Interval Set Cluster Analysis: A Re-formulation

Yiyu Yao[1], Pawan Lingras[2], Ruizhi Wang[3], and Duoqian Miao[3]

[1] Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
yyao@cs.uregina.ca
[2] Department of Mathematics and Computing Science, Saint Mary's University
Halifax, Nova Scotia, Canada B3H 3C3
pawan@cs.smu.ca
[3] Department of Computer Science and Technology
The Key Laboratory of Embedded System and Service Computing
Tongji University, Shanghai, 201804, P.R. China

Abstract. A new clustering strategy is proposed based on interval sets, which is an alternative formulation different from the ones used in the existing studies. Instead of using a single set as the representation of a cluster, each cluster is represented by an interval set that is defined by a pair of sets called the lower and upper bounds. Elements in the lower bound are typical elements of the cluster and elements between the upper and lower bounds are fringe elements of the cluster. A cluster is therefore more realistically characterized by a set of core elements and a set of boundary elements. Two types of interval set clusterings are proposed, one is non-overlapping lower bound interval set clustering and the other is overlapping lower bound interval set clusterings, corresponding to the standard partition based and covering based clusterings.

1 Introduction

Cluster analysis focuses on grouping objects of similar kind into categories and organizing data into meaningful structures [1]. Objects are sorted into groups so that objects in the same group show a high degree of association and objects in different group show a low degree of association. A common assumption underlying many cluster analysis methods is that a cluster can be represented by a set with crisp boundary. The requirement of a sharp boundary leads to easy analytical results, but may be too restrictive for some practical applications. Several proposals have been made to remove such a stringent assumption.

In fuzzy cluster analysis, it is assumed that a cluster is represented by a fuzzy set that models a gradually changing boundary [3]. However, a fuzzy clustering provides a quantitative characterization of the unsharp cluster boundary at the expense of losing the qualitative characterization that better shows the structures provided by a clustering. To resolve this problem, Lingras and his associates [6,7,8,9] propose and systematically study rough clustering and interval set clustering. The basic idea is to derive and describe a cluster by a pair of

H. Sakai et al. (Eds.): RSFDGrC 2009, LNAI 5908, pp. 398–405, 2009.

lower and upper approximations. By describing a cluster in terms of a pair of crisp sets, one recovers the qualitative characterization of a cluster.

There exists a semantic gap in the studies by Lingras and his associates. On the one hand, rough clustering algorithms are explained in rough set terminology. On the other hand, an equivalence relation that is needed for defining approximations is not explicitly referred to. The main objective of this paper is to fill in such a semantic gap by representing a cluster as an interval set defined by a pair of bounds. This leads to the introduction of interval set cluster analysis. Elements in the lower bound of an interval set are typical elements of the cluster and elements between the upper and lower bounds are fringe elements of the cluster. That is, a cluster is more realistically characterized by a set of core elements and a set of fringe elements.

The strategy of interval set cluster analysis does not require an equivalence relation. A set of properties of an interval set clustering is proposed and examined. Based on these properties, two types of interval set clusterings are proposed, one is non-overlapping lower bound interval set clustering and the other is overlapping lower bound interval set clusterings. They correspond to the standard partition based and covering based clusterings.

2 Overview of Interval Sets

In cluster analysis, a cluster may be interpreted as the extension of a concept, that is, the set of objects that are instances of the concept. In some situations, an object may actually be either an instance or not an instance of a concept. On the other hand, due to a lack of information and knowledge, one can only express the state of instance and non-instance for some objects, instead of all objects. That is, one has a partially known concept defined by a lower bound and upper bound of its extension. This leads to the interval set representation of a partially known set [16].

Interval sets are defined and interpreted in a similar way that interval numbers are introduced in interval analysis [10]. The notion of interval sets represents a new kind of sets, defined by a pair of sets, namely, its lower and upper bounds [13,16]. Mathematically, interval sets are defined as follows. Let U be a finite set, called the universe or the reference set, and 2^U be its power set. A subset of 2^U of the form,

$$\mathcal{A} = [A_l, A_u] = \{A \in 2^U \mid A_l \subseteq A \subseteq A_u\}, \tag{1}$$

is called a closed interval set, where it is assumed that $A_l \subseteq A_u$. Being an interval of the power set lattice 2^U, an interval set $\mathcal{A}$ is also a lattice, with the minimum element A_l, the maximum element A_u, and the standard set-theoretic operations. The set of all closed interval sets is denoted by:

$$I(2^U) = \{[A_l, A_u] \mid A_l, A_u \subseteq U, A_l \subseteq A_u\}. \tag{2}$$

A degenerate interval set of the form $[A, A]$ is equivalent to the ordinary set A.

Semantically, an interval set, when interpreted as a family of sets of objects, provides an appropriate means to represent a partially known concept [5,12,13,16,19]. Although the extension of a concept is actually a subset of U, a lack of knowledge makes us unable to specify this subset. We can only provide a lower bound A_l and an upper bound A_u. Any subset A that lies between A_l and A_u, namely, $A_l \subseteq A \subseteq A_u$, can be the actual extension of the concept. The set,

$$\mathrm{BND}([A_l, A_u]) = A_u - A_l, \tag{3}$$

is called the boundary of the interval set $[A_l, A_u]$. For those elements, we are unable to tell if they are instances or non-instances of the concept.

Interval sets are subsets of the universe U. The symbols $\in, \subseteq, =, \cap, \cup$ may be used, in their usual set-theoretic sense, to represent relationships between elements of 2^U and an interval set, and between different interval sets. Thus, $A \in [A_l, A_u]$ means that A is a subset of U such that $A_l \subseteq A \subseteq A_u$. We write $[A_l, A_u] \subseteq [B_l, B_u]$ if the interval set $[A_l, A_u]$ as an ordinary set is contained in $[B_l, B_u]$ as an ordinary set. In other words, by $[A_l, A_u] \subseteq [B_l, B_u]$ we mean that $B_l \subseteq A_l \subseteq A_u \subseteq B_u$. Similarly, two interval sets are equal, written $\mathcal{A} = \mathcal{B}$, if they are equal in set-theoretic sense, that is $\mathcal{A} = \mathcal{B}$ if and only if $A_l = B_l$ and $A_u = B_u$.

Let $\cap, \cup$ and $-$ be the usual set intersection, union and difference defined on 2^U, respectively. Following the results of power algebras [2] and interval analysis [10], we can lift set operations into interval set operations. Specifically, for two interval sets $\mathcal{A} = [A_l, A_u]$ and $\mathcal{B} = [B_l, B_u]$ we have:

$$\begin{aligned}
\mathcal{A} \sqcap \mathcal{B} &= \{A \cap B \mid A \in \mathcal{A}, B \in \mathcal{B}\},\\
\mathcal{A} \sqcup \mathcal{B} &= \{A \cup B \mid A \in \mathcal{A}, B \in \mathcal{B}\},\\
\mathcal{A} \setminus \mathcal{B} &= \{A - B \mid A \in \mathcal{A}, B \in \mathcal{B}\}.
\end{aligned} \tag{4}$$

These operations are referred to as interval set intersection, union and difference. They are closed on $I(2^U)$, namely, $\mathcal{A} \sqcap \mathcal{B}$, $\mathcal{A} \sqcup \mathcal{B}$ and $\mathcal{A} \setminus \mathcal{B}$ are interval sets. They can be explicitly computed by using the following formulas [13,16]:

$$\begin{aligned}
\mathcal{A} \sqcap \mathcal{B} &= [A_l \cap B_l, A_u \cap B_u],\\
\mathcal{A} \sqcup \mathcal{B} &= [A_l \cup B_l, A_u \cup B_u],\\
\mathcal{A} \setminus \mathcal{B} &= [A_l - B_u, A_u - B_l].
\end{aligned} \tag{5}$$

Interval set complement $\neg[A_l, A_u]$ of $[A_l, A_u]$ is defined as $[U, U] \setminus [A_l, A_u]$. It is equivalent to $[U - A_u, U - A_l] = [A_u^c, A_l^c]$, where $A^c = U - A$ denote the usual set complement operation. Clearly, we have $\neg[\emptyset, \emptyset] = [U, U]$ and $\neg[U, U] = [\emptyset, \emptyset]$.

3 Interval Sets, Fuzzy Sets and Rough Sets

Interval sets model concepts that are partially known; they are related to, but different from, fuzzy sets [18] and rough sets [11]. A brief comparison of the three

notions will provide an argument supporting the proposed framework of interval set cluster analysis.

Fuzzy sets model concepts with gradual memberships [18]. Suppose $\mu_A : U \longrightarrow [0,1]$ is a fuzzy membership function. Given a number $\alpha \in [0,1]$, an α-cut of μ_A is defined by:

$$\mu_A^{\alpha} = \{x \in U \mid \mu_A(x) \geq \alpha\}. \tag{6}$$

For a pair of numbers $0 \leq \beta \leq \alpha \leq 1$, the pair of (α, β)-cuts of μ_A gives rise to an interval set $[\mu_A^{\alpha}, \mu_A^{\beta}]$ with $\mu_A^{\alpha} \subseteq \mu_A^{\beta}$. Thus, an interval set may be used as a qualitative approximation of a fuzzy set [13].

Rough sets model the approximations of concepts under indiscernibility [11]. Suppose an equivalence relation on U is used to formally represent the indiscernibility of elements in U. The pair $apr = (U, E)$ is called an approximation space [11]. The equivalence relation E induces a partition of U, denoted by U/E. The equivalence class containing x is given by $[x] = \{y \in U \mid xEy\}$. The equivalence classes of E are the basic building blocks to construct rough set approximations. For a subset $A \subseteq U$, its lower and upper approximations are defined by [11]:

$$\begin{aligned} \underline{apr}(A) &= \{x \in U \mid [x] \subseteq A\}; \\ \overline{apr}(A) &= \{x \in U \mid [x] \cap A \neq \emptyset\}. \end{aligned} \tag{7}$$

The pair $(\underline{apr}(A), \overline{apr}(A))$ is referred to as a rough set generated by A. For a subset $A \subseteq U$, we have $\underline{apr}(A) \subseteq A \subseteq \overline{apr}(A)$. It follows that A induces an interval set $[\underline{apr}(A), \overline{apr}(A)]$. By applying the ideas of (α, β)-cuts of a fuzzy set, one can define probabilistic rough set approximations in a decision-theoretic rough set model [15,17].

Consider now the reverse process of constructing a fuzzy set or a rough set from an interval set. Given an interval set $[A_l, A_u]$, we can define a fuzzy set as follows:

$$\mu_A(x) = \begin{cases} 0, & x \in U - A_u, \\ 0.5, & x \in A_u - A_l, \\ 1, & x \in A_l. \end{cases} \tag{8}$$

If $\min, \max$ and $1 - ()$ are used to define fuzzy set intersection, union, and complement, respectively, we express interval set operations in terms of such three-valued fuzzy sets.

In the case of rough sets, given an interval set $[A_l, A_u]$, in general we may not be able to find a set A so that $A_l = \underline{apr}(A)$ and $A_u = \overline{apr}(A)$. Iwiński [4] suggests another formulation of rough sets, which is closely related to interval sets [14]. Let $\mathrm{Def}(U)$ denote the family of all definable subsets of U given by:

$$\mathrm{Def}(U) = \{A \subseteq U \mid A = \underline{apr}(A) = \overline{apr}(A)\}. \tag{9}$$

For a pair of sets $\underline{A}, \overline{A} \in \mathrm{Def}(U)$ with $\underline{A} \subseteq \overline{A}$, Iwiński refers to the pair $\langle \underline{A}, \overline{A} \rangle$ as a rough set. By definition, it corresponds to the interval set $[\underline{A}, \overline{A}]$. Conversely,

for an interval set $[A_l, A_u]$ with $A_l, A_u \in \text{Def}(U)$, we have an Iwiński rough set $\langle A_l, A_u \rangle$. Thus, the family of all Iwiński rough sets corresponds to a sub-family of all interval sets. Furthermore, their set-theoretic operations are the same [4,14].

Although an interval set may be induced from either a fuzzy set or a rough set, and the reverse is also true under certain conditions, it does have to be interpreted in this way. The interpretation of an interval set as the bounds of a partially known set makes it different from fuzzy sets and rough sets. This interpretation seems to be appropriate for the task of clustering. A cluster may be considered to be a partially known set; we know that certain elements must be in the cluster (e.g., elements in a small neighborhood), and certain elements may be in the cluster (e.g., elements in a large neighborhood).

4 Strategies of Interval Set Clustering

A main task of cluster analysis is to group objects in a universe so that objects in the same cluster are more similar to each other and objects in different clusters are dissimilar. There are two basic strategies of clustering that produce flat non-overlapping and overlapping clusters, respectively.

Suppose

$$\mathbf{C} = (C^1, C^2, \ldots, C^m) \tag{10}$$

is a family of clusters of U, that is, $\mathbf{C}$ is a clustering of the universe. Formally, a non-overlapping clustering is defined by the properties:

$$\begin{aligned} &\text{(i)} && C^i \neq \emptyset, 0 \leq i \leq m, \\ &\text{(ii)} && \bigcup_{C^i \in \mathbf{C}} C^i = U, \\ &\text{(iii)} && C^i \cap C^j = \emptyset, i \neq j. \end{aligned}$$

Property (i) requires that each cluster cannot be empty. Property (ii) states that every $x \in U$ belongs to at least one cluster, and property (iii) states that x belongs to at most one cluster. Together they require that every $x \in U$ belongs to exactly one cluster. In this case, $\mathbf{C}$ is a partition of the universe. On the other hand, an overlapping clustering only requires properties (i) and (ii). For overlapping clustering, it is possible that an element belongs to more than one cluster. The family $\mathbf{C}$ is only a covering of the universe.

An underlying assumption of such a clustering is that one can precisely form a family of clusters with well defined boundary. A questioning of this assumption has led to other clustering strategies. For example, fuzzy clustering produces a family of fuzzy sets, where each cluster is a fuzzy set with gradually changing boundary. Given a pair of numbers $0 \leq \beta \leq \alpha \leq 1$, the (α, β)-cuts of a fuzzy set can be viewed as an interval set [16]. This immediately motivates the introduction of interval set clustering, although in general an interval set clustering can be interpreted without direct reference to a fuzzy clustering.

We assume that each cluster C^i is partially known based on the available information. One may use an interval set to represent such a partially known cluster, namely, C^i is represented by an interval set $[C_l^i, C_u^i]$ satisfying the constraint:

$$C_l^i \subseteq C^i \subseteq C_u^i. \tag{11}$$

The constraint reflects the fact that we do not know the exact cluster C^i but a pair of lower and upper bounds within which C^i lies. Any set in the family $[C_l^i, C_u^i] = \{X \mid C_l^i \subseteq X \subseteq C_u^i\}$ may be the actual cluster C^i. The elements in C_l^i may be interpreted as typical elements of the cluster C^i and elements in $C_u^i - C_l^i$ as fringe elements. With respect to the family of clusters $\mathbf{C} = (C^1, C^2, \ldots, C^m)$, we have the following family of interval set clusters:

$$\begin{aligned}\mathbf{IC} &= ([C_l^1, C_u^1], [C_l^2, C_u^2], \ldots, [C_l^m, C_u^m]) \\ &= \{(C^1, C^2, \ldots, C^m) \mid C_l^i \subseteq C^i \subseteq C_u^i, 1 \leq i \leq m\}.\end{aligned}$$

That is, an interval set cluster is interpreted as a pair of bounds of a family of possible crisp clusters and an interval set clustering is interpreted as bounds of a family of crisp set clusterings.

Based on interval set operations, corresponding to properties (i)-(iii), we adopt the following properties for an interval set clustering:

$$\begin{aligned}&\text{(I)} && C_l^i \neq \emptyset, 0 \leq i \leq m, \\ &\text{(II)} && \bigcup_{[C_l, C_u] \in \mathbf{IC}} C_u = U, \\ &\text{(III)} && C_l^i \cap C_l^j = \emptyset, i \neq j.\end{aligned}$$

Property (I) requires that the lower bound must not be empty. It implies that the upper bound is not empty, namely, $C_u^i \neq \emptyset$. Thus, $C_l^i \neq \emptyset$ may be viewed as a strong version and $C_u^i \neq \emptyset$ as a weak version. It is reasonable to assume that each cluster must contain at least one typical element and hence its lower bound is not empty. We therefore adopt the strong version, instead of the weak version, in order to make sure that an interval set clustering is physically meaningful. Property (II) states that any element of U belongs to the upper bound of a cluster, which ensures that every element is properly clustered. Property (III) demands that the lower bounds of clusters are pairwise disjoint; a typical element of one cluster cannot, as the same time, be a typical element of another cluster.

Additional support for adopting properties (I), (II), and (III) is given by the following theorem that shows the connection of a standard clustering and an interval set clustering.

Theorem 1. *Suppose* $\mathbf{IC} = ([C_l^1, C_u^1], [C_l^2, C_u^2], \ldots, [C_l^m, C_u^m])$ *is an interval set clustering. If* $\mathbf{IC}$ *satisfies properties* (I), (II), *and* (III), *then there exists a family of clusters* $\mathbf{C} = (C^1, C^2, \ldots, C^m)$ *that satisfies the constraint* $C_l^i \subseteq C^i \subseteq C_u^i$ *and properties* (i), (ii), *and* (iii). *If* $\mathbf{IC}$ *satisfies properties* (I) *and* (II), *there exists a family of clusters* $\mathbf{C} = (C^1, C^2, \ldots, C^m)$ *that satisfies the constraint* $C_l^i \subseteq C^i \subseteq C_u^i$ *and properties* (i) *and* (ii).

Proof. The theorem can be proved constructively by building a family of clusters **C** from **IC**. Assume that **IC** satisfies properties (I), (II), and (III), one can construct a **C** as follows. We first construct a family of clusters $\{C^i = C^i_l \mid 1 \leq i \leq m\}$ based on typical elements of clusters. For each element x in the set of the fringe elements, $F = \bigcup\{C^i_u - C^i_l \mid 1 \leq i \leq m\}$, we assign it to only one of the clusters C^i's that satisfies the condition $x \in C^i_u - C^i_l$. By the property (I), it follows that **C** satisfies property (i); by the properties (II) and (III) and the construction procedure, it follows that **C** satisfies properties (ii) and (iii). To prove the second part of the theorem, we follow the same procedure except that we may assign each fringe element to a set of clusters instead of one. It can be easily seen that the resulting **C** satisfies properties (i) and (iii).

Based on the results from the theorem, an interval set clustering **IC** is called a lower bounds non-overlapping interval set clustering if it satisfies properties (I), (II), and (III); it is called a lower bounds overlapping interval set clustering if it only satisfies properties (I) and (II). They suggest different interval set clustering algorithms.

There are several differences between rough set clustering and interval set clusterings. Rough set clustering requires an underlying equivalence and hence is only applicable to non-overlapping clustering. In general, one may use a non-equivalence relation to obtain an overlapping clustering. In this case, it is necessary to refer to this underlying relation in order to properly interpret the rough set lower and upper approximations. In contrast, interval set clustering does not require such an underlying relation. In some earlier studies of rough set cluster analysis, it is assumed that a fringe element must belong to the upper bounds of at least two clusters [6,7,8,9], which is motivated by properties of the upper approximations in the rough set theory. With interval set clustering, we no longer need to impose such a constraint. It is possible that a fringe element belongs to the upper bound of only one cluster.

5 Conclusion

There is a growing interest in rough set cluster analysis. An important issue that has not received enough attention is a semantic interpretation of the derived clusters. Since rough set approximations must satisfy certain properties, their directly application to cluster analysis may be unnecessarily restrictive. In this paper, we outline a framework of interval set cluster analysis, which is motivated by, and different from, rough set cluster analysis.

The clarification of rough set cluster analysis and interval set cluster analysis have both theoretical and practical values. Although the results from both clustering methods are intervals in the power set of a set, they have different semantic interpretations. Rough set approximations are approximation of known sets in an approximation space defined by an underlying equivalence or non-equivalence relation. In order to explain rough set cluster analysis, we need to refer to the relation. In contrast, interval sets are approximations of partially known sets; interval set cluster analysis does not require such a relation. With

interval set clustering, an object can belong to the upper bound of one cluster, which is different from rough set clustering where an object, if in the upper approximation of one cluster, must be in the upper approximation of at least one more cluster.

References

1. Anderberg, M.R.: Cluster Analysis for Applications. Academic Press, New York (1973)
2. Brink, C.: Power structures. Algebra Universalis 30, 177–216 (1993)
3. Höppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. Wiley, Chichester (1999)
4. Iwiński, T.B.: Algebraic approach to rough sets. Bulletin of the Polish Academy of Sciences, Mathematics 35, 673–683 (1987)
5. Marek, V.W., Truszczyński, M.: Contributions to the theory of rough sets. Fundamenta Informaticae 39, 389–409 (1999)
6. Lingras, P.: Rough K-Medoids clustering using GAs. In: Proceedings of the 8th IEEE International Conference on Cognitive Informatics, pp. 315–319 (2009)
7. Lingras, P., Hogo, M., Snorek, M.: Interval set clustering of web users using modified Kohonen self-organizing maps based on the properties of rough sets. Web Intelligence and Agent Systems: An International Journal 2, 217–230 (2004)
8. Lingras, P., Hogo, M., Snorek, M., West, C.: Temporal analysis of clusters of supermarket customers: conventional versus interval set approach. Information Sciences 172, 215–240 (2005)
9. Lingras, P., West, C.: Interval set clustering of web users with rough K-Means. Journal of Intelligent Information Systems 23, 5–16 (2004)
10. Moore, R.E.: Interval Analysis. Prentice-Hall, Englewood Cliffs (1966)
11. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
12. Wang, Y.Q., Zhang, X.H.: Some implication operators on interval sets and rough sets. In: Proceedings of 2009 IEEE International Conference on Cognitive Informatics, pp. 328–332 (2009)
13. Yao, Y.Y.: Interval-set algebra for qualitative knowledge representation. In: Proceedings of the Fifth International Conference on Computing and Information, pp. 370–374 (1993)
14. Yao, Y.Y.: Two views of the theory of rough sets in finite universes. International Journal of Approximation Reasoning 15, 291–317 (1996)
15. Yao, Y.Y.: Probabilistic rough set approximations. International Journal of Approximation Reasoning 49, 255–271 (2008)
16. Yao, Y.Y.: Interval sets and interval-set algebras. In: Proceedings of the 8th IEEE International Conference on Cognitive Informatics, pp. 307–314 (2009)
17. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximating concepts. International Journal of Man-machine Studies 37, 793–809 (1992)
18. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
19. Zhang, X.H., Jia, X.Y.: Lattice-valued interval sets and t-representable interval set t-norms. In: Proceedings of 2009 IEEE International Conference on Cognitive Informatics, pp. 333–337 (2009)