FID3: Fuzzy Induction Decision Tree

Julio Rives

Area Tecnológica Tratamiento de la Información
Inisel
Torrejón de Ardoz, Apdo. Correos 294 , Madrid 28850, Spain

Abstract

An overview of FID3, a technique of fuzzy learning, is presented. The organization of the paper is as follows: first, some basic ideas of uncertainty and information are reviewed; second, a measure of the dependence degree among the variables involved in a possibility distribution is described which suggests a method to select the most informative attributes of a training base; next, a measure of the vagueness of fuzzy sets is defined which suggests a new method of simplifying concepts; finally, the author highlights some FID3 enhancements from which the reader can comprehend its potential advantages.

Introduction

A technique of machine learning based on the fuzzy sets theory, the possibility theory, and classic induction algorithms is elaborated; therefore, the area covered by this article is in the field of the knowledge acquisition from fuzzy-possibilistic domains.

The technique consists of two stages. The first one is the generalization stage, which focuses on common features in the set of instances and tries to develop a new rule or concept. The second one is the abstraction stage, which generates a manageable concept from a multifold conclusion by the elimination of irrelevant attributes. In both phases, concepts are learnt within an informational framework which is presented as formally as possible.

What is intended is to demonstrate a method by which unknown elements of a fuzzy set can be determined on the basis of generalization plus abstraction from a sample of training examples. The learning process could be seen as an alternative approach to "Zadeh's interpolation" based on classical system analysis [16] and "Ruspini's extrapolation" between possible worlds based on similarity functions [10].

The approach presented improves induction machine learning techniques operating in crisp domains [7; 8] since: first, FID3 makes the training examples more user-friendly, second, FID3 measures the amount of information the training base is able to transmit [1] enabling us to select the most informative attributes, and third, FID3 measures the fuzziness of the conclusions learnt and enables us to select the most distinguishable ones. Furthermore, FID3 also copes with typical problems of induction studied by Bratko and Cestnik [2], Niblett [6], and Utgoff [13] such as binarization of attribute values, noisy domains, dependences among the input attributes, and incrementality.

Fundamentals of uncertainty and information

Two categories of uncertainty arise from daily life; they are captured quite well by the terms ambiguity and vagueness. Broadly speaking, ambiguity is associated with situations in which the choice between several alternatives is left unspecified, whereas vagueness represents the difficulty of making sharp distinctions in a world [3]. Some concepts akin to ambiguity are generality, variety and divergence; some concepts related to vagueness are haziness, unclearness, and indistinctiviness.

Given a domain, the lack of information makes you hesitate about the meaning of a concept; for example, if a girl tells you "I like tall men", you do not exactly know whether or not she likes men who are 2.20 m. tall, ... , 1.90 m. tall, 1.80 m. tall ... (at best, you may know the corresponding possibilities), so you get a one-to-many relation. But even after you acquire information, some uncertainty may remain. Suppose you ask her: "I am 1.85 m tall, am I a tall man?", and she answers: "you are more or less tall"; at this point, you get confused because even though you have been given a specific response you cannot delimit her reply by precise boundaries, so it is completely useless.

The situation resembles psychological human processes from the viewpoint of the intention and intension [5]: on one hand, ambiguity (or nonspecificity) is in the field of the intention and affects our decisions, and on the other hand, vagueness (or fuzziness) is in the field of the intension and affects our state.

In connection to this, the phases of generalization and abstraction imply minimization of nonspecificity and fuzziness respectively. Taking uncertainty as the complementary concept of information (I will use the complement to 1), our objective becomes a purpose of specificity and distinguishability maximization.

TH0334-3/90/0000/0457$01.00 © 1990 IEEE

The uniqueness of the U-uncertainty as a possibilistic measure of nonspecificity, and the properties of the Fuzziness as a fuzzy measure of vagueness, leads us to the principles of minimum nonspecificity and minimum fuzziness, possibilistic-fuzzy versions of the most fundamental principle of simplification [3]: "When we want to reduce the descriptive complexity of a system and there are several ways of doing so, we should select that way for which the increase of relevant uncertainty is minimal (or the increase of relevant information is maximal)."

Selecting the most specific attribute

The meaning of the above basic principle has been obviously misunderstood in specialized literature due to overlooking the keyword RELEVANT [4]. To put into practice the complete significance of the principle of simplification, it would be a logical attitude to look at the WHOLE training base as a system able to transmit relevant information. Given the sets A0, A1, ..., AN we can define the function called "information transmission":

$$T(A0,A1,...,AN) = U(A0) + U(A1) + ... + U(AN) - U(A0,A1,...,AN),$$

where U(A0), U(A1), ..., U(AN) are simple U-uncertainties based on marginal possibility distributions, and where U(A0,A1,...,AN) is the joint U-uncertainty of the variables involved. T is used as a measure of the strength of constraint among elements of A0, A1, ..., AN [1; 3]. The more interactive, i.e., the more dependent on each other the variables involved are, the higher the value of T is. When the sets are noninteractive, we get T(A0,A1,...,AN) = 0; otherwise, T(A0,A1,...,AN) > 0. Thereby, from this point and on, we will consider T as the only function able to give us a measure of how much potential information is contained in our base of examples. From the definition of T (when using averaged U-uncertainties 0 =< U =<1), we can deduce its lower and upper bounds: 0 =< T =< N + 1, where N is the number of input attributes of the base of examples, so T can be averaged by defining the ratio T(A0,A1,...,AN)/(N+1). To be precise, a quasi-possibilistic distribution is obtained because all the marginal possibilities are not present (remember that only a sample of cases is available), and therefore, what we are going to use in practice is the "available averaged information transmission Î". As an exercise, the calculation of Î(E) follows:

U(Height) = U(tall,short) = U(1,1) = 1,
U(Weight) = U(fat,average,thin) = U(1,1,.67) = .878,
U(AGe) = U(pantaloon,justice,soldier,lover) =
= U(1,.89,1,.78) = .922,
U(APpearance) = U(good,normal,bad) = U(1,.78,1) =.919,
U(H,W,AG,AP) = U(.78,.33,...,.11,1,.22,.33,.55) =.645,
T(H,W,AG,AP) = U(H) + U(W) + U(AG) + U(AP) -
- U(H,W,A,AP) = 3.074,
Î(E) = I/4 = .768,

where U(v1,v2,...) is the abbreviation of U(poss(v1), poss(v2), ...), and where poss(v1) is the maximum marginal possibility through the examples having v1 as an attribute value.

Example	Height	Weight	AGe	bad APp	norm. APp	good APp
1	tall	fat	lover	.78	.33	0
2	short	thin	soldier	.55	.44	.11
3	short	average	justice	.11	.78	.22
4	tall	fat	pantaloon	1	.11	0
5	tall	average	justice	.11	.11	.89
6	tall	thin	pantaloon	.22	.22	.67
7	short	average	soldier	.11	0	1
8	short	fat	lover	.55	.33	.22
9	tall	average	soldier	0	.11	1
10	tall	thin	lover	.22	.33	.55
Poss.				1	.78	1

Table 1. Initial training base E: quasi-possibilistic distribution. Input attributes: Height {tall, short}, Weight {fat, average, thin}, AGe {pantaloon, justice, soldier, lover*}; output attribute: APpearance {good, normal, bad}.

* "AS YOU LIKE IT" (by W. Shakespeare) Act II, Scene 7. One man in his time plays many parts, his acts being seven ages: the infant, the schoolboy, the lover, the soldier, the justice, the pantaloon, and the second childishness.

Example	Height	Weight	AGe	bad APp	norm. APp	good APp
4	tall	fat	pantaloon	1	.11	0
6	tall	thin	pantaloon	.22	.22	.67
Poss.				1	.22	.67

Table 2. Training-base: $E_{pantaloon}$; Poss(pantaloon): 1; Decomposed-by: Age; $\hat{I}(E_{pantaloon}) = I/3 = .2$

Example	Height	Weight	AGe	bad APp	norm. APp	good APp
3	short	average	justice	.11	.78	.22
5	tall	average	justice	.11	.11	.89
Poss.				.11	.78	.89

Table 3. Training-base: $E_{justice}$; Poss(justice): .89; Decomposed-by: Age; $\hat{I}(E_{justice}) = .344$

Example	Height	Weight	AGe	bad APp	norm. APp	good APp
2	short	thin	soldier	.55	.44	.11
7	short	average	soldier	.11	0	1
9	tall	average	soldier	0	.11	1
Poss.				.55	.44	1

Table 4. Training-base: $E_{soldier}$; Poss(soldier): 1; Decomposed-by: Age; $\hat{I}(E_{soldier}) = .451$

Example	Height	Weight	AGe	bad APp	norm. APp	good APp
1	tall	fat	lover	.78	.33	0
8	short	fat	lover	.55	.33	.22
10	tall	thin	lover	.22	.33	.55
Poss.				.78	.33	.55

Table 5. Training-base: E_{lover}; Poss(lover): .78; Decomposed-by: Age; $\hat{I}(E_{lover}) = .409$

Starting with the training base E, let us see how FID3 generalizes. Known the value of $\hat{I}(E)$, the first attribute must be selected which will be the root of our decision tree. This is done by measuring increments $\partial\hat{I}$ before and after the selection of the input attributes, and maximizing this value over them to acquire as much relevant information as possible from E.

In order to know the value of $\hat{I}$ after the selection of an attribute, we decompose E into new quasi-possibilistic distributions called partitions. Each partition is obtained from its mother by collecting adequate examples according to the attribute being inspected. For example, $\hat{I}(AG)$ must be the possibility of offering information as a function of the transmissible information of the partitions $\hat{I}(E_{pantaloon})$, $\hat{I}(E_{justice})$, ... (see tables 2, 3, 4 and 5), and we axiomatically use the function maximum. But the transmissible information of the obtained partitions is conditioned by the possibility of selecting their corresponding values, so that we use min(poss(pantaloon), $\hat{I}(E_{pantaloon})$), min(poss(justice), $\hat{I}(E_{justice})$), ... as chunks of possible information:

$$\begin{aligned}\hat{I}(AG) = \max\ (\ & \min(poss(pantaloon),\hat{I}(E_{pantaloon})),\\ & \min(poss(justice),\hat{I}(E_{justice})),\\ & \min(poss(soldier),\hat{I}(E_{soldier})),\\ & \min(poss(lover),\hat{I}(E_{lover}))\\ ) = (&.2,.344,.451,.409\ ) = .451\end{aligned}$$

Note that in spite of having used the minimum, another T-norm can be employed. We know the possibility of a pantaloon, a justice ... appearing, and we know the possibility of transmitting information $\hat{I}(E_{pantaloon})$, $\hat{I}(E_{justice})$ ... given the fact of a pantaloon, a justice ... having appeared, so we can apply the following inferential expression [10]:

$$Poss(y) = \sup_{x} [\ T_i\ (Poss(y/x)\ .\ Poss(x)\ )\ ],$$

where $x \in$ AGe, $y \in \{\ \hat{I}(E_{pantaloon}), \hat{I}(E_{justice}), \hat{I}(E_{soldier}), \hat{I}(E_{lover})\ \}$, and T_i is any T-norm.

In accordance with our settings, the final results corresponding to H, W, and AG are:

$\partial\hat{I}(E_{AG}) = \hat{I}(E) - \hat{I}(AG) = .317,$
$\partial\hat{I}(E_{H}) = .116,$
$\partial\hat{I}(E_{W}) = .232.$

As a result, we claim "AGe" as the attribute that contributes most to the strength of the relationship through the variables involved. It is remarkable the fact that $\hat{I}$ also takes into account the dependencies among the input attributes of the training base which affect the total information E is capable of offering. Such a property of $\hat{I}$ marks the difference between classic inductive machine learning techniques like ID3 and the FID3 approach: the former gets "the least deep tree" whereas the latter gets "the most informative tree".

This is the generalization phase of the algorithm which reduces intentional uncertainty: taking E as input, FID3 learns four rules, one per value of AGe; for example, "if age of man_X is lover then appearance of man_X will be bad with poss .78, normal with poss .33, and good with poss .55" (see table 5).

Selecting distinguishable conclusions

Before continuing to generate the fuzzy-possibilistic decision tree it is worth pausing to validate the type of information obtained. Notice that once we know the age of "man_X", the generalization allows us to give the possibility of "man_X" being good, normal and bad. The author claims that these are too many data to transmit: there is redundant and irrelevant information.

Due to the fact "bad"," normal" and "good" appearance being potential fuzzy sets, it would be logical to select the most distinguishable one to reduce the descriptive complexity of the conclusion (we strictly apply the principle of simplification). Trillas [11; 12] and Yager [14; 15] proposed that the relevant fuzziness F of a fuzzy set could be expressed in terms of the lack of distinction between the set and its complement:

$$F_C(A) = |A| - \sum_{a \in A} | \mu(a) - c(\mu(a)) | ,$$

where A is a fuzzy set, $\mu(a)$ is the membership degree of the element a, $c(\mu(a))$ is the membership degree of the complement of a, and $|A|$ is the cardinality of A.

From this function we can get a measure of relevant distinguishability D by first, taking c as the complement to 1, second, defining the ratio F(A)/|A| in order to average fuzziness, and third, thinking of uncertainty and information as complementary concepts (we again use the complement to 1):

$$D(A) = (\sum_{a \in A} |2* \mu(a) - 1|) / |A| .$$

The author wishes to make the following conjecture here: if we require for a measure of relevant fuzziness that

$$U(1,.5) = F_C(1,.5) = .5,$$

then the unique function of relevant distinguishability is D. This desideratum means that a proposition from which it cannot be affirmed whether it is general or specific, in some sense is neither vague nor crisp, i.e., semispecificity implies semiclearness. In other words, we make the logical demand that, for some fuzzy set, semiambiguity and semivagueness coexist.

When calculating distinguishability of fuzzy sets taken out from training bases (indeed quasi-fuzzy sets), D is called "available averaged distinguishability", $\mu(a)$ is the membership degree of an available element a, and $|A|$ is the number of examples of the training base.

This is the abstraction phase of FID3 in which the conclusion is left as sharp as possible, thereby reducing intensional uncertainty. For example, "if the age of man_X is lover ... " then we get:

D(bad) = .407,
D(normal) = .34,
D(good) = (|2*0 - 1|+|2*.22 - 1|+|2*.55 - 1|) / 3 = .553,

and " ... then the appearance of man_X will be good with possibility .55", omitting explicitly information concerning the possibility of appearing normal or bad.

Summing up the FID3 algorithm.

Here is a sample of how FID3 works. Let us suppose we have a set E of examples. Each example consists of NA input attributes, each attribute bound to its respective value, and C fields containing the output classes. In order to simplify the analysis, let us suppose that the variables involved in E are nominal.

1- The amount of transmissible information $\hat{I}(E)$ is calculated.

2- Starting with a root node, FID3 generates the decision tree by selecting attributes to branch the tree, which is partitioned into NVAj subtrees, where NVAj is the number of nominal values of the input attribute A_j (1=< j =<NA). Let A_{jk} correspond to the kth value of A_j. The possible transmissible information after selecting A_j is

$$Poss(\hat{I}(A_j)) = \max_k [\min (Poss(\hat{I}(A_j)/A_{jk}), Poss(A_{jk}))].$$

FID3 chooses the attribute that maximizes the gaining of relevant information $\partial\hat{I}$, where $\partial\hat{I}(E_{Aj}) = \hat{I}(E) - \hat{I}(A_j)$.

3- Let A_{sk} correspond to the kth value of the selected attribute A_s, then FID3 selects the most distinguishable output class C_{skc} for each partition E_{sk}, by maximizing $D(C_{skc})$, where 1=<c=<C.

4- FID3 repeats the last three steps for each resulting partition until all the attributes have been selected.

The fuzzy-possibilistic decision tree generated from E by FID3 is shown in the figure 1.

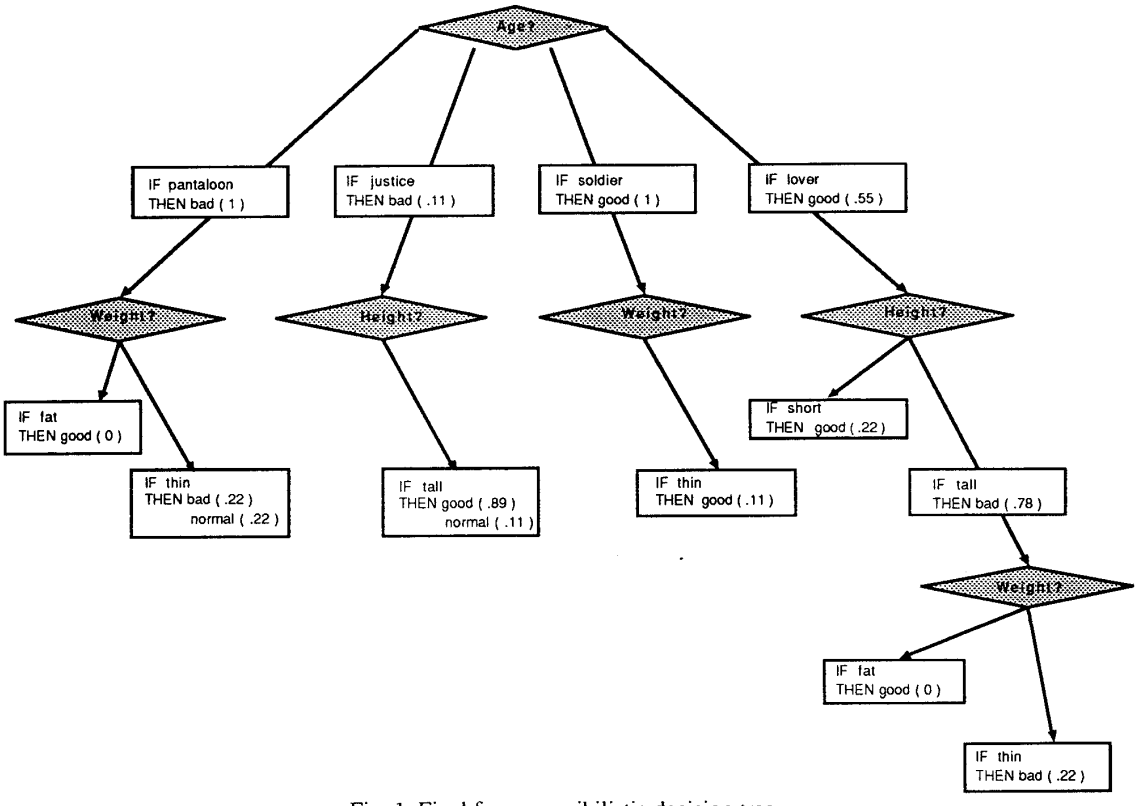

Fig. 1. Final fuzzy-possibilistic decision tree.

Concluding remarks

It is important to emphasise some properties of the function Î. First, its value denotes a reference from which we can decide whether E is valid as a training base or not. Furthermore, FID3 enables the user to know whether to include a branch of the decision tree as a new rule or not by managing increments of Î in relation to the required threshold: even though one attribute is the most informative, we may prefer not to include it if the possible responses do not represent an appreciable increment of relevant information from the current state to the next. Moreover, Î can be used to fuse values of an input or output attribute in clusters (notice that Î is a symmetrical function and does not distinguish between input and output attributes) so that the initial training base is simplified [9].

The approach expressed is able to solve practically all the problems that ID3 has posed. A stream of learning instances could be managed by slightly modifying an incremental algorithm. For example, ID5R [13] could substitute the positive and negative counts on the decision tree nodes for pairs [possibility of positive, possibility of negative]. Noisy domains [2; 6] are not a problem for FID3 because lost or wrong data hardly change the possibilities of the output classes and, therefore, the information transmission of the training base stays similar. Obstacles as continuous attributes and binarization must be also treated by looking for clusters and maximizing Î over them, so that as much relevant information as possible is preserved.

Some other disadvantages of the ID3 algorithm may be also solved. For example, FID3 is able to take structural relations as inputs, and it copes with the practical problem of, through all the possible operators on an input attribute (e.g., >, parity, = , and so on), selecting the most informative.

We have just studied syntactic aspects in the sections above, but the approach of FID3 is also capable of capturing pragmatic aspects of information. Suppose, for example, it is costly (in terms of time, money, uselessness, etc) to get some attribute value. You have to take into account its cost in order to obtain the most economical induction tree versus the most informative one. We can easily modify the FID3 approach to

incorporate this pragmatic aspect by forming information chunks min (poss(pantaloon) , (1 - cost(pantaloon)), $\hat{I}(E_{pantaloon})$) ... versus min (poss(pantaloon), $\hat{I}(E_{pantaloon})$) ..., where "1 - cost(pantaloon)" is a normalized measure of the economy of pantaloon as a possible answer.

Finally, in order to complete the approach and get a rulebase the rules of which have possibility and necessity degrees, we would have to be supplied with the possibilities of nonbad, nonnormal, and nongood besides bad, normal and good for the same input examples. In such a case, the FID3 process would be accomplished in the same way, but comprehensively with 6 output classes instead of 3: bad, normal and good would serve to find possibility measures, and 1-nonbad, 1-nonnormal, 1-nongood would serve to find necessity measures.

FID3, like its probabilistic version PRID3 [9], is currently just a prototype written in "C" implemented on a workstation SUN4. Increasing the cooperation between machine learning and fuzzy logic techniques is necessary. From the author's point of view, more resources have to be provided in order to advance in the field of knowledge acquisition under uncertain domains, at the expense of the approximated reasoning field. The progress of the latter depends on the effort put into the former.

Acknowledgments

I want to express my deepest gratitude to Enric Trillas for his advice and motivation, to Francisco Alcaraz for his comments, to George J. Klir for his lectures, and to Inisel for its support.

References

1. Ashby, W. R. (1972). Systems and Their Informational Measures. Trends in general systems theory. Edited by G. J. Klir, Wiley-Interscience, New York, pp. 78-97.

2. Cestnik,B., Kononenko,I., & Bratko, I. (1987). ASSISTANT 86: A knowledge elicitation tool for sophisticated users. Proceedings of the Second European Working Session on Learning, Bled, Yugoslavia: Sigma Press, pp. 31-45.

3. Klir, George J. & Folger, Tina A. (1988). Fuzzy sets, uncertainty, and information. Prentice-Hall Int. Editions.

4. Mingers, J. (1989). An Empirical Comparation of Selection Measures for Decision-Tree Induction. Machine Learning, Vol. 3, N. 4, pp. 319-342.

5. Mira, J. & Fonseca, J. S. (1970). Neural Nets From the Viewpoint of Signification and Intention. In: Signification and Intention, Ed. by J. S. da Fonseca, Facultade de Medicina, Universidad de Lisboa.

6. Niblett, T. (1987). Constructing decision trees in noisy domains. Proceedings of the Second European Working Session on Learning, Bled, Yugoslavia: Sigma Press, pp. 31-45.

7. Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess and games. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), ML: An AI approach. Los Altos, CA: Morgan Kaufmann.

8. Quinlan, J. R. (1987). Generating production rules from decision trees. Proceedings of the Tenth International Joint Conference on Artificial Intelligence, Milan, Italy: Morgan Kaufmann, pp. 139-145.

9. Rives, J. (1990). PRID3: Probabilistic Induction Decision Tree. Proceedings of International Conference on Fuzzy Logic and Neural Networks, Iizuka, Japan; Vol. 2, pp. 857-862.

10. Ruspini, E. H. (1989). On the semantic of fuzzy logic. SRI International, Technical Note N. 475.

11. Trillas, E. & Sanchís C. (1979). Sobre entropías de conjuntos borrosos deducidas de métricas. Estadística Española, N. 82 and 83.

12. Trillas, E. (1982). Sobre la igualdad de conjuntos borrosos. Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Tomo LXXVI, cuaderno 4.

13. Utgoff, Paul E. (1989). Incremental Induction of Decision Trees. Machine Learning, Vol. 4, N. 2, 161-186.

14. Yager, R. R. (1979). On the measure of fuzziness and negation: membership in the unit interval. International Journal of General Systems, 5, pp. 221-229.

15. Yager, R. R. (1980). On the measure of fuzziness and negation: lattices. Information and Control, 44, pp. 236-260.

16. Zadeh, L. A. (1990). Interpolative Reasoning Based on Fuzzy Logic and its Application to Control and System Analysis. Proceedings of International Conference on Fuzzy Logic and Neural Networks, Iizuka, Japan; Vol.1, p 3.