LINGUISTIC RULE INDUCTION BASED ON A RANDOM SET SEMANTICS

Zengchang Qin Jonathan Lawry
A.I. Group, Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TR, U.K.
Email: {z.qin, j.lawry}@bristol.ac.uk

ABSTRACT: Fuzzy logic based methods have been widely used in linguistic modeling [6]. Here we use a different framework of random set to interpret imprecise concepts. This framework is referred to as label semantics [8]. Within this framework, fuzzy concepts are modeled by quantifying the subjective uncertainty associated with whether or not a label expression is appropriate to describe a particular value. In this paper, a method of modeling data with logical expressions of fuzzy labels is discussed and a simple information based algorithm based on FOIL [10] is proposed for generating a set of linguistic rules for classification.

Keywords: Label semantics, linguistic rule, fuzzy labels, mass assignment, LFOIL.

1 INTRODUCTION

Rule learning has a long history within the field of machine learning and data mining. Many separate-and-conquer rule-based classification methods have been proposed and studied comprehensively. Rule learning has much better transparency compared to other sub symbolic models such as neural networks. A set of intuitively understandable rules can give us a better understanding of how the classification is made. The research of fuzzy rules (e.g, IF-THEN rules) have been widely studied in both fuzzy or machine learning communities because of their good transparency and comparable accuracy to other approaches [2]. Here in this paper, we use a random set framework to interpret linguistic or fuzzy rules.

Label semantics is a random set based framework for using linguistic expressions (or fuzzy labels) to model data. Previous work has been done to apply this framework to decision tree learning [9] and Naive Bayes learning [11]. Empirical results show that the new proposed models based on label semantics have both improved transparency and accuracy. In this paper, a method of using logical expressions of labels as linguistic rules for classification is discussed. FOIL, initially proposed by Quinlan [10], is a system to learn Horn clauses from data expressed as relations. In this paper, we proposed a new label semantics rule learning system based on FOIL and tested on two problems.

This paper is organized as follows: Section 2 gives a short introduction on label semantics and on the use random set descriptions of fuzzy labels to analyze data. In section 3, logical expressions of labels and the relation between random set description and logical expressions are discussed. In section 4, the new algorithm for linguistic rule learning is introduced. In section 5, we tested the new algorithm with an artificial data set and a real-world data set.

2 RANDOM SET SEMANTICS

Label semantics, proposed by Lawry [8], is an approach to modelling fuzzy concepts by quantifying the subjective uncertainty associated with whether or not a label expression is appropriate to describe a particular value or instance. The semantics is based on random set theory but different from earlier work of Goodman and Nguyen [4]. The underlying question posed by label semantics is how to use linguistic expressions to label numerical values. For a variable x into a domain of discourse Ω we identify a finite set of linguistic labels $\mathscr{L} = \{L_1, \cdots, L_n\}$ with which to label the values of x. Then for a specific value $x \in \Omega$ an individual I identifies a subset of $\mathscr{L}$, denoted D_x^I to stand for the description of x given by I, as the set of labels with which it is appropriate to label x. If we allow I to vary across a population V, then D_x^I will also vary and generate a random set denoted D_x into the power set of $\mathscr{L}$. We can view the random set D_x as a description of the variable x in terms of the labels in $\mathscr{L}$. More formally,

Definition 1 (*Label Description*) *For $x \in \Omega$ the label description of x is a random set from V into the power set of $\mathscr{L}$, denoted D_x, with associated distribution m_x, given by*

$$\forall S \subseteq \mathscr{L}, \quad m_x(S) = P_V(\{I \in V | D_x^I = S\})$$

Where P_V is the prior distribution of V (We usually assumes it is a uniform distribution) and $m_x(S)$ is the mass associated with a set of labels S and

$$\sum_{S \subseteq \mathscr{L}} m_x(S) = 1$$

Intuitively $m_x(S)$ quantifies the evidence that S is the set of appropriate labels for x.

For example, given a set of labels defined on the scores of a single throw of a particular dice: $\mathscr{L}_{score} = \{small, medium, large\}$. Suppose 7 of 10 people agree that '*small* is the only appropriate label for the score of 2' and 3 people agree 'both *small* and *medium* are appropriate labels'. According to def.1, $m_2(small, medium) = 0.3$ and $m_2(small) = 0.7$ so that the mass assignment for 2 is $m_2 = \{medium\} : 0.3$, $\{low, medium\}$: 0.7. More details about the theory of mass assignment can be found in [1].

Consider the previous example, can we know how appropriate for single label, say *small*, is to describe 2? In this framework, *appropriateness degrees* are used to evaluate how appropriate a single label is for describing a particular value of variable x.

Definition 2 (*Appropriateness Degrees*)

$$\forall x \in \Omega, \forall L \in \mathscr{L} \quad \mu_L(x) = \sum_{S \subseteq \mathscr{L}: L \in S} m_x(S)$$

Simply, given a particular value α of variable x, the appropriateness degree for labeling this value with the label L, which is defined by fuzzy set F, is the membership value of α in F. The reason we use the new term 'appropriateness degrees' is partly because it more accurately reflects the underlying semantics and partly to highlight the quite distinct calculus based on this framework [8]. This definition provides a relationship between mass assignments and appropriateness degrees. Consider the previous example, we then can obtain

$$\mu_{small}(2) = 0.7 + 0.3 = 1,\ \mu_{medium}(2) = 0.3$$

It is certainly true that a mass assignment on D_x determines a unique appropriateness degree for any functions of μ_L but generally the converse does not hold. For example, given $\mu_{L_1} = 0.3$ and $\mu_{L_2} = 1$. We may obtain the sets of appropriate labels with associated masses as:

$$\begin{array}{c} \{L_2\}:0.7, \{L_1,L_2\}:0.3 \\ \{L_1\}:0.1, \{L_2\}:0.8, \{L_1,L_2\}:0.2 \\ \{L_1\}:0.2, \{L_2\}:0.9, \{L_1,L_2\}:0.1 \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \end{array}$$

There are infinite number of possible representations. That is if we know the appropriateness degrees of the labels, we may not be able to infer a unique underlying mass assignment. This problem can be overcame by making some assumptions.

The first is the *consonance assumption*, according to which we can determine the mass assignment uniquely from the appropriateness degrees as follows. (For the justification of the consonance assumption, see [8])

Definition 3 (*Consonance Assumption*) *Let* $\{\beta_1, \cdots, \beta_k\} = \{\mu_L(x) | L \in LA,\ \mu_L(x) > 0\}$ *ordered such that* $\beta_t > \beta_{t+1}$ *for* $t = 1, 2, \cdots, k-1$ *then:*

$$m_x = M_t : \beta_t - \beta_{t-1}, t = 1, 2, \cdots, k-1,$$

$$M_k : \beta_k, \quad M_0 : 1 - \beta_1$$

where $M_0 = \emptyset$ *and* $M_t = \{L \in \mathscr{L} | \mu_L(x) \geq \beta_t\}$ *for* $t = 1, 2 \ldots, k$.

Based on this assumption, there is a unique mass assignment for a given set of appropriateness degree values. For example, given $\mu_{L_1} = 0.3$ and $\mu_{L_2} = 1$, the only unique consonant mass assignment is $\{L_2\}:0.7, \{L_1,L_2\}:0.3$, but not $\{L_2\}:0.8, \{L_1,L_2\}:0.2, \{L_1\}:0.1$ or others. However, it is undesirable to have mass associated with the empty set. In order to avoid this, we define a *full fuzzy covering* assumption as follows:

Definition 4 (*Full Fuzzy Covering*) *Given a continuous discourse* Ω*, LA is called a full fuzzy covering of* Ω *if:*

$$\forall x \in \Omega, \exists L \in \mathscr{L} \qquad \mu_L(x) = 1$$

Suppose we use N_F fuzzy sets with 50% overlap, so that the appropriateness degrees satisfy: $\forall x \in \Omega$, $\exists i \in \{1, \cdots, N_F - 1\}$ such that $\mu_{L_i}(x) = 1$, $\mu_{L_{i+1}} = \alpha$ and $\mu_{L_j}(x) = 0$ for $j < i$ or $j > i+1$. In this case,

$$m_x = \{L_i\} : 1 - \alpha, \{L_i, L_{i+1}\} : \alpha \qquad (1)$$

In this paper, unless otherwise stated, the fuzzy labels are defined by trapezoidal fuzzy sets with 50% overlap. For example, see figure 1 which shows a full fuzzy covering of Ω with

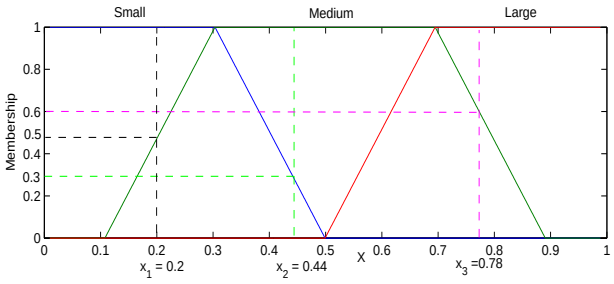

Figure 1: A full fuzzy covering (discretization) with three trapezoidal fuzzy sets with 50% overlap on a continuous universe.

three fuzzy labels: *small*, *medium* and *large*. Based on these assumptions, we can isolate a set of subsets of $\mathscr{L}$ with non-zero mass assignments. These are referred to as *focal sets*:

Definition 5 (Focal Set) *Given a universe* Ω *for variable x, the focal set of* $\mathscr{F}$ *is a set of focal elements defined as:*

$$\mathscr{F} = \{S \subseteq \mathscr{L} | \exists x \in \Omega, m_x(S) > 0\}$$

Figure 1 shows the universe of a variables x which is fully covered by 3 fuzzy sets with 50% overlap. For x, the following focal elements occur: $\{small\}$, $\{small, medium\}$, $\{medium\}$, $\{medium, large\}$ and $\{large\}$. Since *small* and *large* do not overlap, the set $\{small, large\}$ cannot occur as a focal element according to def. 5. We can then always find the unique translation from a given data point to a mass assignment on focal elements, specified by the function μ_L; This is referred to as *linguistic translation (LT)*. For a particular attribute with an associated focal set, linguistic translation is a process of replacing data elements with masses of focal elements of these data. For example in fig. 1,

$$\mu_{small}(x_2 = 0.44) = 0.3,\ \mu_{medium}(x_2 = 0.44) = 1$$

and $\mu_{large}(0.44) = 0$. They are simply the memberships read from the fuzzy sets. We then can obtain the mass assignment of this data element according to eq. 1:

$$m_{0.44}(medium) = 0.7,\ m_{0.44}(small, medium) = 0.3$$

Similarly, the linguistic translations for $x_3 = 0.2$ and $x_3 = 0.78$ are as follows:

$x_1 = \langle \mu_{small}(0.2) = 1,\ \mu_{medium}(0.5) = 1 \rangle \rightarrow$
$m_{0.2}(small) = 0.5,\ m_{0.2}(small, medium) = 0.5$

$x_3 = \langle \mu_{medium}(0.78) = 0.6,\ \mu_{large}(0.78) = 1 \rangle \rightarrow$
$m_{0.78}(large) = 0.4,\ m_{0.78}(medium, large) = 0.6,$

Through the linguistic translation, numerical data can be represented by a random set descriptions on fuzzy labels. This framework provides me a way of using high level knowledge representation language to model data. In next section, we will discuss the relation of logical expression of labels and random set description of labels, while the former is the representation of linguistic rules and the latter provides us a way of quantifying the appropriateness of those rules.

3 LOGICAL EXPRESSIONS OF FUZZY LABELS

In label semantics, linguistic rules are represented by propositional logic sentences. Consider a formal language consisting of the set of labels $\mathscr{L} = \{L_1, \ldots, L_n\}$, we can present compound linguistic descriptions generated recursively by the applications of the connectives:

Definition 6 (*Logical Expressions of Labels*) *The set of logical expressions, LE, is defined by recursively as follows:*

(i) $L_i \in LE$ *for* $i = 1, \ldots, n$.
(ii) If $\theta, \varphi \in LE$ *then* $\neg\theta, \theta \wedge \varphi, \theta \vee \varphi, \theta \rightarrow \varphi \in LE$

Basically, we interpret the main logical connectives as follows: $\neg L$ means that L is not an appropriate label, $L_1 \wedge L_2$ means that both L_1 and L_2 are appropriate labels, $L_1 \vee L_2$ means that either L_1 or L_2 are appropriate labels, and $L_1 \rightarrow L_2$ means that L_2 is an appropriate label whenever L_1 is. If we consider logical label expressions formed from $\mathscr{L}$ by recursive application of the connectives then an expression θ identifies a set of *possible label sets* according to the following λ-function.

Definition 7 (λ-*function*) *Let θ and φ be expressions generated by recursive application of the connectives $\neg, \vee, \wedge$ and $\rightarrow$ to the elements of $\mathscr{L}$ (i.e. $\theta, \varphi \in LE$). Then the set of possible label sets defined by a linguistic expression can be determined recursively as follows:*

$$
\begin{array}{ll}
(i) & \lambda(L_i(x)) = \{S \subseteq \mathscr{F} | \{L_i\} \subseteq S\} \\
(ii) & \lambda(\neg\theta) = \overline{\lambda(\theta)} \\
(iii) & \lambda(\theta \wedge \varphi) = \lambda(\theta) \cap \lambda(\varphi) \\
(iv) & \lambda(\theta \vee \varphi) = \lambda(\theta) \cup \lambda(\varphi) \\
(v) & \lambda(\theta \rightarrow \varphi) = \overline{\lambda(\theta)} \cup \lambda(\varphi)
\end{array}
$$

It should also be noted that the λ-function provides us with notion of logical equivalence for label expressions

$$\theta \equiv_L \varphi \Longleftrightarrow \lambda(\theta) = \lambda(\varphi)$$

Basically, λ-function provides a way of transferring logical expressions of labels (linguistic rules) to random set descriptions of labels (i.e. focal elements). $\lambda(\theta)$ corresponds to those subsets of $\mathscr{F}$ identified as being possible values of D_x by expression θ. $\lambda(\theta)$ corresponds a particular subset of focal set based on the assumptions we made in section 2.

Example 1 *Given a continuous variable x shown in fig. 1 and $\mathscr{L}_x = \{small, medium, large\}$, suppose we are told that "x is **not large** but it is **small or medium**". This constraint can be interpreted as the logical expression*

$$\theta_x = \neg large \wedge (small \vee medium)$$

According to definition 7, the possible label sets of the given logical expression θ_x are

$$\lambda(\theta_x) = \lambda(\neg large \wedge (small \vee medium)) = \{\{small\}, \{small, medium\}, \{medium\}\} \wedge (\{\{small\}, \{small, medium\}\} \vee \{\{small, medium\}, \{medium\}, \{medium, large\}\}) = \{\{small\}, \{small, medium\}, \{medium\}\}$$

3.1 Linguistic Interpretation of Appropriate Labels

Based on the inverse of the λ-function (def. 7), a set of linguistic rules (or logical label expressions) can be obtained from a given set of possible label sets. For example, suppose we are given the possible label sets $\{\{small\}, \{small, medium\}, \{medium\}\}$, which does not have an immediately obvious interpretation. However using the α-function (see below), we can convert this set into a corresponding linguistic expression $(small \vee medium) \wedge \neg large$ or its logical equivalence.

Definition 8 (α-*function*)

$$\forall F \in \mathscr{F} \quad let \quad \mathscr{N}(F) = \left(\bigcup_{F' \in \mathscr{F} : F' \supseteq F} F' \right) - F \qquad (2)$$

$$then \quad \alpha_F = \left(\bigwedge_{L \in F} L \right) \wedge \left(\bigwedge_{L \in \mathscr{N}(F)} \neg L \right) \qquad (3)$$

We can then map a set of focal sets to label expressions based on the α-function as follows:

$$\forall R \in \mathscr{F} \quad \theta_R = \bigvee_{F \in R} \alpha_F \quad where \quad \lambda(\theta_R) = R \qquad (4)$$

The motivation of this mapping is a follows. Given a focal set $\{s, m\}$ this states that the labels appropriate to describe the attribute are exactly *small* and *medium*. Hence, they include *s* and *m* and exclude all other labels that occur in focal sets that are supersets of $\{s, m\}$. Given a set of focal sets $\{\{s, m\}, \{m\}\}$ this provides the information that the set of labels is either $\{s, m\}$ or $\{m\}$ and hence the sentence providing the same information should be the disjunction of the α sentences for both focal sets. The following example gives the calculation of the α-function.

Example 2 *Let $\mathscr{L}$ ={very small (vs), small (s), medium (m), large(l), very large (vl)} and $\mathscr{F}$ ={{vs,s}, {s}, {s,m}, {m}, {m,l}, {l}, {l,vl}}. For calculating $\alpha_{\{l\}}$, we obtain*

$$F' \in \mathscr{F} : F' \supseteq \{l\} = \{\{m, l\}, \{l\}, \{l, vl\}\} = \{m, l, vl\}$$

$$\mathscr{N}(\{l\}) = \left(\bigcup_{F' \in \mathscr{F} : F' \supseteq \{l\}} F' \right) - \{l\} = \{l, vl, m\} - \{l\} = \{vl, m\}$$

$$\alpha_{\{l\}} = \left(\bigwedge_{L \in F} L \right) \wedge \left(\bigwedge_{L \in \mathscr{N}(F)} \neg L \right) = (l) \wedge (\neg m \wedge \neg vl) = \neg m \wedge l \wedge \neg vl$$

Also we can also obtain

$$\alpha_{\{m,l\}} = m \wedge l \quad \alpha_{\{l,vl\}} = l \wedge vl$$

Hence, a set of label sets $\{\{m, l\}, \{l\}, \{l, vl\}\}$ can be represented by a linguistic expression as follows,

$$\theta_{\{\{m,l\},\{l\},\{l,vl\}\}} = \alpha_{\{m,l\}} \vee \alpha_{\{l\}} \vee \alpha_{\{l,vl\}} =$$

$$(m \wedge l) \vee (\neg\, m \wedge l \,\neg\, vl) \vee (l \wedge vl) \equiv_L large$$

where '$\equiv_L$' represents logical equivalence (see def. 7).

Basically, α-function provides a way of obtaining logical expressions from a random set description of labels. It is an inverse process of λ-function.

3.2 Appropriateness Degrees for Linguistic Rules

Based on def. 7, we can easily extend λ-function to the multi-dimensional case, such that the set of *n*-dimensional label expressions $MLE^{(n)}$ is defined by:

Definition 9 (*Multi-dimensional λ-function*) *$\lambda^{(n)}$: $MLE^{(n)} \rightarrow 2^{(2^{\mathscr{L}_1}\times\cdots\times 2^{\mathscr{L}_n})}$ is defined recursively as follows: Let $\mathscr{F}_j$ denote the set of focal elements for $\mathscr{L}_j$: $j = 1,\cdots,n$ then $\forall\theta \in MLE^{(n)}$, $\lambda^{(n)} \subseteq \mathscr{F}_1 \times \cdots, \mathscr{F}_n$.*

Given a particular data, how can we evaluated if a linguistic rule is appropriate for describing it? Based on the one-dimensional case, we now extend the concepts of appropriateness degrees to the multi-dimensional case as follows:

Definition 10 (*Multi-dimensional Appropriateness Degrees*) *Given a set of n-dimensional label expression $MLE^{(n)}$:*

$$\forall\ \theta \in MLE^{(n)}, \forall x_j \in \Omega_j : j = 1,\cdots,n,$$

$$\mu^n_\theta(\mathbf{x}) = \mu^n_\theta(x_1,\cdots,x_n) = \sum_{\langle F_1,\cdots,F_n\rangle\in\lambda^{(n)}(\theta)} (F_1,\cdots,F_n)$$

$$= \sum_{\langle F_1,\cdots,F_n\rangle\in\lambda^{(n)}(\theta)} \prod_{j=1}^{n} m_{x_j}(F_j)$$

The appropriateness degrees in one-dimension are for evaluating a single label for describing a single data element, while in multi-dimensional cases are for evaluating a linguistic rule for describing a data vector.

Example 3 *Consider a modelling problem with two variables x_1 and x_2 for which $\mathscr{L}_1 = \{small\ (s), medium\ (med), large(lg)\}$ and $\mathscr{L}_2 = \{low(lo), moderate(mod), high(h)\}$. Also suppose the focal elements for $\mathscr{L}_1$ and $\mathscr{L}_2$ are:*

$$\mathscr{F}_1 = \{\{s\},\{s,med\},\{med\},\{med,lg\},\{lg\}\}$$

$$\mathscr{F}_2 = \{\{lo\},\{lo,mod\},\{mod\},\{mod,h\},\{h\}\}$$

According to the multi-dimensional generalization of definition 7 we have that

$$\lambda^{(2)}((med\wedge\neg s)\wedge\neg lo) = \lambda^{(2)}(med\wedge\neg s)\cap\lambda^{(2)}(\neg lo)$$

$$= \lambda(med\wedge\neg s)\times\lambda(\neg lo)$$

Now, the set of possible label sets is obtained according to the λ-function:

$$\lambda(med\wedge\neg s) = \{\{med\},\{med,lg\}\}$$

$$\lambda(\neg lo) = \{\{mod\},\{mod,h\},\{h\}\}$$

Hence,

$$\lambda^{(2)}((med\wedge\neg s)\wedge\neg lo) = \{\langle\{med\},\{mod\}\rangle,\langle\{med\},$$

$$\{mod,h\}\rangle,\langle\{med\},\{h\}\rangle,\langle\{med,lg\},\{mod\}\rangle,$$

$$\langle\{med,lg\},\{mod,h\}\rangle,\langle\{med,lg\},\{h\}\rangle\}$$

Given $\mathbf{x} = \langle x_1,x_2\rangle = \langle x_1 = \{med\} : 0.6, \{med,lg\} : 0.4\rangle$, $\langle x_2 = \{lo,mod\} : 0.8, \{mod\} : 0.2\rangle$, we obtain:

$$\mu_\theta(\mathbf{x}) = (m(\{med\}) + m(\{med,lg\}))\times(m(\{mod\})+$$

$$m(\{mod,h\}) + m(\{h\})) = (0.6+0.4)\times(0.2+0+0) = 0.2$$

And according to def. 7:

$$\mu^n_{\neg\theta}(\mathbf{x}) = 1 - \mu_\theta(\mathbf{x}) = 0.8$$

4 LINGUISTIC RULE INDUCTION

In previous sections, we have shown how to evaluate the appropriateness of using a linguistic rule to describe a data vector. In this section, a new algorithm for learning a set of linguistic rules is proposed based on the FOIL algorithm [10]. It is referred to as Linguistic FOIL (LFOIL). Like FOIL, Linguistic FOIL uses an information-based estimate which provides effective guidance for rule construction.

4.1 Information Heuristics

The heuristics are for assessing the usefulness of a literal as the next component of the rule. Here the heuristics used for learning the linguistic rules are modified from the FOIL algorithm [10]. Consider a classification rule of the form:

$$R_i = \theta \rightarrow C_j \ \ where\ \ \theta \in MLE^{(n)}$$

Given a data set $\mathscr{D}$ and a particular class C_j, the data belonging to class C_j are referred to as *positive examples* and the rest of them are *negative examples*. For the given rule R_i, the coverage of positive data is evaluated by

$$T_i^+ = \sum_{k\in\mathscr{D}_j} \mu_\theta(\mathbf{x}_k) \tag{5}$$

and the coverage of negative examples is given by

$$T_i^- = \sum_{k\in\mathscr{D}-\mathscr{D}_j} \mu_\theta(\mathbf{x}_k) \tag{6}$$

where $\mathscr{D}_j$ is the subset of the database which is consisted by the data belonging to class C_j. The information for the original rule R_i [1] can by evaluated by

$$I(R_i) = -\log_2\left(\frac{T_i^+}{T_i^+ + T_i^-}\right) \tag{7}$$

Suppose we then propose to another label expression φ to the body of R_i to generate a new rule

$$R_{i+1} = \varphi\wedge\theta \rightarrow C_j$$

where $\varphi,\theta \in MLE^{(n)}$. By adding the new literal φ, the information becomes:

$$T_{i+1}^+ = \sum_{k\in\mathscr{D}_j} \mu_{\theta\wedge\varphi}(\mathbf{x}_k) \tag{8}$$

$$T_{i+1}^- = \sum_{k\in\mathscr{D}-\mathscr{D}_j} \mu_{\theta\wedge\varphi}(\mathbf{x}_k) \tag{9}$$

Therefore,

$$I(R_{i+1}) = -\log_2\left(\frac{T_{i+1}^+}{T_{i+1}^+ + T_{i+1}^-}\right) \tag{10}$$

Then we can evaluate the information gain from adding expression φ as follows.

$$G(\varphi) = T_{i+1}^+(I(R_i) - I(R_{i+1})) \tag{11}$$

[1] We use two different notations of appropriateness degrees for the rule $R_i = \theta \rightarrow C_j$: μ_θ and μ_{R_i}. The former is used in logical expressions and the latter is used in rule learning algorithm.

We can see that G measure consists of two components. T_{i+1}^{+} is the coverage of positive data by the new rule R_{i+1} and $(I(R_i) - I(R_{i+1}))$ is the increase of information. The probability of C_j given a linguistic rule R_i is evaluated by:

$$P(C_j|R_i) = \frac{\sum_{k\in\mathscr{D}_j}\mu_\theta(\mathbf{x}_k)}{\sum_{k\in\mathscr{D}}\mu_\theta(\mathbf{x}_k)} = \frac{T_i^+}{T_i^+ + T_i^-} \tag{12}$$

when $P(C_j|R_{i+1}) > P(C_j|R_i)$ (i.e., by appending a new literal, more positive examples are covered), we can obtain that $(I(R_i) - I(R_{i+1})) > 0$. By choosing a literal φ with maximum G value, we can form the new rule which covers more positive examples and thus increasing the accuracy of the rule.

4.2 LFOIL

We define a prior knowledge based $KB \subseteq MLE^{(n)}$ and a probability threshold $PT \in [0,1]$. KB consists of fuzzy label expressions based on labels defined on each attribute. For example,suppose we use fuzzy labels $\{small_1\ large_1\}$ to describe attribute 1 and $\{small_2\ large_2\}$ to describe attribute 2. In this case a possible knowledge base is: $KB = \{small_1, \neg small_1, large_1, \neg large_1, small_2, \neg small_2, large_2, \neg large_2\}$.

Generating a Rule

- Given rule $R_i = \theta_1 \wedge \cdots \wedge \theta_d \rightarrow C_j$ be the rule at step i, we find the next literal $\theta_{d+1} \in KB - \{\theta_1, \cdots, \theta_d\}$ for which $G(\theta_{d+1})$ is maximal.
- Replace rule R_i with $R_{i+1} = \theta_1 \wedge \cdots \wedge \theta_d \wedge \theta_{d+1} \rightarrow C_j$
- If $P(C_j|\theta_1 \wedge \cdots \wedge \theta_{i+1}) \geq PT$ then terminate else repeat.

Generating a Rule Base

Let $\Delta_i = \{R_1 \rightarrow C_j, \cdots, R_t \rightarrow C_j\}$ be the rule-base at step i. We evaluate the coverage of Δ_i as follows:

$$CV(\Delta_i) = \frac{\sum_{k\in\mathscr{D}_j}\mu_{R_1\vee\cdots\vee R_t}(\mathbf{x}_k)}{|\mathscr{D}_j|} \tag{13}$$

We define a coverage function $\delta : \Omega_1 \times \cdots \times \Omega_n \rightarrow [0,1]$ according to:

$$\delta(\mathbf{x}|\Delta_i) = \mu_{\neg\Delta_i}(\mathbf{x}) = \mu_{\neg(R_1\vee\cdots\vee R_t)}(\mathbf{x}) \tag{14}$$
$$= 1 - \mu_{(R_1\vee\cdots\vee R_t)}(\mathbf{x}) = 1 - \sum_{w=1}^{t}\mu_{R_w}(\mathbf{x})$$

where $\delta(\mathbf{x}|\Delta_i)$ represents the degree to which $\mathbf{x}$ is *not* covered by a given rule base Δ_i. If CV is less than a predefined coverage threshold $CT \in [0,1]$:

$$CV(\Delta_i) < CT$$

then we generate a new rule for class C_j according to the above rule generation algorithm to form a new rule base Δ_{i+1} but where the entropy calculations are amended such that for a rule $R = \theta \rightarrow C_j$,

$$T^+ = \sum_{k\in\mathscr{D}_j}\mu_\theta(\mathbf{x}_k) \times \delta(\mathbf{x}_k|\Delta_i) \tag{15}$$

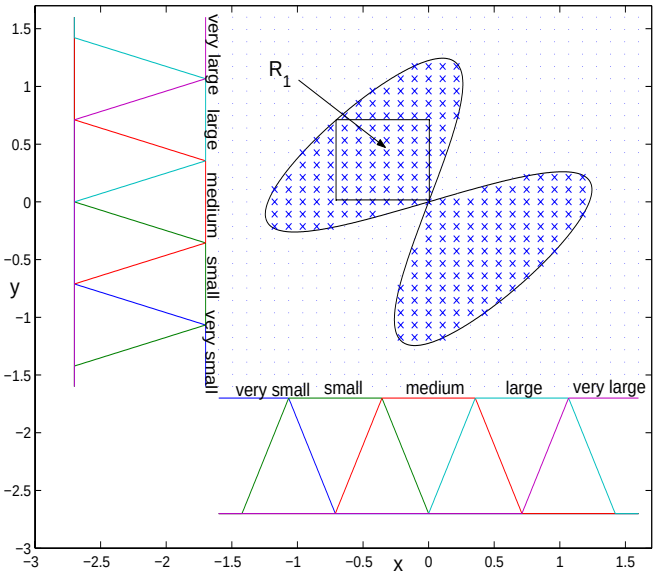

Figure 2: The illustration of the 'eight' problem, where each attribute is discretized by 5 fuzzy sets: *very small, small, medium, large* and *very large*, respectively.

$$T^- = \sum_{k\in\mathscr{D}-\mathscr{D}_j}\mu_\theta(\mathbf{x}_k) \tag{16}$$

The algorithm terminates when $CV(RB_{i+1}) \geq CT$ or $CV(RB_{i+1}) - CV(RB_i) < \varepsilon$ where $\varepsilon \in [0,1]$ is a very small value.

4.3 Class Probabilities Given a Rule Base

Given a rule base $\Delta_i = \{R_1 \rightarrow C_j, \cdots, R_t \rightarrow C_j\}$ and an unclassified data $\mathbf{x}$, we can estimate the probability of C_j, $P(C_j|\mathbf{x})$, as follows: Firstly, we determine the rule $R_{max} \rightarrow C_j$ for which $\mu_{R_k}(\mathbf{x})$ is maximal:

$$R_{max} = \max_{k\in\Delta_i}\mu_{R_k} \tag{17}$$

Given the unclassified data $\mathbf{x}$, rule R_{max} is the most appropriate rule from the rule base we learned. For the rule $R_{max} \rightarrow C_j$ we evaluate two probabilities p_{max} and q_{max} where:

$$p_{max} = P(C_j|R_{max}) \tag{18}$$

and,

$$q_{max} = P(C_j|\neg R_{max}) \tag{19}$$

We then use Jeffrey's rule [5] to evaluate the class probability by:

$$P(C_j|\mathbf{x}) = p_{max} \times \mu_{R_{max}}(\mathbf{x}) + q_{max} \times (1 - \mu_{R_{max}}(\mathbf{x})) \tag{20}$$

5 EXPERIMENTAL STUDIES

In this section we test the new algorithm with a toy problem described as follows: A figure of eight shape was generated according to the equation $x = 2^{(-0.5)}(sin(2t) - sin(t))$ and $y = 2^{(-0.5)}(sin(2t) + sin(t))$ where $t \in [0, 2\pi]$ (see figure 2). Points in $[-1.6, 1.6]^2$ are classified as legal if they lie within

the 'eight' shape (marked with $\times$) and illegal if they lie outside (marked with points). The database is consisted of 961 examples generated from a regular grid on $[-1.6, 1.6]^2$ for training, and 961 unseen examples from the same distribution as the test data set.

The following rules are generated by LFOIL algorithm with $PT = 0.7$, $CV = 0.9$ and $\varepsilon = 0.005$:

R_1 : x is $\neg$ very small $\wedge$ small $\wedge$ medium $\wedge$ $\neg$ large and y is $\neg$ small $\wedge$ medium $\rightarrow$ legal
R_2 : x is $\neg$ small $\wedge$ medium and y is $\neg$ very small $\wedge$ small $\wedge$ medium $\wedge$ $\neg$ large $\rightarrow$ legal
R_3 : x is medium $\wedge$ $\neg$ large and y is large $\wedge$ very large $\rightarrow$ legal
R_4 : x is large $\wedge$ very large and y is medium $\wedge$ $\neg$ large $\rightarrow$ legal
R_5 : x is very small $\wedge$ small $\wedge$ $\neg$ medium and y is medium $\wedge$ $\neg$ large $\rightarrow$ legal
R_6 : x medium $\wedge$ $\neg$ large and y is very small $\wedge$ small $\wedge$ $\neg$ medium $\rightarrow$ legal

These rules are symmetric and as we can see from the fig. 2, the rules capture the legal area very well. The area covered by R_1 is marked by a box shown in fig. 2.

We also tested the LFOIL on the Pima Indianan data which is a benchmark problem from UCI machine repository [3]: the database contains the details of 768 females from the population of Pima Indians living near Phoenix Arizona, USA. The diagnostic binary-valued variable investigated is whether the patient shows sign of diabetes according to the world Health Organisation criteria. We use 3 fuzzy labels: *low*, *medium* and *high* for each of the 8 attributes. Half of the data is used for training and the rest of them for test. We got the test accuracy of 71.67% with the following rules that decide patient have the signs of diabetes:

R_1: *Plasma concentration* (Attribute 2) is low $\wedge$ medium and *the Number of times pregnant* (Attribute 1) is medium $\wedge$ $\neg$ high
R_2: *Plasma concentration* is medium and *age* (Attribute 8) is $\neg$ low
R_3: *Plasma concentration* is low $\wedge$ medium and *the Number of times pregnant* is high
R_4: *Plasma concentration* is $\neg$ medium $\wedge$ high and *Diabetes pedigree function* (Attribute 7) is medium

For this problem, C4.5 has the average accuracy of 72.16%, Naive Bayes has 75.05% and Neural network has 74.64% [9]. Although the accuracy is not better than other models (it is only a slightly worse than C4.5), the transparency is greatly improved by using only 4 rules that give much better understanding this problem.

6 CONCLUSIONS AND DISCUSSIONS

In this paper, we introduce a method for linguistic rule generation based on label semantics. In particular, a new algorithm is proposed based on FOIL algorithm and tested on a toy problem and a real-world problem from UCI repository. The results show that very compact linguistic rules can be learned that reflect the essence of the problem.

The main contribution of this paper is to describe a method of evaluating linguistic rules through label semantics and to propose a new FOIL based algorithm for linguistic rule learning. In the new algorithm, we use a information based heuristics to guide the rule construction. This is not the only way of constructing good rules. Another approach is to search exhaustively through the knowledge base *KB*. Assuming that we do not use too many fuzzy labels for discretization, this approach may also be computational tractable. The rules which covers less positive examples will be discard according to a predefined threshold. Ref [2] reports similar idea for generating simple fuzzy logic (IF-THEN) rules. Future work is needed for testing this approach with more data sets and to study the influence of ranging parameter settings.

ACKNOWLEDGEMENT

We thank the anonymous reviewer for pointing out some related research and useful suggestions. This research is partly funded by ORS, UK.

REFERENCES

[1] J. F. Baldwin, T.P. Martin and B.W. Pilsworth. *Fril-Fuzzy and Evidential Reasoning in Artificial Intelligence*. John Wiley & Sons Inc, 1995.

[2] J. F. Baldwin and D. Xie. Simple fuzzy logic rules based on fuzzy decision tree for classification and prediction problem. *Intelligent Information Processing II*, Z. Shi and Q. He (Ed.), Springer, 2004.

[3] C. Blake and C. J. Merz. UCI machine learning repository. *http://www.ics.uci.edu/ mlearn/ MLRepository.html*

[4] I. R. Goodman. Fuzzy sets as equivalence classes of random sets. *Fuzzy Sets and Possibility Theory*. Ed. R. Yager pp 327-342, 1982.

[5] R. C. Jeffrey, *The Logic of Decision*, Gordon & Breach Inc., New York, 1965.

[6] J. Lawry, J. Shanahan, and A. Ralescu. *Modelling with Words: Learning, fusion, and reasoning within a formal linguistic representation framework*. LNAI 2873. Springer-Verlag 2003.

[7] J. Lawry, J. Hall, R. Bovey. Fusion of expert and learnt knowledge in a framework of fuzzy labels, *International Journal of Approximate Reasoning*, Vol. 36, pp151-198, 2004

[8] J. Lawry. A framework for linguistic modelling, *Artificial Intelligence*, 155: pp. 1-39, 2004.

[9] Z. Qin and J. Lawry. Decision tree learning with fuzzy labels, *Information Sciences*, To appear

[10] J. R. Quinlan. Learning logical definitions from relations, *Machine Learning*, 5, 239-266, 1990.

[11] N. J. Randon and J. Lawry. Classification and query evaluation using modelling with words, *Information Sciences, Special Issue - Computing with Words: Models and Applications*, To appear.