Knowledge-Based Systems 000 (2018) 1–12

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Local neighborhood rough set

Qi Wang[a,b,c], Yuhua Qian[a,b,c,*], Xinyan Liang[a,b,c], Qian Guo[a,b,c], Jiye Liang[b]

[a] Institute of Big Data Science and Industry, Shanxi University, Taiyuan, Shanxi 030006, China
[b] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China
[c] School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China

ARTICLE INFO

Article history:
Received 28 November 2017
Revised 18 April 2018
Accepted 19 April 2018
Available online xxx

MSC:
00-01
99-00

Keywords:
Rough set
Local neighborhood rough set
Concept approximation
Attribute reduction
Limited labeled data

ABSTRACT

With the advent of the age of big data, a typical big data set called limited labeled big data appears. It includes a small amount of labeled data and a large amount of unlabeled data. Some existing neighborhood-based rough set algorithms work well in analyzing the rough data with numerical features. But, they face three challenges: limited labeled property of big data, computational inefficiency and overfitting in attribute reduction when dealing with limited labeled data. In order to address the three issues, a combination of neighborhood rough set and local rough set called local neighborhood rough set (LNRS) is proposed in this paper. The corresponding concept approximation and attribute reduction algorithms designed with linear time complexity can efficiently and effectively deal with limited labeled big data. The experimental results show that the proposed local neighborhood rough set and corresponding algorithms significantly outperform its original counterpart in classical neighborhood rough set. These results will enrich the local rough set theory and enlarge its application scopes.

1. Introduction

Rough set theory was introduced by Pawlak [1–3] as a powerful soft computing tool for modeling and processing uncertainty information. It has been applied to feature selection [4–8], pattern recognition [9,10], uncertainty reasoning [11], granular computing [12–15], data mining and knowledge discovery [16–21]. Over the past decades, it has an enormous impact on the uncertainty management and uncertainty reasoning.

There are two significant notions for rough set. One fundamental notion is concept approximation, in which a general concept represented by a set is always characterized via the so-called upper and lower approximations. Given a data set U and a binary relation R including equivalence relation, tolerance relation, neighborhood relation, dominance relation, and so on, and this given binary relation partitions a data set into a family of concepts, also called a granular structure U/R in granular computing, and each of which is called an information granule used to approximate a target concept [22–24]. One can get a rough set of any subset on the data set via employing information granule from U/R. The other important notion is attribute reduction which can be considered as a kind of specific feature selection [25–28], whose objective is to reduce the number of attributes and to preserve a certain property that we want at the same time. In rough set theory, we are interested in the property of retaining the distinguishing ability provided by the originally whole attribute set [29,30], rather than try to maximize the classification power [31–34]. In other words, based on rough set theory, one can omit irrelevant and redundant attributes that will not influence the discriminability to current recognition tasks [29,35–37] and select useful features from a given data set. Given a set of objects with class labels, some decision rules, which is called a rough classifier, can be obtained by utilizing attribute reduction induced by rough set model. We can predict the class label of an unseen object through using this set of decision rules. Considering this point, classical rough model can be thought as a supervised learning method.

Rough set theory is originally constructed on the basis of an equivalence relation. However, it is limited in many real-world applications. To overcome this limitation, ones extend the equivalence relation to other binary relations, such as similarity relation, tolerance relation, dominance relation and neighborhood relation, to generalize the classical rough sets. Among them, neighborhood rough sets are very important extension to deal with numeric data.

* Corresponding author at: School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China.
E-mail addresses: counter_king@163.com (Q. Wang), jinchengqyh@sxu.edu.cn (Y. Qian), liangxinyan48@163.com (X. Liang), zcguoqian@163.com (Q. Guo), ljy@sxu.edu.cn (J. Liang).

https://doi.org/10.1016/j.knosys.2018.04.023
0950-7051/

Table 1
A data table with limited labeled objects.

Objects	x_1	x_2	$\cdots$	x_p	$\cdots$	x_{n-1}	x_n
a_1	$a_1(x_1)$	$a_1(x_2)$	$\cdots$	$a_1(x_p)$	$\cdots$	$a_1(x_{n-1})$	$a_1(x_n)$
a_2	$a_2(x_1)$	$a_2(x_2)$	$\cdots$	$a_2(x_p)$	$\cdots$	$a_2(x_{n-1})$	$a_2(x_n)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_k	$a_k(x_1)$	$a_k(x_2)$	$\cdots$	$a_k(x_p)$	$\cdots$	$a_k(x_{n-1})$	$a_k(x_n)$
Class labels	d_1	d_2	$\cdots$	d_r			

For convenience, we combine neighborhood rough set [38] with the decision-theoretic rough sets [39,40] into the same rough set model, as a representative, called global neighborhood rough set in this paper. Let (U, N) be a neighborhood approximation space with N being neighborhood relation on U. The lower and upper approximations of the set X are defined as follows.

$$\begin{cases} \underline{N_\alpha}(X) = \{x | \mathcal{P}(X|\delta(x)) \geq \alpha, x \epsilon U\}, \\ \overline{N}_\beta(X) = \{x | \mathcal{P}(X|\delta(x)) > \beta, x \epsilon U\}. \end{cases} \tag{1}$$

where $\mathcal{P}(.)$ is a conditional function, $\delta(x)$ is neighborhood of x and α, β are two parameters from the decision-theoretic rough set:

The existing rough set models have made great achievement in rough data analysis, but they encounter some challenges when handling large-scale data sets. In what follows, we present a detailed description.

(a) Semi-supervised property of big data

Many state-of-the-art algorithms focus on classifiers or regressors from a given training set, where every object must be labeled. With the development of the age of the big data, one can get more data objects than ever. Some methods [41–44] have been proposed to deal with stream data, such as data obtained from all kinds of sensors and that from social media, which increase dynamically. However, these models generally use labeled objects, and these unlabeled objects are not used to construct concept approximation for rough set-based supervised learning, where these algorithms require a large number of labeled data, and labeling these data is expensive and laborious. On the contrary, with the advent of Internet, obtaining unlabeled data becomes easy and cheap. Under the environment of big data, a data set to deal with could be represented as a data table shown in Table 1 (we can call it limited labeled decision table). In the original rough set model, only the data set $\{x_1, x_2, \ldots, x_p\}$ is used, which means that the model cannot use other information provided by unlabeled data. So a semi-supervised learning strategy is necessary, in which it can automatically learn rough classifiers from big data with limited labeled data. This is one motivation of rough data analysis in big data.

(b) Computational inefficiency

From the Eq. (1), we can know that, calculating its lower/upper approximation needs to use all information granules obtained by scanning all objects, which is exceedingly time-costing. And its time complexity is $\mathcal{O}(n^2)$ without pre-ranking and $\mathcal{O}(n \log n)$ with pre-ranking [45,46]. For a large-scale data set, they cannot effectively and efficiently work to satisfy the requirement in real world. How to reduce the time consumption is the second motivation of this study.

(c) Over-fitting in attribute reduction

The over-fitting degree in attribute reduction can be observed by the monotonicity of positive regions of a target decision, which is often measured by the accuracy of approximation in Eq. (6). It is a truth modeling classifier task which is influenced by noise easily [47]. So, we should consider robustness and sensitivity of attribute reduction to noise samples. If the measures used to evaluate significance of attribute in attribute reduction are robust to noisy objects, the performance of the trained classifier would be better. Some existing extended rough set model, such as variable precision rough set [48], decision rough set [39,40], Bayesian rough set [49], probabilistic rough set [40,50,51], etc., can be used to solve this issue. Each of these rough set models can be used to control the degree of uncertainty, misclassification and imprecise information. We can see that for these rough set models, lower/upper approximation of a target concept are often not monotonic with the number of attributes, where objects outside this target concept may be included. How to ensure the monotonicity of an attribute reduction process is also a motivation of this study.

In order to address these three challenges, a new rough set model for rough data analysis in big data, called local neighborhood rough set, is presented. To construct lower/upper approximations of a target concept under the learning framework of the local neighborhood rough set, it is unnecessary to compute information granules of all objects in advance. Only those of objects within a target concept need to be calculated. This saves a great amount of computing time and fully meets the needs of big data analysis. Some interesting properties and measures in the local neighborhood rough set will also be given. Based on the local rough set, the LLAC algorithm for computing a local lower approximation of a target concept and the LARC algorithm for searching a local attribute reduction of a target concept, were designed. Moreover, the LLAD algorithm for calculating a local lower approximation of a target decision and the LARD algorithm for finding a local attribute reduction of a target decision, will be proposed. The one of the advantages of these four algorithms is that their time complexity is linear. Hence, LNRS can fully be apply to rough data analysis in big data. At last, we use four real data sets from UCI and an artificial data set to verify the performance of these four algorithms. Corresponding experiment results show that these algorithms achieve a great success for rough data analysis in big data.

The remainder of this paper is organized as follows. In Section 2, local rough set and neighborhood rough set are reviewed. In Section 3, we first construct the local neighborhood rough set and explore its prime properties and measures. Section 4 provides solutions of how to compute the lower/up local approximation of a target concept and how to find an attribute reduction of a target decision in the local neighborhood rough set. In Section 5, we verify scalability of the local neighborhood rough set on an artificial large-scale data set. Finally, we conclude this paper by outlook for further research and discussion in Section 6.

2. Related work

In this section, we briefly review some basic concepts related to local rough set (LRS) [52] and neighborhood rough set (NRS) [53].

2.1. LRS

For obtaining a rough set $< lower\,approximation, upperapproximation >$ of any subset on sample set, one first computes all the information granules by comparing the difference between any two objects from a given data set. This implies that a global rough set must observe the relationships between a target concept and each of the information granules. However, this is not a good strategy for approximating a target concept $X \subseteq U$. In fact, the information granules $\{[x] : [x] \cap X = \phi, x \in U\}$ are not useful for computing the lower/upper approximation of X. Indeed, we only need to calculate the information granules related to the target concept X. In particular, this kind of large-scale data sets $n >> |X|$ often exist in real applications (even we can have lots of labeled data, we still can obtain more unlabeled data. For

Please cite this article as: Q. Wang et al., Local neighborhood rough set, Knowledge-Based Systems (2018), https://doi.org/10.1016/j.knosys.2018.04.023