Fuzzy-Rough Feature Significance for Fuzzy Decision Trees

Jensen, Richard; Shen, Qiang

Fuzzy-Rough Feature Significance for Fuzzy Decision Trees

Richard Jensen and Qiang Shen
Department of Computer Science,
The University of Wales, Aberystwyth
{*rkj,qqs*}*@aber.ac.uk*

Abstract

Crisp decision trees are one of the most popular classification algorithms in current use within data mining and machine learning. However, although they possess many desirable features, they lack the ability to model vagueness. As a result of this, the induction of fuzzy decision trees (FDTs) has become an area of much interest. One important aspect of tree induction is the choice of feature at each stage of construction. If weak features are selected, the resulting decision tree will be meaningless and will exhibit poor performance. This paper introduces a new measure of feature significance based on fuzzy-rough sets for use within fuzzy ID3. The measure is experimentally compared with leading feature rankers, and is also compared with traditional fuzzy entropy for fuzzy tree induction.

1 Introduction

A decision tree can be viewed as a partitioning of the instance space. Each partition, represented by a leaf, contains the objects that are similar in relevant respects and thus are expected to belong to the same class. The partitioning is carried out in a data-driven manner, with the final output representing the partitions as a tree. An important property of decision tree induction algorithms is that they attempt to minimize the size of the tree at the same time as they optimize a certain quality measure.

The general decision tree induction algorithm is as follows. The significance of features is computed using a suitable measure (in C4.5 this is the information gain metric [11]). Next, the most discriminating feature according to this measure is selected and the dataset partitioned into sub-tables according to the values this feature may take. The chosen feature is represented as a node in the currently constructed tree. For each sub-table, the above procedure is repeated, i.e. determine the most discriminating feature and split the data into further sub-tables according to its values.

This is a similar process in the fuzzy case. However, a measure capable of handling fuzzy terms (instead of crisp values) must be used. Data is partitioned according to the selected feature's set of fuzzy terms. There must also be a way of calculating the number of examples that belong to a node. In the crisp case, this is clear; objects either contain a specific attribute value or they do not. In the fuzzy case this distinction can no longer be made, as objects may belong to several fuzzy terms. A suitable stopping condition must also be chosen that will limit the number of nodes expanded.

Clearly, one important aspect of this procedure is the choice of feature significance measure. This measure influences the organization of the tree directly and will have profound effects on the resulting tree's accuracy.

It has been shown that the fuzzy-rough metric is a useful gauger of (discrete and real-valued) attribute information content in datasets [7]. This has been employed primarily within the feature selection task, to benefit the rule induction that follows this process. It is therefore interesting to investigate how an induction algorithm based on the fuzzy-rough measure would compare with standard algorithms such as fuzzy ID3 [6], both in terms of the complexity of the trees constructed and the resulting accuracy.

The rest of this paper is structured as follows. The second section summarises the theoretical background of the basic ideas of fuzzy-rough sets and their use for feature evaluation. Section 3 provides results of the application of the fuzzy-rough measure to artificial data in order to locate the relevant features. Its performance is gauged in comparison to that of several leading feature rankers. A brief introduction to

fuzzy decision trees is then given in section 4, and experimental results comparing fuzzy ID3 with fuzzy-rough ID3 are presented in section 5. Finally, the paper is concluded and future work is outlined.

2 Fuzzy-Rough Feature Significance

Although FDTs based on the fuzzy entropy selection measure have been successful, there have been few attempts to investigate radically different selection measures. Research has concentrated primarily on fuzzy entropy and its extensions [2, 13]. Here, the focus is on introducing a new measure of feature significance based on fuzzy-rough sets for FDT induction.

2.1 Fuzzy Equivalence Classes

In the same way that crisp equivalence classes are central to rough sets [3], *fuzzy* equivalence classes are central to the fuzzy-rough set approach [4]. For typical applications, this means that the decision values and the conditional values may all be fuzzy. The concept of crisp equivalence classes can be extended by the inclusion of a fuzzy similarity relation S on the universe, which determines the extent to which two elements are similar in S. The usual properties of reflexivity ($\mu_S(x,x) = 1$), symmetry ($\mu_S(x,y) = \mu_S(y,x)$) and transitivity ($\mu_S(x,z) \geq \mu_S(x,y) \wedge \mu_S(y,z)$) hold.

Using the fuzzy similarity relation, the fuzzy equivalence class $[x]_S$ for objects close to x can be defined:

$$\mu_{[x]_S}(y) = \mu_S(x,y) \tag{1}$$

The following axioms should hold for a fuzzy equivalence class F:

- $\exists x,\ \mu_F(x) = 1$
- $\mu_F(x) \wedge \mu_S(x,y) \leq \mu_F(y)$
- $\mu_F(x) \wedge \mu_F(y) \leq \mu_S(x,y)$

The first axiom corresponds to the requirement that an equivalence class is non-empty. The second axiom states that elements in y's neighbourhood are in the equivalence class of y. The final axiom states that any two elements in F are related via S. Obviously, this definition degenerates to the normal definition of equivalence classes when S is non-fuzzy. The family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes [4].

2.2 Fuzzy Lower and Upper Approximations

From the literature, the fuzzy P-lower and P-upper approximations are defined as [4]:

$$\mu_{\underline{P}X}(F_i) = inf_x max\{1 - \mu_{F_i}(x), \mu_X(x)\} \quad \forall i \tag{2}$$

$$\mu_{\overline{P}X}(F_i) = sup_x min\{\mu_{F_i}(x), \mu_X(x)\} \quad \forall i \tag{3}$$

where F_i denotes a fuzzy equivalence class belonging to $\mathbb{U}/P$ which in turn stands for the partition of $\mathbb{U}$ with respect to a given subset P of features.

For an individual feature, a, the partition of the universe by $\{a\}$ (denoted $\mathbb{U}/IND(\{a\})$) is considered to be the set of those fuzzy equivalence classes for that feature. For example, if the two fuzzy sets N_a and Z_a are generated for feature a during fuzzification, the partition $\mathbb{U}/IND(\{a\}) = \{N_a, Z_a\}$.

Note that although the universe of discourse in feature reduction is finite, this is not the case in general, hence the use of *sup* and *inf* above. These definitions diverge a little from the crisp upper and lower approximations, as the memberships of individual objects to the approximations are not explicitly available. As a result of this, the fuzzy lower and upper approximations are redefined as:

$$\mu_{\underline{P}X}(x) = \sup_{F \in \mathbb{U}/P} min(\mu_F(x), \inf_{y \in \mathbb{U}} max\{1 - \mu_F(y), \mu_X(y)\}) \tag{4}$$

$$\mu_{\overline{P}X}(x) = \sup_{F \in \mathbb{U}/P} min(\mu_F(x), \sup_{y \in \mathbb{U}} min\{\mu_F(y), \mu_X(y)\}) \tag{5}$$

In implementation, not all $y \in \mathbb{U}$ are needed to be considered - only those where $\mu_F(y)$ is non-zero, i.e. where object y is a fuzzy member of (fuzzy) equivalence class F. The tuple $< \underline{P}X, \overline{P}X >$ is called a fuzzy-rough set.

If the fuzzy-rough approach is to be useful, it must be able to deal with multiple features, finding the dependency between various subsets of the original feature set. For example, it may be necessary to be able to determine the degree of dependency of the decision feature(s) with respect to feature set $P = \{a, b\}$. In the crisp case,

$\mathbb{U}/P$ contains sets of objects grouped together that are indiscernible according to both features a and b. In the fuzzy case, objects may belong to many equivalence classes, so the cartesian product of $\mathbb{U}/IND(\{a\})$ and $\mathbb{U}/IND(\{b\})$ must be considered in determining $\mathbb{U}/P$. In general,

$$\mathbb{U}/P = \otimes\{a \in P : \mathbb{U}/IND(\{a\})\} \quad (6)$$

For example, if $P = \{a, b\}$, $\mathbb{U}/IND(\{a\}) = \{N_a, Z_a\}$ and $\mathbb{U}/IND(\{b\}) = \{N_b, Z_b\}$, then

$$\mathbb{U}/P = \{N_a \cap N_b, N_a \cap Z_b, Z_a \cap N_b, Z_a \cap Z_b\}$$

Clearly, each set in $\mathbb{U}/P$ denotes an equivalence class. The extent to which an object belongs to such an equivalence class is therefore calculated by using the conjunction of constituent fuzzy equivalence classes, say F_i, $i = 1, 2, ..., n$:

$$\mu_{F_1 \cap ... \cap F_n}(x) = min(\mu_{F_1}(x), \mu_{F_2}(x), ..., \mu_{F_n}(x)) \quad (7)$$

2.3 Positive Region

The crisp positive region in the standard RST is defined as the union of the lower approximations. By the extension principle, the membership of an object $x \in \mathbb{U}$, belonging to the fuzzy positive region can be defined by

$$\mu_{POS_P(Q)}(x) = \sup_{X \in \mathbb{U}/Q} \mu_{\underline{P}X}(x) \quad (8)$$

Object x will not belong to the positive region only if the equivalence class it belongs to is not a constituent of the positive region. This is equivalent to the crisp version where objects belong to the positive region only if their underlying equivalence class does so.

2.4 Dependency Measure

Using the definition of the fuzzy positive region, the new dependency function can be defined as follows:

$$\gamma'_P(Q) = \frac{|\mu_{POS_P(Q)}(x)|}{|\mathbb{U}|} = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_P(Q)}(x)}{|\mathbb{U}|} \quad (9)$$

As with crisp rough sets, the dependency of Q on P is the proportion of objects that are discernible out of the entire dataset. In the present approach, this corresponds to determining the fuzzy cardinality of $\mu_{POS_P(Q)}(x)$ divided by the total number of objects in the universe. It is this fuzzy-rough measure of dependency that will be used in the fuzzy-rough ID3 process.

3 Evaluating the Fuzzy-Rough Metric

In order to evaluate the utility of the new fuzzy-rough measure of feature significance, a series of artificial datasets were generated and used for comparison with 5 other leading feature ranking measures. The datasets were created by generating around 30 random feature values for 400 objects. Two or three features (referred to as x, y, or z) are chosen to contribute to the final boolean classification by means of an inequality. For example, in table 2, if the inequality $(x + y)^2 > 0.25$ holds for an object then it is classified as 1, with a classification of 0 otherwise. The task for the feature rankers was to discover those features that are involved in the inequalities, ideally rating the other irrelevant features poorly in contrast.

The tables presented in the metric comparison section show the ranking given to the features that are involved in the inequality that determines the classification. The final row indicates whether all the other features are given a ranking of zero. For the data presented in table 1, the first feature, x, is used to determine the classification. The values of features y and z are derived from x: $y = \sqrt{x}$, $z = x^2$.

3.1 Compared Metrics

The metrics compared are: the fuzzy-rough measure (FR), Relief-F (Re), Information Gain (IG), Gain Ratio (GR), OneR (1R) and the statistical measure χ^2. Metrics other than the fuzzy-rough measure were obtained from [14]. A brief description of each is presented next.

3.1.1 Information Gain

The Information Gain (IG) [11] is the expected reduction in entropy resulting from partitioning the dataset objects according to a particular feature. The entropy of a labelled collection of objects S is defined as:

$$Ent(S) = \sum_{i=1}^{c} -p_i log_2 p_i \quad (10)$$

where p_i is the proportion of S belonging to class i. Based on this, the Information Gain metric is:

$$IG(S,A) = Ent(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Ent(S_v) \tag{11}$$

where $values(A)$ is the set of values for feature A, S the set of training examples, S_v the set of training objects where A has the value v.

3.1.2 Gain Ratio

One limitation of the IG measure is that it favours features with many values. The Gain Ratio (GR) seeks to avoid this bias by incorporating another term, split information, that is sensitive to how broadly and uniformly the attribute splits the considered data:

$$Split(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \tag{12}$$

where each S_i is a subset of objects generated by partitioning S with the c-valued attribute A. The Gain Ratio is then defined as follows:

$$GR(S,A) = \frac{IG(S,A)}{Split(S,A)} \tag{13}$$

3.1.3 χ^2 Measure

In the χ^2 method [10], features are individually evaluated according to their χ^2 statistic with respect to the classes. For a numeric attribute, the method first requires its range to be discretized into several intervals. The χ^2 value of an attribute is defined as:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{14}$$

where m is the number of intervals, k the number of classes, A_{ij} the number of samples in the ith interval, jth class, R_i the number of objects in the ith interval, C_j the number of objects in the jth class, N the total number of objects, and E_{ij} the expected frequency of A_ij (E_{ij} = R_i * C_j/N). The larger the χ^2 value, the more important the feature.

3.1.4 Relief-F

Relief [8] evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class. Relief-F extends this idea to dealing with multi-class problems as well as handling noisy and incomplete data.

3.1.5 OneR

The OneR classifier [5] learns a one-level decision tree, i.e. it generates a set of rules that test one particular attribute. One branch is assigned for every value of a feature; each branch is assigned the most frequent class. The error rate is then defined as the proportion of instances that do not belong to the majority class of their corresponding branch. Features with the higher classification rates are considered to be more discriminating than those resulting in lower accuracies.

3.2 Metric Comparison

From the results in table 1, it can be observed that all metrics successfully rank the influential features highest. IG, GR, 1R and χ^2 rank these features equally, whereas Re and FR rank feature z higher. Only FR, IG, GR and χ^2 rate all the other features as zero.

As can be seen from these results, feature rankers can discover the influential features but on their own are incapable of determining multiple feature interactions. Table 1 could be reduced to one feature only (either x, y, or z) without any loss of information as only these contribute to the classification. However, the rankers all rate these features highly and would only provide enough information to reduce the data to at least these three attributes. Here, the rankers have found the predictive (or relevant) features but have been unable to determine which of these are redundant.

Table 2 shows the results for the inequality $(x + y)^2 > 0.25$. Both features x and y are required for deciding the classification. All feature rankers evaluated detect this. FR, IG, GR, 1R and χ^2 also rank the tenth feature highly - probably due to a chance correlation with the decision. The results in table 3 are for a similar inequality, with all the feature rankers correctly rating the important features. FR, IG, GR and χ^2 evaluate the remaining features as having zero significance.

In table 4, all metrics apart from 1R locate the relevant features. For this dataset, 1R chooses 22 features as being the most discriminating, whilst ranking features x and y last. This may be due to the discretization process that must precede the application of 1R. If the discretization is poor, then the resulting feature evaluations will be affected.

Tables 5 shows the results for data classified by $x * y * z > 0.125$. All feature rankers correctly detect these variables. However, in table 6 the results can be seen for the same inequality but with the impact of variable z increased. All metrics determine that z has the most influence on the decision, and almost all choose x and y next. Again, the 1R measure fails and chooses features 15, 19 and 24 instead.

This short investigation into the utility of the fuzzy-rough measure has shown that it is comparable with the leading measures of feature importance. Indeed, its behaviour is quite similar to the information gain and gain ratio metrics. This is interesting as both of these measures are entropy-based. Unlike these metrics, the fuzzy-rough measure may also be applied to datasets containing real-valued decision features, hence its application within FDT construction.

4 Fuzzy Decision Trees

As with crisp decision trees, fuzzy decision tree induction involves the recursive partitioning of training data in a top-down manner. The most informative feature is selected at each stage, and the remaining data is divided according to the values of the feature. Partitioning continues until there are no more features to evaluate, or if all the examples in the current node belong to the same class.

4.1 Main Differences

One significant problem that has faced crisp tree induction is how to effectively handle continuous features. A standard approach that aims to address this is C4.5 [11] which considers intervals of values during tree construction. However, some interpretability of the tree is lost as the intervals themselves, although useful for classification purposes, may not have any direct physical relevance or meaning to the problem at hand.

FDTs are able to handle continuous features through the use of *fuzzy sets*. Fuzzy sets and logic allow language-related uncertainties to be modelled and provide a symbolic framework for knowledge comprehensibility. Unlike crisp decision tree induction, FDTs do not use the original numerical feature values directly in the tree. Instead, they use fuzzy sets generated either from a fuzzification process beforehand or expert-defined partitions to construct comprehendible trees. As a result of this, there are several key differences between FDT induction and the original crisp approaches:

- Membership of objects. Traditionally objects/examples belonged to nodes with a membership of $\{0, 1\}$; now these memberships may take values from the interval [0,1]. In each node, an example has a different membership degree to the current example set, and this degree is calculated from the conjunctive combination of the membership degrees of the example to the fuzzy sets along the path to the node and its degrees of membership to the classes.

- Measures of feature significance. As fuzzy sets are used, the measures of significance should incorporate this membership information to decide which features form nodes within the tree. This is particularly important as the quality of the tree can be greatly reduced by a poor measure of feature significance.

- Fuzzy tests. Within nodes, fuzzy tests are carried out to determine the membership degree of a feature value to a fuzzy set.

- Stopping criteria. Learning is usually terminated if all features are used on the current path, or if all objects in the current node belong to the same class. With fuzzy trees, objects can belong to any node with any degree of membership. As a result of this, fuzzy trees tend to be larger in size which can lead to poorer generalisation performance. An additional threshold can be introduced, based on the feature significance measure, to terminate construction earlier in induction. For classification, the decision tree is converted to an equivalent ruleset.

The focus of this paper is the choice of significance measure that determines which attributes form nodes in the resulting decision tree. The natural choice for FDTs is fuzzy entropy [6, 9], a fuzzy extension of the crisp entropy measure used with much success in crisp induction.

4.2 Fuzzy Entropy Measure

Let $I = (\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U}$ is a non-empty set of N finite objects (the universe) and $\mathbb{A}$ is a non-empty finite set of n attributes, $\{A^1, A^2, ..., A^n\}$. An attribute A^k takes m_k values of fuzzy subsets $\{A^k_1, A^k_2, ..., A^k_{m_k}\}$. Based on the attributes, an object is classified into C fuzzy subsets $\omega_1, \omega_2, ..., \omega_C$.

The fuzzy entropy for a subset can be defined as

$$H_i^k = \sum_{j=1}^{C} -p_i^k(j) \log p_i^k(j) \tag{15}$$

where, $p_i^k(j)$ is the relative frequency of the ith subset of attribute k with respect to ω_j ($1 \leq j \leq C$) and defined as

$$p_i^k(j) = \frac{|A_i^k \cap \omega_j|}{|A_i^k|} \tag{16}$$

The cardinality of a fuzzy set is denoted by $|\cdot|$. An attribute s is chosen to split the instances at a given node according to

$$s = \arg \min_{1 \leq k \leq n} E^k \tag{17}$$

where

$$E^k = \sum_{i=1}^{m_k} \frac{|A_i^k|}{\sum_{j=1}^{m_k} |A_j^k|} H_i^k \tag{18}$$

5 Experimentation

To demonstrate the applicability of the proposed approach, both the fuzzy and fuzzy-rough decision tree induction methods were applied to a variety of benchmark datasets obtained from [1]. This section presents the results of experimentation carried out on these datasets.

5.1 Setup

In order for both decision tree inducers to operate, fuzzy sets must first be defined for real-valued attributes that appear in the data. For this, a simple fuzzification was carried out based on the statistical properties of the attributes themselves. It is expected that the classification performance would be greatly improved if the fuzzifications were optimized.

The datasets were then split into two halves of equal size, one for training with the other for testing, whilst maintaining the original class distributions. To show the general applicability of both approaches, the inducers were also applied to non-fuzzy data. Datasets WQ2class and WQ3class are the water treatment datasets with the original thirteen classes collapsed into two or three respectively.

5.2 Results

As can be seen from table 7, both approaches perform similarly for the fuzzy data. The size of the resultant rulesets that produce these accuracies can be found in table 8. In general, FR-ID3 produces slightly larger rulesets than the standard approach. There is a notable difference in performance for the Credit dataset, where FR-ID3 produces a reduced classification accuracy. From table 8 it can be seen that this is the only case where FR-ID3 produces a smaller ruleset. The paired t-tests for training and testing results for F-ID3 and FR-ID3 produce the p-values 0.368 and 0.567 respectively.

Dataset	F-ID3		FR-ID3	
	Train	Test	Train	Test
Iris	0.973	0.947	0.973	0.947
Glass	0.514	0.495	0.523	0.476
Credit	0.890	0.849	0.742	0.641
WQ2class	0.824	0.787	0.824	0.782
WQ3class	0.816	0.696	0.801	0.722
Ionosphere	0.862	0.783	0.852	0.783
Olitos	0.678	0.590	0.695	0.656

Table 7: Classification accuracies for fuzzy ID3 and fuzzy-rough ID3 (real-valued data)

Dataset	F-ID3	FR-ID3
Iris	10	13
Glass	32	44
Credit	51	50
WQ2class	108	140
WQ3class	165	328
Ionosphere	22	28
Olitos	9	17

Table 8: Number of rules produced (fuzzy data)

Dataset	F-ID3		FR-ID3	
	Train	Test	Train	Test
Derm	0.514	0.257	0.838	0.615
Derm2	0.932	0.818	0.972	0.906
DNA	0.575	0.348	0.475	0.386
Heart	0.826	0.772	0.826	0.793
WQ-disc	0.798	0.625	0.698	0.563

Table 9: Classification accuracies for fuzzy ID3 and fuzzy-rough ID3 (crisp data)

The results of the application of fuzzy and fuzzy-rough ID3 to crisp data can be found in table 9, with the resulting ruleset size in table 10. The results show that FR-ID3 outperforms F-ID3 in general, as well as producing smaller rulesets. Here, the paired t-tests for the training and testing results produce the p-values 0.695

and 0.283 respectively.

Dataset	F-ID3	FR-ID3
Derm	53	46
Derm2	20	18
DNA	25	22
Heart	25	30
WQ-disc	28	28

Table 10: Number of rules produced (crisp data)

6 Conclusion

Automated generation of feature pattern-based if-then rules is essential to the success of many intelligent pattern classifiers, especially when their inference results are expected to be directly human-comprehensible. This paper has presented such an approach which utilises a fuzzy-rough measure of feature significance to construct fuzzy decision trees. The results show that the proposed method performs comparably to fuzzy ID3 for fuzzy datasets, and better than it for crisp data. Further experimentation is to be carried out on a fuller range of datasets in the future.

One of the issues raised by the experimentation is the size of ruleset produced by the fuzzy-rough method. Future research will investigate the reasons behind this and how this might be addressed. Decision tree pruning is one promising solution, as its positive impact on performance for crisp tree-based methods is well established. Additional future work includes the application of a feature selection step based on fuzzy-rough sets before tree induction takes place. This in itself will significantly reduce tree complexity and induction runtime.

References

[1] C.L. Blake and C.J. Merz. UCI Repository of machine learning databases. Irvine, University of California. 1998. `http://www.ics.uci.edu/~mlearn/.`

[2] B. Bouchon-Meunier and C. Marsala. Measures of discrimination for the construction of fuzzy decision trees. In Proc. of the FIP'03 conference, Beijing, China, pp. 709–714. 2003.

[3] A. Chouchoulas and Q. Shen. Rough set-aided keyword reduction for text categorisation. Applied Artificial Intelligence, Vol. 15, No. 9, pp. 843–873. 2001.

[4] D. Dubois and H. Prade. Putting rough sets and fuzzy sets together. In [12], pp. 203–232. 1992.

[5] R.C. Holte. Very simple classification rules perform well on most commonly used datasets. Machine Learning, Vol. 11, No. 1, pp. 63–90. 1993.

[6] C.Z. Janikow. Fuzzy Decision Trees: Issues and Methods. IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics, Vol. 28, No. 1, pp. 1–14. 1998.

[7] R. Jensen and Q. Shen. Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches. IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 12, pp. 1457–1471. 2004.

[8] I. Kononenko. Estimating attributes: Analysis and Extensions of Relief. Proceedings of the European Conference on Machine Learning, pp. 171–182. 1994.

[9] B. Kosko. Fuzzy entropy and conditioning. Information Sciences, Vol. 40, No. 2, pp. 165–174. 1986.

[10] H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence, pp. 336–391. 1995.

[11] J.R. Quinlan. C4.5: Programs for Machine Learning. The Morgan Kaufmann Series in Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA. 1993.

[12] R. Slowinski, editor. Intelligent Decision Support. Kluwer Academic Publishers, Dordrecht. 1992.

[13] X. Wang and C. Borgelt. Information Measures in Fuzzy Decision Trees. Proc. 13th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'04, Budapest, Hungary), vol. 1, pp. 85–90. 2004.

[14] I.H. Witten and E. Frank. Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco. 2000.

Feature	FR	Re	IG	GR	1R	χ^2
x	0.5257	0.31758	0.997	1.0	99.5	200
y	0.5296	0.24586	0.997	1.0	99.5	200
z	0.5809	0.32121	0.997	1.0	99.5	200
others	$= 0$	$\neq 0$	$= 0$	$= 0$	$\neq 0$	$= 0$

Table 1: Feature evaluation for $x > 0.5$, $y = \sqrt{x}$, $z = x^2$

Feature	FR	Re	IG	GR	1R	χ^2
x	0.2330	0.1862	0.2328	0.1579	86.75	128.466
y	0.2597	0.1537	0.1687	0.1690	87.75	71.971
others	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$

Table 2: Feature evaluation for $(x + y)^2 > 0.25$

Feature	FR	Re	IG	GR	1R	χ^2
x	0.2090	0.140067	0.241	0.156	79.0	119.562
y	0.2456	0.151114	0.248	0.165	78.25	122.336
others	$= 0$	$\neq 0$	$= 0$	$= 0$	$\neq 0$	$= 0$

Table 3: Feature evaluation for $(x + y)^2 > 0.5$

Feature	FR	Re	IG	GR	1R	χ^2
x	0.2445	0.1486	0.134	0.134	87.75	57.455
y	0.2441	0.1659	0.159	0.164	87.25	73.390
others	$= 0$	$\neq 0$	$= 0$	$= 0$	$\neq 0$	$= 0$

Table 4: Feature evaluation for $(x + y)^3 < 0.125$

Feature	FR	Re	IG	GR	1R	χ^2
x	0.1057	0.0750547	0.169	0.123	64.25	73.653
y	0.0591	0.1079423	0.202	0.226	66.75	88.040
z	0.1062	0.0955878	0.202	0.160	67.50	84.283
others	$= 0$	$\neq 0$	$= 0$	$= 0$	$\neq 0$	$= 0$

Table 5: Feature evaluation for $x * y * z > 0.125$

Feature	FR	Re	IG	GR	1R	χ^2
x	0.1511	0.0980	0.1451	0.0947	76.5	65.425
y	0.1101	0.0557	0.0909	0.1080	78.0	35.357
z	0.2445	0.1474	0.2266	0.2271	79.75	93.812
others	$= 0$	$\neq 0$	$= 0$	$= 0$	$\neq 0$	$= 0$

Table 6: Feature evaluation for $x * y * z^2 > 0.125$