction process, T-neighborhoods can be categorized into complete/partial reducible and irreducible. The classifier based on irreducible T-neighborhoods generalizes data space well and is robust for