Improving Dissimilarity Functions with Domain Knowledge, applications with IKBS system

David Grosser, Noël Conruyt, and Jean Diatta

IREMIA, Université de la Réunion
15, avenue René Cassin – BP 7151
97715 Saint-Denis Messag. Cedex 9, France
{grosser, conruyt, jdiatta}@univ-reunion.fr

Abstract. Some of the fundamental and theoretical issues in *Knowledge Discovery in Database* (KDD) rely on knowledge representation and the use of prior and domain knowledge to extract useful information from data. In many data exploration algorithms, dissimilarty functions do not use domain knowledge for the cases comparison. The *Iterative Knowledge Base System* (IKBS) has been designed to improve generalization accuracy of exploration algorithms through the use of structural properties of domain models. A general mathematical framework for utilizing structural properties of the domain model encompassing the definition of a *Dissimilarity Function for Structured Descriptions* is proposed. Applications are conducted with the help of IKBS on a set of databases from the UCI machine learning repository and on structured domain definition data.
Keywords: KDD, Domain Knowledge, Dissimilarity Functions, Generalization Accuracy

1 Taking advantage of Domain Knowledge in KDD

Representation issues, search complexity, use of prior and Domain Knowledge, and statistical inference are some of the core problems in KDD that are still open and require attention[13]. In Data Mining, developing methods and applications for representing knowledge about data is still a serious challenge.

In many fields of real world applications, we can capture a given aspect of the *domain knowledge* by associating attributes of the problem structure with objects linked by composition and/or specialization relationships. We can also structure the *domain definition* of nominal attributes by a hierarchy of values. These techniques enable the algorithms to take into account mutual dependencies between attributes and to compare case properties with more accuracy. For instance, in biosystematics, the scientific discipline that investigates biodiversity, the descriptions of specimens are often highly structured (composite objects, taxonomic attributes), highly noisy (erroneous or unknown data), and highly polymorphous (variable or imprecise data). To take into account this complexity, we need to define a *domain knowledge* that includes information about objects relationships, attribute types and other semantics aspects: the scope of all values, and the meaning of special values (defaults, exceptions). A *domain model* is defined by the association of a domain knowledge and reference data. It represents a given context for the discovery process concerning the *application domain*. The initial domain model is gradually enriched in the course of knowledge discovery to perfect a *domain theory* (see [14] for definitions). Thus, the *Iterative Knowledge Base System* (IKBS) [10] was developed to manage evolving and shared domain models in an object oriented formalism. It enables users to interactively incorporate objects and relations into the domain knowledge (alse called *descriptive model*) to instantiate it with a case base and to conduct supervised and unsupervised classification tasks. This paper will focus on the way to improve accuracy of data exploration algorithms with the use of Domain Knowledge.

Section 2 presents a general mathematical framework for utilizing structural properties of the domain model encompassing the definition of a *Dissimilarity Function for Structured Descriptions*. In section 3, applications are conducted with the help of IKBS on a set of databases from the UCI machine learning repository [9] and on structured domain definition data dealing with corals and marine sponges systematics. We show how nearest neighbor classifiers can be improved by the use of structural properties.

2 Dissimilarity Function for Structured Descriptions

There are many learning systems that depend upon a good distance function to be successful. Many neural network models make use of dissimilarity functions [2], [7]. Dissimilarity functions are also used in many fields besides machine learning, including statistics [3], pattern recognition [6], [8] or in the symbolic data-analysis area [4] [1]. A common problem with these methods is that they

adopt a syntactical and mathematical viewpoint of the dissimilarity measure that does not take into account background knowledge, and relationships between objects. In such traditional methods, attributes are independent of one another. The following sections propose a mathematical framework for defining new dissimilarity functions which use order relations between domain entities.

2.1 Ordered Sets

An *ordered set* X (or *partially ordered set*, briefly, a *poset*) is a set endowed with a binary relation $\leq$ which is *reflexive* (for all $x \in X$, $x \leq x$), *anti-symmetric* ($x \leq y$ and $y \leq x$ imply $x = y$) and *transitive* ($x \leq y$ and $y \leq z$ imply $x \leq z$). If $x \leq y$ and $x \neq y$, one writes $x < y$.

An element $x \in X$ is a *successor* of $y \in X$ if $x < y$ and $x \leq z < y$ implies $z = x$. y is called predecessor of x. The set of successors of $x \in X$ will be denoted $Suc(x)$ and the set of predecessors $Pred(x)$.

2.2 Attributes

An *attribute* is a mapping into a *poset* which is either the set $\mathbb{R}$ of real numbers or some discrete nonnumeric set. In the former case, the attribute is said *quantitative*, and it is said *qualitative* in the latter.

In this paper, we extend the definition of an attribute in such a way that the values of a quantitative attribute are bounded subsets of $\mathbb{R}$, and, likewise, the values of a qualitative attribute are finite subsets of some discrete nonnumeric poset.

The *domain* of an attribute is the least (in the sense of inclusion) set that contains all the possible values of this attribute ; for an attribute a, it will be denoted by $\mathbf{dom}(a)$. Throughout this paper, the domains of quantitative attributes are assumed bounded, and those of qualitative attributes are assumed finite. Note that according to our definition, a missing or unknown value of an attribute corresponds respectively to the empty set or the whole domain. Note also that a qualitative attribute is said *nominal* if its domain is an antichain (a set of non comparable values).

2.3 Structured descriptions

We define a *structured model* as a nonempty n-element object poset X where each object is characterized by a finite set of attributes. The set of attributes of x will be denoted by A_x. A *structured description* (also called a case) is an instance i of a subset P of X, where for all object $x \in P$, each attribute a in A_x is assigned a value $a(i)$. The objects of P will be said *present* on i whereas those of $X \setminus P$ will be said *absent* on i.

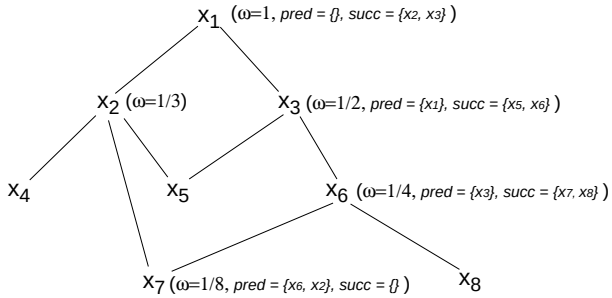

Fig. 1. Example of a structured model with filiation index ω associated to each object $x \in P$. The list of predecessors and successors is associated to each element.

2.4 Global description

The global description of an individual is based on:

(1) the order structure of X ;
(2) the presence/absence of objects on this individual.

To take into account the order structure of X, we will consider the following *filiation index function* ω associated to X, $\omega : X \to \mathbb{R}$ defined by

$$\omega(x) = \begin{cases} 1 & \text{if } x \text{ is maximal} \\ \min\limits_{y \in Pred(x)} \frac{\omega(y)}{|Suc(y)|} & \text{else} \end{cases}$$

See (Fig. 1 for example. To take into account the presence/absence of objects, we also consider situations in which no information is available about the presence/absence of some objets on some individuals. Such objects will be said *unknown* on the corresponding individuals. If an object x is unknown on an individual i, $p_i(x)$ will denote the probability for x to be present on i. If the objects of X are listed in a fixed linear ordering, the global description of an individual i may be identified with the n-vector

$$\big(\omega(x)\, \chi_i(x)\big)_{x \in X}$$

where χ_i is defined on X by

$$\chi_i(x) = \begin{cases} 1 & \text{if } x \text{ is present on } i \\ 0 & \text{if } x \text{ is absent on } i \\ p_i(x) & \text{if } x \text{ is unknown on } i \end{cases}$$

2.5 Dissimilarity measure

The dissimilarity measure we propose in this paper is the Minkowski transform of a 2-vector. The components of this vector are the normalized *global* dissimilarity D_G and the normalized *local* dissimilarity D_L. That is

$$D(i,j) = \left(\left(\mu\, D_G(i,j)\right)^r + \left(\nu\, D_L(i,j)\right)^r \right)^{\frac{1}{r}} \tag{1}$$

where $r \geq 1$ and where μ and ν are normalization coefficients. Following applications are conducted with $r = 1$. On unstructured databases, the component D_G is always null. In that case, the expression (1) is reduced to the local component ($\mu = 0$ and $\nu = 1$).

The local dissimilarity can be for instance the Euclidean metric, or one of those proposed by [15],i.e.:

- Heterogeneous Value Difference Metric (HVDM),
- Discretized Value Difference Metric (DVDM),
- Windowed Value Difference Metric (WVDM),
- Local dissimilarity on Heterogeneous Value defined by position and content factors (DGR) [5]

Any metric can be used in this general equation. Following applications will show how the use of global dissimilarity factor can improve generalization accuracy of data exploration algorithms. The proposed *global dissimilarities* (1) may be divided into two groups. The first group consists of the Minkowski transforms on the n-dimensional vector space of global descriptions :

$$D_G(i,j) = \left(\sum_{x \in X} \omega(x)^r \left| \chi_i(x) - \chi_j(x) \right|^r \right)^{\frac{1}{r}} \tag{2}$$

where $r \geq 1$. The second group consists of the extension of various indices on presence/absence signs, which takes into account the order structure of X as well as the possible unknown objects.

2.6 extensions

The table below shows some extensions of classical dissimilarity indices in Data Analysis on presence/absence boolean vectors, when for any two individuals i and j, we consider the following sets:

- the set E_{ij} of objects present on both i and j,
- the set F_{ij} of objects absent on both i and j,
- the set $K_{ij} = E_{ij} \cup F_{ij}$ of objects either absent or present on both i and j,
- the set D_{ij} of objects present on one but absent on the other,

– the set U_{ij} of objects unknown on at least one of i and j.

We will denote by $p_{ij}(x)$, $f_{ij}(x)$, $k_{ij}(x)$ and $q_{ij}(x)$ the probability fo x to belong to E_{ij}, F_{ij}, K_{ij} and D_{ij}, respectively.

$D_G(i,j)$	Name
$\frac{\sum_{x\in D_{ij}} \omega(x) + \sum_{x\in U_{ij}} \omega(x) q_{ij}(x)}{\sum_{x\in E_{ij}} \omega(x) + \sum_{x\in D_{ij}} \omega(x) + \sum_{x\in U_{ij}} \omega(x)(p_{ij}(x)+q_{ij}(x))}$	Jaccard extended
$\frac{\sum_{x\in D_{ij}} 2\,\omega(x) + \sum_{x\in U_{ij}} 2\,\omega(x) q_{ij}(x)}{\sum_{x\in E_{ij}} \omega(x) + \sum_{x\in D_{ij}} 2\,\omega(x) + \sum_{x\in U_{ij}} \omega(x)(p_{ij}(x)+2\,q_{ij}(x))}$	Rogers and Tamimoto extended
$\frac{\sum_{x\in D_{ij}} \omega(x) + \sum_{x\in U_{ij}} \omega(x) q_{ij}(x)}{\sum_{x\in X} \omega(x)}$	Sokal and Michener extended

3 Applications

In the following section, we shall study a particular case of acyclical oriented graphs which is well adapted to the representation of biological data : the descriptive trees. In this form, $|Pred(y)| = 1$.

3.1 IKBS

The *Iterative Knowledge Base System (IKBS)* [10] is a software that manages evolving and shared knowledge bases. Domain models and data are represented in an object oriented formalism and can be built and transformed through graphical representations. These representations are generalization or composition graphs or trees where nodes are objects of the domain and links are relationships between objects (Fig. 2 shows an example). The system is currently being developed at IREMIA laboratory in the Université de La Réunion in the objet-oriented programming language Java. Classes of the Knowledge Representation Language (KRL) define Knowledge Model properties. Instances of these classes are problem cases. These cases inherite their objects and attributes definition from domain model. The system also provides algorithms to deal with missing, multiple (conjunctive and disjunctive forms), unknown and exception values (those that do not belong to the domain definition).

With IKBS, end-users can define by themselves structured domain models with different kinds of relationships between objects: composition or specialization dependencies. Another way to acquire a domain model consists of importing

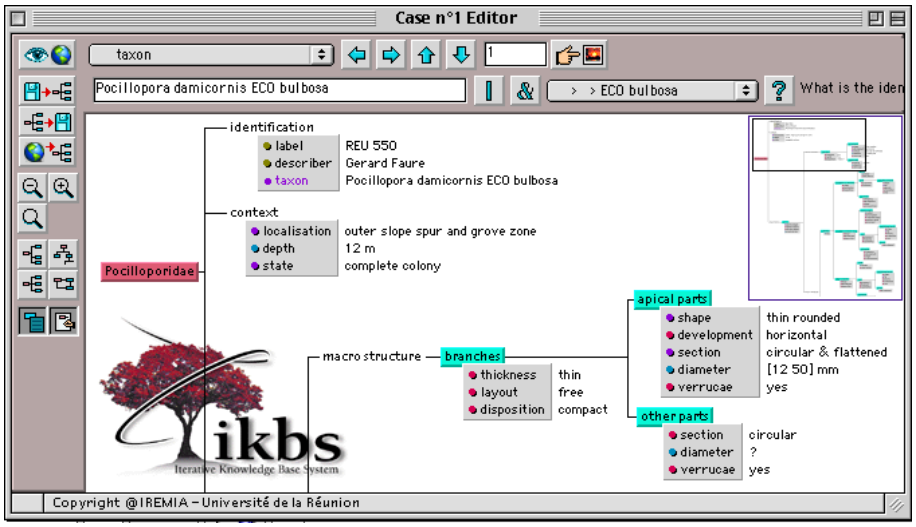

Fig. 2. Part of a descriptive model of *Pocilloporidae* (corals family) in IKBS

external databases or data tables into the internal KRL. We thus obtain an unstructured domain model (Fig. 3) that is automatically generated complete with attributes domain definition and a case base linked to it. IKBS provides tools to interactively define structured descriptive models, hierarchical attributes, and special features such as default or exception values.

Unstructured data definition can be transformed to add composition and/or specialization relationships as shown in Fig. 4.

Moreover contextual knowledge of the domain can be associated with any type of elements pertaining to the descriptive model, such as comments, photographs, pictures, and Web links.

IKBS also provides tools to conduct supervised and unsupervised classifications tasks with different dissimilarity functions. In the following application we will focus on the way to improve accuracy of data exploration algorithms by using domain models. Note that, presently, contextual information are not used by data exploration algorithms. This methodology has been successfully applied to a subset of sponge and corals families. In these cases, domain knowledge plays a central role because as mentioned above, biosystematics data are highly structured, polymorphous and noisy.

3.2 generalization accuracy of dissimilarity functions applied to Databases

For experimention, we tested 17 databases from the *UCI Machine Learning Repository* at the University of California and we used 3 knowledge bases from the IKBS project in marine biology. Fig 5 lists the databases, the number of

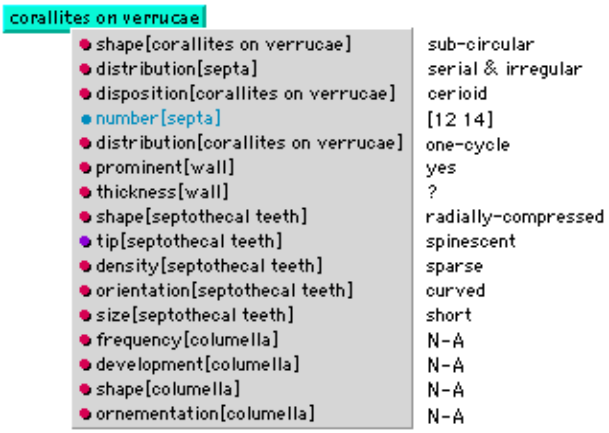

Fig. 3. The *corallites on verrucae* object is described by a list of unstructured properties

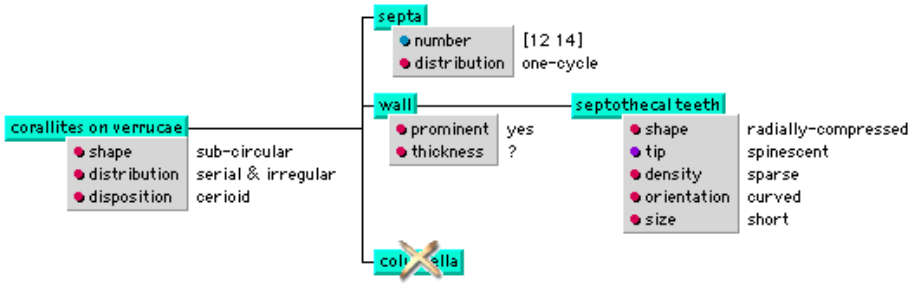

Fig. 4. *corallites on verrucae* object with attributes and composition relationships with *septa*, *wall* and *columella* objects

instances in each database (#cases), the number of quantitative (#quantitative) nominal (#nominal) and ordinal (#ordinal) input attributes. We use the term ordinal for attribute where domain values are structured by specialization relationships.

Two database models have only pure numeric data (Image segmentation and Vehicle) or pure symbolic data (Audiology, Monks). Others are defined by mixed features. Four databases provide additional information on attributes: Bridges, *Pocilloporidae* and *Siderastreidae* coral families, and *Hyalonema* marine sponges. Some attributes of these bases are structured by order relationships (ordinal attributes) and organized by objects. These four structured databases were destructurized (transformed into data tables) in order to highlight the augmentation of Generalization Accuracy between unstructured and structured versions. In the following section, we present results of generalization accuracy of some dissimilarity functions on these databases.

DataBases		Model properties			
	#cases	#Quantitative	#Nominal	#ordinal	#objects
Annealing	798	9	29		
Audiology	200		69		
Audiology test	26		69		
Bridges	108	4	5	2	
* Corals(Pocilloporidae)	113	30	112	15	48
* Corals(Siderastreidae)	60	24	61	10	27
Echocardiogram	132	7	2		
Flag	194	10	18		
Hepatitis	155	6	13		
Images segmentation	420	18	1		
LED+17 noise	10000		24		
Monks-1	432		6		
Monk2-2	432		6		
Monk2-3	432		6		
Mushroom	8124	1	21		
Soybean (large)	307	6	29		
Soybean (small)	47	6	29		
* Sponges (Hyalonema)	60	24	41	10	36
Vehicle	846	18			
Zoo	90		16		

Fig. 5. Databases from UCI ML Repository and IKBS project (*)

We compare the dissimilarity functions previously mentioned. The four dissimilarity measures and a nearest neighbor classifier [12] (with k = 1) were programmed into IKBS. Each function was tested on 20 (+ 4 structured) datasets using cross validation 10-fold. The average generalization accuracy over all 10 trials is reported for each test in (Fig. 6). The highest accuracy achieved for each dataset is shown in bold. This application shows that DGR dissimilarity on average yields improved generalization accuracy on a collection of 24 databases. More important, it shows that using background knowledge and in particular, structures of the domain knowledge, can improve generalization accuracy with regard to any local dissimilarity.

4 Conclusion and future work

It has been shown that no learning algorithm can generalize more accurately than another when called upon to deal with all possible problems [11], unless information about the problem other than the training data is available. It follows then, that no dissimilarity function can be an improvement over another because it possesses a higher probability of accurate generalization. Its accuracy is a factor of its match with the kinds of problems that are likely to occur. Our global dissimilarity function was designed for complex data structures pertaining to the biological domains and is quite well suited for that purpose. Moreover, in

some cases when considering tree-structures, we can obtain better performances in time of execution, attributes pertaining to absent objects are not considered. For the time being, an original version of an inductive algorithm that utilizes background knowledge has been programmed into IKBS [10] and we plan to adapt other algorithms drawn from the area of Case-Based Reasoning.

5 Bibliography

References

1. Guénoche A. Order distance associated with a hierarchy. *Journal of Classification, 14*, pages 101–105, 1997.
2. Broomhead and Lowe. Multi-variable functional interpolation and adaptative networks. *Complex Systems, Vol.2*, pages 321–355, 1988.
3. Atkeson C., Moore A., and Schall S. Locally weighted learning. *Artificial Intelligence Review*, 1996.
4. Diday E. Recent progress in distance and similarity measures in pattern recognition. *Septième Journée Francophone de Classification*, pages 534–539, 1974.
5. Diatta J., Grosser D., and Ralambondrainy H. A general dissimilarity measure for complex data. INF 01, IREMIA, University of Reunion Island, july 1999.
6. Gowda K.C. and Diday E. Symbolic clustering using a new dissimilarity measure. *Pattern Recognition*, 24:567–578, 1991.
7. Kohonen and Teuvo. The self-organizing map. *Proceedings of the IEEE, Vol.78, No.9*, pages 1464–1480, 1990.
8. Ichino M. General metrics for mixed features : the cartesian space theory for pattern recognition. In *IEEE Intl Conf. Syst. Man Cybern.*, 1988.
9. Merz and Murphy. Uci repository of machine learning databases. *Department of Information and Computer Science*, 1996.
10. Conruyt N. and Grosser D. Managing complex knowledge in natural sciences. *Lecture Notes in Computer Science subseries, 1650, Springer Verlag*, pages 401–414, 1999.
11. Schaffer and Cullen. A conservation law for generalization performance. *In Proceedings of the Eleventh International Conference on Machine Learning (ML'94)*, 1994.
12. Cover T. and Hart P. Nearest neighbor pattern classification. *Institute of Electrical and Electronics Enginneers Transactions on Information Theory, Vol.13, No.1*, pages 21–27, 1967.
13. Fayyad U.M., Piatetsky-Shapiro G., Padhraic Smyth, and Ramasamy Uthurusamy, editors. *From Data Mining to Knowledge Discovery: Current Challenges and Future Directions*. Advances in Knowledge Discovery and Data Mining, AAAI Press / MIT Press, 1996.
14. Klösgen W. and Zytkow J.M. *Knowledge Discovery in Databases Terminology*. Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, 1996.
15. Randall W.D. and Martinez T.R. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research, 6*, pages 1–34, 1997.

Databases	*Dissimilarity functions*			
	Euclid	HVDM	WVDM	DGR
Unstructured databases				
Annealing	94.99%	94.61%	95.87%	**98.87%**
Audiology	60.50%	**77.50%**	**77.50%**	76.00%
Audiology test	41.67%	78.33%	78.33%	**88.46%**
Bridges	58.64%	59.64%	56.64%	**60.19%**
* Corals (*Pocilloporidae*)	51.12%	59.6%	59.6%	**61.06%**
* Corals (*Siderastreidae*)	72.80%	85.16%	85.40%	**86.80%**
Echocardiogram	94.82%	94.82%	**100.00%**	82.58%
Flag	48.95%	55.82%	58.74%	46.39%
Hepatitis	77.50%	76.67%	79.88%	78.71%
Images segmentation	92.86%	92.86%	93.33%	**98.10%**
LED+17 noise	42.90%	**60.70%**	**60.70%**	**60.70%**
Monks-1	77.58%	68.09%	68.09%	**79.83%**
Monk2-2	59.04%	**97.50%**	**97.50%**	96.50%
Monk2-3	87.26%	**100.00%**	**100.00%**	**100.00%**
Mushroom	**100.00%**	**100.00%**	**100.00%**	**100.00%**
Soybean (large)	87.26%	90.88%	**92.18%**	89.58%
Soybean (small)	**100.00%**	**100.00%**	**100.00%**	**100.00%**
* Sponges (*Hyalonema*)	49.21%	55.12%	55.12%	**56.8%**
Vehicle	70.93%	70.93%	65.37%	**79.02%**
Zoo	97.78%	**98.89%**	**98.89%**	98.11%
Structured databases				
Bridges	60.20%	56.24%	58.88%	**62.74%**
* Corals (*Pocilloporidae*)	53.48%	60.86%	60.86%	**63.50%**
* Corals (*Siderastreidae*)	77.30%	88.20%	88.20%	**90.00%**
* Sponges (*Hyalonema*)	51.20%	**58.00%**	**58.00%**	56.80%
Average	71.64%	79.31%	79.57%	**80.43%**

Fig. 6. % Generalization Accuracy with different dissimilarity functions, on unstructured and structured databases from UCI Machine Learning Repository and IKBS projects (*). Structured databases are utilized in unstructured and structured versions to show the interest to use global dissimilarity to improve generalization accuracy.