SEARCH AND LINGUISTIC DESCRIPTION OF CONNECTED REGIONS IN QUANTITATIVE DATA

Riquelme J. and Toro M.

Facultad de Informática y Estadística
Avd. Reina Mercedes s/n 41012 SEVILLA (SPAIN)
E-mail: {riquelme|mtoro}@lsi.us.es

Abstract: The aim of this paper is to resume a great volume of quantitative knowledge in a qualitative model formed by linguistic rules. The initial information will be formed by the numeric and quantitative data that a real system spplies from its perfomance. First connected regions that have a similar behaviour will be found, and later every region will be described by means of linguistic terms. A transformation of the parameter space is proposed in order to reduce the number of regions, and so, the number of rules.

Keywords: qualitative analysis, classifiers, genetic algorithm, pattern recognition, dynamic systems.

1. INTRODUCTION.

Qualitative notion concerns to the general properties of the behaviour of a system. These properties are related to the geometric form of the behaviour and their global aspects. There is a tendency to name qualitative to what is not susceptible to have a quantitative treatment.

In the bibliography (Zadeh, 1973), the author suggested a linguistic analysis to avoid "the principle of incompatibility", according to this "as the complexity of a system increases, our ability to make precise and yet significant statements about its behavior disminishes until a threshold is reached beyond which precision and significance (or relevance) become almost mutually exclusive characteristics". This idea about linguistic analysis in relation to modelling is a viewpoint of the qualitative modelling.

According to (Kleer and Brown, 1984), the behaviour of a physical system can be discussed with exact values of its variables in every instant of time. This description is complete, but it fails in order to understand how the system works. The long-term job is to develop an alternative physics with simple concepts which are obtained in a simple way, but with a formal qualitative base. That is to say, reducing the quantitative accuracy in the description of the behaviours but keeping the important differences.

In another paper (Sugeno and Yasukawa, 1993), a qualitative model is derived from a fuzzy model using the linguistic approximation method. They have proposed the use of a fuzzy clustering method for the structure identification of a fuzzy model.

In general, we say that there is a series of knowledge clearly qualitative, for example, if some quantity is enclosed or will be increased without limit, if the integral of some function between two points is independent of the way, the existence of nonhyperbolic points in a system of differential equations, ... All that has been previously explained, is due to the interest of science and engineering by qualitative (no numeric) structures. These structures are the base of a great number of useful techniques to scientifics and engineers. In general qualitative concepts are not used consciously, but they are the base of many useful tools. The following task must be to automate this scientific and technical reasoning.

The aim of the paper is to resume a great volume of quantitative knowledge in a qualitative model formed by linguistic rules. First connected regions that have a similar behaviour will be found, and later every

region will be described by means of linguistic terms. A transformation of the parameter space is proposed in order to reduce the number of regions, and so, the number of rules. The objectives for reaching are:

1. To improve the accuracy of the classifier system.
2. To supply added information about the system, just as determination of the relevant parameters, discovery of relations between them, ...
3. To obtain a more simple qualitative model.

2. PROBLEM STATEMENT.

The initial information will be formed by the numeric and quantitative data that a real system supplies to its environment from its performance or behaviour. This information is made up by a database where each register is composed of two fields: a m-tuple of real values, that represents the co-ordinates of a point in a domain Ω_p, and an associated class o type to this point that takes values in a discrete finite space. Formally it can be represented by:

$$\{R_i\}_{i=1}^{N}=\{(p_i,C_i)/\ p_i\in\Omega_p\subset R^m,\ C_i\in T=\{T_1,,\ldots,T_c\}\}_{i=1}^{N}$$

In order to simplify the denomination of the tuples in Ω_p they are called parameters or features. The different values C_i are called type, class or label. These databases are denominated labelled.

Given a labelled database we want to search the regions in the coordinate space Ω_p that are connected and have a same label. Likewise the points that are inner to these regions and the points that are border or frontier.

In order to find these regions it can be applied techniques of supervised learning as well as the techniques based on the nearest neighbour norms (Dasarathy 1991), or methods that produce a partition in n-orthohedrons as the system C4.5 (Quinlan 1993). The first one has the advantage of adapting to regions without a defined form and the handicap of not producing rules. The other produces rules but only searches separated regions by hiperplanes. Our proposal joins the advantages of both of them, searching regions through a technique of nearness in the parameter space and later assigning its linguistic rules.

The proposed algorithm tries to discover the points that they are in the same region from the information of its types. It is presupposed that if a point and its neighbours are of a same type, it means that they are in the same region and its intermediate points will also be. This means that there will not be very "small" regions of measurement and if they exist the algorithm will not detect them. It will be understood by distance "small" if it is smaller than the maximum of the distance between a point and its neighbours.

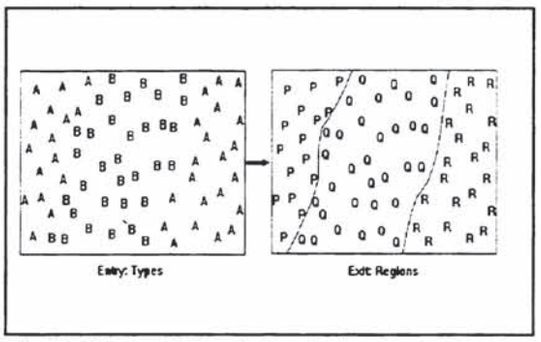

Fig. 1: Connected regions obtainment.

The figure 1 shows what is wished to obtain. On the left is the representation of a database in function of two labels that represent a class: A or B. On the right the database has been classified by regions: P, Q and R, the first two by points of type A and the last one of type B.

3. SEARCH OF CONNECTED REGIONS.

In order to find the connected regions from the database it is proposed the following procedure:

Entry: A labeled database.

Exit: Assignment to each database point of the region to the one which it belongs to, if it inner of frontier, and in this case with what regions.

Notation:
Region: Connected set of points with the same type.
Visited: Point that has assigned a region.
Classified: Point that has visited all its neighbours.

Algorithm:
1. All points in the database are indicated as not classified and not visited.

2. While there are not visited points.
   2.1 Let Z a number of region noone has used.
   2.2 Choose a point not visited and assigne it to the region Z. It is marked as visited.
   2.3 While there are visited points and not classified
      2.3.1 Choose a visited and not classified point P, that is has been already assigned to a region, let this Z_P.
      2.3.2 Let Q_i i=1..N its neighbours, i.e. the nearest points to P in the database according to some metrics.
      2.3.3 For each Q_i do:
         2.3.3.1 If Q_i is not visited and his type is of

region Z_P, being marked as visited.

2.3.3.2 If Q_i is not visited and his label is different of Z_P then choose a new value of region Z_Q which is assigned to Q_i and is marked as visisted. P is noted as frontier between Z_P and Z_Q.

2.3.3.3 If Q_i is visited and his region Z_Q is different of Z_P and its types coincide, then the regions Z_P and Z_Q are the same, therefore substitute Z_Q by Z_P in all database.

2.3.3.4 If Q_i is visited and his region Z_Q is different of Z_P and the types of P and Q_i are different, then both are indicated as frontier points between the zones Z_Q and Z_P.

2.3.4 If all Q_i are of the same region that P, to indicate P as inner.

2.3.5 To label P as classified.

At the end of this process the available information is structured in the following way:

1. The number of regions of the space that show a same type.
2. It is known to what zone each point belongs to and if it is inner or frontier to this region.
3. If a point is frontier, it is known with which zones it limits, and therefore the regions that surround it.

4. TRANSFORMATION OF THE PARAMETER SPACE

If the system to classify is very complex, the number of regions of the previous algorithm can result very high. This would imply that the number of linguistic rules would be excessive as to make them easily intelligible. A possibility to avoid this is to transform the parameters space so that the number of resulting connected regions is sensibly smaller.

It will be understood by a transformation of the space $\Omega_p \subseteq \mathbb{R}^m$ in $\Omega_q \subseteq \mathbb{R}^k$, to a set of k functions $\varphi_i : \mathbb{R}^m \longrightarrow \mathbb{R}$ such that

$$(x_1,\ldots,x_m) \in \Omega_p \rightarrow (\varphi_1(x_1,\ldots x_m),\ldots,\varphi_k(x_1,\ldots x_m)) \in \Omega_q$$

If the dimension of Ω_q is greater than of Ω_p, it is being added new characteristic to the classification space, this would imply an increase in the possibilities of finding a better classifier. However, an increase in the number of attributes also implies a greater computational difficulty for this search. Furthermore the classifier will be more complex and then more difficult to understand from a qualitative point of view. On the contrary, if the dimension of Ω_q is less than that of the original space, the previous characteristics are inverted, with a deterioration of the possibilities to find a better classifier, but possibly easier to understand.

As the principal objetive of this paper is to find simple models that can explain a system in a qualitative way, it is necessary to limit the complexity of the functions. Thereby, the search of k functions φ_i has been limited to arithmetic operations between only two parameters, permiting the possibility of the fact that some parameters stay as coordinated upon going from Ω_p to Ω_q.

4.1 Implementation: function to minimize.

The objective is to minimize the number of regions in those which the parameters space is divided and to obtain an error rate that is possibly smaller upon classifying a new case.

In order to obtain a error rate in the resulting classifier of the previous algorithm it has been used the technique "leaving one out", i.e is eliminated a point from the database, it is applied the previous algorithm and is compared the classification obtained for the point eliminated with the real. If this action is repeated for all the points of the database, it is obtained an estimator of the error rate of the classifier or apparent error rate.

In this way, the function to minimize would be obtained either by maintaining constant the apparent error rate and minimizing the number of regions, or to maintain constant the number of regions and to minimize the error rate. Below the two possible pseudocodes are exposed:

```
1) If the number of regions > prefixed value
     Function = number of regions + 0.01 apparent
                error rate
   else
     If apparent error rate > number of regions
       Function = apparent error rate
     else
       Function = apparent error rate + number of
                  regions
     Endif
   Endif

2) If apparent error rate > prefixed value
     Function = apparent error rate + 0.01 *
                number of regions
   else
     If number of regions > apparent error rate
       Function = number of regions + 0.01 *
                  apparent error rate
     else
       Function = apparent error rate + number of
                  regions
     Endif
   Endif
```

4.2 Implementation: optimization method.

In order to minimize the function above defined, an evolutionary algorithm has been used (Michalewicz, 1994). An individual of the population represents a possible transformation of the parameter space. Each function φ_i is defined as a combination of two parameters operated by an arithmetic operator. Therefore, three integers can define a function. The values of the arithmetic operators can be 1 (sum), 2 (difference), 3 (multiplication) and 4 (division). Besides one value of 0 represents that the function is defined only as the second parameter.

Thus, for example, in a three dimension space the transformation given by:

$$\varphi:\Omega_p \rightarrow \Omega_q$$
$$(x,y,z)\rightarrow(x/y,x-z,z)$$

would be represented by the secuence

1	4	2	1	2	3	#	0	3

The dimension of the transformed space must be prefixed. In this way, all individuals will represent the same complexity in the transformation.

5. OBTAINING OF A LINGUISTIC MODEL.

5.1 Definition of linguistic terms.

For the qualitative model concept to be understood in terms of a linguistic model, it is necessary to be able to transform the previous obtained information. The techniques to obtain a qualitative information of a spatial arrangement, expressed in the previous paragraphs can be in a way simple converted automaticallyn in a model based on linguistic terms.

To express through a linguistic term a value range of a variable, is a relatively easy task. Only it must take into account two considerations: the first one is that the number of terms that are defined must be enough to attempt to cover most of linguistic nuances of the possible value ranges. The second is that the number should not be excessive so as to complicate the understanding of the model. To choose the number of terms together with the nomenclature of these is, then, the only difficulty of this approximation.

In this case the linguistic model would follow the following grammar:

```
<model>          ::= <list_rules>
<list_rules>     ::= <list_rules> <rule>
                   | <rule>
<rule>           ::= IF <list_premises> THEN
                                        <conclusion>
<list_premises>::= <list_premises> AND
                                        <premise>
                   | <premise>
<premise>        ::= PARAMETER IS <list_terms>
<list_terms>     ::= <list_terms> OR <TERM>
                   | <TERM>
<conclusion>     ::= <CLASS>
```

For the assignment between a set of values and what in the grammar has been expressed as <TERM> will have to be followed then steps:

1) The parameter p is supposed to have a range of values that remains defined by the minimal and maximum values of that parameter in the database. Those values will be m and M, respectively.

2) The interval of values is supposed to be divided into L linguistic terms, that may have an equivalence in L ranges of equal length values for the parameter p.

3) The central values of those ranges would come to consider the values succession:

$$m+\frac{M-m}{2L},\quad m+\frac{3(M-m)}{2L},\quad \ldots\ ,\quad m+\frac{(2L-1)(M-m)}{2L}$$

4) From these central values each range would have a value of $\leq(M\text{-}m)/2L$, that is, the first term would come defined by the range $[m,m+(M\text{-}m)/L]$, the second term would give value to the numbers of the range $[m+(M\text{-}m)/L,\ m+2(M\text{-}m)/L]$, and so on until $[m+(L\text{-}1)(M\text{-}m)/L,M]$.

5.2 Linguistic terms assignment to a region.

Given a point subset of the database that has the same type and the property of forming a connected region in Ω_p, it is considered now to assign a decision rule in linguistic terms that represents it.

For this, it is necessary to find the maximum value C and the minimum c for each parameter in the points that form the connected regions. However, to avoid the possible extreme values it seems more appropiate to use adecuate percentiles that for the value c goes from P_{15} to P_{25}, and for C correspond P_{75} and P_{85} .

Once the values C and c have been determined, for the assignment of a rule to each region the following methodology is proposed:

Notation:
C : if it exists superior bound.
c : if it exists inferior bound.
L : number of linguistic terms.
v_i $1\leq i\leq L$: central value of the ith term.

t_i $1 \le i \le L$: ith linguistic term.
TERM: linguistic term assigned.

Method:
If $\exists$ c, let r with $2 \le r \le L$ / $v_{r-1} < c$ & $v_r > c$ else $r=1$.
If $\exists$ C, let s with $1 \le s \le L\text{-}1$ / $v_s < C$ & $v_{s+1} > C$
else $s=L$

TERM is calculated through the equation:

$$TERMINO = \bigcup_{i=r}^{s} t_i$$

where the union symbol indicates the conjunction of the terms by the logical operator or.

It must be taken into account, that if the rule affects a parameter resulted of a transformation, the minimum and maximum values (m and M) of this parameter can be calculated in functions of the original according to the efected transformation.

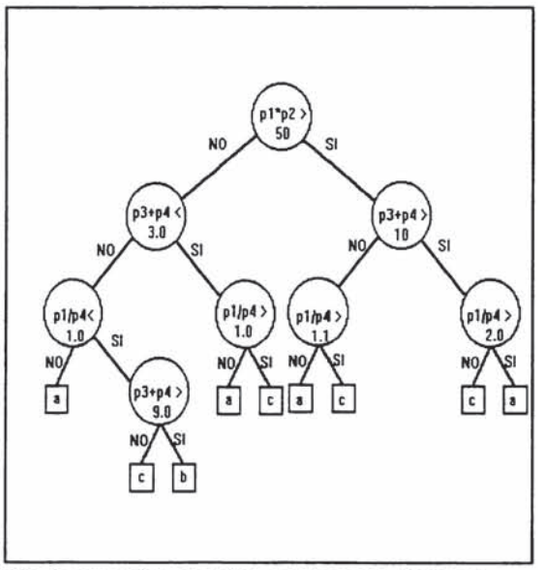

Fig.2: Tree for creation on purpose database.

6. APPLICATION.

In this section the methodology proposed is applied to some databases.

6.1 On purpose database.

It has been built a on purpose test file. The database was built with 200 records and 10 real parameters each record, so that the parameters are uniformly distributed of random way in the interval [0,10]. Each record was classified in one of three possible types attending to the values of 3 hidden characteristics, that is, not present in the database, that were obtained through simple arithmetic operations with the first four original parameters.

The algorithm that gives the type of the values of the parameters is expressed through the tree of the figure 2.

The kindness of the technique can be seen from several points of view:

1) A better classifier is obtained: the original file was classifying wrong by 9 records (a 4.5%). With the new characteristic found, the apparent error rate was of 0%.
2) The classifier is much more simple: originally the number of found regions was 21, while thereinafter come to 7, what means that the set of rules is reduced to a third of the original.
3) Possibility of more than a solution: upon executing the method more than once, or even an alone time, but with more generations, other possible combinations are obtained, in addition to the one which originally gave place to the file. Concretely, the original operations were $p_0 * p_1$, $p_2 + p_3$ and p_0 / p_3, being found furthermore combinations as $p_0 + p_1$, $p_2 + p_3$ and p_3 / p_0 or $p_0 + p_1$, $p_2 + p_3$ and $p_3 - p_0$, with the same mistake rates. This result facilitates the fact of finding explanations with sense for the different solutions that are found.
4) The parameters that actually have relevancy in the classification of the file have been found without no type of difficulty or error. In none of the reached solutions was included the fictitious parameters introduced in the training file.

If it is wanted a linguistic model, though losing precision and obtainning understanding, it can be appealed to four linguistic terms: small, small-middle, middle-large and large:

If p_1 is middle-large and
  p_2 is small-middle and
  p_3 is small-middle or middle-large and
  p_4 is small or small-middle then type a.

If p_1 is small-middle and
  p_2 is small-middle or middle-large and
  p_3 is small-middle or middle-large and
  p_4 is large then type b.

If p_1 is small-middle or middle-large and
  p_2 is middle-large or large and
  p_3 is small or small-middle and
  p_4 is small-middle or middle-large then type c.

It should be emphasized that, in spite of the fact that the data base was created in function of a discrimination based on arithmetic operations of the parameters, it is possible to build a linguistic model that does not possess intersection between their premises. Concretely the parameter p_1 discriminates

the types a and b, the parameter p_2 the types a and c and the parameter p_4 the types a and b and also b and c.

6.2 IRIS file.

A tipical database example for classification problems is IRISDATA since was proposed in (Fisher, 1936). The algorithm produces a set of 6 rules with two errors. Accomplishing a coordinates transformation, the apparent error can to be equal to 0 without need of increasing the number of rules. If a transformation of the space is effected, are obtained various results in function whether is wished to obtain few rules or a error smaller. Thus, with only a new parameter obtained from the result of multiplying the third and fourth original parameter, is obtained a system that only needs a rule for each type with a error rate of the 2.7%. Concretely the rules are:

If $p_3*p_4 <= 0.96$ then type 1
If $0.96 < p_3*p_4 <= 7.35$ then type 2
If $p_3*p_4 > 7.35$ then type 3

If the number of new characteristic is increased to two, the number of rules is duplicated, but the apparent error rate can decrease until 0% with parameters as p_2/p_3 and p_3*p_4. As can be observed, the new characteristic p_3*p_4 is repeated; this has an explanation in the meaning of the parameters p_3 and p_4 that they are respectively, the length and the width of the petals of the flowers, after some form p_3*p_4 is approximating the surface of the petal.

The values for the parameter p_3 and p_4 are in the interval [0,5], then, in principle, the possible range of values for p_3*p_4 be in [0,25]. However, the real distribution of p_3*p_4 is in the interval [0,15.9] and almost 80% in [0,10]. Therefore, the thresholds 0.96 and 7.35, have more sense on this last interval. The model would be:

If p_3*p_4 is very small then type 1
If p_3*p_4 is small or small-middle or middle or middle-large then type 2
If p_3*p_4 is large or very large then type 3

6.3 Analysis of a dynamical system.

This technique can be applied to the analysis of the behaviour in stationary regime of a dynamical system in function of its parameters. For this, it should be to obtain a database that relates the values of the parameters of the system and the corresponding attractor. For more details it can be seen (Toro et al 1991; Riquelme, 1996).

The chosen system is an ecological system proposed in (Aracil and Toro 1988) with 6 parameters (a_1, a_2, b_1, b_2, c_1, c_2) and three possible atracttors: equilibrium, limit cycle and chaos. Upon applying to the database that result of the system simulation the proposed algorithm were found 15 regions with an apparent error rate of about 25%. This number gives an idea of the difficulty that the system presents to classify its atracttor.

Several transformations are obtained according to various objectives:

1. To separate the three types of atracttors, with only four parameters (c_2/b_2, b_2*a_1, c_2+a_2, c_2*a_1) the number of regions is reduced to 7 (2, 4 y 1 of each atracttor) and the error rate to the 12%.
2. To separate the limit cycle and chaotic atracttors the transformation c_2+b_2, a_1/c_1, c_2+a_2, c_1+a_1 procures an error rate of 8.9% with 6 regions (4 y 2).
3. Finally, to separate equilibrium from the other atracttors, it is found that an only parameter (b_2/a_2) is capable of making it in two regions with an error rate of the 3% thought lets a 7.3% the points without classifying. The linguistic model in this case is:

If b_2/a_2 is very small then not equilibrium point

Clearly, this rule would be decomposed in:

If b_2 is small and a_2 is large then not equilibrium point.

REFERENCES

Aracil J. and M. Toro, (1988). Bifurcations and chaos in a predator-prey-food ecological model. In *Proc. of the Conf. on Synergetics, order and chaos*, pp. 448-459. World Scientific, Singapore.

Dasarathy B.V. (1991). *Nearest Neighbour (NN) Norms: NN pattern classification techniques*, IEEE Computer Society Press, Los Alamitos, CA.

Fisher, R. (1936) The use of multiple measurements in Taxonomic problems. *Annals of Eugenics*, **7**, 179-188.

Kleer J. de and J.S. Brown (1984). A qualitative physics based on confluences. *Artificial Intelligence*, **24**, 7-83.

Michalewicz Z. (1994). *Genetic algorithms + data structures = evolution programs*. Springer-Verlag, New York.

Quinlan J.R. (1993). *C4.5: Programs for machine learning*, Morgan Kaufmann Publiher, San Mateo, CA.

Riquelme J., (1996). *Obtención de información cualitativa a partir de datos cuantitativos:*

aplicación al análisis cualitativo de sistemas complejos, Tesis Doctoral, U. de Sevilla.

Sugeno, M. and T. Yasukawa, (1993). A fuzzy-logic-based approach to qualitative modeling, *IEEE Transactions on fuzzy systems*, **1(1)**, 7-31.

Toro, M., J. Riquelme and J. Aracil (1992). Classifying system behaviour model by statistical search in the parameter space. In: *Proc. of the European Simulation Multiconference*, (P. Geril, Ed.), pp. 181-185, SCSI.

Zadeh, L.A. (1973). Outline of a new approach to the analysis of complex systems and decision processes, *IEEE Trans. on Systems, Man and Cybernetic*, **3**, 28-44.