Knowledge acquisition from multiple experts

Aïda Valls Mateu, Vicenç Torra Reventós

Dpt. d'Enginyeria Informàtica
Escola Tècnica Superior d'Enginyeria (ETSE)
Universitat Rovira i Virgili
Carretera de Salou, s/n
E-43006 Tarragona (Spain)

aidav@goliat.upc.es, vtorra@etse.urv.es

Abstract

Synthesis of information from multiple experts is needed in the knowledge acquisition process to obtain an accurate model of the application domain. This paper presents a methodology to analyse two synthesis methods at different levels: synthesis of matrices and synthesis of classifications. The paper describes our work in this area and the tools that have been developed to apply this methodology.

1. Introduction

One of the most difficult tasks in the construction of knowledge based systems (KBS) [14] is the knowledge acquisition (KA) process. KA pretends to express the structure underlying a certain domain (e.g. to build a model). This is achieved by means of an expert or directly from raw data. Due to the difficulty of building this model, knowledge acquisition used to be considered the bottleneck in the development of KBS. Although that nowadays this is considered as a misleading metaphor [10] it underlines the importance of the problem. An area of Artificial Intelligence is devoted to build systems to ease in this process, some help the expert to express his knowledge while others extract knowledge from raw data. When knowledge has to be extracted from data supplied by multiple experts, a synthesis method to combine information is needed. In this paper we introduce a methodology that can be used to study two different kinds of synthesis methods. Our final goal is to characterize these methods, in order to establish when should one method be preferred to the other.

Among KA systems we consider those based on Kelly's Personal Construct Theory (PCT) [17]. In these systems the expert is questioned about the attributes (i.e. the properties or characteristics) that discriminate a set of objects. In this way, a numerical grid is built. This grid represents functions of the form f: Object*Attribute -> Value. This grid is used to discover implicit rules in the domain. As these tools are interactive, the expert is allowed at any moment to analyse the data (to reconsider decisions already made and to develop incomplete aspects) and to generate rules from it (e.g, GAR [22]).

Usually, when it is intended to build a KBS for ill-structured domains, a previous step to the generation of rules is needed. It consists on the definition of a structure for the elements of the domain. This is obtained by classifying the objects according to the values of their attributes. Therefore, in general, there is a three-step process: (1) the data matrix is built (i.e. the definition of the numerical object*attribute grid), (2) the classification is obtained, and (3) rules are generated. The process described here as linear has feedback. Among the systems based on Personal Construct Theory we can outline ETS [6], AQUINAS [7,11] and KSSn [11].

However, there are still some open questions in the development of knowledge acquisition tools. Some of them are the following:

1. Value types other than quantitative ones are needed for object*attribute pairs. In classical systems based on PCT only the ordinal reduced scale (1 to 5) is considered. Such situation is not very appropriate in other fields (e.g. in medicine). See, e.g. KAFES [16], EGAC [26], LINNEO$^+$ [5], EAR* [21] for some attempts to use other kinds of mappings.

2. Dealing with information from several experts is needed. The synthesis of the knowledge collected from a group of experts rather than from a single one can improve in the process of modeling. ETS, AQUINAS and EAR* encompass tools for synthesis of judgements. EGAC is a tool designed with a similar goal.

3. Other kind of knowledge representation than mappings for pairs object*attribute is needed. Some knowledge is expressed more easily using of other formalisms: rules (LINNEO$^+$ [5]), predicates (Klass [12]), relations between attributes (e.g. causality relations).

4. Friendly interfaces are an essential requirement. Since KBSs have increasing complexities and experts have to deal with more information, friendly interfaces help the user when manipulating data. Some tools (e.g. AQUINAS) have already been developed in this way.

The work introduced here is focused on points 1, 2 and 4 above. The focal point is the study of two methods of synthesis of information in the KA process (point 2). Tools are assumed to be based on PCT. However, grids are not limited to numeric values but also qualitative values are considered (point 1).

2. Synthesis in Personal Construct Theory based systems

As it has been said, considering several experts we can get a more accurate model. When a system is built, it is required that the knowledge spreads over the whole domain. However, the information supplied by a single expert is often not reliable enough or too narrow in relation to the working domain. In this case, the information provided by several experts can be combined to improve data reliability and accuracy. Nevertheless, if the information from a single expert can cause problems, this is also the case when several experts are considered (e.g. contradictory opinions, different meanings of similar terms). See [6] and [11] about some of the problems concerning KA (they are reviewed in [26]). Sometimes the difficulties keep from defining a synthesis function when some conditions are required (a classical example are Arrow's impossibility theorems [3]). In any system, the synthesis process can be performed at several levels. Usually systems based on PCT (e.g., ETS [6]) generate rules for the matrices obtained [27]. In our case, we have studied the synthesis at two different levels:
a) at the matrix level, building a consensued matrix from those supplied by the experts, and proceeding with the KA process as if this new matrix was the only one defined by an expert;
b) at the classification level, obtaining a classification for each matrix, synthesizing these classifications into a consensed one and then generating from it the rules for the KBS.

EGAC is an Automatic Consensus Generator Tool which is able to synthesize information at these two levels. The methods used are described in [25,26].

The goal of this work is to compare and characterize both synthesis methods. Following the classification of [6], two grids may be disjoint (no sharing objects nor attributes), grids may have the same set of attributes but disjoint elements, they may have the same set of objects but disjoint attributes, or they may share some subset of objects or attributes (they are overlaped). Our work is restricted to the case in which all experts work with the same set of attributes and the same set of objects. In this framework, when there is coincidence of attributes in all the experts, we consider coincidence or non coincidence of their terminology [11]. Terminology stands for the set of values that an expert uses to refer to an attribute (its domain). When there is a coincidence of terminology it is said [11] that there is *consensus* among experts, while when there is no coincidence of terminology it is said that there is *correspondence* among experts.

3. The methodology

To perform the analysis we need a set of matrices in a unified framework so that their synthesis is possible, and also we need an *ideal* classification of the data, which could be used as *the best* model to achieve. However, in usual domains it is neither possible to decide which of the classifications built is the best, nor to know the *ideal* classification of the data. For this reason the measure of the distance between an ideal result and the one obtained from this method is not possible.

To overcome this problem, we start our analysis from a single quantitative data matrix, that is considered the *ideal* matrix. It is used to generate the set of matrices involved in the process by means of a discretization method, that changes quantitative domains into qualitative ones, reducing the accuracy of the information kept in each data matrix at different levels. With the reduction of accuracy we simulate data elicited from experts in some fields as medicine.

Once a set of slightly different matrices has been generated, we can synthesize them and compare the resulting classification with the *ideal* one. In the same way, we can classify each matrix, synthesize these classifications and compare the agreed one with the one obtained from the *ideal* matrix. These comparisons are going to be used to establish the properties of each synthesis methods.

According to this description, this analysis requires several tools: a discretization method that helps in the definition of new matrices from a quantitative one, a classifier, synthesis methods for both levels and the definition of a distance between classifications. Some of these tools have already been developed, while work is in progress with the last two ones. We comment now these tools with more detail.

A discretization method

To construct new matrices (with both qualitative and quantitative values) from a quantitative one, we have built a tool that translate quantitative domains into qualitative ones. This tool, described in [29], pretends to define linguistic labels over a quantitative domain as an expert would do. Although based on methods described in [1,18,31] some modifications have been

introduced because our goal is in some way different: in our case, the method should make possible to establish similarities among objects in the relevant part of the domain instead of gathering them all in a single partition as the others do (they are oriented to classification).

A classification system

As classification is required in many fields in AI (e.g. in knowledge elicitation as in ETS, LINNEO$^+$, and in Expert Systems validation [20]), we have developed a general classifier, Sedàs [28,30], that deals with quantitative, qualitative (ordered and unordered) and boolean attributes. Sedàs, with a friendly interface (point 4 in section 1), is a SAHN [9] (sequential, agglomerative, hierarchic, non-overlapping) classifier. The main contribution of this system is that it introduces a new semantics for ordered linguistic labels which is used to calculate distance between qualitative values [24]. This new definition allows the labels to be non equally informative. The semantics is based on the generalization of the classical negation function. Sedàs is going to be used in this classification process.

A tool for synthesis at matrix level

EGAC builds an agreed matrix from a set of matrices defined over the same set of objects and attributes. The method uses functional equations (using the theory of Aczél and Alsina on synthesis of judgements [2]), fuzzy sets and classification methods. To obtain this matrix, we assume the independence of irrelevant alternatives condition [3], on each attribute, i.e. for each attribute the consensus column is calculated using only the columns corresponding to the same attribute in experts' matrices. This condition avoids the exponential cost of considering the synthesis of each attribute with all the other ones, but it restricts the matrices to have the same set of attributes.

The treatment of each attribute depends on its characteristics. A different treatment is done when the experts have used the same domain or when they have used different ones (which are called consensus and correspondence respectively). In the latter case the synthesis is achieved by means of a translation into a classification problem [26]. For each attribute a matrix with objects as rows and experts as columns is built. Then the matrix is classified obtaining a set of classes, each one referring to a new term of the attribute domain in the synthesized matrix. The values in the agreed matrix are then assigned according to the class to which each object belongs. Sedàs is also going to be applied in this process.

A tool for synthesis at classification level

About synthesis of classification we intend to apply the functions studied in the framework of biology [8,13,19] to combine hierarchical classifications or partitions. Some preliminary work has been done [25] but a more detailed analysis is in progress.

A distance between classifications

To compare the results obtained from both techniques on synthesis of information, a distance between classifications is needed [4,8,12]. These functions are highly related to the characteristics of the representation model, and to the goal to be achieved with the

measurement of the distance. We are going to study distances that fit our particular requirements.

4. Conclusions and future work

We have presented in this paper our methodology to study tools for synthesis of information from multiple experts in the framework of knowledge acquisition. We have described some of our results (the discretization method, the classifier system and the tool for synthesis at matrix level). We are now working in the synthesis at classification level and the method to calculate distances between classifications.

5. Acknowledgements

The help of Dr. Lluís Godo is gratefully acknowledged. We would also like to thank the referees for their valuable suggestions and comments, that will improve our future work. Aïda Valls acknowledges the support of the Department of Software at UPC.

6. References

[1] Anderberg, M.R., (1973), *Cluster Analysis for Applications*, Academic Press, Inc.
[2] Aczél, J., Alsina, C., (1986), On synthesis of judgements, *Socio-Econom. Plann. Sci.*
[3] Arrow, K.J., (1951), *Social choice and individual values*, Wiley, New York.
[4] Barthélemy, J.P, McMorris, F.R., (1986), The median procedure for n-trees, *J. of Classification*, 3 329-334.
[5] Béjar, J., (1994), *Automatic Knowledge Acquisition in ill-structured domains*, (in spanish), Master Thesis, U. P.C.
[6] Boose, J.H., (1986), *Expertise transfer for expert system design*, Elsevier Science Publishers B.V.
[7] Boose, JH., Bradshaw J.M., (1987), Expertise transfer and complex problems: using AQUINAS as a knowledge-acquisition workbench for knowledge-based systems, *Int. J. of Man-Machine Studies*, 26 3-28.
[8] Day, W.H.E., (1986), Special issue on comparison and consensus of classifications, *J. of Classification*, 3.
[9] Everitt, B., (1977), *Cluster Analysis*, Heinemann Educational Books Ltd.
[10] Ford, K.M., Bradshaw, J.M., (1993), Special Issue on "Knowledge Acquisition as Modeling", Parts I and II, *Int. J. of Intelligent Systems*, 8:1 and 8:2.
[11] Gaines, B.R., Shaw, M.L.G., (1993), Knowledge Acquisition Tools Based on Personal Construct Psychology, *The Knowledge Engineering Review*, 8:1 49-85.
[12] Gibert, K., Cortés, U., (1994), Combining a knowledge-based system and a clustering method for a construction of models in ill-structured domains, Chap.V *Selecting Models from Data. Lecture Notes in Statistics*, 89, P.Cheeseman, R.W. Oldford (Eds.).
[13] Gordon, A.D, (1987), A review of Hierarchical Classification, *J.of the R. Statistical Society. A*, 150:2 119-137.
[14] Guida, G., Tasso, C., (1994), *Design and development of Knowledge-Based Systems*, UK:John-Wiley.
[15] Hernández, C., Sancho, J.J., Belmonte, M.A., Sierra, C., Sanz, F., (1995), Validation of the Medical Expert System RENOIR, *Computers and Biomedical Research*, 27 457-471.
[16] Hwang, G.J., (1995), Knowledge Acquisition for Fuzzy Expert Systems, *Int. J. of Intel. Syst.*, 10 541-560
[17] Kelly, G.A., (1955), *The Psychology of Personal Construct*, W.W. Norton & Co.: New York.
[18] Kohonen, T., (1988), *Self-Organization and Associative Memory*, Springer, Berlin.
[19] Margush, T., McMorris, F.R., (1981), Consensus n-Trees, *Bulletin of Mathematical Biology*, 43 239-244.
[20] Meseguer, P., Verdaguer, A., (1993), Verification of Multi-Level Rule-Based Expert-Systems: Theory and Practise, *Int. J.of Expert Systems*, 6:2 163-192.
[21] Plaza, E., Alsina, C., López de Mántaras, R., Aguilar, J., Agustí, J., (1986), Consensus and knowledge acquisition, In R.R. Yager, B. Bouchon (Eds), *Uncertainty Management in Knowledge-Based Systems*, LNCS, Springer-Verlag.
[22] Riaño, D., (1994), Knowledge abstraction using heuristic search I: A specialization approach, Report CS-R9469 of the Centrum voor Wiskunde en Informatica (CWI), Amsterdam, Netherlands, December.
[23] Riaño, D, Torra, V, Valls, A, (1995) A semantics for ordered linguistic labels in KBS. An introduction, *Abstracts of the third hispano-polish symposium on system analysis and computer science*, Cerdanyola del Vallès, Spain.
[24] Torra, V., (1996), Negation functions based semantics for ordered linguistic labels, *Int. J. of Intelligent Systems*, Accepted for publication.
[25] Torra, V., Cortés, U., (1993), Consensus and classification for data matrices, Proceedings of the *V conferencia de la asociación española para la inteligencia artificial*, 216-225, Madrid, Spain.
[26] Torra, V., Cortés, U., (1995), Towards an Automatic Consensus Generator Tool: EGAC, *IEEE Trans. On Systems, Man and Cybernetics*, 25:5 888-894.
[27] Trice, A., Davis, R., (1989), *Consensus Knowledge Acquisition*, MIT, AI Memo No 1183, December.
[28] Valls, A., (1995), *Sedàs. General Classification Module for ill structured domains*, (in Catalan), Bsc Thesis, ETSE, Universitat Rovira i Virgili.

[29] Valls, A., Torra, V., (1996), Quantitative attribute discretization in Knowledge Based Systems, (in Spanish), accepted at the *Sexto congreso español de tecnologias y lógica fuzzy*, Oviedo, Spain.
[30] Valls, A., Riaño, D., Torra, V., (1996), Sedàs: A semantic based general classifier system, submitted to Mathware & Soft Computing.
[31] Yuan, Y., Shaw, M.J., (1995), Induction of fuzzy decision trees, *Fuzzy Sets and Systems*, 69 125-139.