Improving Dissimilarity Functions with Domain Knowledge, applications with IKBS system

David Grosser, Jean Diatta, and Noël Conruyt

IREMIA, Université de la Réunion
15, avenue René Cassin – BP 7151
97715 Saint-Denis Messag. Cedex 9, France
{grosser, jdiatta, conruyt}@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, dissimilarity 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 [6]. 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,

D.A. Zighed, J. Komorowski, and J. Zytkow (Eds.): PKDD 2000, LNAI 1910, pp. 409–415, 2000.

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 [7] for definitions). Thus, the *Iterative Knowledge Base System* (IKBS) [3] 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 [2] and on structured domain definition data dealing with corals and marine sponges systematics. As example, 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. Dissimilarity functions are used in many fields besides machine learning, including statistics, pattern recognition and in the symbolic data-analysis area. 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 Structured descriptions

We define a *structured model* as a nonempty n-element object *partially ordered set* X where each object is characterized by a finite set of attributes (Fig. 1). 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.

2.2 Global description

The global description of an individual is based on (1) the order structure of X and (2) the presence/absence of objects on this individual.

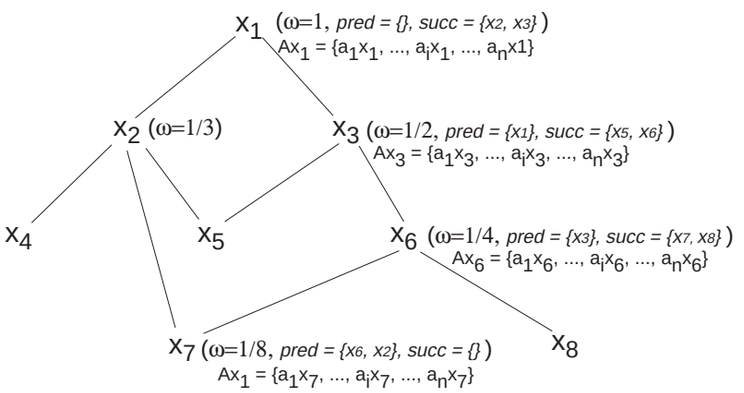

Fig. 1. Example of a structured model with filiation index ω associated to each object $x \in P$. The list of predecessors and successors is associated to each element.

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

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

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 $\left(\omega(x)\, \chi_i(x)\right)_{x \in X}$ where χ_i is defined on X by

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

2.3 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 (1).

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

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 [8],i.e.:

- Heterogeneous Value Difference Metric (HVDM),
- Discretized Value Difference Metric (DVDM),
- Windowed Value Difference Metric (WVDM),
- Local dissimilarity on Heterogeneous Value (DGR) [1]

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

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

Possible extensions of various indices on presence/absence signs can be derived from this expression, which takes into account the order structure of X as well as the possible unknown objects.

3 Applications with IKBS

The *Iterative Knowledge Base System* [3] 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).

With IKBS, end-users can define structured domain models with different kinds of relationships between objects: composition or specialization dependencies. Another way to acquire a domain model consists of importing external databases or data tables. We thus obtain an unstructured domain model 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. 3.

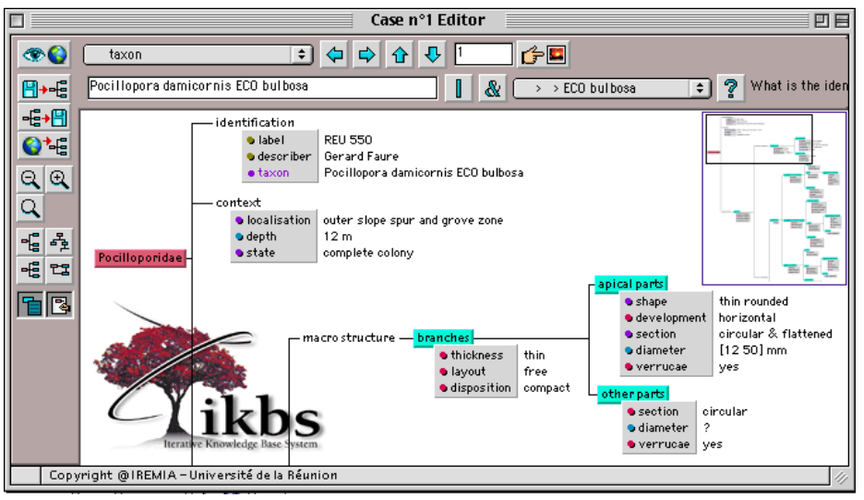

Fig. 2. Part of a structured representation pertaining to *Pocilloporidae* family (corals) in IKBS

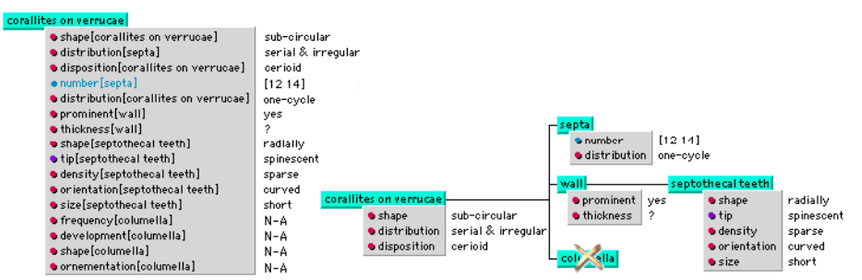

Fig. 3. The *corallites on verrucae* object transformed into a structured representation

3.1 generalization accuracy of dissimilarity functions

We present results of generalization accuracy of some dissimilarity functions on 17 databases from the *UCI Machine Learning Repository* and 3 knowledge bases from *IKBS projects* in marine biology (Fig. 4). 2 database models have only pure numeric data (Image segmentation and Vehicle) or pure symbolic data (Audiology, Monks). Others are defined by mixed features. 4 databases provide additional information on attributes: Bridges, *Pocilloporidae* and *Siderastreidae* coral families, and *Hyalonema* marine sponges. Some attributes 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. We compare the dissimilarity functions previously men-

tioned. The 4 dissimilarity measures and a nearest neighbor classifier [5] (with k = 1) were programmed into IKBS. Each function was tested on 20 (+ 4 structured) datasets using 10-fold cross validation. The average generalization accuracy over all 10 trials is reported for each test in (Fig. 4). 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 [4], 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 and is particularly well suited for data pertaining to the biological domains. Moreover, in some cases when considering tree-structures, we can obtain better performances in time of execution because 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 [3] and we plan to adapt other algorithms drawn from the area of Case-Based Reasoning.

References

1. Diatta J., Grosser D., and Ralambondrainy H. A general dissimilarity measure for complex data. INF 01, IREMIA, University of Reunion Island, july 1999.
2. Merz and Murphy. Uci repository of machine learning databases. *Department of Information and Computer Science*, 1996.
3. Conruyt N. and Grosser D. Managing complex knowledge in natural sciences. *LNCS 1650, Springer Verlag*, pages 401–414, 1999.
4. Schaffer and Cullen. A conservation law for generalization performance. *In Proceedings of ML'94*, 1994.
5. 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.
6. 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.
7. Klösgen W. and Zytkow J.M. *Knowledge Discovery in Databases Terminology.* Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.
8. 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				
Anncaling	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. 4. % 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.