


Article

Intuitionistic and Interval-Valued Fuzzy Set Representations for Data Mining

Fred Petry ^{1,*},[†]  and Ronald Yager ^{2,†}¹ Naval Research Laboratory, Stennis Space Center, Hancock County, MS 39556, USA² Machine Intelligence Institute, Iona College, New Rochelle, NY 10801, USA; yager@panix.com

* Correspondence: frednavy3@gmail.com

† These authors contributed equally to this work.

Abstract: Data mining refers to a variety of techniques in the fields of databases, machine learning and pattern recognition. The intent is to obtain useful patterns and associations from a large collection of data. In this paper we describe extensions to the attribute generalization process to deal with interval and intuitionistic fuzzy information. Specifically, we consider extensions for using interval-valued fuzzy representations in both data and the generalization hierarchy. Moreover, preliminary representations using intuitionistic fuzzy information for attribute generalization are described. Finally, we consider how to use fuzzy hierarchies for the generalization of interval-valued fuzzy representations.

Keywords: data mining; attribute generalization; concept hierarchies; interval-valued fuzzy sets; intuitionistic-valued fuzzy sets

**Citation:** Petry, F.; Yager, R.Intuitionistic and Interval-Valued Fuzzy Set Representations for Data Mining. *Algorithms* **2022**, *15*, 249. <https://doi.org/10.3390/a15070249>

Academic Editors: Eugene Levner and Milan Vlach

Received: 10 June 2022

Accepted: 12 July 2022

Published: 19 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Imprecise or uncertain information and data need to be taken into account for databased organization and management. This sort of data occurs in many diverse areas, including ecological data, economics and forensic information. In order to make use of such information, we must settle on how the various pieces of data can be used to make a decision or to take an action. This can involve some sort of summarization and generalization of the pieces of data regarding what conclusions they can support [1–3]. A currently emerging issue is the management of uncertain information arising from multiple sources and of many forms that appear in the everyday activities and decisions of humans. This can include sensor information and data ranging to the subjective interpretations obtained from expert individuals and analysts. Currently, increasingly massive amounts of heterogeneous data and information from multiple sources are prevalent where the problems of Big Data are being managed [4–7]. However, although effective decision making should be able to make use of all the available and relevant information about such combined uncertainty, an assessment of the value of a generalization result is critical. There have been a number of approaches to using soft computing in data mining [8,9]. One possible approach for such a generalization process can be found in the use of concept hierarchical generalization [10,11]. In previous research, the problem of evidence resolution was studied for crisp concept hierarchies [12].

In this paper, we first briefly review a number of approaches to modeling uncertainty, including for the interest of this paper, interval-valued fuzzy sets (IVFS) and intuitionistic fuzzy sets (IFS). Then, we describe the general process of attribute generalization used for data mining. For the purposes of generalization, we further describe concept hierarchies and their relationship to attribute generalization. Specifically, we consider extensions for using interval-valued fuzzy representations in both data and the generalization hierarchy. Moreover, preliminary representations using intuitionistic fuzzy information for attribute

generalization are described. Finally, we consider how to use fuzzy hierarchies for the generalization of interval-valued fuzzy representations.

2. Background

2.1. Uncertainty Representations

In this section, we briefly overview common uncertainty representations [13], including fuzzy sets, interval-valued sets, intuitionistic fuzzy sets and others for approaches to the generalization of such data to concept hierarchies. The other types can be used in similar approaches.

2.1.1. Fuzzy Set Theory

Fuzzy set representations [14,15] provide membership degrees of data values in a set, as opposed to crisp sets. For a domain D , a fuzzy set, FS , is

$$FS(D) = \{ \langle a_i, m(a_i) \rangle \mid 0 \leq m(a_i) \leq 1, a_i \in D \} \tag{1}$$

where a_i is a data value, and $m(a_i)$ is the membership of the data value.

2.1.2. Interval-Valued Fuzzy Sets

Interval values are used in many areas to capture the imprecision and uncertainty of data. We first provide the formalisms for interval arithmetic [16–18] as needed for interval-valued fuzzy sets. We let D be the domain, and intervals are represented by the values of the lower bound, $z_+ = lb(a_i)$, and an upper bound, $z^\dagger = ub(a_i)$, of an interval $I(a_i)$ for the data value $a_i \in D$

$$I(a_i) = [z_+, z^\dagger] = \{ z \in D \mid z_+ \leq z \leq z^\dagger \} \tag{2}$$

Now, an interval-based fuzzy set representation, $IVF(D)$, is based on using upper, $mu(a_i)$, and lower bounds, $ml(a_i)$, on fuzzy memberships

$$IVF(D) = \{ \langle a_i, I(a_i) \rangle \mid I(a_i) = [ml(a_i), mu(a_i)] \} \tag{3}$$

For an interval $I(a_i)$, the size or length of the interval, IW , is the difference of the lower and upper bounds,

$$IW(I(a_i)) = |ml(a_i) - mu(a_i)| \tag{4}$$

IW is often used as a representation of the uncertainty of a data value a_i in an IVF as an information measure [19,20].

2.1.3. Intuitionistic Fuzzy Sets

Intuitionistic fuzzy set theory extends ordinary fuzzy set theory by allowing both positive and negative memberships to be specified. Recall that an ordinary fuzzy set $FS(D) = \{ \langle a_i, m(a_i) \rangle \}$ has only one membership value for a data element a_i . An intuitionistic fuzzy set $IFS(D)$ [21] allows both positive, $m_S(a_i)$, and negative membership values, $m_S^*(a_i)$.

$$IFS(D) = \{ \langle a_i, m_S(a_i), m_S^*(a_i) \rangle \mid a_i \in D \} \tag{5}$$

where $m_S(a_i), m_S^*(a_i) \in [0, 1]$.

Specifically, the sum of the membership, $m_S(a_i)$, and non-membership, $m_S^*(a_i)$, is not necessarily one, then: $0 \leq m_S(a_i) + m_S^*(a_i) \leq 1$. Additionally, the hesitation $h_S(a_i)$

$$h_S(a_i) = 1 - (m_S(a_i) + m_S^*(a_i)) \tag{6}$$

is the degree of indeterminacy (hesitation).

2.1.4. Type-2 Fuzzy Sets

A type-2 fuzzy set $TY2(D)$ [22,23] is one in which the membership values, $m_T(a_i,r)$, are themselves uncertain and can be represented by a fuzzy set itself. Therefore, if there is no uncertainty in the membership function, this reduces to ordinary fuzzy sets.

$$TY2(d) = \{ \langle (a_i, r), m_T(a_i,r) \rangle \mid a_i \in D \} \tag{7}$$

where $r \in P_x \subseteq [0, 1]$.

2.1.5. Rough Set Theory

The core concept of rough sets is an indiscernibility relation IR on the domain D [24]. A rough set N is specified by using the upper, R^uN , and lower approximations, R_lN , of N .

- The lower approximation of N is the set $R_lX = \{a_i \in D / [a_i]_R \subseteq N\}$.
- The upper approximation of N is the set $R^uX = \{a_i \in D / [a_i]_R \cap N \neq \emptyset\}$.

where $[a_i]_R$ denotes the equivalence class of the indiscernibility relation R containing a_i .

In summary, the lower approximation of a set is a conservative approximation comprising only elements which can definitely be determined to be members of the set. The upper approximation is a *liberal* approximation, including all elements that may be members of the set.

2.1.6. Dempster–Shafer Uncertainty Theory

The Dempster–Shafer (D–S) theory is an established approach to modeling uncertainty [25] by providing representations of non-specific forms of uncertainty. A Dempster–Shafer belief structure consists of a collection of non-empty crisp subsets of a space D , called focal elements: R_1, R_q . The mass or basic probability, bp , is used to assign every member of the power set a belief, bp :

$$bp: 2^D \rightarrow [0, 1]$$

Therefore, here, our knowledge of the value of a variable is inexact, where for the focal set, $R_i \subset D$, $bp(R_i)$ indicates the probability that the value is in R_i . Two important properties of bp are: the basic probability of the empty set is zero,

1. $bp(\Phi) = 0$, and the basic probabilities of the rest of the power set’s elements sum to 1,
2. $\sum_{R_i \in 2^D} bp(R_i) = 1$.

Two commonly used measures for a Dempster–Shafer belief structure are measures of belief (best case) and plausibility (worst case). The belief for a specific set W , $Bel(W)$, is the sum of the basic probabilities of all subsets of W . The plausibility, $Pl(W)$, is the sum of the bps of the sets R_i that intersect W .

3. Attribute Generalization and Concept Hierarchies

In this section, we overview the overall process and objectives of data mining using the attribute generalization approach. Then, concept hierarchies are introduced, and the relationship of the hierarchies is used for attribute generalization.

3.1. Attribute Generalization

The objective of attribute generalization in databases is both to reduce the number of tuples in a relation and to have values of some of the attributes to be more general or a higher [26–28]. This aids in the user interpretation and analysis of the data. Attribute generalization has been applied in a number of uncertainty representations, including fuzzy [29,30] and rough databases [31]. In this section, we consider a number of combinations of uncertainty in hierarchies as well as in the data being generalized.

We consider the database to consist of relations with attributes, $R_k (A_1, A_2, \dots, A_n)$, where each relation is a set of tuples, t_j , that contains the actual data values of the attributes, a_i . A tuple t_j is of the form

$$t_j (a_1, a_2, \dots, a_n) \in R_k \tag{8}$$

where a_i is the specific data value attribute A_i in tuple t_j . Now, after generalization for a specific relation, the tuples have the original data replaced by the more general values, a_i' .

After generalization, we must check to see if tuples have become similar enough to be merged, which assists the desired reduction in the number of tuples. For simplicity, consider two tuples that have been generalized, and let $\text{Sim} (t_1, t_2)$ be the similarity of the two tuples. If two generalized tuples become similar enough, they are merged, and a multiple count attribute, Mul , is added to keep track of how many tuples have been merged to form the current generalized tuple. Therefore, if $\text{Sim} (t_1, t_2) = 1$, where a_i' denotes the generalized value, then

$$t_1 (a_1', \dots, a_i', \dots) = t_2 (a_1', \dots, a_i', \dots) \tag{9}$$

and we then have the merger of these tuples:

$$t_{12} (a_{12}' = a_1', \dots, \text{Mul} = 2) \tag{10}$$

The value of the multiple count of a tuple should be carried to its generalized tuple, and the counts should be accumulated when merging identical tuples in generalization.

The overall objective of generalization is to reduce the data into forms that are more easily analyzed and interpreted. This can be achieved by using thresholds; therefore, there is not an over-generalization. An attribute threshold is compared to the distinct values in the tuples for attribute domain $P(A_i)$, and if they are more than the threshold, additional generalization on this attribute is needed. Moreover, the number of tuples in a generalized relation should be lower than this tuple threshold, and if not, generalization continues.

A common approach to analyzing a class of the data is by using characteristic rules from generalized data. For this, tuples that are relevant to the class correspond to a disjunct of a rule. The count attribute formulates the rule condition's strength. These rules provide the conditions characterizing the particular class in which a user is interested.

3.2. Concept Hierarchies

We consider a simple or ordinary concept hierarchy, CH_i , associated with A_i , an attribute variable in the database where $P(A_i)$ is the domain of the data values of A_i . Then, CH_i has a number of levels which partition the domain space $P(A_i)$.

In a concept hierarchy (Figure 1), each level k specifies a relationship $\Gamma_k: P(A_i) \times P(A_i) \rightarrow \{0, 1\}$ [32]. This relationship is reflexive, symmetric and transitive, i.e., an equivalence relationship. This relationship is a many-to-one relationship, which means that many data values are, in general, related to the same concepts at the next level of the hierarchy. We also consider in this section the idea of complex hierarchies for which the data values may be related to more than one concept at a higher level, i.e., a many-to-many relationship. This was used to represent fuzzy hierarchies in previous research [33].

The meaning of this relationship, $\Gamma_k(x, y) = 1$, is that two values, x and y , are essentially the same. Therefore, Γ partitions $P(A_i)$ into m_k disjoint subsets [34]. We can denote Q_i , the i th equivalence class at the k th level of a concept hierarchy.

As CH is ascended, partitions coarsen, and if data values are in same class Q for the level j of CH, they are in the same class in any higher-level k . Therefore, the concept hierarchy, CH, at each level, consists of the partitioning of the data domain $P(A_i)$ into m_k categories (equivalence classes):

$$Q_1, Q_2, \dots, Q_{m_k} \tag{11}$$

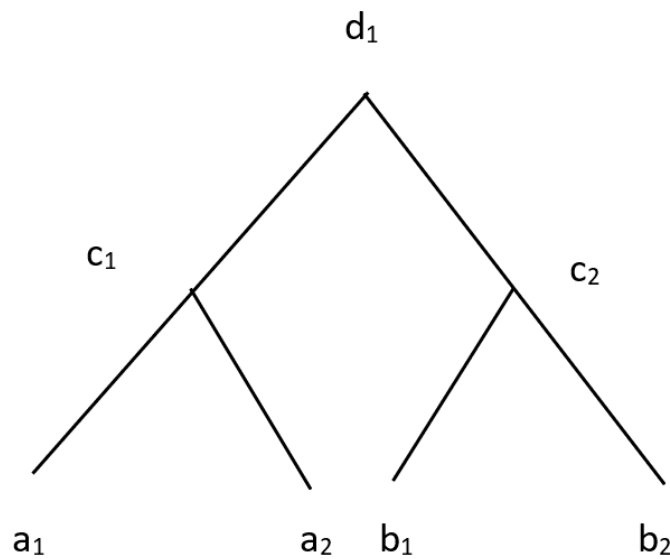


Figure 1. Simple concept hierarchy illustration.

If we have m levels, then the concept hierarchy is a collection of m partitions of the space $P(A_i)$. In particular, the concept hierarchy consists of

Partition 1: $Q_{1|i} \quad i = 1 \dots m_1$

Partition 2: $Q_{2|i} \quad i = 1 \dots m_2$

Partition t : $Q_{t|i} \quad i = 1 \dots m_t$

Each category $Q_{t|i}$ corresponds to a subset of $P(A_i)$, and it has an associated label or value describing the category. Therefore, when performing a generalization of the data for a given attribute in a tuple, we use a linguistic label c' for the generalized data value in the tuple.

We note that the partition of data values for any category may be an imprecise set, as we shall use in following sections.

3.3. Generalization with Respect to Partitions

We can describe attribute generalization with respect to these hierarchy concepts. For simplicity, we consider a one-attribute relation $R(A_1)$, where $P(A_1) \subseteq P(A)$ is the domain of values for A_1 . At level k in a concept hierarchy, we have the categories $Q_{k|1}, Q_{k|2}, \dots, Q_{k|n}$. Therefore, in the relation $R(A_1)$, the subset of data is $P(A_1)' \subseteq P(A_1)$, in general. Then, for this data subset, there is $Q'_{k|i}$

$$Q'_{k|i} \subseteq Q_{k|i} \tag{12}$$

In some cases, $Q'_{k|i}$ may be empty, i.e., none of the specific data values, A_1 , in R generalize to the concept $Q_{k|i}$.

Now, in this simplified situation of a single-attribute relation, $Q'_{k|i}$ corresponds to the tuples merged after generalization. If the label corresponding to $Q'_{k|i}$ is c'_k , the merged tuple is

$$t'(c'_k, |Q'_{k|i}|) \tag{13}$$

and the cardinality of $Q'_{k|i}$ is the merge count, i.e., the number of tuples merged to form t' . The number of tuples in the single-attribute relation R , after generalization to a level k , is

$$N = \sum_{i=1}^{n_k} \{ 1 \text{ if } Q'_{k|i} \neq \emptyset ; 0 \text{ otherwise} \} \tag{14}$$

For a relation with additional attributes—for example, two attributes— $R(A_1, A_2)$, then, after generalizing on A_1 , it is completely possible that not all tuples are merged. This arises

since there may be two generalized tuples in which the values of A_2 are different, even after generalization, such as

$$t'(c'_{k|i}, c'_{k|p}) \neq t'(c'_{k|i}, c'_{k|q}) \tag{15}$$

which cannot be merged, since $Q_{k|p} \neq Q_{k|q}$. If R' is the relation after merging, then we can see that N is a lower bound on the size (number of tuples) of R'

$$|R'| \geq N \tag{16}$$

4. Generalization Involving Interval-Valued and Intuitionistic Fuzzy Information

This section proposes a number of extensions to attribute generalization, in which the forms of imprecise information may be in both the data values as well as being represented in a concept hierarchy. We describe concept hierarchies as either ordinary or complex. By this, we mean that, in an ordinary hierarchy, data generalize to a single concept in the hierarchy. For a complex hierarchy, we allow multiple concepts that correspond to data values. This allows the degrees of the memberships of the relationship of the data values to concepts and has been used to model fuzzy hierarchies in previous research [12,33].

4.1. Generalization with Ordinary Hierarchies

We first consider the generalization of the values of the imprecise attributes using a crisp concept hierarchy. Again, after generalization, we must check to see if tuples become similar enough to be merged. For simplicity, we use an example of two tuples that are being considered for generalization.

4.1.1. IVF Data

For interval-valued fuzzy data, we use the following notation, $I(a')$, for a data value's membership $m(a')$:

$$m(a') = I(a') = [ml(a'), mu(a')]. \tag{17}$$

Then, consider the two tuples

$$t_1(a', I_1(a')) \text{ and } t_2(b', I_2(b')) \tag{18}$$

Here, the data values a' and b' generalize to one concept in CH_i , labeled as c' . This produces two tuples for which the membership values of the data are not affected by the generalization when using a crisp hierarchy:

$$t_1(c', I_1(a')) \text{ and } t_2(c', I_2(b')) \tag{19}$$

Now, to assess the similarity of the tuples, $Sim(t_1, t_2)$, we must consider the relationships possible between the membership intervals in the tuples. There are several cases to consider for the generalization of interval-valued fuzzy data:

1. Data intervals are contained: $I_j \subseteq I_k$.
2. Data interval overlap: $I_j \cap I_k \neq \emptyset$.
3. The intervals are disjoint, so tuples cannot be combined: $I_j \cap I_k = \emptyset$.

For the first two cases, we must combine the membership intervals of the generalized tuples that are being merged. The first approach uses an arithmetic averaging of interval bounds, F_{avg} :

$$F_{avg}(I(a') = [ml(a'), mu(a')]; I(b') = [ml(b'), mu(b')]) \\ = [(ml(a') + ml(b'))/2, (mu(a') + mu(b'))/2]$$

The second approach is a melding process of the intervals, F_{md} :

$$F_{md}(I(a') = [ml(a'), mu(a')]; I(b') = [ml(b'), mu(b')])$$

$$= [\min(\text{ml}(a'), \text{ml}(b')), \max(\text{mu}(a'), \text{mu}(b'))]$$

Here is an example of three tuples that have been generalized:

$$t_1(c', [0.2, 0.6]); t_2(c', [0.3, 0.5]); t_3(c', [0.4, 0.8]) \tag{20}$$

Using F_{avg} , we can obtain the merged tuple $t'(c', [0.3, 0.635])$, and for F_{md} , there is the tuple $t'(c', [0.2, 0.8])$.

To compare these results, we can use the interval width $IW(a') = |\text{ml}(a') - \text{mu}(a')|$. Note that $0 < IW(a') < 1$, where a greater interval width implies more uncertainty in the interval. Then, the interval width of the merged tuples for F_{avg} is $IW(c') = 0.335$, and for F_{md} , $IW(c') = 0.6$. Therefore, F_{avg} has preserved more information in the merged tuples. We also note that the average IW for the three original tuples is 0.333, so the information content has essentially been preserved.

4.1.2. IVF Concepts in Hierarchies

Next, we consider the case where the concept sets, Q_i s, in the hierarchy have associated uncertainty. In this case, for interval-valued fuzzy memberships, $I(c'_i)$,

$$Q_1: \langle c'_1, I(c'_1) \rangle, Q_2: \langle c'_2, I(c'_2) \rangle, \dots, Q_m: \langle c'_m, I(c'_m) \rangle$$

For cases where n of the data values a_i, a_j, \dots in tuples generalize to some c' , we have n tuples of the form $t(c'_i, I(c'_i))$, and these can be merged as in the previous section. Now, if the data values themselves have interval-valued fuzzy memberships, then a fusion of the intervals from the hierarchy and in the data must be considered. Depending on the semantics of the application, there are several possibilities. Either interval-valued fuzzy memberships may be used in the generalized and combined values by applying F_{avg} or F_{md} as discussed above, or a weighted combination may be used. Note that the resultant generalized tuples must still be combined based on their similarity. Again, we must consider the possible interval relationships:

$$I(c') \subseteq I(a') \text{ or } I(c') \cap I(a') \neq \emptyset$$

If a concept interval, $I(c')$, does not overlap a particular data interval, $I(a')$, then a reasonable semantic decision is that the tuple $t(a')$ should not be generalized. For the other cases, F_{avg} or F_{md} can be used. The information about the interval-valued membership of c' is likely more significant, since the value from the hierarchy, c' , is the value in the generalized tuple. Therefore, for example, in F_{avg} , we can weigh $I(c')$ appropriately.

$$F_{\text{avg}}(I(a'), w(c') * I(c')) \tag{21}$$

This captures the general trend of the case of the relationships between the intervals. When there is no overlap, $w = 0$ corresponds to there being no generalization. As the overlap increases, w increases to 1 when $I(c')$ is totally contained in $I(a')$.

4.1.3. Evaluation of Generalization

In order to make effective usage of generalizations for decision-making criteria, measures and metrics must be used to provide an analysis of the generalizations. We can consider measures such as granularity and the overlap of data associated with different concept hierarchies. The most granular partitioning of data occurs when all values are lumped into one set, and the finest partition is where each data value is in a separate set. A measure that is termed as coarseness or granularity was used to characterize partitioning, where the coarseness of the maximum partition was the greatest, 1, and the minimum was the finest, with 0 granularity.

Let $R_i(A_1, A_2, \dots, A_n)$ be a relation with attributes that have interval-valued fuzzy membership values for their data values. Depending on specific applications, there may

be more than just one concept hierarchy relevant that can be used to generalize these data values. As a consequence, we would like to compare the effectiveness of generalization using m different hierarchies, CH_1, \dots, CH_m . For this, we compare the sets of generalized data values, S_1, S_2, \dots, S_m , with the granularity measure. The formulation of granularity for crisp sets uses their cardinality, Cd , but must be extended for imprecise data, such as for fuzzy sets [15]. Here, as above, we use the information measure IW of the intervals in the generalized data for Cd .

$$Cd(S_j) = |S_j| = \sum_{c_i \in S_j} IW(c_i) \tag{22}$$

Then, the expression for the granularity G , extended from the form of Yager (2008), is

$$G(S_1, S_2 \dots S_m) = \frac{\sum_{j=1}^m (Cd(S_j)^2) - \Delta}{Z} \tag{23}$$

where

$$Z = (\Delta^2 - \Delta) \text{ and } \Delta = \sum_{j=1}^m Cd(S_j) \tag{24}$$

Now, we consider the use of three potential concept hierarchies, CH_1, CH_2 and CH_3 , to generalize some of the data in a relation R_i . This example corresponds to three sets of the data values in the final generalized tuples, S_1, S_2 and S_3 .

$$S_1 = \{c_1, IW(c_1) = 0.7; c_2, IW(c_2) = 0.6 \ c_3, IW(c_3) = 0.6\}$$

$$S_2 = \{c_1, IW(c_1) = 0.5; c_2, IW(c_2) = 0.4\}$$

$$S_3 = \{c_2, IW(c_2) = 0.8; c_3, IW(c_3) = 0.7\}$$

Therefore, the granularity of the possible choices, S_1 and S_2, S_2 and S_3 , and S_1 and S_3 must be considered. First, we have

$$G(S_1, S_2) = (Cd(S_1)^2 + Cd(S_2)^2) - \Delta / Z = ((3.61 + 0.81) - 2.8) / 0.04 = 0.32 \tag{25}$$

Similarly,

$$G(S_2, S_3) = 0.66 / 3.36 = 0.196; \ G(S_1, S_3) = 2.46 / 8.16 = 0.30$$

For the purposes of data mining, this provides a criterion to provide multiple generalizations that may be useful to provide alternative data representations for evaluation and interpretation. Therefore, the best choice for this example is using CH_1 , with CH_2 corresponding to S_1 and S_2 , although S_1 and S_3 derived from CH_1 and CH_3 are acceptable.

4.1.4. IFS Data

The structure of the intuitionistic fuzzy data and concept hierarchies has complications that are not found in the generalization of the interval-valued fuzzy data. Therefore, in this section, we describe some representations for IFS data that can potentially facilitate attribute generalization techniques.

The issue is that positive and negative memberships that are found in IFS data may have different semantics for each membership, which is not what we encounter with the interval approach. Next, we consider how to generalize tuples in which the attributes for the data have fuzzy intuitionistic memberships. We must consider how the crisp concept hierarches for such attributes can be specified for this case. We propose to derive separate tuples in which the positive and negative memberships are separated before generalization. This can allow a simpler process for merging tuples after generalization.

Therefore, for an attribute A_i , we partition the domain $P(A_i)$ of its IFS values into two sets of tuples: tp positive for ordinary memberships and tn negative for non-membership degrees.

For example, then, we can have a data value a' in both sets, although, for some applications, there may not be a negative membership value for a' .

$$tp_j = \{ \dots a' / m_j(a') \dots \} \quad tn_j = \{ \dots a' / m^*_j(a') \dots \} \tag{26}$$

When a value a' is generalized to a concept C_j with label c' , since it can have both positive and negative membership values, this can indicate support or the lack of support for the concept. Based on this, as a further refinement, we can consider dividing the data generalized by considering data for which a' has a membership in which the positive membership $m_j(a')$ is greater or less than $m^*_j(a')$.

$$P1_j = \{ a' \mid (m_j(a') > 0 \wedge m_j(a') \geq m^*_j(a')) \} \tag{27}$$

$$N1_j = \{ a' \mid (m_j(a') > 0 \wedge m_j(a') \leq m^*_j(a')) \}$$

Therefore, then, the generalized tuples based on this sort of data value decomposition can be merged more meaningfully. We can consider a further approach to the structure of the intuitionistic generalization of data in the concepts. We have shown a structure of sets where either the positive or negative membership values are greater for the data. However, for evaluations, if the memberships are relatively small, then the structure has less usefulness. For example, if $m_j(a') = 0.2$ and $m^*_j(a') = 0.1$, the distinction of $m_j(d_i)$ being greater is less important. In these cases, a structure can be introduced in which the larger positive or negative membership is greater than some threshold T , such as 0.4, and the analysis can proceed from this sort of data organization.

$$P11_j = \{ a' \mid (m_j(a') > 0 \wedge m_j(a') \geq m^*_j(a') \wedge m_j(a') \geq T) \}$$

$$P12_j = \{ a' \mid (m_j(a') > 0 \wedge m_j(a') \geq m^*_j(a') \wedge m_j(a') < T) \}$$

$$N11_j = \{ a' \mid (m_j(a') > 0 \wedge m_j(a') \leq m^*_j(a') \wedge m^*_j(a') \geq T) \} \tag{28}$$

$$N12_j = \{ a' \mid (m_j(a') > 0 \wedge m_j(a') \leq m^*_j(a') \wedge m^*_j(a') < T) \}$$

4.2. Generalization with Complex Hierarchies

In complex hierarchies, a data value can generalize to more than one concept in the hierarchy, as seen in Figure 2. The data value b_1 generalizes to more than one value, c_1 and c_2 . We note that this is not true in general, as a_2 generalizes to only c_1 .

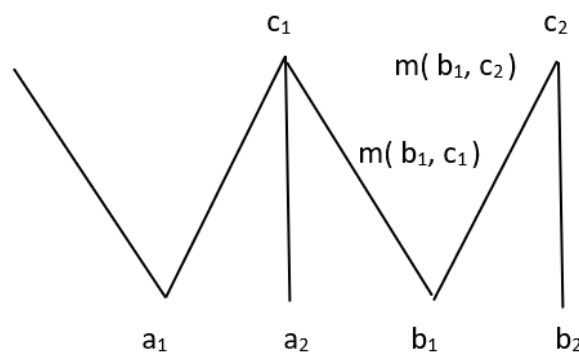


Figure 2. Complex Hierarchy: CH_C .

Now, we can describe a complex hierarchy following the simple concept hierarchy with respect to partitions, as in Section 3. Therefore, at each level k , we have a defining many-to-many relationship:

$$\Gamma_k: P(A) \times P(A) \rightarrow [0, 1] \tag{29}$$

Such a relationship entails many-to-many equivalence classes of concepts at each level. However, these sets do not form a formal set partitioning of a domain as for crisp equivalence classes in simple concept hierarchies. This means that, instead of a partitioning of the domain $P(A)$, there is set decomposition. At each level, this is $Sd_{k|1}, Sd_{k|2}, \dots$, and, in general,

$$Sd_{k|i} \cap Sd_{k|j} \neq \emptyset \tag{30}$$

This means that there may not be a unique concept at level r to which a value at level $r - 1$ generalizes, as seen in Figure 2, which then allows fuzzy concept hierarchies.

Fuzzy Hierarchies

Lastly, in this section, we describe the extensions for the generalization of interval-valued fuzzy data representations using fuzzy hierarchies. The generalization of crisp and fuzzy data for fuzzy hierarchies has been previously considered [12,33,35]. For crisp and fuzzy data, the only factor is the modification of the count kept when similar tuples are merged after generalization. Now, consider the relationship of the membership found in the fuzzy hierarchy concepts and the interval-valued fuzzy data. As shown in Figure 2, a data value may generalize to exactly one concept c_j in CH_C , implicitly with a membership degree of 1, $m(a, c_j) = 1$. In general, a data value can generalize to two or more concepts with associated membership values. Let us consider a specific data value a' which generalizes to two concepts, denoted c_j' and c_k' . There are then memberships $m(a', c_j')$ and $m(a', c_k')$, where $m(a', c_j') + m(a', c_k') = 1$. In particular, we must consider the relationship to the interval-valued fuzzy membership of the data being generalized and how this changes the interval after generalization. First, we analyze where the data value generalizes to the first one of the values in the hierarchy, i.e., the value a' generalizes to c_j' with membership $m(a', c_j')$.

$$t_1(a', I_1(a')) \rightarrow t_1(c_j', I_1^*(c_j')) \tag{31}$$

where $I_1^* = F(m(a', c_j'), I_1(a'))$.

How does this function, F , produce the new interval-valued membership I_1^* ? We must consider three possibilities.

First, if

$$m(a', c_j') \in I_1(a'), \text{ then } I_1^*(c_j') = I_1(a') \tag{32}$$

Otherwise, we use an application-derived threshold, $0 < h < 1$. Therefore, if $m(a', c_j')$ is within this threshold of the upper or lower bounds of the interval, we modify the interval by the membership. Therefore, if

$$ml(a') - m(a', c_j') < h, \text{ then } ml^*(a') = m(a', c_j')$$

or if

$$m(a', c_j') - mu(a') < h, \text{ then } mu^*(a') = m(a', c_j') \tag{33}$$

where we allow the respective interval bound to be modified.

Finally, if a' generalizes to multiple values in the hierarchy, we must use the above approach for multiple fuzzy membership values in the generalized tuples. However, since this is a generalization to different concepts, there are multiple tuples that cannot be merged. Recall that our objective in attribute generalization is to reduce the number of tuples in a given relation; therefore, in general, only one of these tuples should be maintained. We need to establish criteria to select which tuple to keep. We can proceed by comparing the intervals using our previous interval measure IW .

Let the multiple concepts be indexed $c_1 \dots c_n$. Then, for a' with $I_{k=1, n}(a')$ in the m tuples, we have the intervals $I^*_1(c_1'), \dots I^*_m(c_n')$, as above. Then, our criteria is

$$\text{Min}_{j=1}^n(IW(I * (c_j))) \tag{34}$$

Let us illustrate this with an example for a data value a' , $I(a') = [lb = 0.2, ub = 0.6]$, for two tuples where a' generalizes to two concepts, c_1 and c_2 . Let the memberships be

$$m(a', c_1) = 0.3; \quad m(a', c_2) = 0.7 \tag{35}$$

First, $m(a', c_1) \in I(a')$, so $I^*_1(c_1') = I(a')$. If we let the threshold $h = 0.15$, then we can extend the upper bound for c_2' , so $I^*_1(c_2') = [lb = 0.2, ub = 0.7]$.

$$IW(I^*_1(c_1')) = 0.4; \quad IW(I^*_1(c_2')) = 0.5 \tag{36}$$

Now, to compare, we use the interval width.

Then, with this evaluation, we retain the tuple $t(a', I^*_1(c_1'))$ as the generalization result. If the membership is not within the threshold, then the semantics of the relationship between a' and c_j' is such that the tuple with the data value a' cannot be generalized in a consistent manner.

5. Conclusions

We have developed a number of approaches for attribute generalization on interval-valued and intuitionistic fuzzy data. Simple hierarchies for which generalization is single-valued were first considered. The merging of tuples after generalization was developed both for crisp concept hierarchies and concept hierarchies with uncertainty.

There are a number of future topics to be developed, including other representations that can be used in data generalization, such as rough set theory and Dempster–Shafer theory. The more complex structures of these require extensions for generalization to be developed for these representations.

There are other uncertainty approaches that have been recently developed and that can be considered for further research. An extension to intuitionistic sets is called Pythagorean membership functions: PFS [36,37]. The key concept is extending the membership negation value $(1 - m_k)$ by introducing a Pythagorean negation

$$(\text{not}(m_k))^2 = 1 - m_k^2$$

Since these PFS membership functions allow extended values $(m_k + m^*_k > 1)$, the space of such memberships is then larger, i.e.,

$$\text{IFS} \subseteq \text{PFS}.$$

Therefore, Pythagorean membership functions can be used as extensions to intuitionistic fuzzy memberships in some applications (Saeed et al., 2022). Consider an application in which an analyst needs to develop an evaluation of data values with memberships such as $m_k(a_i) = 0.7$ and $m^*_k(a_i) = 0.5$. However, the restriction of IFS is exceeded as $0.7 + 0.5 = 1.2 > 1.0$. However, using PFS, the restriction is not violated, because by using the Pythagorean condition, we have

$$(0.7)^2 + (0.5)^2 = 0.49 + 0.25 = 0.74 < 1.0$$

This allows the analyst more freedom to make use of the significant positive and negative membership values found for the data value a_i . Therefore, such an analysis better reflects the actual assessments of the analyst.

Author Contributions: Conceptualization, F.P. and R.Y.; methodology, F.P. and R.Y.; validation, F.P. and R.Y.; formal analysis, F.P. and R.Y.; writing—original draft preparation, F.P. and R.Y.; writing—review and editing, F.P. and R.Y. All authors have read and agreed to the published version of the manuscript.

Funding: Fred Petry was supported by the Naval Research Laboratory Base Program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yager, R. On linguistic summaries of data. In *Knowledge Discovery in Databases*; Piatetsky-Shapiro, G., Frawley, W., Eds.; MIT Press: Cambridge, MA, USA, 1991; pp. 347–363.
2. Kacprzyk, J. Fuzzy logic for linguistic summarization of databases. In Proceedings of the FUZZ-IEEE'99, 1999 IEEE International Fuzzy Systems, Seoul, Korea, 22–25 August 1999; pp. 813–818.
3. Dubois, D.; Prade, H. Fuzzy sets in data summaries—outline of a new approach. In Proceedings of the 8th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU 2000), Madrid, Spain, 3 July 2000; pp. 1035–1040.
4. Miller, M.; Miller, K. Big data: New opportunities and new challenges. *IEEE Comput.* **2013**, *46*, 22–25.
5. Richards, D.; Rowe, W. Decision-making with heterogeneous sources of information. *Risk Anal.* **1999**, *19*, 69–81. [[CrossRef](#)]
6. Belcastro, L.; Cantini, R.; Marozzo, F.; Orsino, A.; Talia, D.; Trunfio, P. Programming big data analysis: Principles and solutions. *J. Big Data* **2022**, *9*, 1–50. [[CrossRef](#)]
7. Vranopoulos, G.; Clarke, N.; Atkinson, S. Addressing big data variety using an automated approach for data characterization. *J. Big Data* **2022**, *9*, 1–28. [[CrossRef](#)]
8. Hirota, K.; Pedrycz, W. Fuzzy computing for data mining. *Proc. IEEE* **1999**, *87*, 1575–1599. [[CrossRef](#)]
9. Laurent, A. A new approach for the generation of fuzzy summaries based on fuzzy multidimensional databases. *Intell. Data Anal.* **2003**, *7*, 155–177. [[CrossRef](#)]
10. Raschia, G.; Mouaddib, N. Mouaddib. SAINTETIQ: A fuzzy set-based approach to database summarization. *Fuzzy Sets Syst.* **2002**, *129*, 137–162. [[CrossRef](#)]
11. Yager, R.; Petry, F. A multi-criteria approach to data summarization using concept ontologies. *IEEE Trans Fuzzy Syst.* **2006**, *14*, 767–780. [[CrossRef](#)]
12. Petry, F.; Yager, R. Fuzzy concept hierarchies and evidence resolution. *IEEE Trans Fuzzy Syst.* **2014**, *22*, 1151–1161. [[CrossRef](#)]
13. Kruse, R.; Mostaghim, S.; Borgelt, C.; Braune, C.; Steinbrecher, M. *Computational Intelligence: A Methodological Introduction*, 3rd ed.; Springer Nature: Cham, Switzerland, 2022.
14. Zadeh, L. Fuzzy sets. *Inf. Control.* **1965**, *8*, 338–353. [[CrossRef](#)]
15. Klir, G.; St. Clair, U.; Yuan, B. *Fuzzy Set Theory: Foundations and Applications*; Prentice Hall: Hoboken, NJ, USA, 1997.
16. Moore, R. *Interval Analysis*; Prentice Hall: Englewood Cliffs, NJ, USA, 1966.
17. Moore, R.; Kearfott, B.; Cloud, M. *Introduction to Interval Analysis*; SIAM: Philadelphia, PA, USA, 2009.
18. Deschrijver, G. Arithmetic operators in interval-valued fuzzy set theory. *Inf. Sci.* **2007**, *177*, 2906–2924. [[CrossRef](#)]
19. Reza, F. *An Introduction to Information Theory*; McGraw Hill: New York, NY, USA, 1961.
20. Burillo, P.; Bustince, H. Entropy on intuitionistic fuzzy sets and on interval-valued fuzzy sets. *Fuzzy Sets Syst.* **1996**, *78*, 305–316. [[CrossRef](#)]
21. Atanassov, K. Intuitionistic fuzzy sets. *Fuzzy Sets Syst.* **1986**, *20*, 87–96. [[CrossRef](#)]
22. Mendel, J. Type-2 fuzzy sets and systems: An overview. *IEEE Comput. Intell. Mag.* **2007**, *2*, 20–29. [[CrossRef](#)]
23. Bustince, H.; Fernandez, J.; Hągras, H.; Herrera, F.; Pagola, M.; Barrenechea, E. Interval type-2 fuzzy sets are generalization of interval-valued fuzzy sets: Toward a wider view on their relationship. *IEEE Trans. Fuzzy Syst.* **2015**, *23*, 1876–1882.
24. Pawlak, Z. Rough sets and fuzzy sets. *Fuzzy Sets Syst.* **1985**, *17*, 99–102. [[CrossRef](#)]
25. Shafer, G. *A Mathematical Theory of Evidence*; Princeton University Press: Princeton, NJ, USA, 1976.
26. Carter, C.L.; Hamilton, H.J. Efficient attribute-oriented generalization for knowledge discovery from large databases. *IEEE Trans. Knowl. Data Eng.* **1998**, *10*, 193–208. [[CrossRef](#)]
27. Hilderman, R.J.; Hamilton, H.J.; Cercone, N. Data mining in large databases using domain generalization graphs. *J. Intell. Inf. Syst.* **1999**, *13*, 195–234. [[CrossRef](#)]
28. Han, J.; Kamber, M. *Data Mining: Concepts and Techniques*, 4th ed.; Morgan Kaufmann: San Francisco, CA, USA, 2022.
29. Buckles, B.; Petry, F. A fuzzy representation for relational data bases. *Int. J. Fuzzy Sets Syst.* **1982**, *7*, 213–226. [[CrossRef](#)]
30. Angryk, R.; Petry, F. Attribute-oriented Generalization in proximity and similarity-based relational database systems. *Int. J. Intell. Syst.* **2007**, *22*, 763–779. [[CrossRef](#)]

31. Beaubouef, T.; Buckles, B.; Petry, F. An attribute-oriented approach for knowledge discovery in rough relational databases. In Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference, Key West, FL, USA, 7–9 May 2007; pp. 507–509.
32. Bachman, R. What IS-A is and isn't: An analysis of the taxonomic links in semantic networks. *IEEE Comput.* **1983**, *16*, 30–36. [[CrossRef](#)]
33. Petry, F.; Zhao, L. Data mining by ttribute generalization with fuzzy hierarchies in fuzzy databases. *Fuzzy Sets Syst.* **2009**, *160*, 2206–2223. [[CrossRef](#)]
34. Ćirić, M.; Ignjatović, J.; Bogdanović, S. Fuzzy equivalence relations and their equivalence classes. *Int. J. Fuzzy Sets Syst.* **2007**, *158*, 1295–1313. [[CrossRef](#)]
35. Lee, D.H.; Kim, M.H. Database summarization using fuzzy ISA hierarchies. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **1997**, *27*, 68–78.
36. Yager, R. Pythagorean membership grades in multicriteria decision making. *IEEE Trans. Fuzzy Syst.* **2014**, *22*, 958–966. [[CrossRef](#)]
37. Li, H.; Cao, Y.; Su, L. Pythagorean fuzzy multi-criteria decision-making approach based on Spearman rank correlation coefficient. *Soft Comput.* **2022**, *26*, 3001–3012. [[CrossRef](#)]