Artif Intell Rev
DOI 10.1007/s10462-014-9426-2

Rough set-based approaches for discretization: a compact review

Rahman Ali · Muhammad Hameed Siddiqi · Sungyoung Lee

Abstract The extraction of knowledge from a huge volume of data using rough set methods requires the transformation of continuous value attributes to discrete intervals. This paper presents a systematic study of the rough set-based discretization (*RSBD*) techniques found in the literature and categorizes them into a taxonomy. In the literature, no review is solely based on *RSBD*. Only a few rough set discretizers have been studied, while many new developments have been overlooked and need to be highlighted. Therefore, this study presents a formal taxonomy that provides a useful roadmap for new researchers in the area of *RSBD*. The review also elaborates the process of *RSBD* with the help of a case study. The study of the existing literature focuses on the techniques adapted in each article, the comparison of these with other similar approaches, the number of discrete intervals they produce as output, their effects on classification and the application of these techniques in a domain. The techniques adopted in each article have been considered as the foundation for the taxonomy. Moreover, a detailed analysis of the existing discretization techniques has been conducted while keeping the concept of *RSBD* applications in mind. The findings are summarized and presented in this paper.

Keywords Rough set theory (RST) · Rough set discretization · Data reduction · Real values · Knowledge discovery · Categorization · Taxonomy

R. Ali · M. H. Siddiqi · S. Lee (✉)
Ubiquitous Computing Lab, Department of Computer Engineering, Kyung Hee University, Global Campus, 1 Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, South Korea
e-mail: sylee@oslab.khu.ac.kr

R. Ali
e-mail: rahmanali@oslab.khu.ac.kr

M. H. Siddiqi
e-mail: siddiqi@oslab.khu.ac.kr

Published online: 04 February 2015

1 Introduction

Due to the rapid growth of organizational data, the manual extraction of hidden patterns and knowledge in the form of rules is almost an impossible task. To overcome this problem, researchers have preferred to focus on automatic ways of extracting rules. Usually, for automatic rule extraction, data preprocessing may be performed, such as the transformation of continuous values to discrete intervals and completing missing values. The former is termed discretization (Maimon and Rokach 2010; Tsaia et al. 2008; Luengo et al. 2012). Discretization is an essential task that must be performed over real values in real world applications, such as medical (Paul and Maji 2011), industry (Wu et al. 2004), and power systems (Yan et al. 2004). The motivation for discretization is that real values may not be directly processed by all data mining techniques. Likewise, in most of the knowledge discovery systems, continuous values need to be transformed to discrete intervals before applying the knowledge extraction process. In data mining applications, data are usually represented in decision tables and when they are used for processing, they need to be expressed in discrete value form (Jiang and Zhang 2011).

Discretization can be viewed as a dimensionality reduction problem. In this problem, the range of continuous value attributes is expressed as smaller discrete intervals. This leads to generalization of patterns in large quantities of data, which ultimately results in a more generalized classifier for a particular domain. Similarly, discretization is also necessary for rough set feature selection; this is a way of reducing dimensionality in terms of features and improving the performance of a classifier (Tian et al. 2011). Other advantages of discretization are the classification of unseen instances (Kurgan and Cios 2004; Wu et al. 2006), simplification of data, faster learning, improved accuracy, compact and shortened results and reduced noise (Luengo et al. 2012). Discretization is a popular area of research and different techniques have been proposed so far; these include information entropy (Fayyad and Irani 1993; Dougherty et al. 1995), the statistical χ^2 test (Kerber 1992; Liu and Setiono 1997), likelihood (Wu 1996; Boulle' 2006) and class-tailored (Shehzad 2012) and rough sets (Zhang et al. 2005; Xu et al. 2012; Tian et al. 2011).

In the literature, researchers have reviewed existing discretization techniques from time to time with a focus on the different aspects of discretization. For example, (Dougherty et al. 1995) have reviewed and empirically evaluated the binning method for entropy-based and purity-based methods. They showed that the performance of the Naïve Bayes algorithm significantly improves when entropy-based discretization methods are used. Similarly, (Liu et al. 2002) have reviewed static/dynamic, global/local and univariate/multivariate discretization algorithms and provided a systematic study in terms of the history of development, effects on classification and trade-off between speed and accuracy. They have summarized the existing algorithms in a hierarchical framework. In the same way, a four level taxonomy of the existing discretizers has been created in previous studies (Bakar et al. 2009; Maimon and Rokach 2010; Blajdo et al. 2008). In this taxonomy, the authors have classified discretizers into hierarchical and non-hierarchical; splitting, merging and combination; supervised and unsupervised; and binning, statistic, entropy and some other related techniques. Recently, Luengo and his group (Luengo et al. 2012) reviewed the existing discretization algorithms and implemented them as a part of *KEEL* software (Alcalá-Fdez et al. 2009). They have empirically analyzed the algorithms over *KEEL* datasets and produced empirical results. These reviews are general in nature and not specific to any particular method, such as rough set theory (*RST*). *RST*, proposed for the first time by Pawlak (1982), is an efficient mathematical tool that can be best used for data analysis and knowledge discovery. A number of *RST-based* discretization techniques, such as those created by (Xu and

Yingwu 2009; Jiang and Zhang 2011; Jian-Hua 2004; Hong et al. 2005) can be found in literature. These discretizers perform even better in cases with inconsistent data. In the existing discretization literature, Luengo and his co-authors (Luengo et al. 2012) mentioned some rough set *(RS)-based* discretizers from a theoretical perspective in their review. Similarly, (Blajdo et al. 2008) have also discussed some *RS-based* discretizers and empirically analyzed their performance on *UCI* Machine Learning Repository datasets (Frank and Asuncion 2010). In recent years, many other rough set methods have been proposed for discretization and are still disregarded as reviews. Moreover, out of previous reviews, no review exists that is solely focused on *RST-based* discretization methods and hence no formal categorization has been made so far. This may provide a road map for the community for further research in this area. To come up with solutions to these problems, this study is motivated by the objectives described below, which can also be regarded as the main contributions of the paper.

- To propose a complete taxonomy of rough set-based discretization (*RSBD*) techniques and describe the key features of each method observed in it. The taxonomy will help to guide researchers towards the future trends of research in this field.
- To formalize the *RSBD* process in four steps and to briefly explain these steps with the help of a case study.
- In addition to the taxonomy and case study, the proposed study presents a detailed analysis of the existing discretization techniques while keeping the concept of *RSBD* applications in mind in order to depict a domain-wise distribution of the techniques.

The rest of the paper is structured such as follows. Section 2 presents the preliminaries of *RST* and its use in discretization. Section 3 presents a case study to fully explain the process of *RSBD*. Section 4 defines a criterion for the survey and then reviews each article in detail and summarizes the findings in a tabular format. After the detailed theoretical study, a taxonomy is proposed to the readers. At the end of this Section, an analysis of the existing techniques is made while keeping the applications of each method in mind in order to pictorially represent them. Section 5 discusses the outcomes of the review and recommends some guidelines to the readers. Section 6 describes the concluding remarks of this review.

2 Rough set theory and discretization

RST was proposed by (Pawlak 1982) and it is a powerful mathematical tool for analyzing inconsistent data. *RST* is widely used in applications such as machine learning, pattern recognition, data mining and decision support systems (Jian-Hua 2004). *RST-based* approaches to data analysis perform better with discrete data compared to continuous values. However, real world problems contain continuous values that cannot be processed directly by *RST*. To cope with such situations, the transformation of continuous values to discrete values is needed. This process is called discretization (Kurgan and Cios 2004; Luengo et al. 2012; Tsaia et al. 2008). In literature, the problem of discretization has been addressed using different techniques. A brief introduction to *RST* is presented here, and the notations used are based on the work of Pawlak (Pawlak 1982, 1992). Let us consider an information system (*IS*), represented as follows.

$$IS =< OBJ,\ ATTRIB > \tag{1}$$

where *OBJ* is a non-empty set of n objects known as examples or training instances (experience) and *ATTRIB* is a non-empty set of m conditional attributes. The set of objects, *OBJ,* is represented as follows.

$$OBJ = \{Obj_1,\ Obj_2, \ldots,\ Obj_{n-1}, Obj_n\} \tag{2}$$

Similarly, the set of conditional attributes, *ATTRIB,* is represented as follows.

$$ATTRIB = \{CAttrib_1,\ CAttrib_2,\ CAttrib_3,\ \ldots, CAttrib_{m-1},\ CAttrib_m\} \tag{3}$$

In Eq. (3), *CAttrib* represents a conditional attribute. For each *IS,* there is a corresponding decision system (*DS*); the only difference is that it has an additional attribute termed the decision attribute. Therefore, for the *IS* in Eq. (1), the corresponding *DS* is represented as follows.

$$DS =< OBJ,\ ATTRIB\ \cup\ \{Dec\} > \tag{4}$$

In Eq. (4), *DS* represents a training dataset with n training examples, each with m conditional attributes and one decision attribute. The decision attribute is represented by Dec which is always non-empty and has more than one value such as, $Dec = \{d_1, d_2, \ldots, d_n\}$. The values $d_1, d_2, \ldots, d_n$ represent classes. Each Obj_i belongs to a decision class (i.e., $Obj_i \in d$). The domain of the decision attribute is represented as follows.

$$Dom_{Dec} : OBJ \rightarrow Val_{Dec} \text{ where } Val_{Dec} = \{d_1,\ d_2, \ldots,\ d_n\} \tag{5}$$

Similarly, each conditional attribute ($CAttrib_i$) has a set of values associated with all of the objects (Obj_i) of the object set *OBJ*. This set of values represents the domain of the conditional attribute and is denoted as follows.

$$Dom_{CAttrib_i} : OBJ \rightarrow Val_{CAttrib_i} \tag{6}$$

If Dom_{CAttrib_i} is a set of continuous values R (i.e., $Dom_{\text{CAttrib}_i} \sqsubset R$), then the rough set needs a discretization system to transform the continuous values to a set of discrete intervals. This transformation must guarantee the discernibility and consistency of the original *DS*.

To convert continuous value attributes to discrete intervals, data is structured in the *DS*. After applying discretization, the *DS* is transformed into a new discretized decision system (*DDS*), which is represented as follows.

$$DDS =< OBJ, ATTRIB^d \cup \{Dec\} > \tag{7}$$

Here, $ATTRIB^d$ represents the discretized attributes obtained after discretization. To discretize continuous value attributes, the domain of each attribute Dom_{CAttrib_i} is split into discrete intervals using cut-points. According to Eq. (6), the domain of a conditional attribute is the set of values of all of the objects in the *DS*.

Let's assume l represents the lower value of a continuous value attribute (*CAttrib*) and u represents its upper value, then, in terms of intervals, the domain of the continuous value attribute ($CAttrib_i$) can be represented as follows.

$$Dom_{CAttrib_i} = [l_{CAttrib_i}, u_{CAttrib_i}] \tag{8}$$

For discretization, the domain $[l_{CAttrib_i}, u_{CAttrib_i}]$ is split into a set of i intervals (*INTR*) using the set of n cut-points represented by C, as shown in Fig. 1.

The set of all possible cut-points for a continuous value attribute $CAttrib$ is represented by C, as shown below.

$$C = \{c_1, c_2, \ldots, c_{n-1}, c_n\} \quad where \quad c_1 < c_2 < \cdots < c_n \tag{9}$$

The set of initial candidate cut-points (C) can be calculated by first sorting the values of *CAttrib* in ascending order, $v_1 < v_2 <, \ldots, < v_{n-1} < v_n$, and then taking the average of every two consecutive values as follows.

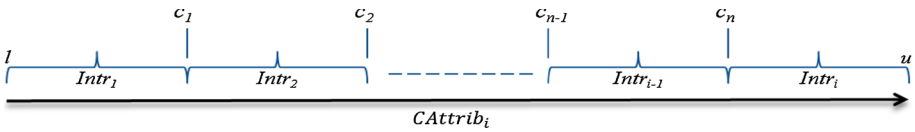

Fig. 1 The method by which real values of a continuous value attribute are divided into discrete intervals

$$C = \left\{ \left(\frac{v_1 + v_2}{2} \right), \left(\frac{v_2 + v_3}{2} \right), \ldots, \left(\frac{v_{n-2} + v_{n-1}}{2} \right), \left(\frac{v_{n-1} + v_n}{2} \right) \right\} \tag{10}$$

The cut-points of a continuous value attribute *CAttrib* are represented as ordered pairs, such as $(CAttrib, c_i\ )$. Therefore, all of the cut-points of an attribute *CAttrib* are represented as $\{(CAttrib, c_1), (CAttrib, c_2), (CAttrib, c_3), \ldots, (CAttrib, c_n)\}$. Using the values of all of the cut-points from Eq. (10), the set of all cut-points of an attribute *CAttrib* is shown by Eq. (11).

$$C_{CAttrib} = \left\{ \left(CAttrib, \frac{v_1 + v_2}{2} \right), \left(CAttrib, \frac{v_2 + v_3}{2} \right), \ldots, \left(CAttrib, \frac{v_{n-1} + v_n}{2} \right) \right\} \tag{11}$$

Similarly, the set of all possible cut-points on all attributes of a given *DS* is represented as follows.

$$C_{CAttrib} = \bigcup_{CAttrib_i \in ATTRIB} C_{CAttrib_i} \tag{12}$$

The set of cut-points *C in* Eq. (10) splits $Dom_{CAttrib_i}$ into a set of i intervals (partitions) represented by *INTR*. The total number of intervals is one more than the number of cut-points.

$$INTR = \{Intr_1, Intr_2, \ldots, Intr_{i-1}, Intr_i\} \tag{13}$$

Here, *Intr* represents a sub-interval and *INTR* represents the set of all sub-intervals of a continuous value attribute $CAttrib_i$. *INTR* is shown below.

$$INTR_{CAttrib_i} = \left[l^{CAttrib_i}, c_1^{CAttrib_i} \right) \cup \left[c_1^{CAttrib_i}, c_2^{CAttrib_i} \right) \cup \ldots \cup \left[c_n^{CAttrib_i}, u^{CAttrib_i} \right] \tag{14}$$

Similarly, the set of all possible intervals of all attributes of the *DS* is represented as follows.

$$INTR_{ATTRIB} = \bigcup_{CAttrib_i \in ATTRIB} INTR_{CAttrib_i} \tag{15}$$

One of the main objectives of discretization algorithms is to optimize Eq. (12) in such a way that the discernibility and consistency of the *DS* may not be disturbed. For this purpose, different techniques are used to optimize C_{ATTRIB} and $INTR_{ATTRIB}$ to their optimized sets. The optimum set of cut-points is then used to discretize the original *DS* to a new *DDS*, as represented in Eq. (7).

3 RSBD process: a case study

Discretization is a multi-step process to transform continuous values into discrete intervals. Some researchers have declared it a two-step process (Jiang and Zhang 2011), while others have declared it a three-step (Jiang and Zhang 2011; Luengo et al. 2012; Liu et al. 2008) or

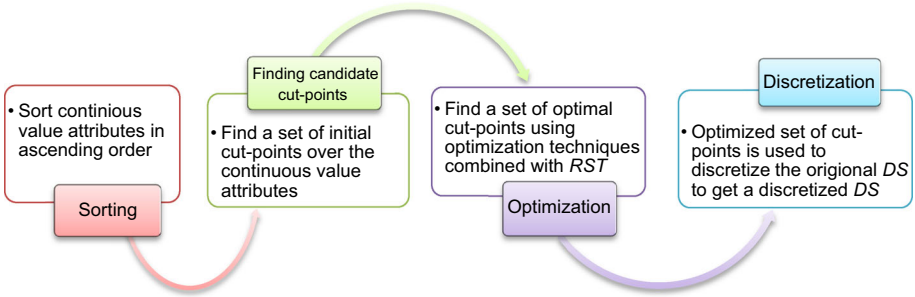

Fig. 2 Working steps of a rough set-based discretizer

Table 1 Decision table showing three continuous value attributes

OBJ	ATTRIB			Dec
	$CAttrib_1$	$CAttrib_2$	$CAttrib_3$	
Obj_1	1.8	3	0.5	1
Obj_2	2	2.5	2.1	0
Obj_3	2.3	4	2	0
Obj_4	2.4	2	2.5	1
Obj_5	2.4	3	2.1	0
Obj_6	2.6	4	2	1
Obj_7	2.3	2	1.5	1

four-step process (Liu et al. 2002; Kotsiantis and Kanellopoulos 2006). The two-step process proposed by Jiang and Zhang (Jiang and Zhang 2011) first finds the initial candidate cut-points and then validates them for consistency. In the three-step process (Luengo et al. 2012; Liu et al. 2008), the initial cut-points are first calculated, then a subset of optimal cut-points is chosen from the list of available cut-points and in the final step discretization is performed. In the four-step discretization process (Liu et al. 2002; Kotsiantis and Kanellopoulos 2006), the steps are as follows: sort the values of the continuous value attributes, set and evaluate the cut points, split/merge the intervals of the continuous values and stop at certain point. To simplify the understanding of these steps, we also formalize discretization into four steps. These steps are as follows: sorting, finding candidate cut-points, optimization and discretization, as shown in Fig. 2.

Example 1 Consider a *DS* with three continuous value attributes, $CAttrib_1$, $CAttrib_2$ and $CAttrib_3$, and seven objects, $Obj_1 \ldots Obj_7$, as shown in Table 1.

In the first step, the values of attributes $CAttrib_1$, $CAttrib_2$ and $CAttrib_3$ are sorted in ascending order, as shown in columns 1, 3 and 5 of the decision table in Table 2. In the second step, the sets of candidate cut-points, shown in columns 2, 4 and 6 of the decision table in Table 2, are computed for each attribute using Eq. (10). The total number of cut-points produced by $CAttrib_1$ is 4, by $CAttrib_2$ is 3 and by $CAttrib_3$ is 4. Consequently, the set of all cut-points C_{ATTRIB} has 11 cut-points; it is shown in last row of the decision table in Table 2 and is calculated using Eq. (12). The process is given below.

$$C_{ATTRIB} = \bigcup_{CAttrib_i \in ATTRIB} C_{CAttrib_i}$$
$$C_{ATTRIB} = C_{CAttrib_1} \cup C_{CAttrib_2} \cup C_{CAttrib_3}$$
$$C_{ATTRIB} = \{1.9, 2.15, 2.35, 2.5\} \cup \{2.25, 2.75, 3.5\} \cup \{1.0, 1.25, 2.05, 2.3\} \quad (16)$$

Optimization of the initial cut-points is one of the main tasks of any discretization algorithm. Therefore, after computing the initial set of cut-points, shown in Eq. (16), the next step is to turn them into an optimized set such that the discernibility/consistency of the *DS* remains intact. For this purpose, researchers have come up with a number of optimization techniques. In this survey, we review different methods of optimization that have been proposed in the literature for *RSBD*. For the optimization step of the *RSBD* process, we use the Rough Set Exploration System (RSES) version 2.2 (Bazan and Szczuka 2005) to turn the initial set of cut-points into an optimized set of cut-points, as shown in eqs. 16 and 17, respectively. The *RSES* uses a global discretization technique proposed by Polkowski et al. (Polkowski et al. 2000) for optimization. As a result, the possible eleven possible cut-points (shown in Eq. (16)) are optimized to only four cut-points, two for $CAttrib_1$ and one each for $CAttrib_2$ and $CAttrib_3$.

$$C = \{(CAttrib_1, 1.9), (CAttrib_1, 2.35), (CAttrib_2, 2.25), (CAttrib_3, 2.05)\} \quad (17)$$

The last step is to discretize the original decision table (i.e., Table 1, in this case) using the optimized set of cut-points shown in Eq. (17). The decision table in Table 3 is the final *DDS* in interval format while the decision table in Table 4 is the *DDS* in the discrete value format.

Table 4 is the output of the discretization process. This output can be best used for further data analysis purposes. The subsequent section, Sect. 4, is focused on the detailed survey of these techniques that combine *RST* and optimization approaches together for discretization. The most commonly used techniques so far for *RSBD* are statistical measures, entropy, fuzzy sets, Boolean reasoning (*BR*) and genetic algorithms.

4 RSBD: techniques, taxonomy and applications

This section presents a categorization of *RSBD* methods and sets a criterion for constructing a taxonomy. A criterion is first defined for the review and then each article is studied accordingly. The review focuses on the main features of each article and then categorizes them. At the end, each article is also reviewed for its application in different domains.

4.1 Key features for evaluation

It is very difficult to determine the effectiveness of a discretization technique; however, there are some measurement parameters that can grade a discretization technique as good or bad. The following subsections briefly describe these parameters.

4.1.1 Consistency

The analysis of a large volume of real world data sometimes suffers from the problem of inconsistencies, which can be characterized as having conflicting decisions for the same conditional attributes. When discretization takes place, it may produce some inconsistencies in the *DDS*. Therefore, while discretizing real values into discrete intervals, consistency must be guaranteed in the discretized *DS* (Jian-Hua 2004; Nguyen and Skowron 1995). The

Table 2 Real values of attributes in sorted format and calculation of the initial cut-points (constructed from Table 1)

$C_{ATTRIB} = \bigcup_{CAttrib_i \in ATTRIB} C_{CAttrib_i}$

$CAttrib_1$		$CAttrib_2$		$CAttrib_3$	
Sorting values	$C_{CAttrib_1} = \frac{v_i + v_{i+1}}{2}$	Sorting values	$C_{CAttrib_2} = \frac{v_i + v_{i+1}}{2}$	Sorting values	$C_{CAttrib_3} = \frac{v_i + v_{i+1}}{2}$
$v_1 = 1.8$	$(1.8 + 2)/2 = 1.9$	$v_1 = 2.0$	$(2 + 2.5+)/2 = 2.25$	$v_1 = 0.5$	$(0.5 + 1.5)/2 = 1.0$
$v_2 = 2.0$	$(2 + 2.3)/2 = 2.15$	$v_2 = 2.5$	$(2.5 + 3)/2 = 2.75$	$v_2 = 1.5$	$(1.5 + 2)/2 = 1.25$
$v_3 = 2.3$	$(2.3 + 2.4)/2 = 2.35$	$v_3 = 3.0$	$(3 + 4)/2 = 3.5$	$v_3 = 2.0$	$(2 + 2.1)/2 = 2.05$
$v_4 = 2.4$	$(2.4 + 2.6)/2 = 2.5$	$v_4 = 4.0$	–	$v_4 = 2.1$	$(2.1 + 2.5)/2 = 2.3$
$v_5 = 2.6$	–	–	–	$v_5 = 2.5$	–
# of cut-points = $C_{CAttrib_1} = \mathbf{4}$		# of cut-points = $C_{CAttrib_1} = \mathbf{3}$		# of cut-points = $C_{CAttrib_1} = \mathbf{4}$	
Total number of candidate cut-points = $C_{ATTRIB} = \mathbf{11}$					

Table 3 Discretized values in interval format

OBJ	ATTRIB			Dec
	$CAttrib_1$	$CAttrib_2$	$CAttrib_3$	
Obj_1	(−Inf,1.9)	(2.25,Inf)	(−Inf,2.05)	1
Obj_2	(1.9,2.35)	(2.25,Inf)	(2.05,Inf)	0
Obj_3	(1.9,2.35)	(2.25,Inf)	(−Inf,2.05)	0
Obj_4	(2.35,Inf)	(−Inf,2.25)	(2.05,Inf)	1
Obj_5	(2.35,Inf)	(2.25,Inf)	(2.05,Inf)	0
Obj_6	(2.35,Inf)	(2.25,Inf)	(−Inf,2.05)	1
Obj_7	(1.9,2.35)	(−Inf,2.25)	(−Inf,2.05)	1

Table 4 Discretized values in discrete format

OBJ	*ATTRIB*			Dec
	$CAttrib_1$	$CAttrib_2$	$CAttrib_3$	
Obj_1	0	1	0	1
Obj_2	1	1	1	0
Obj_3	1	1	0	0
Obj_4	2	0	1	1
Obj_5	2	1	1	0
Obj_6	2	1	0	1
Obj_7	1	0	0	1

number of inconsistencies caused by a discretization technique must be less than the number of inconsistencies caused by the original *DS* (Kotsiantis and Kanellopoulos 2006). Thus, inconsistency is an important evaluating parameter for discretization.

4.1.2 Minimal cut-points

Discretization can be viewed as a data reduction technique which reduces the range of values of a continuous values attribute into a minimum number of discrete intervals (Kurgan and Cios 2004). The number of cut-points can determine the level of data reduction. The fewer the number of cut-points the more the data will be reduced and hence a generalized classifier will be possible. There is, however, a tradeoff between the number of cut-points and the consistency and understandability of a *DS*. The greater the number of cut-points, the smaller the intervals will be; therefore the understandability will be lower, but the consistency will be higher. Conversely, the fewer the number of cut-points, the larger the intervals will be and the better the understanding will be, but the consistency will be lower (Kotsiantis and Kanellopoulos 2006). Accordingly, minimal cut-points are favored and therefore considered important in evaluating parameters for discretization.

4.1.3 Classification rate

Any classification method needs discretization as one of its preprocessing step (Lustgarten et al. 2008; Zhu et al. 2011). Therefore, a discretization algorithm is considered efficient if it

improves the classification accuracy. Thus, the classification rate is another important factor for evaluating discretization.

4.1.4 Discretization techniques and complexities

RST is combined with statistical techniques (Singh and Minz 2007), information entropy (Tian et al. 2011), fuzzy sets (Paul and Maji 2011), *BR* (Nguyen 1998) and genetic algorithms (Jian-Hua 2004), to compute an optimal set of cut-points. An optimal number of cut-points results in a consistent *DDS* that can be better for solving rule induction and classification problems. This survey primarily focuses on the techniques adopted in each article and their key features. In addition to the above criteria, time and space complexities are also considered important evaluation criterion.

4.2 RSBD techniques and their features

To write this survey, we performed an extensive review of the existing *RSBD* techniques proposed in the recent literature. For retrieval purposes, the keywords/phrases used were 'rough set discretization', 'quantization and rough set', 'real value attributes and rough set', 'continuous values and rough set' and 'rough set quantization'. The time frame was set between 2000 and 2013, but due to the importance of Nguyen's initial work on *RSBD* (Nguyen and Skowron 1995), we have included their articles published in 1995 and 1998.

This section is focused on the enumeration and design of a detailed study of *RSBD* techniques from a theoretical perspective. We review each article based on the evaluation features discussed in Sect. 4.1. Based on this review, the existing techniques are categorized into statistical, entropy-based, genetic algorithms, fuzzy sets and *BR-based* approaches. Findings from the review are summarized in the tables. In the tables, some cells are marked as '*nil*' which means that the required information for these features was not evident in the article. Likewise, we have used abbreviations as short names for the titles of the articles to make the representation in the tables and taxonomical chart simple and concise.

4.2.1 Statistical

In the literature, statistical techniques, such as distributional index, clustering, correlation coefficient and interval similarity, are the most commonly used methods that have been combined with *RST* to optimize the initial set of cut-points for discretization. Probability distribution (Zhang et al. 2005) is used to find the class-separability function for the discretization in the radar emitter signal processing domain. Similarly, the information entropy and statistical distribution index are combined together in a hybrid approach that finds clusters in continuous data (Wu et al. 2006). In this approach, the minimal distributional index determines the border value for splitting an interval and the maximum of the index decrement is applied in order to select the new intervals to split further. In clustering, the centroid linkage method of agglomerative hierarchical clustering (Wu et al. 2004) is used to find the center of the closest cluster. Dendrograms and statistical plots are used for the hierarchical representation of the dataset and the method is applied for the inspection of defects in wood veneer.

Dynamic clustering techniques (Xu et al. 2012; Jiang and Zhang 2011) are combined with *RST* to cluster candidate cut-points dynamically with the help of their entropy value; the optimized clustered cut points are then obtained using the importance algorithm of cut-points. Density-based clustering (Singh and Minz 2007) has also been combined with *RST*,

where *RST* offers itself as a tool for measuring the degree of significance of the attributes and the dependencies among them. After this, the approach is evaluated by using the total number of intervals produced and the class-attribute interdependence redundancy (*CAIR*) as the parameters. In the same way, correlational approaches, such as correlation coefficients, rough entropy and stability (Xu and Yingwu 2009), and the grey correlation degree analysis algorithm (*RSRGCDMD*) (Tinghui et al. 2009) have also been used for *RSBD* and applied for the diagnosis of aultsin steam turbines.

Similarly, a static, incremental, supervised and bottom-up quantization method has also been used in literature (Jing et al. 2013) which applies chi-square statistics to discover accurate merging intervals. This algorithm is heuristic in nature and results in best quantization and classification accuracy. Zou and his co-authors (Zou et al. 2013) have used an interval similarity-based algorithm (*SIM*) for discretizing real values. The technique used is to first calculate the similar values of the intervals using a similarity function and then check for those cut-points which have maximum similar values for merging them together. If multiple similar values exist then adjacent intervals are merged together. A similar approach can also be found in literature (Jia et al. 2013; Radwan and Assiri 2013). Jia (Jia et al. 2013) have proposed a heuristic-based quantization method which attains satisfactory results that significantly improves the performance of inductive learning. On the other hand, (Radwan and Assiri 2013) have used *RS-based* modified similarity relation (*RS+MSIR*) method for the discretization of continuous value attributes of thyroid disease patients so that to automatically induce rules from the data. These methods are enumerated and reviewed in Table 5.

4.2.2 Information entropy

Information entropy is a heuristic-based technique that defines the significance of the candidate cut-points of a continuous value attribute to obtain an optimized set of cut-points (Xu and Yingwu 2009). It is used in combination with *RST* to improve the discretization results. Minimum, granular and conditional entropy have been used for *RSBD*, which obtain an optimized set of cut-points using an approximation of the information. Discretization algorithms, such as *C-GAME* (Tian et al. 2007, 2011), use an approximation of the minimum entropy and constraint satisfaction to select a set of cut-points capable of generating discrete data with non-empty cores. Minimum information entropy has also been used for discretizing physicochemical parameters of blood stasis syndrome in Traditional Chinese Medicine (*TCM*) (Zheng and Xi 2009). In the same way, granular minimum entropy and rough set equivalence relations are combined and applied for *RSBD* that produces improved results (Zhou 2009). Similarly, the articles (Liu et al. 2008; Hong et al. 2005) also use entropy for the optimization of cut-points and *RSBD*. Information gain that is based on conditional entropy has also been combined with the discernibility relation of *RST* for the discretization of power system data (Yan et al. 2004). Recently, Grzymala-Busse (2013) has proposed a new entropy-based discretization technique which enhances the original entropy-based approach by introducing two new options of dominant attribute selection and multiple scanning. Using the first option of dominant attribute selection, an attribute with the smallest conditional entropy is selected for discretization and then the best cut point is determined. In the second option, all the attributes are scanned multiple times, and at the end, the best cut points are selected for all the attributes. Table 6 lists and summarizes features of previous studies that use information entropy and *RST* for discretization.

Table 5 Statistical *RSBD* techniques and their features

Discritizer Abbreviated name	Technique	Features			Application domain of the approach
		Comparison	# Intervals/ cut-points, Consistency/accuracy	Classification rate and time/comp. complexities	
DA-RSC(Wu et al. 2004)	Clustering: centroid linkage method of agglomerative hierarchical clustering	Fuzzy and k-mean clustering	Minimal rules are extracted	Classification & rules extraction accuracy is very high Time complexity < other clustering methods	Wood veneer:Wood veneer dataset with 246 sample
DRSTA(Zhang et al. 2005)	Probability distribution.	Nil	Nil	Classifier accuracy is improved Higher recognition rate	Radar emitter signals: 10 Radar emitter signal with 8 attributes
NDKD-RS (Wu et al. 2006)	Distributional index: entropy and statistical distributional index	Multi-knowledge approach without the discretizer and unsupervised 5-identical-interval discretizer	Nil	Average decision accuracy improved	General: UCI ML datasets Sonar, Horse-colic,Innosphere, Wine,Crx-data, Heart, Bupa, Ecoli, SPECTF
DRST(Singh and Minz 2007)	Clustering: density based clustering	EW, EF, Patterson and Niblett, IEM, Maximum Entropy, CADD, CAIM	Get minimum number of intervals Increase class-attribute interdependency	Nil	General: UCI ML datasets Iris, ion, hea and pid
HGD-RS(Xu and Yingwu 2009)	Correlation: correlation coefficients, rough entropy and stability	Equal Width (EW), Entropy, DIBD, CAIM	Minimizes number intervals Increase consistency	Achieves optimal classification accuracy	General: UCI ML datasets Iris, Heart, Pima, Ionosphere
RSRGCDM (Tinghui et al. 2009)	Grey correlation degree analysis	Nil	Nil	Nil	Steam turbine: Fault diagnosis dataset

Table 5 continued

Discritizer Abbreviated name	Technique	Features			Application domain of the approach
		Comparison	# Intervals/ cut-points, Consistency/accuracy	Classification rate and time/comp. complexities	
TSID (Jing et al. 2011)	Dynamic clustering and entropy	IE, greedy, breakpoint importance	Nil	Classification/recognition rate is high Time complexity is significantly reduced	General: UCI ML datasets Sona,Pima,Austra, Ecoli, Glass, Iris, Wine, Machine, Breast_w, Wareform,Segment
TSD-RS (Xu et al. 2012)	Dynamic clustering	Nugyen greedy discretization, Breakpoint importance and attribute importance algorithm	Nil	Time complexity is less than the global discretization techniques	General: UCI ML datasets Sona, pima, austral, ecoli, glass, iris, wine, letter
SSDQ-ML (Jing et al. 2013)	A supervised chi-square statistics- method is used	Built-in, Mod-Chi2, Ext-Chi2, CAIM Boolean, MDLP, EF	Nil	Highest classification accuracy than the compared methods and hence high quality of quantization	General: UCI ML Datasets CPS, Iris, Auto, Breast, Inosphere, Pima, Glass, Wine, Machine, Heart, Sonar
SIM (Zou et al. 2013)	Interval similarity-based method is used	EXT and Boolean	Predictive accuracy has increased as compared to EXT and Boolean approach	Less number of classification rules and nodes are produced on C4.5	General: UCI ML Datasets Iris, Auto, Breast, Inosphere, Pima, Glass, Wine, Machine, Bupa
SIAR (Jia et al. 2013)	Interval similarity criterion function is used	C4.5, Mod-chi2 and Ext-Chi2, CAIM, Boolean, Ent-MDLP, EQF	Produces high quality of quantization results than the compared approaches	Highest classification accuracy	General: UCI ML Datasets CSP, Iris, Auto, Breast, Inosphere, Pima, Glass, Wine, Machine, Heart, Sonar

Table 5 continued

Discritizer Abbreviated name	Technique	Features			Application domain of the approach
		Comparison	# Intervals/ cut-points, Consistency/accuracy	Classification rate and time/comp. complexities	
RS-MSIM(Radwan and Assiri 2013)	Modified Similarity Relation is used	C4.5	Nil	Minimal sets of reducts Minimal sets of decision rules Better classification accuracy (371/414)	Medical: Thyroid dataset with 414 patient's data
RS-MFD-FPP(Wang 2013)	Here, the discretization is based on maximum covariance between the classes	Nil	Nil	Effective identification of different operating states of the pump	Petroleum Industry: Dataset of five-plunger pump is used for testing

Table 6 Entropy and *RSBD* techniques with their features

Discritizer Abbreviated name	Technique	Features			Application domain of the approach
		Comparison	# Intervals/cut-points Consistency/accuracy	Classification rate and time/comp. complexities	
C-GAME:RSFS-Exp (Tian et al. 2007)	Approximate minimum entropy	RMEP, DT	Reducts produced using C-GAME are smaller than RMEP	C-GAME based rough set features selection algorithms run faster than RMEP algorithms	General: UCI ML datasets SPECTF, Ionosphere
AD-RSIE (Liu et al. 2008)	Information entropy	Greedy algorithm, attribute importance algorithm	Smaller Reduct size Minimum cut points than the compared algorithms	Time complexityis less than the compared algorithms	General: UCI ML datasets Ecoli, glass, pima, Iris
AD-RSIE-TCM (Zheng and Xi 2009)	Minimum information entropy	Nil	Nil	Nil	Medical: Traditional Chinese Medicine data
AD-GE (Zhou 2009)	Granular entropy (minimum) and equivalence relation of RST	Nil	Nil	Nil	Others: testing example
C-GAME: RSFS (Tian et al. 2011)	Approximation of minimum entropy and constraint satisfaction problem	RMEP, ChiMerge, BR and EF	Produces smaller core for features selection as compared to the compared discretizers	Classification results are better than the compared approaches and classifiers C4.5 and MLPs.	General: UCI ML datasetsSonar, SPECTF, water, ionospshere, wdbc, wpbc, Australian, cleveland, hungarian, housing
DERS (Yan et al. 2004)	Normalized information gain (calculated usingentropy) and indiscernibility relation of RST	EF	Better performance than EF based discretization	Improved classification results as compared to results produced with the help of EF discretizer	Power System (generators): 400 instances of power system data

Table 6 continued

Discritizer Abbreviated name	Technique	Features			Application domain of the approach
		Comparison	# Intervals/cut-points Consistency/accuracy	Classification rate and time/comp. complexities	
DRST-IE (Hong et al. 2005)	Information entropy	Nil	Nil	Computational complexity decreases for large number of candidate cut-points.	General: UCI ML datasets Iris,glass, heart,, ecoli, austral, pima
DBE-MS (Grzymala-Busse 2013)	Information entropy	EW and EF	Error rate for both Dominant Attribute and Multiple Scanning is smaller than EW and EF	Both are faster than EW and EF	General: UCI ML datasets 17 publically available datasets

4.2.3 Genetic algorithm (GA)

A genetic algorithm (*GA*) is an effective way of searching and optimizing that approaches a global optimum solution for non-linear and high dimensional functions (Rangel-Merino et al. 2005). The combined effect of a *GA* and *RST* may therefore produce improved results for optimizing candidate cut-points. Jian-Hua (2004) has used *GA* and *RST* for discretization, which performs better than Nguyen's (Nguyen and Skowron 1995) and the ChiMerge (Kerber 1992) algorithms in producing an optimal set of cut-points. In this study, chromosomes are represented by strings with a length equal to the number of candidate cut-points; a fitness function is used for representing the consistency. Similarly, Chen and his co-authors (Chen et al. 2003) have proposed a *GA*-based approach for discretization. They have used optimization strategies, such as elitist selection and father-offspring combined selection, restart and penalty, to optimize the candidate cut-points. Chebrolu and Sanjeevi (2012) used the discernibility relation of *RST* and proposed a 2-step discretization algorithm in which the first step obtains the candidate cut-points with the help of an *MD*-heuristic algorithm and the second step optimizes the number of cut-points with the help of a *GA*. Table 7 shows the results of the review of those discretization algorithms that use *GA* and *RST* together.

4.2.4 Fuzzy set

Fuzzy discretization is characterized by the membership value, group or interval number and affinity corresponding to an attribute value, unlike crisp discretization, which only considers the interval number (Roy and Pal 2003). In problem domains where input data for a rough set classifier overlaps, crisp discretizers may not produce satisfactory results. Therefore, fuzzy rough set discretization is favored over crisp discretization. Fuzzy sets assign discrete intervals to continuous values using a fuzzy membership function. A fuzzy membership function and *RST* may produce improved results for overlapped data. For the direct computation of the relevance and significance of continuous valued genes, Paul and Maji (2011) have proposed fuzzy discretization that improves the classification results over a crisp set. Similarly, Wang et al. (2012) proposed a two-step fuzzy-rough set discretization technique in which the first step calculates the K most important cuts using an "*MD* heuristic" algorithm and the second step obtains the fuzzy sets for discretization. The trapezoidal membership function of the fuzzy set has been used for fuzzy rough set discretization by Roy and Pal (2003). Table 8 describes the main features of the fuzzy rough set discretization literature reviewed in this study.

4.2.5 Boolean reasoning (BR)

Boolean reasoning can be best used for computing prime implicants, therefore its combination with *RST* may produce improved discretization results. In the literature (Nguyen 2005, 1998; Dai and Li 2002), heuristic-based approaches have been used for discretization and rules generation for classification problems. In this literature, *BR* is proposed to determine the discernibility formula, discernibility Boolean function and prime implicants in order to find an optimal set of cut-points for the best discretization results. In the case of a large quantity of data, their approach may not perform better (Jiang and Zhang 2011), therefore the phenomenon of bound cuts has been proposed by Dai and Li (2002), which produces optimized results for a large quantity of data. Table 9 summarizes the results of the review we have done for the articles that use *BR* and *RST* for discretization.

Table 7 Genetic algorithms and *RSBD* techniques with their features

Discritizer Abbreviated name	Techniques	Features			Application domain of the approach
		Comparison	#Intervals/cut-points and consistency/accuracy	Classification rate and time/comp. complexity	
GA-RS (Chen et al. 2003)	Genetic algorithm optimization strategies: elitist selection and father-offspring combined' selection strategy, restart and penalty strategy	Quantization of real value attributes.	Minimum # of cuts as compared to the compared approach Higher accuracy than the traditional discretizations	Faster than traditional approaches	General: UCI ML datasets Glass, iris, pima and wine
GA-DDS (Jian-Hua 2004)	Genetic algorithm and Rough Sets Theory	ChiMerge & Nguyen's method	Minimal interval (almost half of the compared approaches) High consistency	Nil	General: UCI ML datasets Iris
DR-RST-GA (Chebrolu and Sanjeevi 2012)	Discernibility relation of ***RST***, GA and MD-Heuristics	Compared with (Chen et al. 2003)	Nil	Time complexity is less than the compared approach	General: UCI ML datasets Iris, glass, wine, liver-disorders, ecoli

Table 8 Fuzzy set and *RSBD* techniques with their features

Discritizer Abbreviated name	Technique	Features			Application domain of the approach
		Comparison	#Intervals/cut-points and consistency/ accuracy	Classification rate and time/comp. complexity	
FDFS (Roy and Pal 2003)	Trapezoidal membership function of fuzzy set, RS classifier and pattern recognition tools, MLP	Crisp discretization	Nil	Best for classification	Medical: Speech and hepatobiliary Vowel and Hepatobiliary disorder data
FD-RSGS (Paul and Maji 2011)	Fuzzy set	Mean-standard deviation and EF	Nil	Classification results for genes selection is better than crisp classification approaches	Medical: Cancer Breast, Leukemia, Colon, Lung and Prostate, (RAOA and RAHC) microarray data sets
FRSF-AR (Wang et al. 2012)	MD heuristic algorithm and fuzzy set.	Fuzzy c-means (FCM) based fuzzification method	Nil	Higher classification accuracy Improved features reduction results	General: UCI ML datasets Crx, Ecoli, Glass, Iono, Labor, Sonar, WDBC, WPBC, Wine

Table 9 Boolean reasoning and *RSBD* techniques with their features

Discritizer Abbreviated name	Technique	Features			Application domain of the approach
		Comparison	# Intervals/cut-points and consistency/accuracy	Classification rate and time/comp. complexity	
RS-BR (Nguyen and Skowron 1995)	BR for finding discernibility formula and prime implicants.	N.B, C4.5, EW, Holt's 1R, recursive minimal entropy partitioning	Smaller number of cut-points than the compared algorithms	Increases quality of classification	General: UCI ML datasets Airline, vehicle, iris, heart, glass, Australian, inverted pendulum, calling pattern 1
				Cannot be applied to a large quantity data (Jiang and Zhang 2011)	
DPRS (Nguyen 1998)	BR and discernibility Boolean function	Nil	Nil	Nil	Others: test example
SD-RST (Dai and Li 2002)	Bound cuts using Boolean reasoning	Compared with the algorithm of (Nguyen and Skowron 1995)	Nil	Space complexity and time complexity decline obviously	General: UCI ML datasets Iris and BUPA

4.2.6 Complementary approaches

Along with the above techniques, other approaches have also been attempted for discretization. Jiang et al. (2010) have proposed a rough set-based supervised and multivariate approach. This approach takes multiple attributes at the same time as the class label and computes the discernibility function to find an optimal set of cut-points. The results produced by their method are better than the results of the *EF*, naive, semi-naive, *BR* and *EM* algorithms. Similarly, for the discretization of interval-valued attributes, Xin et al. (2007) proposed a supervised *RS*-based approach. Likewise, Jiang and Zhang (2011) proposed a three-step discretization algorithm using *RST*. The first step finds the candidate cut-points, the second step computes the importance of each cut-point using a top-down approach and the last step uses a heuristic-based approach for selecting the best cut-points for discretization. Furthermore, in the area of fighters, field programmable gate array (*FPGA*) devices are used for faster computation and an algorithm has been proposed to diagnose faults in fighters (Sun et al. 2013). They have designed a system that consists of eight-modules. *FPGA* devices are combined with the attributes dependency degree in *RST* to improve the processing speed of discretization. Similarly, for the mechanical fault diagnosis of five-plunger pump, in the petroleum industry, a maximum covariance-based discretization method has been proposed in literature (Wang 2013). Table 10 summarizes the review results of these techniques.

4.3 Taxonomy

Tables 5, 6, 7, 8, 9 and 10 present a summary of the review of rough set-based discretizers. In this study, the main focus of the review is on the techniques adopted by each discretizer for discretization. Column 2 of the aforementioned tables contains detailed information about the methods used for discretization in the existing literature. A classification of the *RSBD* techniques has been done on two levels and as a result, a taxonomy has been drawn (Fig. 3). The taxonomy has two levels. The first level represents the main techniques. The main techniques are categorized into five groups: statistical, entropy, *GA*, fuzzy and *BR* methods. The second level of the taxonomy represents the sub-techniques within the main techniques. Each technique is mapped with its corresponding example(s) of a discretizer in its abbreviated form.

The motivation for selecting this order is to clearly represent the *RSBD* techniques for better understanding. The proposed taxonomy helps the researchers to classify new rough set-based discretizers into one of these categories based on its features. This taxonomy depicts the existing *RSBD* techniques and clarifies the similarities and dissimilarities among them.

In summary, this categorization delivers a broad picture of the state-of-the-art *RSBD* for early-stage researchers of the topic or for those who want to discretize data for an application.

4.4 Applications

Apart from a detailed analysis of the existing literature in *RSBD* and taxonomy, the survey has also explored the distribution of these techniques on the basis of their application fields. An outlook of these fields is depicted in Fig. 4 which shows the scope of the topic. The *RSBD* techniques are used in the fields of radar systems for discretizing the radar signals, power generator systems, steam turbines for diagnosing different faults in turbines, timber industry for inspecting defects in wood plates, petroleum engineering for analyzing faults in the pumps, aircraft manufacturing to analyze fighters' data, and medical domain for different prognosis tasks.

Table 10 Other approaches to *RSBD* along with their features

Discritizer Abbreviated name	Technique	Features			Application domain of the approach
		Comparison	# Intervals/cut-points and consistency/accuracy	Classification rate and time/comp. complexity	
DRSTA (Xin et al. 2007)	Supervised approach of RST for discretization of interval-valued attributes.	Nil	Nil	Enhances the precision of classification and accurate recognition rate in pattern recognition.	Radar system: 17 radar emitter signals dataset having 3 features
SMD (Jiang et al. 2010)	A supervised and multivariate approach for multiple attributes at the same time with the class label and computes discernibility function	EF, Naive, Semi-naive, BR, EM	Minimum interval than Naïve algorithm	Computational complexity is low	General: UCI ML datasets Iris, glass, ecoli, heart, Pima
NMD-RST (Jiang and Zhang 2011)	Three steps discretization algorithm using RST.	Nil	Nil	Nil	General: UCI ML datasets Iris, glass, coli, austral
FPGA-based- Dis(Sun et al. 2013)	FPGA-based algorithm is used and a system consisting of eight modules is designed	Nil	Nil	Save much time of pretreatment in rough set and improve operation efficiency	Aircraft:Dataset used is the output of a nonlinear aircraft model

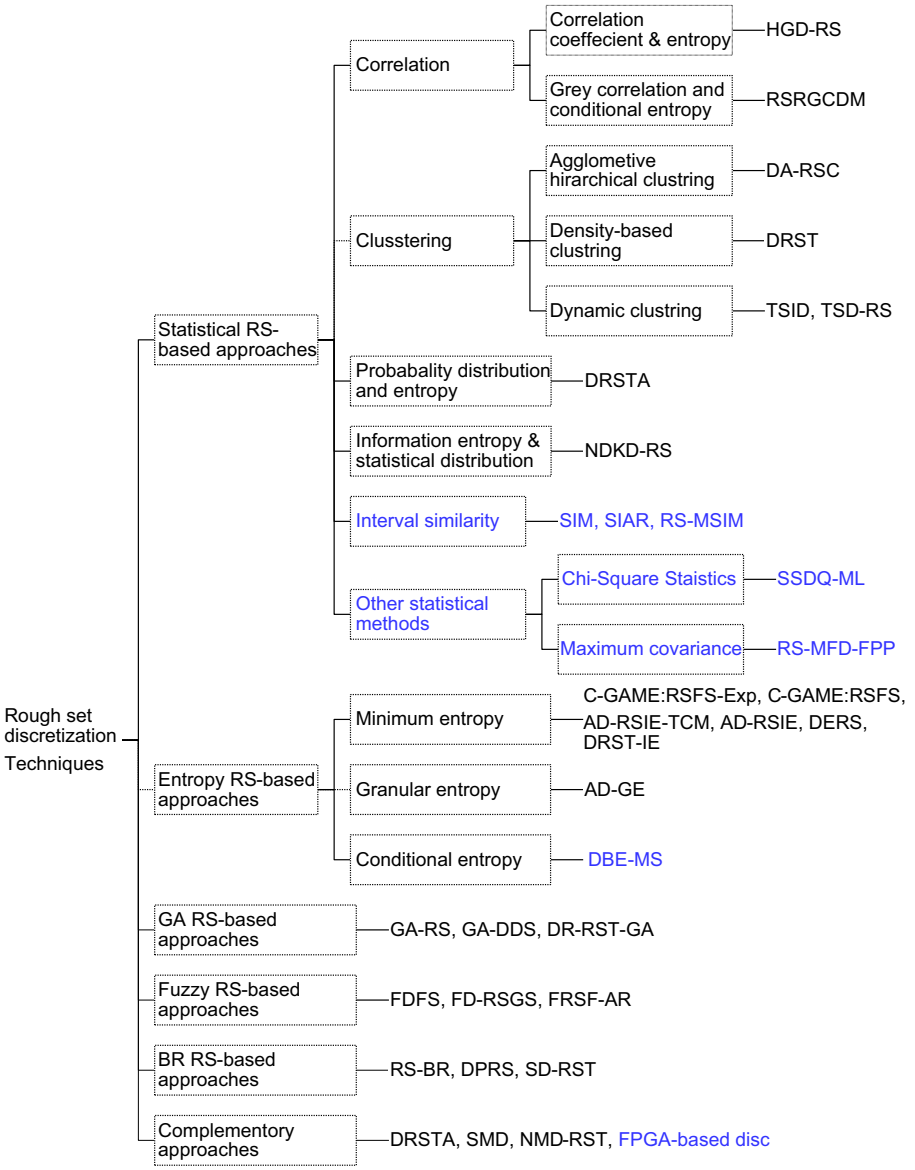

Fig. 3 Taxonomical chart of RSBD techniques

Most of these algorithms are general in nature and are therefore not applied to any specific field. This general category of algorithms is evaluated by the benchmark datasets from the *UCI ML* Repository (Frank and Asuncion 2010). In Fig. 5, this group of algorithms is labeled as 'General' and has twenty articles in this survey. In the same figure, the domain labeled 'Other' represents the set of algorithms that have neither been applied in a specific domain nor evaluated by the *UCI ML* benchmark datasets; they have been evaluated by simple example considered by the authors. In this survey, two articles have been reviewed under this label.

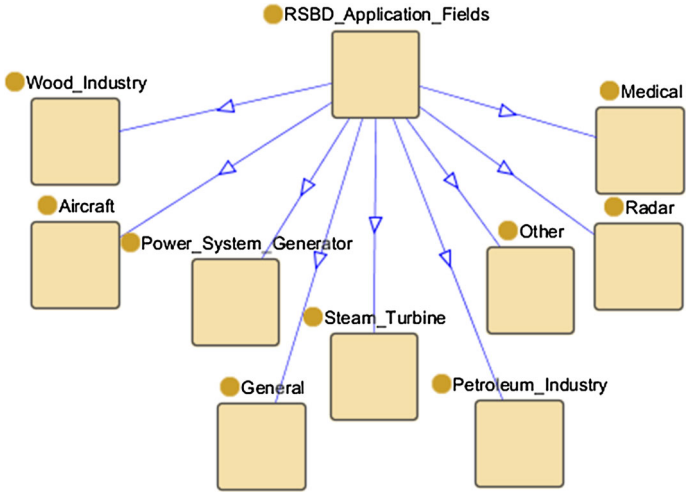

Fig. 4 Outlook of the RSBD application fields

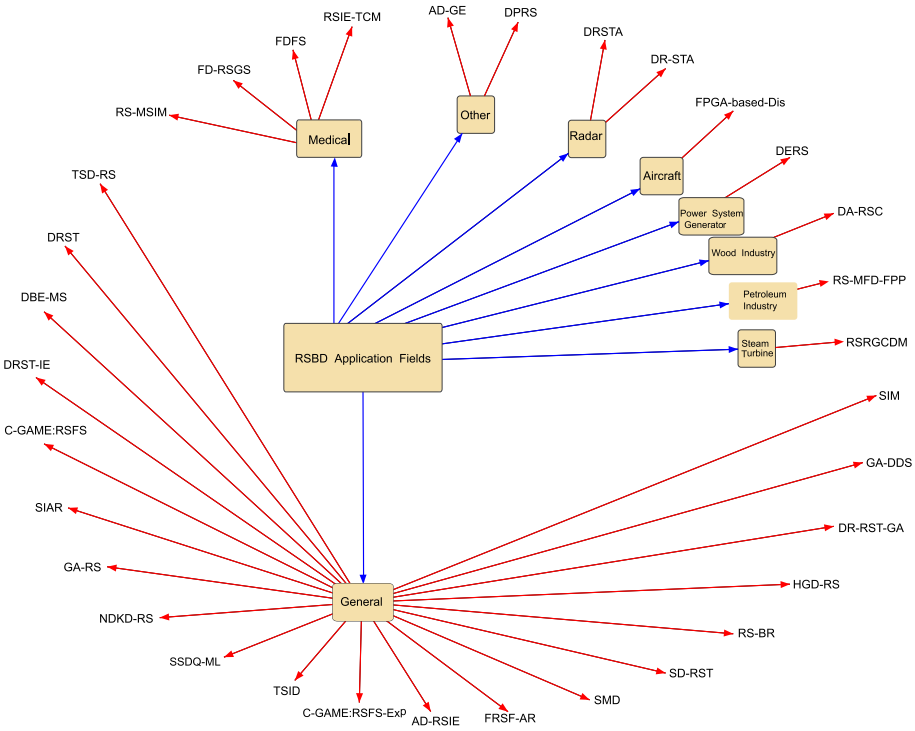

Fig. 5 Domains-wise distribution of the RSBD techniques

In the medical domain, four rough set-based discretizers were found in which two use fuzzy sets concept. In radar systems, two articles were found, while for the rest of domains, such as wood industry, steam turbines, aircraft, petroleum, and power system generator only a single article was found.

From the technical application perspective, discretization has effectively been used for generalization and data reduction (Hu and Cercone 1999), increasing the learning efficiency, classification accuracy and clarity of different data analysis processes (Shehzad 2012), rough set-based features selection (Tian et al. 2011) and induction of decision rules (Shang et al. 2005). In regard to features selection, the core of discrete datasets can be best calculated by applying *RSBD* to select the prominent features out of a high dimensional feature space in different datasets (Tian et al. 2011). Similarly, the hybrid of *RSBD* with evolutionary computation method can generate a smaller number of high quality decision rules to classify the new instances in different domains (Shang et al. 2005).

5 Discussion

Based on the survey and taxonomy developed above, several guidelines can be recommended to early-stage researchers who are interested in work within this field. This survey and taxonomy also help experts in the area of *RSBD* to select an appropriate method for their application. Researchers who are interested in applying a particular *RSBD* method (depicted in the taxonomy above) must be aware of the properties that define the methods in order to choose the most appropriate one for their specific application/s. In this paper, rough set-based discretizers have been categorized into statistical, entropy, *GA*, fuzzy and *BR*-based rough set discretizers, but one cannot provide a single concluding statement as to which approach, out of these, is the optimum method. This depends on the nature of the problem and the size of the dataset under consideration. Discretization is a *NP-hard* problem (Chlebus and Nguyen 1998; Nguyen 1998), which requires an optimization technique to optimize the initial set of candidate cut-points in such a way that the *DS* remains consistent. The following remarks help researchers to limit the set of candidate approaches and to select an appropriate method for their application according to their requirements.

- *Statistical:* For *RSBD*, the most frequently used statistical techniques are clustering (Wu et al. 2004), correlation (Xu and Yingwu 2009) and distributional index (Wu et al. 2006). These can be easily combined with *RST* to obtain an optimized set of cut-points in discretization. Due to the unsupervised nature of clustering methods, they may produce good discretization results for data with no classification label.
- *Information Entropy (IE):* is the measure of the randomness in the distribution of the data (Batu et al. 2005). This is a heuristic-based technique that approximates the values of random variables to define the significance of the candidate cut-points in order to select an optimal set for rough set discretization. Entropy is the local treatment of attributes to find the significance of the cut-point values; therefore, it may compromise consistency (which is lower in this case) over the computational complexity (which is better) (Xu et al. 2012).
- *Fuzzy sets:* are based on membership functions to accurately map the overlapped data boundaries. In cases with any high-dimensional data from a physical process, where data is overlapped and has sensitive boundaries, crisp discretization may not result in an optimal set of cut-points. Therefore, fuzzy *RSBD* is a good choice to put an optimal set of cut-points and obtain a consistent discretized *DS* (Roy and Pal 2003; Wang et al. 2012).
- *Genetic algorithm:* is an effective searching and optimization technique that achieves a global optimum solution for non-linear and high dimensional functions (Rangel-Merino et al. 2005). *GAs* can obtain consistent and minimal discretization results and are best applied for problems with high dimensional feature space with continuous values. For a

huge volume of data records, the only compromise a user has to make is the time and computational complexity (Chen et al. 2003).

- *Boolean Reasoning*: is a greedy search method for solving problems that uses heuristics that may get stuck at local optima. In general, due to the high dimensional features space, it has high time and space complexity (Nguyen 2005), which may not result in an optimal set of cut-points for discretization in linear time. There is a tradeoff between the consistency of the *DS* and the complexities (i.e., time and space) in comparison with a *GA*. However, in low dimensional feature space, heuristic-based *BR* may determine an optimal set of cut-points for rough set discretization in linear time. As a result, for an improved quality of *RSBD* and a suboptimal set of cut-points in cases with high dimensional feature space, *GA* is favored over *BR*; however, for low dimensional feature space, heuristic-based *BR* is favored due to its fewer time and space complexities (Nguyen and Skowron 1995).
- For domains with overlapping value boundaries (e.g., medical), fuzzy rough set discretization may produce better results than other approaches in terms of consistency of the *DS*.

At present, with various *RSBD* methods available in the literature, researchers are facing difficulties in choosing the best discretization method that could be applied in specific domain. The scientists need such *RSBD* methods, which obtain the optimal set of cut-points for the continuous values with minimum inconsistency and higher accuracy. The literature studied has revealed that from the perspective of future research trends in this area, consideration of the applications of hybridized rough set with other machine learning techniques, such as probability and statistics, heuristic approach like entropy, genetic algorithms, fuzzy logic, etc. can lead to new and interesting opportunities for future research. The hybrid of different statistical techniques, such as probability, clustering and covariance and correlation with the core of rough set theory may add more into this area in terms of robustness. The statistical techniques may find accurate and optimal set of cut-points while the *RST* can take care of the inconsistency issues so that to preserve the fidelity of the original dataset. Similarly, the hybrid of entropy with *RST* may approximate the values of random variables for defining the significance of candidate cut-points to select the optimal set and thus can be more looked in the future for better performance.

6 Conclusion

This paper provides a detailed survey of the *RSBD* methods proposed in literature. We have presented a brief literature survey on discretization in general and the published research in this field. The process of *RSBD* has been elucidated with a four-step model and presented with the help of a case study. In this survey, we have described the techniques used for discretization and the approaches adopted for optimizing candidate cut-points. We have discussed each article from the perspective of the domain, datasets used for evaluation, list of algorithms for comparison, consistency, effects on the classification rate and complexities. Based on the techniques used for the optimization of cut-points and discretization, we have designed a taxonomy that can be helpful for readers working in this area. Moreover, the discretization algorithms have also been theoretically analyzed for the domain areas in which they have been used. In order to support the study, a discussion section has been added to suggest some general guidelines to the readers who are new to this field and who desire to continue their research in it.

Acknowledgments This work was supported by the Industrial Core Technology Development Program (10049079, Develop of mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea)". This work was supported by the National Research Foundation of Korea(NRF) Grant funded by the Korea government (MSIP) NRF-2014R1A2A2A01003914. This work was supported by a Grant from the Kyung Hee University in 2011." (KHU-20111209). This work was supported by a Grant from the Kyung Hee University in 2013." (KHU- 20131712).

References

Alcalá-Fdez J, Sánchez L, García S, Jesus MJd, Ventura S, Garrell JM, Otero J, Romero C, Bacardit J, Rivas VM, Fernández JC, Herrera F (2009) KEEL: a software tool to assess evolutionary algorithms to data mining problems. Soft Comput 13(3):307–318. doi:10.1007/s00500-008-0323-y

Bakar AA, Othman ZA, Shuib NLM (2009) Building a new taxonomy for data discretization techniques. In: 2nd conference on data mining and optimization DMO'09, IEEE, pp 132–140

Batu T, Dasgupta S, Kumar R, Rubinfeld R (2005) The complexity of approximating the entropy. SIAM J Comput 35(1):132–150

Bazan J, Szczuka M (2005) The rough set exploration system. Trans Rough Sets III:25–42

Blajdo P, Grzymala-Busse J, Hippe Z, Knap M, Mroczek T, Piatek L (2008) A comparison of six approaches to discretization—a rough set perspective. Rough Sets Knowl Technol, 31–38

Boulle' M (2006) MODL: a Bayes optimal discretization method for continuous attributes. Mach Learn 65(1):131–165

Chebrolu S, Sanjeevi SG (2012) Rough set theory for discretization based on boolean reasoning and genetic algorithm. Int J Comput Corp Res 2(1):75–86

Chen C-Y, Li Z-G, Qiao S-Y, Wen S-P (2003) Study on discretization in rough set based on genetic algorithm. In: IEEE international conference on machine learning and cybernetics, pp 1430–1434

Chlebus BS, Nguyen SH (1998) On finding optimal discretizations for two attributes. In: Polkowski L, Skowron A (eds) Rough sets and current trends in computing. Springer, Berlin, pp 537–544

Dai J-H, Li Y-X (2002) Study on discretization based on rough set theory. In: International conference on machine learning and cybernetics. IEEE, pp 1371–1373

Dougherty J, Kohavi R, Sahami M (1995) Supervised and unsupervised discretization of continuous features. In: Prieditis A, Russell S (eds) 12th International Conference on Machine Learning, Lake Tahoe, CA. Morgan Kaufmann Publishers Inc, pp 194–202

Fayyad U, Irani K (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the thirteenth international joint conference on artificial intelligence, San Francisco, CA. Morgan Kaufmann, pp 1022–1027

Frank and Asuncion (2010) UCI Machine Learning Repository University of California. School of Information and Computer Science, Irvine, CA [http://archive.ics.uci.edu/ml]

Grzymala-Busse JW (2013) Discretization based on entropy and multiple scanning. Entropy 15(5):1486–1502. doi:10.3390/e15051486

Hong X, Zhong CH, Xiao ND (2005) Discretization of continuous attributes in rough set theory based on information entropy. Chin J Comput 28(9):1570–1573

Hu X, Cercone N (1999) Data mining via discretization, generalization and rough set feature selection. Knowl Inf Syst 1(1):33–60

Jia Z, Xie H, Liang Y (2013) Interval similarity-based quantization method for continuous data. J Netw 8(11):2614–2619

Jiang B, Zhang Y (2011) A novel method of continuous attributes discretization in rough sets theory. J Comput Sci Eng 2:85–91

Jiang F, Zhao ZX, Ge Y (2010) A supervised and multivariate discretization algorithm for rough sets. In: Proceeding of the 5th international conference on Rough Set and Knowl Technol, vol 6401. LNCS, Springer Berlin Heidelberg, pp 596–603

Jian-Hua D (2004) A genetic algorithm for discretization of decision systems. In: Proceedings of international conference on machine learning and cybernetics, pp 1319–1323. doi:10.1109/ICMLC.2004.1381977

Jing W, Haohan XIE, Xiangzheng Z, Liang Y, Zhang X, Zhang Y (2013) A supervised statistical data quantization method in machine learning. J Multimed 8(4)

Jing L, Wei-ming L, Jing-bo L, Zhen Y (2011) A two-step integration algorithm for discretization based on rough set theory. In: 2011 6th IEEE joint international information technology and artificial intelligence conference (ITAIC). IEEE, pp 418–420

Kerber R (1992) Chimerge: discretization of numeric attributes. In: Paper presented at the national conference on artifical intelligence American association for artificial intelligence (AAAI)
Kotsiantis S, Kanellopoulos D (2006) Discretization techniques: a recent survey. GESTS Int Trans Comput Sci Eng 32(1):47–58
Kurgan LA, Cios KJ (2004) CAIM discretization algorithm. IEEE Trans Knowl Data Eng 16(2):145–153. doi:10.1109/TKDE.2004.1269594
Liu H, Setiono R (1997) Feature selection via discretization. IEEE Trans Knowl Data Eng 9(4):642–645
Liu H, Hussain F, Tan CL, Dash M (2002) Discretization: an enabling technique. Data Min Knowl Discov 6(4):393–423
Liu H, Liu D-Y, Shi X-H, Gao Y (2008) An attribute discretization algorithm based on rough set and information entropy. In: International conference on machine learning and cybernetics. IEEE, pp 206–211
Luengo J, Saez J, Lopez V, Herrera F (2012) A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. IEEE Trans Knowl Data Eng X (Y)
Lustgarten JL, Gopalakrishnan V, Grover H, Visweswaran S (2008) Improving classification performance with discretization on biomedical datasets. In: AMIA annual symposium proceedings, american medical informatics association, p 445
Maimon O, Rokach L (2010) Data mining and knowledge discovery handbook, vol 14. Springer Incorporated, New York
Nguyen H (1998) Discretization problem for rough sets methods. In: Rough sets and current trends in computing.Springer, pp 545–552
Nguyen HS (2005) Approximate Boolean reasoning approach to rough sets and data mining. In: Rough sets, fuzzy sets, data mining, and granular computing. Springer, pp 12–22
Nguyen SH, Skowron A (1995) Quantization of real value attributes—rough set and boolean reasoning approach. In: Proceedings of the second joint annual conference on information sciences, Wrightsville Beach, North Carolina, pp 34–37
Paul S, Maji P (2011) Fuzzy discretization for rough set based gene selection algorithm. In: Second international conference on emerging applications of information technology (EAIT), pp 317–320
Pawlak Za (1982) Rough sets. Int J Parallel Program 11(5):341–356
Pawlak Z (1992) Rough sets: theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht
Polkowski L, Tsumoto S, Lin T, Bazan J, Nguyen H, Nguyen S, Synak P, WrÃ^{3}blewski J (2000) Rough set algorithms in classification problem. In: Rough set methods and applications, vol 56. Studies in fuzziness and soft computing. Physica-Verlag HD, pp 49–88. doi:10.1007/978-3-7908-1840-6_3
Radwan E, Assiri AMA (2013) Thyroid diagnosis based technique on rough sets with modified similarity relation. Int J Adv Comput Sci Appl 4(10):115–119
Rangel-Merino A, López-Bonilla JL, y Miranda RL (2005) Optimization method based on genetic algorithms. Apeiron 12(4):393–406
Roy A, Pal SK (2003) Fuzzy discretization of feature space for a rough set classifier. Pattern Recognit Lett 24(6):895–902. doi:10.1016/S0167-8655(02)00201-5
Shang L, Wan Q, Zhao Z-H, Chen S-F (2005) Evolutionary computation and rough set-based hybrid approach to rule generation. In: Advances in natural computation. Springer, Berlin, pp 855–862
Shehzad K (2012) EDISC: a class-tailored discretization technique for rule-based classification. IEEE Trans Knowl Data Eng 24(8):1435–1447
Singh GK, Minz S (2007) Discretization using clustering and rough set theory. In: International conference on computing: theory and applications ICCTA'07, pp 330–336
Sun G, Wang H, He X, Lu J (2013) Continuous attributes discretization algorithm based on FPGA. TELKOMNIKA Indones J Electr Eng 11(7):3656–3664
Tian D, Zeng X, Keane J (2011) Core-generating discretization for rough set feature selection. Trans Rough Sets XIII:135–158
Tian D, Keane J, Zeng X-J (2007) Core-generating approximate minimum entropy discretization for rough set feature selection: an experimental investigation. In: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp 1–6
Tinghui L, Liang S, Qingshan J, Beizhan W (2009) Reduction and dynamic discretization of multi-attribute based on rough set. In: WRI world congress on software engineering, WCSE'09. IEEE, pp 50–54
Tsaia C-J, Leeb C-I, Yang W-P (2008) A discretization algorithm based on class-attribute contingency coefficient. Inf Sci 178(3):714–731. doi:10.1016/j.bbr.2011.03.031
Wang J (2013) A rough set approach of mechanical fault diagnosis for five-plunger pump. Adv Mech Eng. doi:10.1155/2013/174987
Wang X, Han D, Han C (2012) Fuzzy-rough set based attribute reduction with a simple fuzzification method. In: 24th Chinese control and decision conference (CCDC). IEEE, pp 3793–3797

Wu X (1996) A Bayesian discretizer for real-valued attributes. Comput J 39(8):688–691
Wu Q, Cai J, Prasad G, McGinnity TM, Bell DA, Guan J (2006) A novel discretizer for knowledge discovery approaches based on rough sets. In: Paper presented at the proceedings of the first international conference on rough sets and knowledge technology (RSKT)
Wu C, Li M, Han Z, Zhang Y, Yue Y (2004) Discretization algorithms of rough sets using clustering. In: IEEE international conference on robotics and biomimetics. ROBIO, IEEE, pp 955–960
Xin G, Xiao Y, You H (2007) Discretization of continuous interval-valued attributes in rough set theory and its application. In: International conference on machine learning and cybernetics. IEEE, pp 3682–3686
Xu T, Yingwu C (2009) Half-global discretization algorithm based on rough set theory. J Syst Eng Electron 20(2):339–347
Xu C, Xu Y, Xiao D (2012) A two-step discretization algorithm based on rough set. In: International conference on computer science and electronics engineering (ICCSEE), pp 178–182. doi:10.1109/ICCSEE.2012.14
Yan L, Xueping G, Jun L (2004) A novel discretization scheme combined entropy with rough set theory for transient stability assessment based on ANN. In: International conference on power system technology, PowerCon 2004. IEEE, pp 119–122
Zhang G, Hu L, Jin W (2005) Discretization of continuous attributes in Rough set theory and its application. In: Computational and information science, vol 3314. Lecture notes in computer science. Springer Berlin Heidelberg, pp 1020–1026. doi:10.1007/978-3-540-30497-5_157
Zheng R, Xi G (2009) The application of discretization based on rough set and information entropy in TCM. In: World congress on nature & biologically inspired computing, NaBIC. IEEE, pp 1695–1701
Zhou Y (2009) Research of attributes discretization based on granular entropy considering the discrimination of rough set. In: Second international conference on intelligent networks and intelligent systems ICINIS'09, pp 526–529
Zhu Q, Lin L, Shyu M-L, Chen S-C (2011) Effective supervised discretization for classification based on correlation maximization. In: IEEE international conference on information reuse and integration (IRI), pp 390–395
Zou L, Yan D, Karimi HR, Reza, Shi P (2013) An algorithm for discretization of real value attributes based on interval similarity. J Appl Math 2013:8. doi:10.1155/2013/350123