An Uncertainty Representation for Set-Valued Attributes with Hierarchical Domains

Frank Rügheimer
University of Magdeburg
ruegheim@iws.cs.uni-magdeburg.de

Rudolf Kruse
University of Magdeburg
kruse@iws.cs.uni-magdeburg.de

Abstract

In collected data information about a single property may be presented with variable resolution and focus. The present paper describes how hierarchically structured attribute domains support the transfer of knowledge between alternative frames of discernment allowing to flexibly serve information needs and facilitate the processing of inhomogeneous data. The approach is later extended to accommodate set-valued attributes, which have previously been employed to represent imprecision and have recently gained attention in text processing, hierarchy learning or multi-label classification.

Keywords: Information Fusion, Random Sets, Knowledge Representation

1 Introduction

One of the major steps in solving classification and prediction tasks consists in the analysis and representation of interaction patterns between attributes. To distinguish genuine and reproducible relationships from random variations such statistical analyses rely on a sufficient number of sample cases. For discrete distribution, what constitutes a "sufficient number" is considerably influenced by the cardinality of the underlying sample space. Consequently, to ensure a minimum number of cases per instance, a lower granularity sample space may be preferred over a finer one. In other cases samples are supplied by an inhomogeneous collection of information sources that provide observations on different levels of detail. Hierarchically structured attribute domains provide a robust and interpretable approach to dealing with such problems.

Using probability trees as a starting point, the present paper investigates data representation with hierarchical attributes. After recapitulating the concept, frames of discernment and their relation to structured attribute domains are discussed. In section 2 we suggests a method to efficiently represent practically relevant sets of frames and manage their interaction. That approach is extended in section 4 to account for set-valued attributes, which, for instance may reflect multi-label descriptions, sets of alternatives or imprecise data. The suggested information-compressed representation differs from the more general random-set approaches in providing a scalable solution when a large number of focal sets is admitted.

2 Frames of Discernment

Let O be a universe of objects and $\Lambda = \{\lambda_1, \ldots \lambda_n\}$ a finite set of labels used as attribute values for characterizing individual objects w. r. t. a certain property. Ideally, that property can be described for each object $o \in O$ by associating exactly one of the attribute values from Λ. An attribute $\mathcal{A}$ is then identified with a function $\mathcal{A} : O \to \Lambda$

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 197–203
Torremolinos (Málaga), June 22–27, 2008

that assigns the correct[1] descriptive label to each of the considered objects.[2].

The above definition of an attribute assumes the existence of a generally accepted set of mutually exclusive attribute values that is suitable for recording, as well as processing and presenting the information. Yet, in certain situations it may be favorable to use several complementing views to represent the available information. Such situations may be marked by

- information sources that differ w. r. t. observation capabilities,
- information needs that differ between the individual users or
- specific requirements of processing steps, e.g., w. r. t. sample size.

For instance, leaving out detail that is irrelevant to a particular user may actually contribute to a better understanding of relevant pieces of information. Applying the idea of alternative frames, an attribute must be reimagined as a collection of frame-specific mappings that assign labels to objects; but in contrast to the more general interaction between attributes, label assignments for frames referring the same attribute should closely correspond to each other. Distractive distinctions in a knowledge representation, for instance, are suppressed, by applying a surjection from the given set of lables onto a set of less specific one. This reduction of detail reflects a conversion to a coarser, i.e. less informative, *frame of discernment* [8].

The existence of such a mapping also defines a partial ordering on the set of frames. However, that idea is more commonly expressed using the notion of a refinement.

Definition 1 (Shafer 1976) *A set Λ' is a refinement of Λ if there is a mapping* $\mathrm{ref} : 2^{\Lambda} \rightarrow 2^{\Lambda'}$ *such that* $\forall \lambda_1, \lambda_2 \in \Lambda$:

1. $\forall \lambda \in \Lambda : \mathrm{ref}(\{\lambda\}) \neq \emptyset$
2. $(\lambda_1 \neq \lambda_2) \Rightarrow \mathrm{ref}(\{\lambda_1\}) \cap \mathrm{ref}(\{\lambda_2\}) = \emptyset$
3. $\bigcup \{\mathrm{ref}(\{\lambda\}) \mid \lambda \in \Lambda\} = \Lambda'$
4. $\mathrm{ref}(A) = \bigcup \{\mathrm{ref}(\{\lambda\}) \mid \lambda \in A\}$

Condition (i) ensures that any label in the original frame is still represented by at least one label in the refined frame, whereas condition (ii) guarantees the preservation of existing distinctions. Conditions (iii) and (iv) ensure correspondence of the considered attribute domains and provide a set extension for mapping operations, respectively. A set A is called a coarsening of a set B if there is a refinement ref, such that $\mathrm{ref}(A) = B$. The refinement relation structures the set of reference frames as a lattice.

Although there are usually several meaningful ways to subdivide the equivalence class associated with a label during refinement, a hierarchical organization of the attribute domain is advantageous in that it permits to easily find, summarize or arrange objects and information. Libraries, for instance, are organized according to a fixed hierarchy of topics, even though several equally suited taxonomies may be conceivable. The selected hierarchy generates a family of related frames, which differ only with regard to the level of detail employed for corresponding subframes. The set of a label λ's direct children in the hierarchy H is called its *direct refinement* w.r.t. H (written $\mathrm{children}_H(\lambda)$). In extension of that, the set of all descendants, including indirect ones, of a label λ in H is denoted by $\mathrm{desc}_H(\lambda)$. Similarly, the functions $\mathrm{parent}_H(\lambda)$ and $\mathrm{anc}_H(\lambda) = \{\lambda' | \lambda \in \mathrm{desc}_H(\lambda')\}$ are defined, which permit to link labels with their parents and ancestors in the label hierarchy[3].

A simple label hierarchy is shown in Figure 1. The attribute value hierarchy reflects a sub-

[1] The definition does not require that function to be known.

[2] For a number of problems a more servicable description is usually achieved by using a number of attributes that reflect properties relevant to the current information needs. The interaction of such attributes is addressed in a different publication [7]

[3] To enforce that the parent is defined for all admissible labels, an auxiliary root symbol is introduced in the label hierarchy.

division of the classes labeled a_1 and a_3 with

$$\begin{aligned}\mathrm{desc}_H(a_1) &= \mathrm{children}_H(a_1) = \{a_{11}, a_{12}, a_{13}\},\\ \mathrm{desc}_H(a_3) &= \mathrm{children}_H(a_3) = \{a_{31}, a_{32}\}.\end{aligned}$$

Starting from the coarsest frame $\{a_1, a_2, a_3\}$ (repeatedly) replacing labels with their direct refinements H produces three new frames of discernment $\{a_{11}, a_{12}, a_{13}, a_2, a_3\}$, $\{a_{11}, a_{12}, a_{13}, a_2, a_{31}, a_{32}\}$, $\{a_1, a_2, a_{31}, a_{32}\}$.

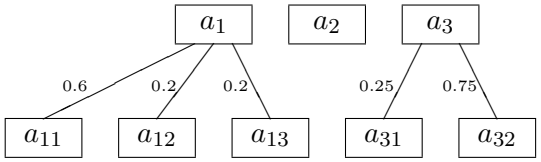

Figure 1: Attribute value hierarchy with attached conditional probabilities for sublabels

3 Modeling Uncertainty with Hierarchical Attribute Domains

Having discussed the formalization of attributes and their domains, let us now consider the representation of uncertain knowledge on a family of frames. Uncertainty arises, for instance, from limited observation capabilites potentially resulting in wrongly assigned lables. Quantitative modelings for this type of uncertainty are usually based on assumptions about the measuring process or on estimates from representative reference data. They are formally expressed using a family of conditional probability distributions $P_\Lambda(\mathcal{A}_\Lambda = \lambda_i \mid \mathcal{O}_\Lambda = \lambda_j)$, which statistically relate the unknown matching labels on a given frame Λ with observed ones. In combination with observed instantiations, this determines a distribution and the associated measure $P_\Lambda : \Lambda \to [0, 1]$ over Λ.

When switching between alternative frames of discernment information may also be lost due to non-corresponding lables. To alleviate the effects of that problem a model can again be supplemented with generic information about the statistical interaction between data representations on pairs of frames. Such interaction patterns are expressed, for instance, using conditional distributions.

A general obstacle to this approach is the amount of storage required to encode frame interaction. Fortunately if the admissible frames are restricted to those generated from a single hierarchically structured attribute domain, it suffices to store branching probabilities for each of the child-lables. For an object that is correctly described by a label λ, the branch probabilities $P_H(\lambda_i \mid \lambda)$ quantify the uncertainty w.r.t. which of the sublabels $\lambda_i \in \mathrm{children}_H(\lambda)$ provides the matching description on a frame, where label λ is expanded. Presuming the sub-label distributions are conditionally independent given the given the parent labels, P_H may be applied for all descendants of λ as the value of the measure is obtained by multiplying branch probabilities along a path of serial refinements. The probability tree representation, which is illustrated in Figure 1, thus supplements the information required for converting distributions to frames with locally higher resolution. The suggested approach has the additional advantage, that the uncertainty component introduced due to frame conversion is contained locally. As a result of the imposed restrictions only three cases have to be considered when mapping an element λ_1 from a frame Λ_1 to a frame Λ_2 generated by the same hierarchy H of attribute values:

- λ_1 is an element of Λ_2 as well,
- λ_1 summarizes a subframe $A \subseteq \Lambda_2$ consisting only of its (possibly indirect) descendents in the hierarchy
- λ_1 is itself part of a subframe associated with a unique element of Λ_2.

Neither of the frames is marked out so Λ_1 and Λ_2 can be interchanged in that statement (Figure 2). Applied to the conversion of probability distribution between frames of the same family we obtain:

Definition 2 *Let Λ_1 and Λ_2 be two frames of discernment generated from the same hierarchy H and P_{Λ_1} a probability function over Λ_1. The mapping $T_{\Lambda_1 \to \Lambda_2} : \mathrm{Prob}(\Lambda_1) \to \mathrm{Prob}(\Lambda_2)$ that converts a probability distribution from*

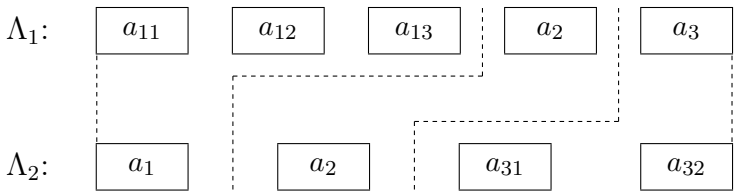

Figure 2: Correspondence of subframes and single labels

frame Λ_1 *to a probability distribution over* Λ_2 *is computed as:*

$$P_{\Lambda_2}(\lambda_2) = T_{\Lambda_1 \to \Lambda_2}(\lambda_2) = \begin{cases} P_{\Lambda_1}(\lambda_2) \\ \quad \textit{if } \lambda_2 \in \Lambda_1, \\ \sum\limits_{\lambda \in \mathrm{desc}_H(\lambda_2) \cap \Lambda_1} P_{\Lambda_1}(\lambda) \\ \quad \textit{if } \mathrm{desc}_H(\lambda_2) \cap \Lambda_1 \neq \emptyset, \\ P_H(\lambda_2 \mid \lambda) \cdot P_{\Lambda_1}(\lambda) \\ \quad \textit{if } \exists \lambda \in \Lambda_1 : \lambda_2 \in \mathrm{desc}_H(\lambda), \end{cases} \quad (1)$$

where $\mathrm{Prob}(\Lambda)$ *denotes the set of all possible probability functions on a frame* Λ*.*

Note, that the probability assigned to a given label does not depend on the composition of the frame Λ_2 under consideration. Thus, equation 1 may be used to condition the probabilities for all labels in the hierarchy with new information from observations on a specific frame.

As an example, consider a conversion from Λ_1 to Λ_2 as given by Figure 2 with an original distribution P_{Λ_1} on Λ_1; $P_{\Lambda_1}(a_{11}) = P_{\Lambda_1}(a_2) = 0.2$, $P_{\Lambda_1}(a_{12}) = P_{\Lambda_1}(a_{13}) = 0.1$ and $P_{\Lambda_1}(a_3) = 0.4$. Equation 1 permits to compute $P_{\Lambda_2} = T_{\Lambda_1 \to \Lambda_2}(P_{\Lambda_1})$. The subframe $\{a_{11}, a_{12}, a_{13}\} = A \subseteq \Lambda_1$ is represented in Λ_2 by the single attribute value a_1. The probability originally assigned to the elements of A is now attributed to a_1, i.e $P_{\Lambda_2}(a_1) = P_{\Lambda_1}(a_{11}) + P_{\Lambda_1}(a_{12}) + P_{\Lambda_1}(a_{13}) = 0.4$. Label a_2 appears in both frames, so $P_{\Lambda_2}(a_2) = P_{\Lambda_1}(a_2)$. The two remaining probabilities are computed using the estimated sub-label distribution, i.e., $P_{\Lambda_2}(a_{31}) = P_H(a_{31} \mid a_3) \cdot P_{\Lambda_1}(a_3) = 0.1$ and $P_{\Lambda_2}(a_{32}) = P_H(a_{32} \mid a_3) \cdot P_{\Lambda_1}(a_3) = 0.3$, which fully determines the probability function P_{Λ_2}.

4 Extension to Set-Valued Data

So far it was assumed that all objects in O could be described using no more than one label per object. However, given that often only a subset of those objects would actually have been observed by the time the attribute hierarchy is chosen, that idealization may turn out too optimistic. Additionally if composite objects are considered (e.g. texts), a single label per attribute may not provide the best possible specification. This means, that the attribute $\mathcal{A}$ is actually set-valued with mappings ($\mathcal{A}^*_\Lambda : O \to 2^\Lambda \setminus \{\emptyset\}$, where 2^Λ denotes the power set of Λ.) Unfortunately, as labels are not mutually exclusive, uncertainty may no longer be represented by a probability distribution over the frame.

Several noteworthy approaches to dealing with uncertainty w.r.t. set-valued entities may be expressed in terms of *random sets* [6], that is, set-valued random variables. A very similar concept has previously been used by Dempster [2], who investigated upper and lower probabilities induced by set-valued mappings from a probability space and focused on sets as means to express imperfect knowledge about the distribution of pieces of probability mass. Under a subjectivist interpretation of probability that representation also gives rise to the mass distributions of Shafer's *theory of evidence* [8, 5].

The introduction of random sets formally reduces the problem of uncertainty representation for set-valued attributes to the probabilistic case, the only difference being that the distributions are defined on the power set of the frames instead of the frames themselves. Yet the reduction is only achieved at the cost of an increased cardinality of the sample space, which, unless the admissible set-

valued outcomes are restricted by favorable conditions, would render an approach implementing that strategy very resource-intensive. Moreover, as argued before, the larger sample space is disadvantageous for estimating distributions from data. Still the approach provides a reference modeling with interpretable aggregation operators, by which information may be converted between frames.

Possibility distributions in the sense of [3] provide a compact representation of uncertainty and imprecision (the latter being a special interpretation of set-valuedness) on a given reference frame, but rely on consonant, that is nested, focal sets. Unfortunately that assumption cannot be justified for the more general setting at hand. The consonance requirement is overcome by the context model interpretation given in [4], though at the expense of reduced representational power and the lack of a consistent aggregation operator [1]. Without a meaningful aggregation operator, the context model cannot support the conversions required for a frame-spanning representation. Nevertheless it contributes to the solution of the problem at hand in suggesting one-point-coverages i.e., the combined probability assigned to the sets that contain a given element of the reference frame, as information summaries. One-point-coverages can be interpreted as the probability of a particular element $\lambda \in \Lambda$ being among the acceptable labels for an object. Conversely in Dempster's framework [2] or under an interpretation of set-valuedness as imprecision, the one-point-coverage corresponds to an upper probability bound.

Like with the information compressed approaches discussed above, we suggest a representation that only aims at preserving properties of the distribution that are relevant in the uncertainty interpretation. In particular, we selected the following pieces of information that should be recoverable for each element λ in H from an extended version of the data structure presented in section 3:

- The estimated probability for the singleton $\{\lambda\}$ i.e., the probability that λ is the *only* correct label;
- The one-point-coverage of λ i.e., the probability that λ is *among* the accepted labels.

One may also be interested in the most specific single label expected to summarize the true class of an object $o \in O$ with probability of at least p, where $p \in (0.5, 1]$. That goal can be reduced to the first one, because finding an adequate summary amounts to searching the hierarchy of labels with their respective assigned probabilities.

To meet the listed objectives a detailed distribution over the power set of the original frame is not required. Instead the suggested representations uses a distribution over a coarser sample space that does not distinguish between multi-valued attribute instantiations. Formally this sample space is reflected as an extended frame

$$\Lambda' = \mathrm{Ext}(\Lambda) = \Lambda \cup \{\lambda_s\} \tag{2}$$

associated with each frame Λ. The new symbol λ_s is used to denote any multi-valued outcome. The distribution defined w.r.t. that frame, is induced by the set-valued observations from 2^Λ. For an extended frame $\Lambda' = \mathrm{Ext}(\Lambda)$, the probabilities assigned to the elements of the extended frame given uncertain set-valued descriptions from 2^Λ are

$$P'_\Lambda(\lambda) = \begin{cases} P_\Lambda(\mathcal{A}^*_\Lambda = \{\lambda\}) & \text{if } \lambda \neq \lambda_s \\ P_\Lambda(\mathcal{A}^*_\Lambda \in A_s & \text{if } \lambda = \lambda_s, \end{cases} \tag{3}$$

where $A_s = \{A \mid A \in 2^\Lambda \wedge |A| > 1\}$ represents the set of multi-valued outcomes and $\lambda \in \mathrm{Ext}(\Lambda)$. With probabilities still assigned to disjoint groups of (set-valued) outcomes, aggregation is based on addition. The resulting distribution over Λ' directly supplies probabilities assigned to the singleton values of $\mathcal{A}^*_\Lambda$, yet to restore the one-point-coverages $\mathrm{opc}(\lambda)$ additional parameters are required. In order to identify those parameters, we rewrite the one-point-coverage as $\forall(\lambda) \in \Lambda$:

$$\mathrm{opc}(\lambda) = \sum_{\{A \mid A \in 2^\Lambda \wedge \lambda \in A\}} P_\Lambda(\mathcal{A}^*_\Lambda = A)$$

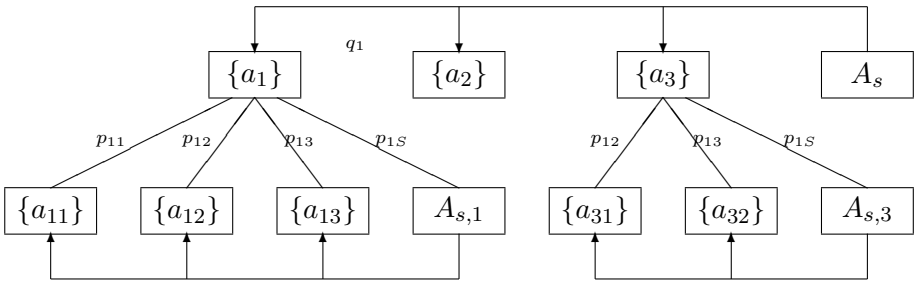

Figure 3: Attribute value hierarchy extended to accommodate one-point-coverages

$$\begin{aligned} &= P_\Lambda(\mathcal{A}^*_\Lambda = \{\lambda\}) \\ &+ \sum_{\{B|B\in A_s \wedge \lambda\in B\}} P_\Lambda(\mathcal{A}^*_\Lambda = B) \\ &= P'_\Lambda(\lambda) + q_\lambda \cdot P'_\Lambda(\lambda_s). \end{aligned} \tag{4}$$

The factor q_λ denotes the fraction of set-valued outcomes which contain λ among those represented by λ_s. For positive values of $P'_\Lambda(\lambda_s)$ it is defined as the proportion

$$q_\lambda = \frac{P_\Lambda(\lambda_i \in \mathcal{A}^*_\Lambda \wedge \mathcal{A}^*_\Lambda \in A_s)}{P_\Lambda(A^*_\Lambda \in A_s)}.$$

Like the (conditional) probability functions complemented by them, the values q_λ can be determined empirically. With hierarchical attributes, the above representation is applied for all subframes that result from direct refinements[4]. Together, these adaptations result in a modified attribute value hierarchy, which is illustrated in Figure 3.

Because the labels in each direct refinement still denote disjoint events, equation 1 may be applied to lables of the extended frame hierarchy as well. Concerning the calculation of one-point-coverages it is useful to recapitulate the structure of a frame in that modified hierarchy. Starting from the set $\{\rho\}$ that only contains the root label, each refinement step substitutes a label with its direct refinement and adds an auxiliary label for the summarized multi-valued elements. Thus, for any frame that arose from a series of refinement operations, the one-point-coverage of a label $\lambda \in H$ includes partial contributions from the higher levels of the label hierarchy. With the initial constraint $\mathrm{opc}_H(\rho) = P'_H(\rho) = 1$, the generic one-point-coverage opc_H for the labels can be computed using the recursion formula given in equation 5. In that equation $\mathrm{m}(\lambda)$ denotes the special label for the multi-valued cases within the direct refinement of λ's parent label, which was introduced along with λ and its siblings in a refinement operation, whereas q_λ is the associated coverage factor with respect to $\mathrm{m}(\lambda)$.

$$\begin{aligned} &\mathrm{opc}_H(\lambda) \\ &\quad = \mathrm{opc}_H(\mathrm{parent}_H(\lambda)) \\ &\quad \cdot (P'_H(\lambda \mid \mathrm{parent}_H(\lambda)) \\ &\qquad + P'_H(\mathrm{m}(\lambda) \mid \mathrm{parent}_H(\lambda)) \cdot q_\lambda) \end{aligned} \tag{5}$$

If the goal is to convert case-specific information on one-point-coverages and probabilities for certain lables, the observed values have precedence over the generic ones, and the recursion is broken early. To efficiently compute one-point-coverages for several elements of a frame, the implementation reuses partial results. Due to shared ancestors in the hierarchy, the recursion may then be stopped early.

We remark, that equation 5 assumes the local distribution within direct refinements to be invariant w.r.t. single or set-valued instantiations on coarser frames. Depending on the interpretation of set-valuedness, this assumption may require justification. It can be avoided though, by introducing separate sets of conditional probabilities.

5 Summary

The hierarchically structured attribute domain permits to fuse information from sources

[4]For that purpose the coarsest frame is considered a direct refinement of an abstract root label ρ.

that differ w.r.t. resolution, reliability and focus. Uncertain knowledge is described using probability distributions on a group of frames of discernment that were generated using a common attribute value hierarchy. We provided operations to map distributions between frames allowing comparisons or the integration of information from different sources. The suggested operations combine apriori knowledge on the general distribution of sub-labels with case specific information from observations w.r.t. to specific frames, and can be used to support decisions when only partial information is available.

The compressed representation of uncertain set-valued information introduced in section 4 is related to the more general framework of random sets but trades some representation capabilities in favor of storage efficiency. We argue that the reduction in representational power is acceptable when models have to be induced from to data as the detailed interaction structure potentially available with a full representation would often be masked by sampling effects.

The proposed ideas have been implemented in a C library and are currently applied in the development of measures for hierarchy learning from text data.

References

[1] Christian Borgelt and Rudolf Kruse. *Graphical Models—Methods for Data Analysis and Mining*. J. Wiley & Sons, Chichester, 2002.

[2] Arthur P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, 38:325–339, 1967.

[3] Didier Dubois and Henri Prade. *Possibility Theory*. Plenum Press, New York, New York, 1988. Translation of: Théorie des possibilités.

[4] Jörg Gebhardt and Rudolf Kruse. The context model – an integrating view of vagueness and uncertainty. *International Journal of Approximate Reasoning*, (9):283–314, 1993.

[5] Rudolf Kruse, Detlef Nauck, and Frank Klawonn. Reasoning with mass distributions. In B. D. D'Ambrosio, P. Smets, and P. P. Bonissone, editors, *Uncertainty in Artificial Intelligence*, San Mateo, California, 1991. Morgan Kaufmann.

[6] Hung T. Nguyen. On random sets and belief functions. *Journal Math. Anal. Appl.*, 65:531–542, 1978.

[7] Frank Rügheimer. A condensed representation for distributions over set-valued attributes. In *Proc. 17. Workshop Computational Intelligence*, Karlsruhe, Germany, 2007. Universitätsverlag Karlsruhe.

[8] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.