Exemplar Learning in Fuzzy Decision Trees

C. Z. Janikow
Mathematics and Computer Science
University of Missouri
St. Louis, MO 63121

Abstract

Decision-tree algorithms provide one of the most popular methodologies for symbolic knowledge acquisition. The resulting knowledge, a symbolic decision tree along with a simple inference mechanism, has been praised for comprehensibility. The most comprehensible decision trees have been designed for perfect symbolic data. Over the years, additional methodologies have been investigated and proposed to deal with continuous or multi-valued data, and with missing or noisy features. Recently, with the growing popularity of fuzzy representation, a few researchers independently have proposed to utilize fuzzy representation in decision trees to deal with similar situations. Fuzzy representation bridges the gap between symbolic and non-symbolic data by linking qualitatitive linguistic terms with quantitative data. In this paper, we overview our fuzzy decision tree and propose a few new inferences based on exemplar learning.

1. Introduction

With the increasing amount of data readily available, automatic knowledge acquisition capabilities are becoming more and more important. When all data elements are preclassi ed, no extensive domain knowledge is available, and the objective is to acquire knowledge describing those classes (as to reason about them or simply to classify future data), the knowledge acquisition process is called *supervised learning from examples*. Because the objective is to infer new knowledge, *induction* must be employed. When both the language describing the training data and that describing the resulting knowledge use symbolic features, we speak of *symbolic* learning.

Decision-tree algorithms provide one of the most popular methodologies for symbolic knowledge acquisition from feature-based examples. Among them, Quinlan's ID3 is the most widely known. It was originally designed for symbolic data when all the needed information is available. The acquired knowledge is expressed with a highly comprehensible symbolic decision tree (a *model*), which paired with a simple inference mechanism assigns symbolic decisions (class assignments) to new data. Because of the natural interpretation of the knowledge, symbolic decision trees can be easily translated to a set of rules suitable for use in rule-based systems [10].

An alternative learning scenario may involve a simple statistical inference on the data, or selection of data elements to be used in a *proximity* model. Quinlan has recently proposed to combine standard decision-tree reasoning with such *instance-based* scenarios [9]. *Exemplar-based* learning [1] is a di erent combination of instance- and model-based learning. There, special examples (exemplars) are selected from data to be used with a proximity measure, but these examples can also be generalized.

In real-world applications, data is hardly ever perfectly tted to a given algorithm. This imperfectness can be manifested in a number of ways. Symbolic inductive learning requires symbolic domains, while some or all attributes may be described by multi-valued or continuous features. Some features may be missing from data descriptions, and this may happen both in training or in decision-making. In training, such incomplete data can be disregarded, but this may unnecessarily overlook some available information. In decision-making, a decision must be made based on whatever information is available (or a new test may be suggested). Data can also be noisy or simply erroneous. While the latter problem can be minimized, the former must be addressed, especially in continuous domains from some sensory data. Finally, features may involve inherently *subjective* linguistic terms, without unambiguous de nitions.

Some of such problems have been explored in the context of decision trees, resulting in the proposal of a number of methodological advancements. To deal with continuous data, CART algorithms have been proposed [2]. Unfortunately, these trees su er from reduced comprehensibility, which is not always a welcome trade-o . In symbolic decision trees involving multi-valued or continuous domains, it has been proposed to use domain partition blocks as features. In cases when such blocks overlap (*coverings*), a probabilistic inference can be used. Methods for dealing with missing features and noise have also been studied [6, 8, 10].

In recent years, an alternative representation has grown in popularity. This representation, based on fuzzy sets and used in approximate reasoning, is especially applicable to bridging the conceptual gap between subjective/ambiguous features and quantitative data [3, 13]. Because of the gracefulness of gradual fuzzy sets and approximate reasoning methods used, fuzzy representation is also well suited for dealing with inexact and noisy data. Fuzzy rules, based on fuzzy sets, utilize those qualities of fuzzy representation in a comprehensible structure of rule bases.

Recently, a few researchers have proposed to combine fuzzy representation with the popular decision tree algorithms. When the objective is high comprehensibility rather than "best" *fuzzy partitioning* of the description space, the combination involves symbolic decision trees [4, 5, 11]. The resulting fuzzy decision trees exhibit high comprehensibility, yet fuzzy sets and approximate reasoning methods provide natural means for dealing with continuous domains, subjective linguistic terms, and noisy measurements. Since methodologies for dealing with missing features are readily available in symbolic decision trees, such can also be easily incorporated into those fuzzy decision trees. And since trees can be interpreted as rule-bases, fuzzy decision trees can be seen as a means for learning fuzzy rules. This makes fuzzy decision trees an attractive alternative to other recently proposed learning methods for fuzzy rules (for example, [12]).

A decision-tree learning algorithm has two major components: tree-building and inference. Our fuzzy decision tree is a modi cation of the ID3 algorithm, with both components adapting fuzzy representation and approximate reasoning. In this paper, we overview fuzzy decision trees. Then, following Quinlan's proposition to combine instance- with model-driven learning, we employ ideas borrowed from exemplar-based learning to de ne new inferences for the fuzzy decision tree.

2. Symbolic decision trees

In decision-tree algorithms, examples, described by features of some descriptive language and with known decision assignments, guide the tree-building process. Each branch of the tree is labeled with a condition. To reach a leaf, all conditions on its path must be satis ed. A decision-making inference procedure (class assignments in this case) matches features of new data with those conditions, and classi es the data based on the classi cation of the training data found in the satis ed leaf. Figure 1 illustrates a sample tree.

Tree-building is based on recursive partitioning, and for computational e ciency it usually assumes independence of all attributes. ID3 and CART are the two most popular such algorithms. While ID3 aims at knowledge comprehensibility and is based on symbolic domains, CART is naturally designed to deal with continuous domains but lacks the same level of comprehensibility.

The recursive partitioning routine selects one attribute at a time, usually the one which maximizes some information measure for the training examples satisfying the conditions leading to the node. This attribute is used to split the node, using domain values of the attribute to form additional conditions leading to subtrees. Then, the same procedure is recursively repeated for each child node, with each node using the subset of the training examples satisfying the additional condition. A node is further split unless all attributes are exhausted, when all examples at the node have the same classi cation, or when some other criteria are met. For example, in the shaded node of Figure 1, another attribute could be used to separate its examples if they still represented con icting class assignments and more attributes were available.

The recursive tree-building can be described as follow:

1. Compute the information content at node N, $I_N = -\sum_k^{|C|} p_k \cdot log(p_k)$, where C is the set of decisions and p_k is the probability (estimated from data) that an example found present in the node has classi cation k.

2. For each remaining attribute a_i (previously unused on the path to N), compute the information gain based on this attribute splitting node N. The gain $G_i = I_N - \sum_j^{|D_i|} w_j \cdot I_{N_j}$, where D_i denotes the set of features associated with a_i, I_{N_j} is the information content at the j^{th} child of N, and w_j is the proportion of N's examples that satisfy the condition leading to that node.

3. Expand the node using the attribute which maximizes the gain.

The above tree-building procedure in fact creates a partition of the description space, with guiding principles such as having "large blocks" and unique classi cations of training data in each of the blocks. It is quite natural to make classi cation decisions based on those partitions in such a way that a new data element is classi ed the same way as the training data from the same partition block. Of course, problems arise if a partion block contains training data without unique classi cations. This may result from a number of factors, such as an insu cient set of features, noise or errors. Another potential problem arises if a block has no training data. This may result from an insu cient data set.

Following the above intuitive decision procedure, in the inference stage a new sample's features are compared against the conditions present of the tree. This of course corresponds to deciding on the partition block that the new example falls into. The classi cation of the examples of that leaf whose conditions are satis ed by the data is returned as the algorithm's decision. For example, assuming that the shaded node contains samples with a unique decision, any other sample described in particular by the same two features "young-age" and "blond-hair" would be assigned the same unique classi cation.

ID3 assumes symbolic features, and any attempt to avoid this assumption trades its comprehensibility. Quinlan has extensively investigated ID3 extensions to deal with missing features, inconsistency (when a leaf contains examples of di erent classes), and incompleteness (when a branch for a given feature is missing out of a node).

Quinlan has suggested that in tree-building, when an attribute has its information contents computed in order to determine its utility for splitting a node, each example whose needed feature is missing be partially matched, to the same normalized degree, to all conditions of the attribute. However, the overall utility of the attribute should be reduced according to the proportion of its missing features [8]. During inference, he suggested that assuming similar partial truthness of conditions leading to all available children gives the best performance [8]. This, however, leads to inconsistencies (multiple, possibly con icting leaves could be satis ed), and a probabilistic combination must be used [10]. Quinlan has also investigated the behavior of such a probabilistic approach on noisy features [6, 10].

3. Fuzzy sets, rules, and approximate reasoning

Given a universe of discourse (domain) and some sets of elements from the domain, in classical set theory any element either belongs to a set or it does not. Then, such sets can be described with binary *characteristic functions*. However, in the real world there are many sets for which it is di cult or even impossible to determine if an element belongs to such sets. For example, this information might not be available. Moreover, some sets are inherently *subjective* without unambiguous de nition.

Consider the set of *Tall* people and the universe of one's friends. First problem is that some of those friends may fall "too close to call" so that they can neither be categorically included nor excluded in this set. Thus, partial memberships may be used. However, even then someone else may disagree with those memberships, since this concept is not universely de ned.

Fuzzy sets have been proposed to deal with such cases. A fuzzy set is represented by a membership function onto the real interval $[0, 1]$. A fuzzy linguistic variable is an attribute whose domain contains fuzzy terms { labels for fuzzy sets over the actual universe of discourse [13]. Accordingly, one may extend the notion of proposition to *fuzzy proposition* of the form 'V is A', where V is a fuzzy variable and A is a fuzzy term from the linguistic domain, or a fuzzy set on the universe of V [3, 13]. With that, classical logic must be extended to *fuzzy logic*.

One may extend a classical "rule" to a *fuzzy rule* R of the form

$$R = \text{ if } V_1 \text{ is } A_1 \text{ and... } V_j \text{ is } A_j \text{ then } D \text{ is } A_D$$

In this case, these fuzzy propositions are also called *fuzzy restrictions*. To determine the meaning of such a fuzzy rule, one may extend the notion of relation to *fuzzy relation*, and then apply a *composition* operator. Alternatively, fuzzy logic can be used to determine the meaning of the antecedent and the implication [3, 13]. For such reasoning involving fuzzy logic and fuzzy propositions, *approximate reasoning* was proposed [13].

Fuzzy logic extends classical conjunction and implication to fuzzy terms. While in classical logic conjunction and implication are unique, in fuzzy logic there is an in nite number of possibilities. *T-norm*, which generalizes intersection in the domain of fuzzy sets, is usually used for fuzzy conjunction (and often for implication). *T-norm* is a binary operator with the following properties [3]:

- $T(a, b) = T(b, a)$
- $T(T(a, b), c) = T(a, T(b, c))$
- $a \leq c$ and $b \leq d$ implies $T(a, b) \leq T(c, d)$
- $T(a, 1) = a$

The most popular *T-norms* are min and $product$.

S-norm, from generalized union, is similarly used for fuzzy disjunctions. It is de ned similarly except for the last property [3]: $S(a, 0) = a$. The most popular *S-norm* is max.

A set of fuzzy rules $\{R^k\}$ is a knowledge representation means based on fuzzy rather than classical propositions. In general, one would have to combine all the rules into a fuzzy relation, and then use composition for inference. However, this is computationally very expensive [3]. When a crisp number from the universe of the class variable is sought (this simpli cation can be used even if a linguistic response is sought), and a number of other restrictions on fuzzy sets and applied operators are placed, rules can be evaluated individually and then combined [3]. This approach, called a *local* inference, gives a computationally simple alternative, with good approximation characteristics even if not all the necessary conditions are satisfied. The simplest local inference, and the one we will use here, is defined as follows: $\delta = \sum(\beta_l \cdot b_l) / \sum \beta_l$, where the summation is taken over all rules in the rule base, β_l denotes the degree of satisfaction (from conjunction and implication) of rule l, and b_l is a numerical representation of the fuzzy set in the consequent of the rule. b_l will be the crisp number from the universe of discourse of the consequent's variable if such are used. Otherwise, it may the *center of gravity* of the fuzzy set associated with the used linguistic fuzzy term.

Let us denote f_0 as an operator matching two fuzzy sets, or a crisp element with a fuzzy set. Thus, f_0 can be used to determine the degree to which a piece of data satisfies a single fuzzy restriction. Let us denote f_1 as the operator to evaluate fuzzy conjunctions in antecedents of fuzzy rules (*i.e.*, this is a *T-norm*). Let us denote f_2 as the operator for the fuzzy implication of a rule (this is often a *T-norm*).

4. Fuzzy decision trees

Fuzzy decision trees aim at high comprehensibility, normally attributed to ID3, with the gradual and graceful behavior attributed to fuzzy systems (we assume that a crisp class assignment in the class domain is sought). Thus, they extend the symbolic ID3 procedure (Section 2), using fuzzy sets and approximate reasoning both for the tree-building and the inference procedures. At the same time, they borrow the rich existing decision tree methodologies for dealing with incomplete knowledge (such as missing features), extended to utilize new wealth of information available in fuzzy representation [5].

ID3 is based on classical crisp sets, so an example satisfies exactly one of the possible conditions out of a node and thus falls into exactly one subnode – and consequently into exactly one leaf (except for the cases of missing features). In fuzzy decision trees, an example can match more than one condition, for these conditions are now fuzzy restrictions based on fuzzy sets. Because of that, a given example can fall into many children nodes (typically up to two) of a node. Because this may happen at any level, the example may eventually fall into many of the leaves. Because a fuzzy match is any number in $[0, 1]$, and approximate reasoning is closed in $[0, 1]$, a leaf node may contain a large number of examples, many of which can be found in other leaves as well, each to some degree in $[0, 1]$. This fact is actually advantageous as it provides more graceful behavior, especially when dealing with noisy or incomplete information, while the approximate reasoning methods are designed to handle the additional complexity.

The tree-building routine follows that of ID3, except that information utility of individual attributes is evaluated using fuzzy sets, memberships, and reasoning methods [5]. More practically, it means that the number of examples falling into a node (that is, those satisfying conditions leading to the node) is now based on fuzzy matches of individual features to appropriate fuzzy sets. These are the fuzzy sets associated with the linguistic terms used in those conditions, which now become fuzzy restrictions. To reach a leaf a number of fuzzy restrictions must be matched – *T-norms* are used to evaluate the degree to which a given example falls into into deeper nodes (f_1). Moreover, because the decision procedure of a single leaf can be described as the implication "if fuzzy restrictions leading to a leaf are satisfied then infer the classification of the training data in the leaf", then the previous level of satisfaction must be passed through this implication (f_2).

In particular, the difference lies in the way the probabilities p_k (see Section 2) are estimated. These probabilities are estimated from example counts in individual nodes. In fuzzy trees, an example's membership in a given node is a real number $[0, 1]$, indicating the example's satisfaction of the fuzzy restrictions leading to that node. Thus, we need f_0 and f_1 to evaluate this. Denote μ_e^N as the accumulated membership for example e at node N. Clearly, $\mu_e^{Root} = 1$ and $\mu_e^{N_j} = f_1(\mu_e^N, f_0(e, A_j))$, where N_j is the j^{th} child of N, and A_j is the fuzzy term associated with the fuzzy restriction leading to N_j. Then, at node N the event count for the fuzzy linguistic class k is $\sum_{e \in E} f_1(\mu_e^N, f_0(e, A_k))$. Given that, the probability p_k can be estimated as

$$p_k = \frac{\sum_{e \in E} f_1(\mu_e^N, f_0(e, A_k))}{\sum_{k' \in C} \sum_{e \in E} f_1(\mu_e^N, f_0(e, A_{k'}))}$$

and a tree can be constructed using the same algorithm as that of Section 2.

In the inference routine, the first step is to find to what degree a new example satisfies each of the leaves. In other words, it is necessary to find its membership in each of the partition blocks of the now fuzzy partition. Thus, the same f_0 and f_1 mechanisms must be used. The next step is to determine the response of each of the leaves individually, through f_2. As indicated earlier, in fuzzy decision trees the number of satisfied leaves dramatically increases. Such conflicts are handled by f_3 operators.

However, there are likely to be conflicts as well in individual leaves. This is so because the fuzzy partition blocks are very likely to contain training data with different classifications. In cases where the class assignment is not just a fuzzy (or strict) linguistic term but a value from the classification domain (*e.g.*, when the domain is continuous), then there are no unique classifications whatsoever, and we may only speak of the grade of membership of such examples in the linguistic classes. Such class information can be interpreted in a number of di erent ways. For example, this may be treated as a union of weighted fuzzy sets, or as a disjunction of weighted fuzzy propositions relating to class assignments, with the weights proportional to the proportion of examples of that leaf found in fuzzy sets associated with the linguistic classes.

Multiple leaves which can be partially matched by a sample provide the second con ict level. These con icts result from overlapping fuzzy sets, and it is aggravated by cases of missing features in the sample to be recognized and by incomplete knowledge (missing branches in the tree). Because each leaf can be interpreted as a fuzzy rule, then multiple leaves have the disjunctive interpretation.

A number of possible fuzzy inferences were described in [5]. Here, we describe a few di erent inferences, following the idea of exemplars [1] and implementing the local center-of-gravity inference with various interpretations of those two levels of con icts.

In exemplar-based learning, selected training examples are retained, and a decision procedure returns the class assigned to the "closest" exemplar. Moreover, these examples can be generalized, and multiple exemplars can be used for making the decision [1].

The same ideas can lead to a number of new inferences from fuzzy decision trees. Consider a fuzzy decision tree built as above. Suppose that we interpret all individual leaves as generalized exemplars. Obviously path-restrictions determine the generalized forms of the exemplars. However, in fuzzy decision trees such exemplars would not have unique classi cations, while an exemplar, like an example, is supposed to have. Thus, we need methods to create exemplars from leaves (in addition to methods for combining information from such exemplars).

As mentioned before, for practical reasons we follow the idea of local inferences from individual rules (leaves) [3]. We rst de ne two coe cients R and r, two cases each, which will subsequently be used to de ne the inferences. Assume that e_o is the numerical class assignment, that is a value from the universe of the classi cation, of example e (if the classi cations are symbolic or linguistic fuzzy terms these may be centers of gravity of such sets). These coef- cients can be computed by taking all individual examples falling into a leaf l:

$$R_l^1 = \sum_{e \in E} (\mu_e^l \cdot e_o),\ r_l^1 = \sum_{e \in E} \mu_e^l$$

or by favoring examples associated with the best κ decision at each leaf l (*i.e.* κ is the fuzzy linguistic class which has the highest accumulative membership of training data in l):

$$R_l^2 = \sum_{e \in E} f_1(\mu_e^l \cdot e_o, f_0(e, A_\kappa)),\ r_l^2 = \sum_{e \in E} f_1(\mu_e^l, f_0(e, A_\kappa))$$

Those coe cients give an obvious interpretation to R/r. For example, if $\forall_{(e \in E)} e_o = a$ then $R^1/r^1 = a$. In general, this quotient implements the idea of center of gravity of a leaf. These coe cients can be used to de ne our inferences. We present them along with their intuitive interpretations. Note that di erent interpretations actually implicitly refer to di erent f_3 operators { f_2, in addition to previous f_0 and f_1, must be assumed explicitly.

1. Find center of gravity of all the examples in a leaf. This represents an exemplar. Then, nd the classi cation of a super-exemplar taken over all leaves, while individual leaf-based exemplars are weighted by how well they match the new event and how much training data they contain.

$$\delta = \sum_{l \in Leaves} f_2(\mu_e^l, R_l^1) / \sum_{l \in Leaves} f_2(\mu_e^l, r_l^1)$$

2. The same as 1, but the individual exemplars are not weighted by how many examples they contain.

$$\delta = \sum_{l \in Leaves} f_2(\mu_e^l, R_l^1/r_l^1)$$

3. The same as case 1, but individual exemplars represent only examples of the best decision (fuzzy linguistic class) in the leaf.

$$\delta = \sum_{l \in Leaves} f_2(\mu_e^l, R_l^2) / \sum_{l \in Leaves} f_2(\mu_e^l, r_l^2)$$

4. The same as 2, but also restricted as 3.

$$\delta = \sum_{l \in Leaves} f_2(\mu_e^l, R_l^2/r_l^2)$$

5. Suppose that τ denotes the leaf whose data is "most" similar to the new example (*i.e.*, maximizes μ_e^l). Here, the decision is based on the exemplar from such a leaf.

$$\delta = R_\tau^1 / r_\tau^1$$

6. The same as 5, but the selected exemplar represents only the best class from the leaf.

$$\delta = R_\tau^2 / r_\tau^2$$

What the properties of these speci c inferences are remains to be investigated. Some preliminary experiments on this subject, for example assuming that the goal is noise ltration and concept recovery from incomplete data, were reported in [4] for a number of similar inferences.

An apparent advantage of fuzzy decision trees is that they use the same routines as symbolic decision trees (but with fuzzy representation). This allows for utilization of the same comprehensible tree structure for knowledge understanding and veri cation. This also allows more robust processing with continuously gradual outputs. Moreover, one may easily incorporate rich methodologies for dealing with missing features and incomplete trees. For example, suppose that the inference descends to a node which does not have a branch (maybe due to tree pruning, which often improves generalization properties) for the corresponding feature of the sample. This dandling feature can be fuzzi ed, and then its match to fuzzy restrictions associated with the available branches provides better than uniform discernibility among those children [5].

Because fuzzy restrictions are evaluated using fuzzy membership functions, this process provides a linkage between continuous domain values and abstract features { here, fuzzy terms. However, numerical data is not essential for the algorithm since methods for matching two fuzzy terms are readily available. Thus, fuzzy decision trees can process data expressed with both numerical values (more information) and fuzzy terms. Because of the latter, fuzzy trees can also process inherently symbolic domains. Because the same processing has been adapted for fuzzy domains, processing purely symbolic data would result in the same behavior (given proper inference) as that of a symbolic decision tree. Thus, fuzzy decision trees are natural generalizations of symbolic decision trees.

Most research and empirical studies remain to be completed. For illustration, [5] presents two sample applications: one dealing with learning continuous functions, and the other for learning symbolic concepts. The same publication also details the complete algorithm.

5. Summary

We brie y introduced fuzzy decision trees. They are based on ID3 symbolic decision trees, but they utilize fuzzy representation. The tree-building routine has been modi ed to utilize fuzzy instead of strict restrictions. Following fuzzy approximate reasoning methods, the inference routine has been modi ed to process fuzzy information. The inference can be based on a number of di erent criteria, resulting in a number of di erent inferences. In this paper, we de ned new inferences based on exemplar learning | tree leaves are treated as learning exemplars.

Fuzzy representation allows decision trees to deal with continuous data and noise. Unknown-feature methodologies, borrowed from symbolic decision trees, allow fuzzy trees to deal with missing information. The symbolic tree structure allows fuzzy trees to provide comprehensible knowledge. We are currently testing a new prototype, which will subsequently be used to determine various properties and trade-o s associated with fuzzy decision trees.

This research is in part being supported by grant NSF IRI-9504334.

References

[1] E.R. Baires, B.W. Porter & C.C. Wier. PROTOS: An Exemplar-Based learning Apprentice. *Machine Learning III*, Kondrato . Y. & Michalski, R. (eds.), Morgan Kaufmann, pp. 112-127.

[2] L. Breiman, J.H. Friedman, R.A. Olsen & C.J. Stone. *Classi cation and Regression Trees*. Wadsworth, 1984.

[3] D. Driankov, H. Hellendoorn & M. Reinfrank. *An Introduction to Fuzzy Control*. Springer-Verlag 1993.

[4] C.Z. Janikow. Fuzzy Processing in Decision Trees. *Proceedings of the Sixth International Symposium on Arti cial Intelligence*, 1993, pp. 360-367.

[5] C.Z. Janikow. Fuzzy Decision Trees: Issues and Methods. To appear *IEEE Transactions on Systems, Man, and Cybernetics*.

[6] J.R. Quinlan. The E ect of Noise on Concept Learning. *Machine Learning II*, R. Michalski, J. Carbonell & T. Mitchell (eds), Morgan Kaufmann, 1986.

[7] J.R. Quinlan. Induction on Decision Trees. *Machine Learning*, Vol. 1, 1986, pp. 81-106.

[8] J.R. Quinlan. Unknown Attribute-Values in Induction. *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann, 1992.

[9] J.R. Quinlan. Combining Instance-Based and Model-Based Learning. *Proceedings of International Conference on Machine Learning 1993*, Morgan Kaufmann 1993, pp. 236-243.

[10] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[11] M. Umano, H. Okamoto, I. Hatono, H. Tamura, F. Kawachi, S. Umedzu & J. Kinoschita. Fuzzy Decision Trees by Fuzzy ID3 Algorithm and Its Application to Diagnosis Systems. *Proceedings of the Third IEEE Conference on Fuzzy Systems, 1994*, pp. 2113-2118.

[12] L. Wang & J. Mendel. Generating Fuzzy Rules by Learning From Examples. *IEEE Transactions on Systems, Man and Cybernetics*, v. 22 No. 6, 1992, pp. 1414-1427.

[13] L.A. Zadeh. Fuzzy Logic and Approximate Reasoning. *Synthese* 30 (1975), pp. 407-428.