A Granular Signature of Data

Witold Pedrycz*, Michael H. Smith**, Andre Bargiela***

*Dept. of Electrical and Computer Engineering
University of Alberta, Edmonton, Canada T6G 2G7
&
Systems Research Institute, Polish Academy of Sciences
01-447 Warsaw, Poland

** Department of Mechanical & Manufacturing Engineering
University of Calgary
Calgary, AB T2N 1N4 Canada

*** Department of Computing
Nottingham Trent University
Nottingham NG1 4BU, UK

Abstract In this paper, we discuss an issue of description of highly dimensional data realized in terms of fuzzy sets. The underlying idea is to granulate numeric data using fuzzy sets and afterwards reveal and quantify relationships between these granules. This naturally impacts the dimensionality of any original dataset under discussion and provides with its nonlinear transformation (through the corresponding membership functions). These information granules give rise to the notion of associations -- multidimensional information granules. Being fuzzy relations, these constructs are *direction - free*. The directionality arises when one defines inputs and outputs and in this way confines himself to some sort of rules capturing a directional nature of main relationships within the data. Rules arising from associations may be in conflict. The essence of data is then captured via a granular signature regarded as a mixture of associations and rules.

Keywords information granulation, associations, relevance, linkage, granular signature, data mining

1 INTRODUCTORY COMMENTS

Intelligent Data Analysis (IDA) has emerged as a new and promising direction of research that intends to make sense of multivariable data. The agenda of IDA (as well as data mining) is very broad. IDA provides with an insight into the data by revealing main trends and relationships between variables. By analyzing data in the IDA environment, one derives useful patterns that can be refined and used afterwards in the form of detailed models. In this way, these patterns help make predictions and making decisions. In a nutshell, IDA attempts to operate at a level that is comfortable to the human user operating at the level of specificity that allows the user to build a global picture about the nature of relationships occurring within the data set.

The duality of the quantitative-qualitative approach is particularly appealing when dealing with complex systems. They are abstract to a very high extent and usually not guided by any specific laws of physics. Furthermore they exhibit a high level of variability that calls for a careful treatment. In particular, modeling has to be carried out at various levels of specificity. These levels give rise to models of varying level of specificity (granularity) by concentrating on and revealing different levels of details.

In this study, we pursue a concept of associations formed at the level of information granules [3][4][5] regarded as basic building blocks used in data analysis. Roughly speaking, associations are highly homogeneous and experimentally strongly justifiable information granules capturing the essence of given data at hand. To illustrate the concept, refer to Figure 1 illustrating various clouds of two-dimensional experimental data. When analyzing this data set from such perspective, we may easily identify several regions of concentration of data emerging in a form of some clusters (associations).

0-7803-6274-8/00/$10.00 © 2000 IEEE

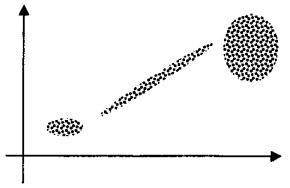

Figure 1. An example data set visualizing some well-formed regions of high density of data along areas of low density and poorly visible structure

As seen there, in some regions, the data points exhibit clear high density and take on some well-defined shape (say, hyperellipsoidal or single-line structure). In some other regions, there are few scattered points with no clear and strong dependency. Intuitively, one can delineate some highly populated regions and view them as highly representative to the entire data set. Subsequently, all such associations can be subject to further detailed analysis and form a basis for a construction of a detailed model quantifying specific dependencies between the variables. It becomes apparent that association analysis can be viewed as a preliminary step in data analysis that may be afterwards refined by building detailed models. In this sense, constructing data associations may serve as blueprint of further, more detailed and mathematically specific models (say, regression lines, correlation coefficients, neural networks and alike).

2 THE DESIGN OF GRANULAR ASSOCIATIONS

In this section, we define a concept of information granules and reveal how these are used in the design of associations. The proposed development methodology involves two key phases: (a) building basic information granules for each individual variable, and (b) forming associations with the use of these information granules.

2.1. Building information granules for individual variables

For each variable (measure), we specify a certain number of information granules. By information granules we mean a collection of elements that are collected together owing to their similarity or functional cohesiveness. Information granules exhibit a well-defined semantics and easily comprehended by humans and therefore could serve as useful descriptors of the problem. For instance, when talking about lines of code of a certain software product, it is convenient to talk about *small*, *medium*, and *large* size of code rather than confining to a single number, say 700K code. These terms like small, medium and alike are examples of information granules: they are easy to understand, highly descriptive, and handy when communicating findings about the given data set. Similarly, they are helpful in making design decisions. The way in which we proceed at the formal end may vary. Such information granules can be represented as sets, fuzzy sets, rough sets, shadowed sets, etc. Now an important question arises as to the determination of the fuzzy sets (information granules). First, we have to decide on the form of the membership functions that describe a way in which the membership grades vary over the space. The simplest model is a triangular or trapezoidal fuzzy set. While each fuzzy set comes with a transparent semantics, its parameters need to be adjusted and should reflect the nature of the experimental data these fuzzy sets have to granulate. Fuzzy equalization [2] forms a simple yet efficient algorithm of determining the parameters of the fuzzy sets. In a nutshell, the concept of fuzzy equalization originates from the idea of fuzzy events [1][3]. Consider a fuzzy set (A) defined over some universe of discourse **X**. In this universe, we have a family of experimental numeric data that are conveniently described by a certain probability density function (pdf) p(x). Then the probability of the fuzzy event (fuzzy set A) is computed in the form

$$P(A) = \int_{X} A(x)p(x)dx \quad (1)$$

Essentially, P(A) is just an expected value of A. Interestingly, the probability of the fuzzy event carries a straightforward interpretation. Fuzzy set can be viewed as meaningful if its probability P(A) is equal or exceeds a certain critical value α, namely $P(A) \geq \alpha$. If this holds, we say A is *experimentally* justified (valid). The monotonicity property holds: if $A \subset A'$ (that is the granularity of A is higher that the one of A') then the corresponding probabilities satisfy the monotonicity condition: $P(A) \leq P(A')$.

Now let us consider a family of fuzzy sets, say $\mathbf{A} = \{A_1, A_2, \ldots, A_c\}$ is given. To make each A_i meaningful (in the above sense of experimental justification), we request that there is an equal level of experimental evidence equal to 1/c, that is

$$P(A_i) = 1/c \quad (2)$$

According to the above equalization condition, fuzzy sets become more specific (detailed) in the

regions of **X** where the pdf attains some local maximum. On the other hand, in the areas of low values of pdf, we need fuzzy sets of broader support to gain sufficient experimental evidence (see Figure 2).

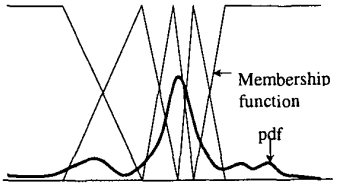

Figure 2. The idea of fuzzy equalization

2.2 Building associations as multidimensional experimental structures

Fuzzy sets defined in the individual coordinates (spaces) are generic components forming associations. Consider "n" variables in the data set. The fuzzy sets in each coordinate is denoted by A_1, A_2, ..., A_c (first variable), B_1, B_2, ..., B_p (second variable), etc. Formally speaking, an association **A** is a Cartesian product of any combination of the fuzzy sets for each variable. Formally, the association **A** comes in the form

$$\mathbf{A} = A_i \times B_j \tag{3}$$

The operation of combining the membership grades is realized using a t-norm. In particular, one may consider a minimum operation as one of the plausible options. This gives rise to the expression

$$\mathbf{A}(x, y) = (A_i \times B_j)(x, y) = \min(A_i(x), B_j(y)) \tag{4}$$

An illustration of the associations in the case of two variables is illustrated in Figure 3.

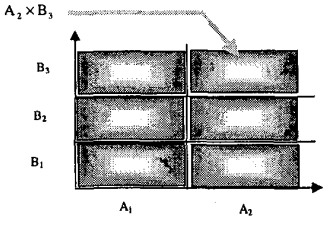

Figure 3. Associations for two dimensional case; note that the variables we granulated by two (A_1 and A_2) and three fuzzy sets (B_1, B_2, and B_3), respectively.

The association **A** is uniquely specified by a sequence of indexes identifying the coordinates of the fuzzy sets. Thus, we may characterize A by a sequence of indexes (i_1, i_2, ..., i_n) where i_k describes an index of the fuzzy set in the k-th coordinate (variable) contributing to this specific association. Say, for n = 3, a certain association **B** comes in the form **B** = (2, 4, 1).

In light of the introduced design procedure, several interesting features are worth emphasizing:
Associations are *relations* (more specifically, fuzzy relations) defined in the Cartesian product of the space of all variables. As relations, they do not assume any direction (meaning that we do not make any statement as to a possible implication between the software measures, say measure "a" implies some values of measure "b"). By being direction-free, the construct is far more general than any rule (that is a conditional statement of the form "if -- then"). In the sense of this lack of directionality, associations resemble correlation coefficient (or correlation matrices) that do not stipulate any implication between the variables. The main difference between correlation analysis and association analysis lies in the fact that the first deals with a global analysis while the latter is aimed at the analysis carried out at a local level of some well-defined and experimentally justified information granules.
Global analysis may be appealing but it could be also dangerous, especially when there are some "gaps" in a data set (and this could be quite visible in cases where there is no justification for continuity of underlying phenomena and the ensuing data). Association analysis helps avoid this pitfall. Figure 4 contrasts these two general approaches. The global analysis would support moving into a single regression model. The local association-based analysis promotes a series of highly justifiable local models built around individual associations and in this manner avoids drawing conclusions in the regions where data are almost nonexistent or very sparse. In the sequel, one may carry out a local type regression analysis.

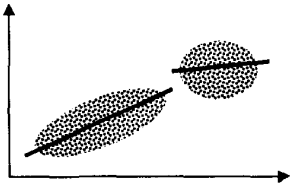

Figure 4. Local regression models constructed on the basis of associations

From the formal standpoint, associations are a better starting point of data analysis. Evidently, not all associations (relations) are functions. It could

well be that there are associations in the data but not necessarily functions. One should identify associations first and then start revealing functions (that are definitely constructed on the basis of the given associations)

3 BUILDING AN AGENDA OF ASSOCIATIONS

The number of all possible associations is tremendous. With "n" variables and "p" information granules (fuzzy sets or sets) defined for each of them, we end up with p^n possible associations. Only a small fraction of these is meaningful that is justified (supported) by the experimental evidence (data at hand). For a uniform distribution of data (which is usually not a case), the probability of data supporting each association is $\frac{1}{p^n}$. Most of the associations are void meaning that there is no data behind them. Naturally, a straightforward criterion to select meaningful associations would be to quantify a level of experimental evidence behind them. Two measures of experimental support are of interest here

σ-count For association **A**, see (8) we compute

$$\sigma(\mathbf{A}) = \sum_{k=1}^{N} \min(A_i(x_k), B_j(y_k))$$

cardinality This applies to the support of **A** and involves counting all elements of the data set falling under the support of **A**. No membership grades are used in this case. The computations follow the formula

$$\text{card}(\mathbf{A}) = \sum_{k=1}^{N} \min(\chi_{A_i}(x_k), \chi_{B_j}(y_k))$$

where χ_{A_i} and χ_{B_j} are the characteristic functions of the supports of the corresponding fuzzy sets, namely

$$\chi_{A_i}(x) = \begin{cases} 1 & \text{if } x \in \text{supp}(A_i) \\ 0 & \text{otherwise} \end{cases}$$

The most meaningful associations are combined in a form of an agenda, that is a collection of associations with the highest values of the cardinality and arranged in a decreasing order with respect to the cardinality value. The size of the agenda is selected in such a way so that all associations there cover a certain percentage of the overall data set (say, 70%). As emphasized, the number of all possible associations is high. A brute-force enumeration can work for 10 - 12 variables and 3 - 4 fuzzy sets in each variable. Quite quickly, this approach is not feasible. Fortunately, there is a simple solution that helps overcome the problem. A complete enumeration can be combined by a simultaneous pruning technique, which avoids exploring combinations of fuzzy sets not supporting enough data. The pruning is possible because the cardinality (and σ-count) decreases its value once more variables come into play. Moreover, at each stage of expanding the association, see Figure 5, some combinations are pruned and the related expansions are not pursued. This immensely reduces the number of possibilities that need to be investigated.

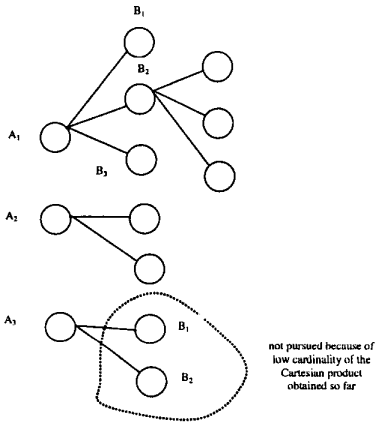

Figure 5. The pruning of associations in their process of successive development; note that at each level when adding one extra variable, only a portion of the most meaningful Cartesian products is retained

4 FROM ASSOCIATIONS TO RULES

Associations are direction-neutral local structures in which we do not distinguish between input (independent) and output (dependent) variables. This obviously leads to a significant level of generality of such constructs. In addition to being direction-free, the associations are structure-free. They do not confine themselves to any specific type of analytical relationships between the software measures. Treating associations, and the agenda, in particular, as a starting point for any detailed modeling pursuit, several main directions can be envisioned

(a) identifying rules
(b) constructing detailed parametric models for each association

In contrast to associations, rules are direction-sensitive structures. Variables have to be declared as independent or dependent. The independent variables come as a condition part of the rule. The dependent variables appear in the conclusion part. Once we have identified the variables, each association can be viewed as an individual rule

- if condition$_1$ *and* condition$_2$ *and* ... *and* condition$_p$ *then* conclusion$_1$ *and* conclusion$_2$ *and* conclusion$_r$

where r + p = n. In contrast to associations, rules cannot be treated individually but have to analyzed and utilized *en block*. Some rules resulting from the above transformation of the associations could be in conflict. Two rules are said to be in conflict if they have *identical* condition parts and have *different* conclusion parts.

5 REDUCTION OF RULEBASES

The rules come as a direct transformation of the associations. Their number could be reduced through some merging process. For instance two rules

- if A1 and B1 then C1
- if A2 and B1 then C1

can be easily merged into a single, more general rule in which one of the conditions is a disjunction of two information granules (assuming that A1 and A2 are two adjacent fuzzy sets)

-if (A1 *or* A2) and B1 then C1

The generalization of this nature reduces the number of all rules and could be automated to some extent. The crux of the optimization approach dwells on a well-known Quine-McCluskey reduction scheme commonly encountered in digital systems. To illustrate how this optimization is carried out, let us confine ourselves to a simple illustrative example. Consider only two variables and four fuzzy sets (information granules) defined in each of them, say A_1, A_2, A_3, and A_4 as well as B_1 to B_4. The rules are

-if A2 and B2 then C1
-if A2 and B3 then C1
-if A2 and B4 then C1

Each fuzzy set is coded in a binary fashion. The binary assignment is carried out in such a way that the neighboring fuzzy sets (say, small - medium, medium-large) are made adjacent on the Karnaugh map (K-map), say A_1 =00, A_2 = 01, A_3 =11, and A_4 =10. Using the above coding scheme, the original rules occupy single entries in the K-map.

The Quine-McCluskey method determines all prime implicants and thus helps us carry out the simplification (generalization) of the rules in an automatic fashion. The prime implicants lead to the reduced rules

- if A2 and (B2 or B3) then C1
- if A2 and (B3 or B4) then C1

Some further generalization is still possible (even though the original simplification method does not cope with this phenomenon). We can merge the two rules by generalizing the first condition in the following form

- if A2 and (B2 or B3 or B4) then C1

6 CONCLUSIONS

Intelligent data analysis (data mining) is definitely a human-centered endeavor with a high level of user/designer interaction. Granularity of information is a key conceptual and algorithmic notion that makes this interaction effective and leads to transparent results. Two fundamental modeling concepts - associations and rules have been introduced and discussed in detail along with a comparative analysis.

ACKNOWLEDGMENTS

The support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and ASERC (Alberta Software Engineering Research Consortium) is gratefully acknowledged.

7 REFERENCES

1. Pedrycz, W., F. Gomide (1998), *An Introduction to Fuzzy Sets: Analysis and Design*, MIT Press, Cambridge, MA.

2. Pedrycz, W., Fuzzy equalization in the construction of fuzzy sets, *Fuzzy Sets and Systems*, to appear.

3. Zadeh, L.A. (1979), Fuzzy sets and information granularity, In: M.M. Gupta, R.K. Ragade, R.R. Yager, eds., *Advances in Fuzzy Set Theory and Applications*, North Holland, Amsterdam, 3-18.

4. Zadeh, L.A. (1996), Fuzzy logic = computing with words, *IEEE Trans. on Fuzzy Systems*, vol. 4, 2, 103-111.

5. Zadeh, L.A. (1997), Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *Fuzzy Sets and Systems*, 90, 111-117.