Linguistic Decision Tree Induction

Zengchang Qin and Jonathan Lawry
A.I. Group, Department of Engineering Mathematics
University of Bristol
Bristol BS8 1TR, United Kingdom
{*Z.Qin, J.Lawry*}*@bristol.ac.uk*

Abstract

Label Semantics is a random set based framework for modeling imprecise concepts where the degree of appropriateness of a linguistic expression as a description of a certain value is measured in terms of how the set of appropriate labels for that value varies across a population. An approach to decision tree induction based on this framework was studied. A new decision tree learning algorithm was proposed and its performance applied in real-world data sets was compared with the C4.5 algorithm.

1 Introduction

Traditionally machine learning and data mining research have focused on learning algorithms with high classification or prediction accuracy. However, this is not always sufficient for some application areas. We may require good algorithm transparency which means that models need to be easily understood and provide information regarding underlying trends and relationships. The research area of *Computing with Words (CW)*, proposed by Zadeh [6], provides us with a framework in which to develop such a system. Linguistic expressions such as *small*, *medium* and *large* whose meaning can be represented by fuzzy sets, can be used to for modeling and computing.

Here we present an alternative framework for CW which was proposed by Lawry [2]. Label Semantics, the new framework, is a random set based semantics for modeling imprecise concepts where the degree of appropriateness of a linguistic expression as a description of a value is measured by mass assignment on labels. Linguistic expressions are labels such as *small*, *medium* and *large*. Such labels are defined by overlapped fuzzy sets which are used to cover the continuous universe of the variables. Based on this semantics, a new tree-structured model, *Linguistic Decision Tree (LDT)* is proposed. A linguistic decision tree expands with *focal elements* from level to level guided by information heuristics. For each branch, the class probabilities given this branch will be evaluated based on *linguistic data set*. If one of the class probabilities reaches a given threshold probability, this branch will be terminated, otherwise, it will continue splitting at the next level until a given maximum depth is reached.

2 Label Semantics

The underlying question posed by label semantics is how to use linguistic expressions (defined by fuzzy sets) to label numerical values. For a variable x into a domain of discourse Ω we identify a finite set of linguistic words (or labels) $LA = \{L_1, \cdots, L_n\}$ with which to label the values of x. Then for a specific value $\alpha \in \Omega$ an individual I identifies a *subset* of LA, denoted D^I_α to stand for the description of α given by I, as the set of words with which it is appropriate to label α. If we allow I to vary across a population V, then D^I_x will also vary and generate a random set denoted D_x into the power set of LA. The frequency of occurrence of a particular label, say S, for D_x across the population then we obtain a distribution on D_x referred to as a mass assignment (see [1]) on labels, more formally:

Definition 1 (*Mass Assignment*)

$$\forall S \subseteq LA, \quad m_x(S) = \frac{|\{I \in V | D^I_x = S\}|}{|V|}$$

In this framework, *appropriateness degree* is used to evaluate how appropriate a label is for describing a particular value of variable x. It can be defined as:

Definition 2 (*Appropriateness Degrees*)

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

This definition provides a relationship between mass assignments and appropriateness degrees. Clearly μ_L is a function from Ω into [0,1] and therefore can technically be viewed as a fuzzy set. 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 belonging to F. The reason we use the new term 'appropriateness degree' is partly because it more accurately reflects the underlying semantics and partly to highlight the quite distinct calculus based on this framework.

2.1 Label Semantics for Data Analysis

Based on the underlying semantics, we can translate a set of numeric data into a set of linguistic data, where each data value is replaced by a mass assignment label set. It is certainly true that a mass assignment on D_x determines a unique appropriateness degree for any function but generally the converse does not hold. 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 overcome by the consonance assumption, according to which we can determine the mass assignment uniquely from the appropriateness degrees as follows: Let $\{y_1, y_2, \cdots, y_k\} = \{\mu_L(x) | L \in LA, \mu_L(x) > 0\}$ ordered such that $y_t > y_{t+1}$ for $t = 1, 2, \cdots, k-1$ then:

$$m_x = M_t : y_t - y_{t-1}, t = 1, 2, \cdots, k-1,$$
$$M_k : y_k, \quad M_0 : 1 - y_1$$

where $M_0 = \emptyset$ and $M_t = \{L \in LA | \mu_L(x) \geq y_t\}$ for $t = 1, 2 \ldots, k$. However, it is undesirable to have mass associated with the empty set. In order to avoid this, we need to make a *full fuzzy covering* of the continuous universe.

Definition 3 (*Full Fuzzy Covering*) *Given a continuous discourse Ω, LA is called a full fuzzy covering of Ω if:*

$$\forall x \in \Omega, \exists L \in LA \qquad \mu_L(x) = 1$$

Suppose we use N_F fuzzy sets with 50% overlap, so that the appropriateness degrees satisfy that $\forall x \in \Omega$, $\exists i \in \{1, \cdots, N_F - 1\}$ 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,

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

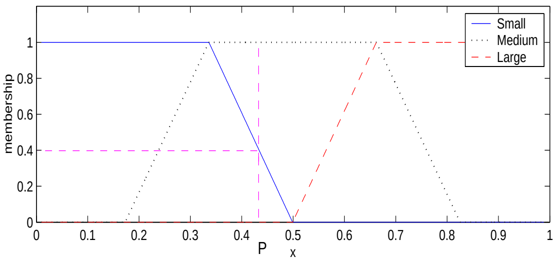

Figure 1: Full fuzzy covering with 3 trapezoidal fuzzy sets with 50% overlap.

For example, Figure 1 shows a full fuzzy covering of the universe with three fuzzy labels: **small**, **medium** and **large**. For the data point P, the appropriate labels are **small** and **medium**, and the appropriateness degrees of these labels are:

$$\mu_{small}(P) = 0.4, \qquad \mu_{medium}(P) = 1$$

We can then obtain the mass assignments as follows:

$$m_P = \{medium\} : 0.6, \{small, medium\} : 0.4$$

It is also interesting to note that given definitions for the appropriateness degrees on labels we can isolate a set of subsets of LA as the only values of D_x with non-zero probability. These are referred to as *focal sets*:

Definition 4 (*Focal Sets*) *The focal set of LA is a set of focal elements defined as follows:*

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

Based on our assumption of a full fuzzy covering with 50% overlap, 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. Then we can always find the unique translation from a given data point to mass assignment on focal elements, specified by the function of μ_L. The new data set after translation is called *linguistic data set.*

Definition 5 (*Linguistic Data Set*) *A linguistic data set is a translation of real value data set based on label semantics. Suppose we are given a given data set $D = \{x_1(i), \cdots, x_n(i) | i = 1, \cdots, N\}$ and focal set on attribute j: $\mathcal{F}_j = \{F_{1_j}, \cdots, F_{h_j} | j = 1, \cdots, n\}$, then the linguistic data set is defined as follow:*

$$LD = \{A_1(i), \cdots, A_n(i) | i = 1, \cdots N\}$$

$$A_j(i) = \{< m_{x_{j_1(i)}}(F_{1_j}), \cdots, m_{x_{j_h(i)}}(F_{h_j}) >\}$$

where $m_{x_{j_r(i)}}(F_{r_j})$ is the associated mass of focal element F_{r_j}, $r = 1, \cdots, h$

3 Linguistic Decision Tree

The ID3 [4] algorithm for decision trees induction has proved to be an effective and popular algorithm for building decision trees from discrete valued data sets. However, it cannot cope with classification problems with continuous attribute values. Here we propose a new decision tree induction algorithm based on label semantics. Consider a real valued database $D = \{x_1(i), \cdots, x_n(i) | i = 1, 2, \cdots, N\}$, with N instances. Each instance has n attributes and is labeled by one of the classes: $\{C_1, \cdots, C_m\}$. A linguistic decision tree is a decision tree where the nodes are the random sets and the branches correspond to particular focal elements. More formally:

Definition 6 (*Linguistic Decision Tree*) *A linguistic decision tree is a set of branches with associated class probabilities of the following form:*

$$LDT = \{< B_1, Pr(C_1|B_1), \cdots, Pr(C_m|B_1) >,$$
$$\cdots, < B_s, Pr(C_1|B_s), \cdots, Pr(C_m)|B_s) >\}$$

A branch is defined as following:

$$B_i = \{< D_{x_{1_i}}, F_{1_i} >, \cdots, < D_{x_{k_i}}, F_{k_i} >\}$$

where, $k_i \leq n$ and $F_{j_i} \in \mathcal{F}_{j_i}$ where $j = 1, \cdots, k$.

A LDT is based on the form of the linguistic data set. Each node splits into branches according to the focal elements of this node(attribute). Each branch has associated class probabilities. For example, consider the branch:

$$<< D_{x_1}, \{s, m\} >, < D_{x_2}, \{m\} >, 0.3, 0.7 >$$

in a binary classification problem. This means the probability of being class C_1 is 0.3 and C_2 is 0.7 if given attribute 1 can be described as *small & medium* and attribute 2 can be described as *medium.*

3.1 Evaluating Class Probabilities for a Given Branch

A branch B can be assumed to have the form:

$$B = \{< D_{x_1}, F_1 >, \cdots, < D_{x_k}, F_k >\}$$

where $k \leq n$ and $F_i \in \mathcal{F}_i$. The probability of Class C_j given B can then be evaluated from D and D_j as follow:

$$Pr(C_j|B) = \frac{S(B|D_j)}{S(B|D)}$$

where, $S(B|D_j) = \sum_{i \in D_j} \prod_{r=1}^{k} m_{x_r(i)}(F_r)$ and $S(B|D) = \sum_{i \in D} \prod_{r=1}^{k} m_{x_r(i)}(F_r) \neq 0$. D_j is the subset consisting of instances belong to class j. In the case of $S(B|D) = 0$, which happens when we use a small scale data base for training LDT, the given branch has no corresponding non-zero linguistic data. We obtain no information from the given dataset so equal probabilities are assigned to each class according to Laplace correction.

$$Pr(C_j|B) = \frac{1}{m} \qquad if: \quad S(B|D) = 0$$

where m is the number of classes.

3.2 Evaluating Class Probabilities Given a Data Element

Consider a given data vector for classification in the form of $\vec{y} =< y_1, y_2, \cdots, y_n >$ which may not be contained in the training data set D. Firstly, we need to translate $\vec{y}$ to linguistic data based on the fuzzy covering of the training data. One problem we may encounter is that, the data element may be beyond the range of training data set. Suppose the attribute j is in the range of $[p_{min}, p_{max}]$, which are covered by N_F fuzzy sets: $F_1, \cdots, F_{N_F}$. Then, we assign the appropriateness degrees of y_j as follows:

$$\mu_{F_i}(y_j) = \mu_{F_i}(p_{min}) \quad if \quad y_j < p_{min}$$
$$\mu_{F_i}(y_j) = \mu_{F_i}(p_{max}) \quad if \quad y_j > p_{max}$$

where, $i = 1, \cdots, N_F$. Then, by Jeffrey's rule we can evaluate for the probabilities of class C_j given a LDT, where $j = 1, 2, \cdots, m$,

$$Pr(C_j|\vec{y}) = \sum_{v=1}^{s} Pr(B_v|\vec{y}) Pr(C_j|B_v)$$

and

$$Pr(B|\vec{y}) = \prod_{r=1}^{k} m_{y_r}(F_r)$$

4 LID3 Algorithm

Linguistic ID3 (LID3) is the learning algorithm for building a linguistic decision tree, and is an extension of ID3. As ID3 search is guided by an information based heuristics, but the information measurements of LDT are modified from classical ones and are based on the label semantics model.

4.1 Searching Heuristics of LID3

The underlying search heuristic is based on the measure of information defined for a branch B and can be viewed as an extension of entropy equation of ID3 in [3] :

Definition 7 (*Branch Entropy*) *The entropy of branch B is given by*

$$E(B) = \sum_{j=1}^{m} Pr(C_j|B) \log_2(Pr(C_j|B))$$

Now, given a particular branch B, suppose we want to expand it with attribute x_i, The evaluation of this attribute will be given by the expected entropy defined as follows:

Definition 8 (*Expected Entropy*)

$$EE(B, x_i) = \sum_{F_i \in \mathcal{F}_i} E(B \cup \{< D_{x_i}, F_i >\}) \\ \cdot Pr(\{< D_{x_i}, F_i >\}|B)$$

where, the probability of node $< D_{x_i}, F_i >$ given B can be calculated as follows:

$$Pr(\{< D_{x_i}, F_i >\}|B) = \frac{S(B \cup \{< D_{x_i}, F_i >\}|D)}{S(B|D)}$$

We can now define the *Information Gain (IG)* obtained when expanding branch B with attribute x_i as:

$$IG(B, x_i) = E(B) - EE(B, x_i)$$

As with ID3 learning, the most informative attribute will form the root of a linguistic decision tree, and the tree will expand into branches associated with all possible focal elements of this attribute. For each branch, the free attribute with maximum information gain will form next node, from level to level, until the tree reaches the maximum depth. If the maximum class probability given this branch is equal or greater than a given threshold probability T, the branch will be terminated at the current depth.

5 Experimental Studies

We applied LID3 algorithm to the Pima Indians Diabetes database and Sonar database (See [5]), respectively. Each attribute of the Pima data is discretized uniformly by 3 full covering fuzzy sets (e.g. see figure 1), and Sonar data with 2 such fuzzy sets, in order to obtain corresponding linguistic data sets. Both linguistic databases are split into two parts with the same number of instances, one is used for training and the other for testing. Table 1 shows the training accuracy (A_{tr}) and test accuracy (A_{ts}) from 100 cross-validation tests on the Pima data set with different maximum depth M_{dep} and threshold probabilities T. Table 2 shows the results on the Sonar data base on a particular split of original database.

$(T)/M_{dep}$	1	2	3	4
$(0.7)A_{tr}$	0.7442	0.7646	0.7833	0.8022
A_{ts}	0.7488	0.7560	0.7563	0.7542
$(0.8)A_{tr}$	0.7442	0.7739	0.8014	0.8323
A_{ts}	0.7488	0.7474	0.7552	0.7499

Table 1: Results on Pima data set with 100 cross-validation.

$(T)/M_{dep}$	5	6	7	8
$(0.9)A_{tr}$	0.9515	1.0000	1.0000	1.0000
A_{ts}	0.7981	0.7855	0.7981	0.7885
$(1.0)A_{tr}$	0.9515	1.0000	1.0000	1.0000
A_{ts}	0.8365	0.8462	0.8654	0.8173

Table 2: Results on Sonar data set with a particular split.

The best result so far for Pima data is $A_{ts} = 0.7563$, when $T = 0.7$ and $M_{dep} = 3$. The test accuracy of C4.5 on the Pima data with 100 cross validation is 0.7422. We can not say that results of LID3 are significantly better than the C4.5. But, in Sonar data, the best result, $A_{ts} = 0.8654$, is significantly better than C4.5 algorithm with the test accuracy of 0.7259. In the Pima test, compared to the decision tree built from C4.5 which has the maximum depth of 8, LDT needs only 2 or 3 levels to obtain comparable (even better) accuracy. Therefore, we can say that LDT has better transparency in this experiment.

References

[1] T. M. J.F. Baldwin and B. Pilsworth. *Fril-Fuzzy and Evidential Reasoning in A.I.* Wiley Inc, New York, 1995.

[2] J. Lawry. Label semantics: A formal framework for modelling with words. In *Symbolic and Quantitative Approaches to Reasoning With Uncertainty, Lecture Notes in A.I.*, pages 374–384. Springer-Verlag, 2001.

[3] T. Mitchell. *Machine Learning.* McGraw-Hill, New York, 1997.

[4] J. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[5] WWW. *UCI Machine Learning Repository: http://www.ics.uci.edu/mlearn/MLrepository.html.*

[6] L. Zadeh. Fuzzy logic = computing with words. *IEEE Transaction on Fuzzy Systems*, 4, 1996.