An Instantiation of Hierarchical Distance-based Conceptual Clustering for Propositional Learning*

A. Funes[1,2], C. Ferri[1], J. Hernández-Orallo[1], M. J. Ramírez-Quintana[1],

[1] DSIC, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, España
[2] Universidad Nacional de San Luis, Ejército de los Andes 950, 5700 San Luis, Argentina
{afunes, cferri, jorallo, mramirez}@dsic.upv.es

Abstract. In this work we analyse the relationship between distance and generalisation operators for real numbers, nominal data and tuples in the context of hierarchical distance-based conceptual clustering (HDCC). HDCC is a general approach to conceptual clustering that extends the traditional algorithm for hierarchical clustering by producing conceptual generalisations of the discovered clusters. This makes it possible to combine the flexibility of changing distances for several clustering problems and the advantage of having concepts which are crucial for tasks as summarisation and descriptive data mining in general. In this work we propose a set of generalisation operators and distances for the data types mentioned before and we analyse the properties by them satisfied on the basis of three different levels of agreement between the clustering hierarchy obtained from the linkage distance and the hierarchy obtained by using generalisation operators.

Keywords: conceptual clustering, hierarchical clustering, generalisation, distances, propositional learning.

1 Introduction

One issue related to some data mining techniques is the lack of comprehensibility. Although several learning techniques have been tested as useful in the way that they offer good predictions, they do not give a description, pattern or generalisation which justifies the decision made for a given individual. For instance, it is useful to know that a given molecule belongs to a cluster according to a certain distance measure, but it is even more interesting to know what the chemical properties shared by all the molecules in that cluster are.

Lack of comprehensibility is a common issue to clustering and classification techniques based on distances. The source of this problem is the dichotomy between distances and generalisations. It is well known that distances and generalisations give

* This work has been partially supported by the EU (FEDER) and the Spanish MEC/MICINN under grant TIN2007-68093-C02 and the Spanish project "Agreement Technologies" (Consolider Ingenio CSD2007-00022). A. Funes was supported by a grant from the Alfa Lernet project and the UNSL.

rise to two different approaches in data mining and machine learning. On the one hand we have distance-based techniques, where we only need to count on a distance function for the data we are working with. However, distance-based techniques (such as [11, 12, 13]) do not provide patterns or explanations justifying the decisions made. On the other hand we have symbolic techniques [7, 8, 9, 10] that, unlike distance-based methods, are founded on the idea that a generalisation or pattern discovered from old data can be used to describe new data covered by this pattern.

An important issue when combining both techniques is to know whether the patterns discovered for each cluster by a distance-based technique are consistent with the underlying distance used to construct the clusters. Inconsistencies can arise when the notion of distance and generalisation are considered independently. That is, given a set of examples and a generalisation of them, it is expected that those examples that are close in a metric space according to its distance are covered by the generalisation, while those examples that are far away are expected to be outside the generalisation coverage. This problem has been extensively treated in [6]. In the present work we focus on the relationship between distances and generalisations in the context of HDCC [1], a general approach for agglomerative hierarchical clustering [2, 3]. HDCC, that stands for Hierarchical Distance-based Conceptual Clustering, constructs a cluster hierarchy by using a distance at the same time that in produces a hierarchy of patterns resulting in an extended dendrogram referred as conceptual dendrogram. The main aspect considered in [1] and that has been ignored by other conceptual clustering methods that use distances is knowing a priori whether the hierarchy of clusters induced by the underlying distance is consistent with the discovered patterns, i.e. how much the cluster elements covered by a given pattern reproduce the distribution of the elements in the metric space. Accordingly, in [1] three different levels of consistency between a distance and a generalisation operator have been defined.

The present work is an instantiation for the propositional learning case of the general framework presented in [1]. Here, we give the results of a formal analysis carried out for a set of distances and generalisation operators useful for propositional clustering, where we prove that intervals and absolute difference distance for real numbers, and the union set and discrete distance for nominal data work well together in HDCC. More importantly, we have also shown that it is also the case when using them as generalisation operators and distances for tuples of real numbers and nominal data. This rounds up the approach for propositional learning. But, additionally, this composability result for tuples is obtained independently from the base data types. The property of composability allows our framework to be directly extended to tuples of any complex data type provided that the generalisation operators associated to the component data types satisfy the property wanted for tuples. For instance, we can assert properties of tuples of graphs, strings and numbers provided we know the properties for the underlying data types. Besides these theoretical results we also present some experiments.

The paper is organised as follows. Due to space limitations, all necessary preliminary concepts about the HDCC approach can be found in [1] and the proposition proofs can be found in [14]. In Section 2 we propose pairs of generalisation operators and distances for numerical and nominal data, which are used in turn to define generalisation operators and distances for tuples. In Section 3 we present some experiments by applying the operators and distances proposed in

Section 2, and we also compare the results obtained in HDCC wrt. traditional hierarchical clustering. Finally, Section 4 closes the paper with the conclusions and future work.

2 Instantiation for Propositional Learning

In this section, we present an instantiation of HDCC for propositional clustering where flat data are expressed in terms of attributes and instances. We propose generalisation operators for numerical and categorical data and also for tuples, which are the data types typically used in propositional learning. In all cases the different levels of consistency defined in [1] between the proposed operators and distances have been verified through a satisfability analysis of the strong and weak boundedness and acceptability properties given in [1].

2.1 Nominal Data

A nominal data type, also referred as enumeration or categorical data type denotes a finite set of possible values that an attribute can take, e.g. gender, days of the week, colours, etc. A Boolean data type is a special case where there are only two possibilities. The metric space for nominal data type is composed of a set X, which is just a finite set of symbolic values, and a distance d.

There are many distances defined for nominal values. Some of the most commonly used distances are the discrete distance –that returns 0 when both values match and 1 otherwise– and the VDM (Value Difference Metric) distance [4], among others. In some cases, a distance defined by the user can be useful. For instance, in the metric space (X, d) where X = {XXL, XL, L, M, S, XS, XXS}, the distance d defined as d(XXL, XL) = 1, d(XXL, L) = 2, d(XXL, M) = 3, d(XXL, S) = 4, d(XXL, XS) = 5, d(XXL, XXS) = 6, d(XL, L) = 1, d(XL, M) = 2, d(XL, S) = 3, d(XL, XS) = 4, d(XL, XXS) = 5, d(L, M) = 1, d(L, S) = 2, d(L, XS) = 3, d(L, XXS) = 4, d(M, S) = 1, d(M, XS) = 2, d(M, XXS) = 3, d(S, XS) = 1, d(S, XXS) = 2, d(XS, XXS) = 1 organizes the points into a line where XXL and XXS are the extreme points.

Typical patterns for nominal data are expressed as conditions over the values of the attributes, e.g. *attributeName* = XL or *attributeName* ≠ XL. However, since X is finite, the coverages[1] of the possible patterns are also finite and they can be expressed extensionally as subsets of X. Thus, the pattern language $\mathcal{L}$ for nominal data can reduce to 2^X.

We propose for the generalisation of a pair of nominal values the set that contains both values.

Proposition 1. *Let* (X, d) *be a metric space,* X *a set of nominal data, and* 2^X *the pattern language. The function* $\Delta: X \times X \to 2^X$ *defined by* $\Delta(e_1, e_2) = \{e_1, e_2\}$ *is a binary generalisation operator*[2] *for nominal data.*

[1] See [1] for definition.of *coverage*.
[2] See [1] for definition.of *binary generalisation operator*.

Given that patterns are sets of nominal values, we propose the set union as the generalisation of two patterns.

Proposition 2. *Let* (X, d) *be a metric space,* X *a set of nominal data, and* 2^X *the pattern language. The function* $\Delta^*: 2^X \times 2^X \to 2^X$ *defined by* $\Delta^*(s_1, s_2) = s_1 \cup s_2$ *is a pattern binary generalisation operator*[3] *wrt.* 2^X.

Proposition 3 gives the properties satisfied by the proposed operators Δ and Δ^*.

Proposition 3. *Let* Δ *and* Δ^* *be the generalisation operators given in Proposition 1 and Proposition 2, d a distance between nominal data and* d_L *a linkage distance.* Δ *and* Δ^* *are* (*i*) *strongly bounded*[4] *by d and* d_L*, respectively;* (*ii*) *weakly bounded*[5] *by d and* d_L*, respectively;* (*iii*) *acceptable*[6].

The example in Figure 1 (left) shows the use of HDCC for the evidence E = {XXS, S, M, XXL}. We have used the discrete distance, and the generalisation operators given in Proposition 1 and Proposition 2 to compute the patterns. Note that applying the user-defined distance given above, the dendrogram changes to that shown in Figure 1 (right).We can also affirm by Proposition 1 in [1] and Proposition 3 that both conceptual dendrograms are equivalent to the corresponding traditional dendrograms.

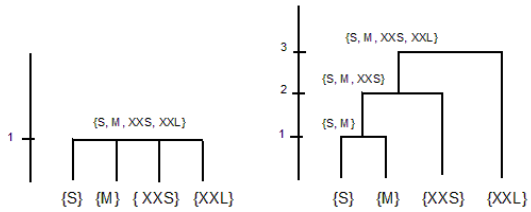

Figure 1. Two applications of HDCC to nominal data under the single linkage distance, using the discrete distance (left) and a user-defined distance (right).

3.1 Numerical Data

Numerical data are widely used to express amounts and measures and many attributes of real word objects. A well known metric space for numeric data is $(\Re, d)$ where d is the distance defined as the absolute difference of two real numbers, i.e. $d(e_1, e_2) = |e_1 - e_2|$. A usual generalisation for a set of numbers is the minimal interval whose extreme values are the least and the greatest values in the set. Thus the pattern language $\mathcal{L}$ we consider here is the set of all the finite closed intervals in $\Re$. We propose for the generalisation in $\mathcal{L}$ of two elements in $\Re$ the minimal interval that includes both elements.

Proposition 4. *Let* $\mathcal{L}$ *be the set of all the finite closed intervals in* $\Re$. *For all* e_1, e_2 *in* $\Re$ *such that* $e_1 \leq e_2$, *the function* $\Delta: \Re \times \Re \to \mathcal{L}$ *defined by* $\Delta(e_1, e_2) = [e_1, e_2]$ *is a binary generalisation operator for real numbers.*

---

[3] See [1] for definition.of *pattern binary generalisation operator*.
[4] See [1] for definition of *strongly bounded.*
[5] See [1] for definition of *weakly bounded.*
[6] See [1] for definition of *acceptable.*

Next we propose for the generalisation of two intervals the minimal interval that covers both.

Proposition 5. *Let $\mathcal{L}$ be the set of all the finite closed intervals in $\Re$. The function Δ^*: $\mathcal{L} \times \mathcal{L} \rightarrow \mathcal{L}$ defined by $\Delta^*([e_{i1}, e_{f1}], [e_{i2}, e_{f2}]) = [e_i, e_f]$, where e_i is the least value in $\{e_{i1}, e_{i2}\}$ and e_f is the greater value in $\{e_{f1}, e_{f2}\}$, is a pattern binary generalisation operator wrt. $\mathcal{L}$.*

Proposition 6. *Let Δ and Δ^* be the generalisation operators of Proposition 4 and Proposition 5, d the absolute difference between numbers and d_L a linkage distance. Δ and Δ^* are (i) strongly bounded by d and d_L, respectively; (ii) weakly bounded by d and d_L, respectively; (iii) acceptable.*

Figure 2 shows a simple application of HDCC under single linkage using the proposed operators and distance for real numbers. By Proposition 1 in [1] and by Proposition 6 the conceptual dendrogram is equivalent to the traditional one.

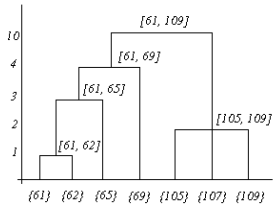

Figure 2. Conceptual dendrogram using single linkage distance for a set of real numbers.

3.2 Tuples

A tuple is a widely-used structure for knowledge representation in propositional learning since examples are represented as tuples of nominal and numerical data.To define a generalisation operator for tuples, unlike the previous data types, we base it on the properties of the basic types from which the tuple type is constructed. We assume they are embedded in metric spaces, therefore we can use the distances defined over each space to define distances between tuples. Analogously, to define the pattern language for tuples, we also use the pattern languages defined for each space.

Let (X_i, d_i) be a collection of metric spaces and $\mathcal{L}_i$ a collection of pattern languages ($i=1,\ldots,n$) corresponding to each of the n dimensions of a tuple. We denote X the space $X_1 \times \ldots \times X_n$. Therefore, if $x \in X$ then x is a n-tuple $(x_1,\ldots, x_n)$, where $x_i \in X_i$.

Let $d_i(\cdot,\cdot)$ be a distance function defined over X_i ($i=1,\ldots,n$). The expressions shown in Table 1 are distance functions in X. In that follows, we denote as d_T any of them.

Table 1. Some distance functions for tuples.

$d(x,y) = \sum_{i=1}^{n} d_i(x_i, y_i)$ *Manhattan distance*	$d(x,y) = \sqrt{\sum_{i=1}^{n} d_i(x_i, y_i)^2}$ *Euclidean distance*	$d(x,y) = \max_{1 \le i \le n} d_i(x_i, y_i)$ *Box or Chebyshev distance*
$d(x,y) = \sum_{i=1}^{n} \alpha_i . d_i(x_i, y_i)$ *Weighted Manhattan*	$d(x,y) = \sqrt{\sum_{i=1}^{n} \alpha_i . d_i(x_i, y_i)^2}$ *Weighted Euclidean distance*	$d(x,y) = \max_{1 \le i \le n} \alpha_i . d_i(x_i, y_i)$ *Weighted Box distance*

We define the pattern language for tuples $\mathcal{L}$ by using the basic pattern languages $\mathcal{L}_i$ ($i=1,\ldots,n$) as $\mathcal{L}=(\mathcal{L}_1,\ldots,\mathcal{L}_n)$. Thus, the generalisation Δ of two tuples x and y (formalized by Proposition 7 below) can be defined as the tuple whose components are the generalisations of the respective components in x and y, while the coverage of a pattern in $\mathcal{L}$ is given by Definition 1.

Definition 1. *Given $p = (p_1,\ldots,p_n) \in \mathcal{L}$, the coverage $Set(p)$ of the pattern p over $\mathcal{L}$ is defined as $\{(x_1,\ldots,x_n) \in X \mid x_i \in Set(p_i),\ i = 1,\ldots,n\}$.*

For example, given the pattern p = ([34, 54], {XXL, XL, XS, XXS}, [0, 130]), the examples e_1 = (54, XXL, 100) and e_2 = (36, XS, 60) are covered by the pattern. However, the tuple (40, M, 70) is not covered by p since M $\notin$ *Set*({XXL, XL, XS, XXS}).

Proposition 7. *Let $X = X_1 \times \ldots \times X_n$ be the space of tuples, $\mathcal{L}_i$ ($i=1,\ldots,n$) a pattern language on the basic type X_i; Δ_i: $X_i \times X_i \rightarrow \mathcal{L}_i$ a binary generalisation operator in X_i and $\mathcal{L} = (\mathcal{L}^1,\ldots,\mathcal{L}^n)$ the pattern language of tuples. The function Δ: $X \times X \rightarrow \mathcal{L}$ defined by $\Delta((x_1,\ldots,x_n), (y_1,\ldots,y_n)) = (\Delta_1(x_1,y_1),\ldots,\Delta_n(x_n,y_n))$ is a binary generalisation operator for X.*

Given that patterns in $\mathcal{L}$ are tuples whose elements are patterns in $\mathcal{L}_i$, the generalisation of two tuples of patterns p and q can be defined as the tuple whose components are the generalisations of the respective components in p and q. This is formalised in Proposition 8.

Proposition 8. *Let $\mathcal{L} = (\mathcal{L}^1,\ldots,\mathcal{L}^n)$ be a pattern language for tuples, with $\mathcal{L}_i$ ($i=1,\ldots,n$) a pattern language on a basic type X_i; Δ^*_i: $\mathcal{L}_i \times \mathcal{L}_i \rightarrow \mathcal{L}_i$ a pattern binary generalisation operator in $\mathcal{L}_i$. The function Δ^*: $\mathcal{L} \times \mathcal{L} \rightarrow \mathcal{L}$ defined by $\Delta^*((p_1, \ldots, p_n), (q_1, \ldots, q_n)) = (\Delta^*_1(p_1, q_1),\ldots,\Delta^*_n(p_n, q_n))$ is a pattern binary generalisation operator wrt. $\mathcal{L}$.*

In HDCC, generalisations of unitary sets are computed as the generalisation of the element with itself. Therefore, the pattern associated to a cluster with only one tuple $\{(x_1, \ldots, x_n)\}$ is given by $\Delta((x_1,\ldots, x_n), (x_1,\ldots, x_n))$, i.e. $(\Delta_1(x_1, x_1),\ldots, \Delta_n(x_n, x_n))$.

Proposition 9. (Composability of Δ) *The binary generalisation operator Δ for tuples given by Proposition 7 when applied to tuples in the space $X = X_1 \times \ldots \times X_n$, where (X_i,d_i) ($i = 1,\ldots,n$) is a metric space equipped with a binary generalisation operator Δ_i, is:*

(*i*) *Strongly bounded by d_T if Δ_i is strongly bounded by d_i, $\forall$ i: i = 1,...,n.*

(*ii*) *Weakly bounded by d_T if Δ_i is strongly bounded by d_i, $\forall$ i: i = 1,...,n.*

(*iii*) *Acceptable if Δ_i is acceptable, $\forall$ i: i = 1,...,n.*

The dendrograms shown by Figure 3 (d) and (e) can be seen as instantiations of propositional clustering in $X = \Re \times \Re$ for the evidence given in Figure 3 (a). We have used the language of closed intervals in $\Re$ as pattern language for each dimension in X, and the absolute difference as the distance between real numbers. Note that a tuple pattern, in this case, describes an axis-parallel rectangle. Figure 3 (d) shows the conceptual dendrogram resulting from the application of HDCC using the single linkage distance d^s_L while Figure 3 (e) using the complete linkage distance d^c_L. We can see that the conceptual dendrogram is not equivalent to the traditional one under single linkage given that although the binary generalisation operator Δ for tuples given in Proposition 7 is strongly bounded by d_T by Proposition 9, Δ^* is not strongly bounded by d^s_L since the generalisation of two rectangles *p1* and *p2* associated to

clusters C_1 and C_2 is a rectangle p that cover points that can fall outside the balls with centre in the linkage points of C_1 and C_2 and radius $d^s_L(C_1, C_2, d_T)$, as it happens for instance with $\{i\}$ which is covered by *p4* (see Figure 3 (b)).

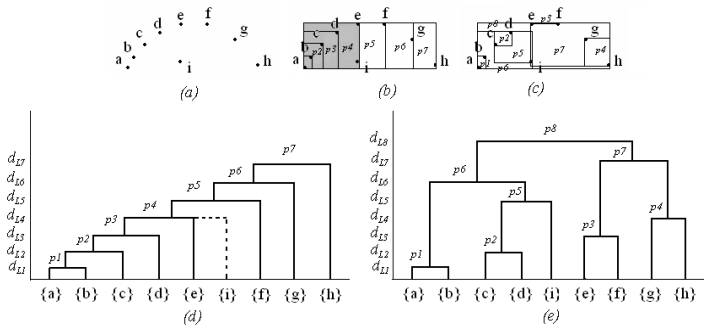

Figure 3. (*a*) A set of points in $\Re\times\Re$. (*b*) Discovered patterns under d^s_L. (*c*) Discovered patterns under d^c_L. (*d*) Application of HDCC for tuples using d^s_L and (*e*) using d^c_L.

Note that the same could happen for tuples in $X_1\times\ldots\times X_n$ when at least two domains X_i are instantiated to $\Re$. Let us consider the following example. C_1={(0, 0, x_3,…, x_n), (1, 1, x_3,…, x_n), (2, 2, x_3,…, x_n), (4, 4, x_3,…, x_n)} and C_2 = {(5.1, 5.1, x_3,…, x_n)} with patterns *p1* = ([0, 4], [0, 4], p_3,…, p_n) and *p2* = ([5.1, 5.1], [5.1, 5.1], p_3, …, p_n), respectively. We have that $\Delta^*(p1, p2) = p$ = ([0, 5.1], [0, 5.1], p_3,…, p_n) and $d^s_L(C_1, C_2, d_T)$ = 1.55 where d_T is the Euclidean distance. However, there exists x = (4.5, 0.5, x_3,…,x_n) that is covered by p but $d^s_L(\{x\}, C_1, d)$ = 2.91 > 1.55, and $d^s_L(\{x\},C_2,d)$ = 4.63 > 1.55.

In fact, the composability property of Δ^* can only be proved wrt. the complete linkage distance d^c_L, as the next proposition establishes.

Proposition 10. (Composability of Δ^*) *The pattern binary generalisation operator Δ^* for tuples in the space $X = X_1\times\ldots\times X_n$ given by Proposition 8 when applied to patterns in the space $\mathcal{L} = \mathcal{L}_1\times\ldots\times\mathcal{L}_n$, where $\mathcal{L}_i$ (i = 1,…,n) is a pattern language for elements in X_i and (X_i, d_i) is a metric space equipped with a pattern binary generalisation operator Δ^*_i, is:*

(*i*) *Strongly bounded by d^c_L if Δ^*_i is strongly bounded by d^c_L, $\forall$ i: i = 1,…,n.*

(*ii*) *Weakly bounded by d^c_L if Δ^*_i is strongly bounded by d^c_L, $\forall$ i: i = 1,…,n.*

(*iii*) *Acceptable if Δ^*_i is acceptable, $\forall$ i: i = 1,…,n.*

We can see in Figure 3 (c) that the application of HDCC to $X = \Re\times\Re$ under complete linkage produces a conceptual dendrogram that is equivalent to the traditional dendrogram as Proposition 1 in [1] establishes given that Proposition 10 (i) and Proposition 9 (i) hold.

4 Experimental Results

In the previous section we proposed a set of generalisation operators and distances for tuples that applied to HDCC under complete linkage distance produces equivalent

conceptual dendrograms with the additional advantage of providing a description of each cluster in the hierarchy. We have also seen through an example that the same operators and distances when used under single linkage distance can produce dendrograms that are not equivalent. The experiments described in this section are aimed to (i) empirically illustrate the first result with a real dataset and (ii) show that the new conceptual clustering, coming from the on-line re-arrangement of the dendrogram, although not equivalent to the traditional dendrogram does not undermine cluster quality when applied under single linkage.

A first experiment was conducted on the Iris Dataset [5]. The dataset consists of three classes, 50 instances each and four numeric attributes. Each class refers to a type of iris plant namely Iris Setosa, Iris Versicolor and Iris Virginica. The numeric attributes refers to the sepal and petal lengths and widths in cms.

To assess the quality of the clustering we employed two different measures: (i) One internal measure, called *S,* which reflects the mean scattering over k clusters with n_i (i = 1, …, k) instances each. This measure is given by eq. (1) where d denotes the Euclidean distance. The lower S is the better the clustering is. (ii) One external measure, the purity P given by eq. (2), where k is the number of clusters, n is the total number of instances and n_i^j the number of instances in cluster i of class j. Purity can be interpreted as classification accuracy under the assumption that all the objects of a cluster are classified to be members of the dominant class for that cluster. Although the class was considered for obtaining purities, it was removed from the dataset to build the clusters.

$$S = \frac{1}{k}\sum_{i=1}^{k}\sqrt{\sum_{j=1}^{n_i}\sum_{l=j+1}^{n_i} d(\vec{x_j},\vec{x_l})^2} \qquad (1) \qquad\qquad P = \frac{1}{n}\sum_{i=1}^{k}\max_j\left(n_i^j\right) \qquad (2)$$

Table 2 shows the patterns discovered by HDCC considering complete and single linkage. Each pattern is a 4-tuple where the component i is also a pattern that provides a description of attribute i.

Table 2. Patterns discovered by HDCC for three clusters.

		Pattern
C1	**Single**	([4.3,5.8],[2.3,4.4],[1.0,1.9],[0.1,0.6])
	Complete	([4.3,5.8],[2.3,4.4],[1.0,1.9],[0.1,0.6])
C2	**Single**	([4.9,7.7],[2.0,3.6],[3.0,6.9],[1.0,2.5])
	Complete	([4.9,6.1],[2.0,3.0],[3.0,4.5],[1.0,1.7])
C3	**Single**	([7.7,7.9],[3.8,3.8],[6.4,6.7],[2.0,2.2])
	Complete	([5.6,7.9],[2.2,3.8],[4.3,6.9],[1.2,2.5])

In cluster C1 the dominant class was Iris Setosa, in C2 was Iris Versicolor and in C3 was Iris Virginica.

In fact, each of these patterns can be seen as a rule. For instance the discovered pattern for C1 under complete linkage and single linkage is ([4.3, 5.8], [2.3, 4.4], [1.0, 1.9], [0.1, 0.6]) that can be interpreted as the rule

```
(sepallength ≥ 4.3 AND sepallength ≤ 5.8 AND sepalwidth ≥ 2.3
AND sepalwidth ≤ 4.4 AND petallength ≥ 1.0 AND petallength ≤ 1.9
AND petalwidth ≥ 0.1 AND petalwidth ≤ 0.6)
```

where `sepallength`, `sepalwidth`, `petallength` and `petalwidth` are the 1st to 4th attributes in the dataset, respectively.

Table 3 shows the values of S and P for HDCC and the traditional hierarchical clustering algorithm under complete distance d^c_L and single linkage distance d^s_L for k = 3 that corresponds to the number of classes in the Iris dataset. As we can see the quality of the conceptual clustering does not differ from that of traditional hierarchical clustering even under single linkage and it provides useful descriptions that allow interpreting the meaning of each group of instances. This result, i.e. cluster quality preserved by HDCC, was confirmed by four experiments carried out on 100 artificial datasets each. Datasets were formed by 600 points drawn from 3 Gaussian distributions in $\Re^2$. In each of the four experiments, means and standard deviations were set to the values reported in Table 4. In these experiments the average values of S over the 100 experiments were obtained for HDCC and the traditional algorithm under single and complete linkage. These values are also reported in Table 4.

Table 3. Values of S and P for the traditional and conceptual dendrograms under d^c_L and d^s_L.

Linkage distance	$S_{\text{Traditional}}$	$S_{\text{Conceptual}}$	$P_{\text{Traditional}}$	$P_{\text{Conceptual}}$
Single (d^s_L)	46.56	46.56	0.68	0.68
Complete (d^c_L)	37.44	37.44	0.84	0.84

Table 4. Values of S averaged over 100 experiments each for HDCC (Conc.) and the traditional hierarchical algorithm (Trad.) for 3 Gaussian distributions with (i) $\sigma = 1$ and $\mu \in [0, 10] \times [0, 10]$; (ii) $\sigma = 1$ and $\mu \in [0, 200] \times [0, 200]$; (iii) $\sigma = 5$ and $\mu \in [0, 100] \times [0, 100]$; (iv) $\sigma = 5$ and $\mu \in [0, 200] \times [0, 200]$.

	Trad. (i)	**Conc. (i)**	**Trad. (ii)**	**Conc. (ii)**	**Trad. (iii)**	**Conc. (iii)**	**Trad. (iv)**	**Conc. (iv)**
d^s_L	524,820	514,417	282,605	282,605	1830,421	1851,406	1607,842	1595,194
d^c_L	285,622	285,622	282,605	282,605	1401,350	1401,350	1410,499	1410,499

5 Conclusions

Hierarchical distance-based conceptual clustering provides an integration of hierarchical distance-based clustering and conceptual clustering. It can be easily seen that for complex datatypes (sequences, graphs, etc.) the original dendrograms are usually different to the dendrograms obtained by applying the generalisation operators. In order to cope with these (negative) results, the notion of conceptual dendrogram and three consistency properties that should be analysed for every pair of distance and generalisation operator have been proposed. Some pairs of distances and generalisation operators are compatible at some degree resulting in equivalent, order-preserving or acceptable conceptual dendrograms while some other pairs are not, so showing that some distances and generalisation operators should not be used together.

In this work, however, we have shown a much more positive picture. In a propositional world, and using the most common distances and generalisation operators for nominal data, numerical data and tuples, we have found out that the

strongest properties (in fact all of them) hold. From these results, we can affirm that the integration of hierarchical distance-based clustering and conceptual clustering for propositional data (i.e., tables, which are still the bulk of most data mining applications) is feasible, congruent and relatively straightforward.

Additionally, the composability result obtained with the tuple datatype and several distances, allow the handling of more elaborate information in the form of tables, where some attributes can have structure, provided that the distance and generalisation operators used for every attribute have some degree of consistency.

In this regard, our immediate future work is focussed on finding operative pairs of distances and generalisation operators for common datatypes in data mining applications, such as sequences, graphs and multimedia objects.

References

1. Funes, A., Ferri, C., Hernández-Orallo, J., Ramirez-Quintana, M.J.: Hierarchical Distance-based Conceptual Clustering. ECML PKDD 2008, Part II, LNAI 5212, pp. 349–364. Springer (2008).
2. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Survey 31(3), 264–323 (1999)
3. Berkhin, P.: A Survey of Clustering Data Mining Techniques, Grouping Multidimensional Data, pp. 25–71. Springer, Heidelberg (2006)
4. Stanfill, A. and Waltz, D.: Toward memory-based reasoning, Comm. of the ACM, 29:1213–1228, (1986)
5. Black C.L.; Merz C. J. UCI Repository of Machine Learning Databases, (1998)
6. Estruch, V.: Bridging the gap between distance and generalisation: Symbolic learning in metric spaces. PhD thesis, DSIC-UPV (2008), `http://www.dsic.upv.es/~vestruch/thesis.pdf`
7. Fisher, D.: 1987, Knowledge acquisition via incremental conceptual clustering, in Machine Learning, pp. 139–172.
8. Michalski, R.S.: Knowledge Acquisition Through Conceptual Clustering: A Theoretical Framework and an Algorithm for Partitioning Data into Conjunctive Concepts. Policy Analysis and Information Systems 4(3), 219–244 (1980)
9. Michalski, R.S., Stepp, R.E.: Learning from Observation: Conceptual Clustering. In: Michalski, et al. (eds.) Machine Learning: An Artificial Intelligence Approach, pp. 331–363. TIOGA Publishing Co. (1983)
10. Talavera, L., Béjar, J.: Generality-Based Conceptual Clustering with Probabilistic Concepts. IEEE Transactions on Pattern Analysis & Machine Intelligence 23(2) (2001)
11. Fisher, R.: The use of multiple measurements in taxonomic problems, in Ann. Eurgenics, Vol. 7, Part II, pp. 179–188, (1936)
12. MacQueen, J. B.: Some methods for classification and analysis of multivariate observations, Proc. of the 5th Berkeley Sym. on Math. Statistics & Probability, pp. 281-297. Univ. of California Press, (1967)
13. Cover, T. M.; Hart, P. E.: Nearest neighbour pattern classification, IEEE Trans. Info. Theory, IT-13, 21-27, January, (1967)
14. Funes A.: Agrupamiento Conceptual Jerárquico Basado en Distancias, Definición e Instanciación para el Caso Proposicional. Master thesis. DSIC-UPV (2008). `http://www.dsic.upv.es/~afunes/masterThesis.pdf`