
Early version, also known as pre-print

Link to publication record in Explore Bristol Research
PDF-document

University of Bristol - Explore Bristol Research
General rights
This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/about/ebr-terms
Abstract

In this paper we propose a random set framework for learning linguistic models for prediction problems. We show how we can model prediction problems based on learning linguistic prototypes defined using joint mass assignments on sets of labels. The potential of this approach is then demonstrated by its application to a model and by benchmark problem and comparing the results obtained with those from other state-of-the-art learning algorithms.

1 Introduction

The idea of using fuzzy sets to represent words was first proposed by Zadeh [15], who stated that fuzzy memberships could be used to model the imprecision and ambiguity of natural language terms such as small, medium and large. However, this generates a number of problems in terms of semantics and computational complexity (see [9] for a discussion).

Here we propose an alternative framework introduced by Lawry (see [10]). This approach uses fuzzy sets to partition an attributes domain into linguistic labels. Random sets (see [3]) and mass assignment are then used as a method for evaluating the appropriateness of the labels for a given value. Prediction is carried out using mass assignment prototypes representing relationships between input and output attributes at the label level. These prototypes are obtained by aggregating linguistic descriptions of examples on the prediction space from a database. The models are then used in conjunction with a Naive or Semi-Naive-Bayes classifier (see [8] and [11]) together with a defuzzification method to perform prediction.

2 Label Semantics

Suppose we have an attribute $x$ with domain $\Omega$ and we ask a set of experts $V$ to provide a finite set of labels $LA$ with which to describe $x$. For $x \in \Omega$ we ask each of the experts $E$ to supply us with a subset of $LA$ that they deem as appropriate to describe $x$. This generates a set of labels describing $x$ denoted $D_x^E$. As each of the experts is likely to have a different subset of appropriate labels to describe the situation, we obtain a random set $D_x$ across the power set of $LA$ as we vary between experts. By combining the label description provided by the experts we can determine a mass assignment on the power sets of $LA (2^{LA})$ representing the distribution of the random set $D_x$.

Definition 1 (Mass Assignment) A mass assignment on $2^\Omega$ is a function $m : 2^\Omega \rightarrow [0,1]$ such that:
\[ \sum_{S \subseteq \Omega} m(S) = 1 \]

Definition 2 (Value Description) Let $V$ be the set of experts. For $x \in \Omega$ the label description of $x$ is a random set from $V$ into the power set of $LA$, denoted $D_x$, with associated mass assignment $m_x$:
\[ \forall S \subseteq LA \ m_x(S) = P_V(\{E \in V : D_x^E = S\}) \]
where \( P_V \) is the prior probability distribution over the population \( V \).

For any mass assignment on \( 2^{LA} \) it is likely to be the case that only a subset of \( 2^{LA} \) will have non-zero mass. These sets are referred to as focal sets of \( LA \).

**Definition 3 (Focal Sets)** The focal sets for the labels \( LA \) are defined as the union of the focal sets for the mass assignment \( m_x \) as \( x \) varies across \( \Omega \).

\[
F_{LA} = \{ S \subseteq LA | \exists x \in \Omega, m_x(S) > 0 \}
\]

We can formally define a measure for the appropriateness of a label \( L \) for a value \( x \), denoted \( \mu_L(x) \), by evaluating the mass of those label sets containing \( L \).

**Definition 4 ( Appropriateness Degrees)**

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

Notice that \( \mu_L : \Omega \rightarrow [0, 1] \) and hence corresponds to a fuzzy set on \( \Omega \). However, the term fuzzy set does not seem entirely suitable in this context since we are not measuring a degree of membership but rather a degree of appropriateness.

Here we have assumed that we have knowledge of the underlying expert behaviour but in many situations this is not the case. Hence, we need to define a mapping from appropriateness degrees to mass assignments. To achieve this we make the assumption that individuals in \( V \) differ regarding what labels are appropriate for a value only in terms of generality and specificity. This is referred to as the consonance assumption. Also, we make the further assumption \( \forall x \in \Omega \max_{L \in LA} \mu_L(x) = 1 \).

**Definition 5 (Consonance Mapping)** Let \( \{\mu_L(x) : L \in LA\} = \{y_1, \ldots, y_n\} \) be ordered such that \( y_i > y_{i+1} \) for \( i = 1, \ldots, n-1 \) then for \( S_i = \{ L \in LA : \mu_L(x) \geq y_i \} \),

\[
m_x(S_i) = y_i - y_{i+1} \quad \text{for} \quad i = 1, \ldots, n-1
\]

\[
m_x(S_n) = y_n, \quad m_x(\emptyset) = 1 - y_1
\]

Clearly, the framework described in this section is related to the random set semantics for fuzzy memberships proposed by Goodman and Nguyen [4]. However, the latter defines random sets on subsets of the attribute universe while for the current framework they are defined on subsets of labels. This provides an interesting new perspective and allows for a more straightforward treatment of continuous domains.

### 3 Label Prototypes for Modelling Prediction Problems

Consider a prediction problem where the objective is to model the relationship between input attributes \( x_1, \ldots, x_{n-1} \) and output attribute \( x_n \). Label sets \( LA_j \) are defined on input universes \( \Omega_j : j = 1, \ldots, n-1 \) and a set of labels \( LC \) is also defined on the output universe \( \Omega_n \). Each \( L \in LC \) is represented by a trapezoidal fuzzy set on the prediction space. The focal sets of \( LC \) are given by \( F_{LC} = \{ S \subseteq LC | \exists x_n \in \Omega_n, m_{x_n}(S) > 0 \} = \{ F_j \} \).

Suppose we have a training set of examples \( DB = \{(x_1(i), \ldots, x_n(i)) | i = 1, \ldots, N\} \). The input attributes \( x_1, \ldots, x_{n-1} \) are now partitioned into subsets \( S_1, \ldots, S_w \) where \( w \leq n-1 \) and for each \( F_j \in F_{LC} \) a joint mass assignment \( m_{i,j} \) is determined as follows: Suppose, w.l.o.g. that \( S_i = \{x_1, \ldots, x_v\} \) then the joint mass assignment on \( 2^{LA_1} \times \cdots \times 2^{LA_v} \) conditional on \( F_j \in F_{LC} \) is defined by:

\[
\forall T_r \in 2^{LA_r} : r = 1, \ldots, v \forall F_j \in F_{LC}
\]

\[
m_{i,j}(T_1, \ldots, T_v) = \frac{\sum_{k \in DB} m_{x_k}(F_j) \prod_{r=1}^v m_{x_r}(T_r)}{\sum_{k \in DB} m_{x_k}(F_j)}
\]

Hence, the prototype describing \( F_j \) is the vector: \( \langle m_{1,j}, \ldots, m_{w,j} \rangle \).

### 4 Prediction Using Prototypes on Linguistic Class

We now give details of how prediction can be performed using linguistic class prototypes, together with a Semi-Naïve-Bayes (see [8]) learning algorithm. We use Semi-Naïve-Bayes in this context to weaken the independence assumption of Naïve-Bayes (see [11]). This is achieved by defining joint mass assignments to model dependences between attributes in variable groupings and then assuming independence between groupings. We then
carry out a defuzzification step to obtain a prediction value from this model.

Bayes theorem is used here to evaluate the probability of each of the focal elements $F_j$ given a vector of input values $(x_1, \ldots, x_{n-1})$ as follows:

$$Pr(F_j|x_1, \ldots, x_{n-1}) = \frac{Pr(F_j) \prod_{r=1}^{w} p(S_r|F_j)}{\sum_k Pr(F_k) \prod_{r=1}^{w} p(S_r|F_k)}$$

Where $Pr(F_j) = \frac{1}{|DB|} \sum_{k \in DB} m_{x_n(k)}(F_j)$. There is now the problem of how to estimate the density function $p(x_1, \ldots, x_{n-1} | F_j)$. Consider the joint mass assignment for grouping $S_i$ given $F_j$. If we assume that there is a uniform prior distribution on $X_{r=1}^{v} \Omega_r$, then the joint prior mass assignment on $X_{r=1}^{v} = 2^{k-1}$ is: $\forall T_i \subseteq L_A_i : i = 1, \ldots, v$

$$pm(T_1, \ldots, T_v) = \prod_{i=1}^{v} \int_{T_i} \Omega_i m_{x_i}(T_i) u_i(x_i) dx_i$$

Where $u(x_1, \ldots, x_v) = \prod_{i=1}^{v} u_i(x_i)$ is the uniform distribution on $X_{r=1}^{v} \Omega_r$ and $u_i(x_i)$ is the uniform distribution on $\Omega_i$. From this we can define the joint density on $x_1, \ldots, x_v$ conditional on $m_{i,j}$:

$$p(S_i|m_{i,j}) = p(x_1, \ldots, x_v|m_{i,j}) =$$

$$u(x_1, \ldots, x_v) \sum_{T_1 \times \ldots \times T_v} \frac{m_{i,j}(T_1, \ldots, T_v)}{pm(T_1, \ldots, T_v)} \prod_{r=1}^{v} m_{x_r}(T_r)$$

The calculation here is motivated by the theorem of total probability (see [12] for a full description). We now define a defuzzification method to determine a predicted value for $x_n$ as follows: Assuming there is a uniform prior distribution on $x_1, \ldots, x_{n-1}$, then, by evaluating $Pr(F_j|x_1, \ldots, x_{n-1})$ for all $F_j$ we obtain a mass assignment on $F_{LC}$. This can then be mapped to a distribution on $x_n$ as follows:

$$p(x_n|x_1, \ldots, x_{n-1}) = \sum_j Pr(F_j|x_1, \ldots, x_{n-1}) p(x_n|F_j)$$

where: $p(x_n|F_j) = \frac{m_{x_n}(F_j)}{\int_{m_{x_n}(F_j)} dx_n}$

We then take our estimate of $x_n$, denoted $\hat{x}_n$, to be the expected value of the distribution:

$$\hat{x}_n = \int_{\Omega_n} x_n p(x_n|x_1, \ldots, x_{n-1}) dx_n = \sum_j Pr(F_j|x_1, \ldots, x_{n-1}) E(x_n|F_j)$$

An alternative defuzzification method can be obtained by replacing $E(x_n|F_j)$, by the mode of the distribution $p(x_n|F_j)$ (i.e. $\arg\max(m_{x_n}(F_j))$).

5 Grouping Methods

In this section we introduce a number of methods for automatically finding attribute groupings that increase discrimination in the model. In general it is too computationally expensive to search the complete problem space of all attribute groupings and then partition to see if discrimination can be increased, as the search space would be exponential. To counter this problem a heuristic search has been proposed, based on the order of importance of each of the attribute groupings $S_i$. The heuristic used to estimate the importance is defined as follows:

Definition 6 (Importance Measure) Let the joint mass assignment for $S_i$ given $F_j$ be denoted $m_{i,j}$. For any input vector $S_i$ the probability of the focal set $F_j$ can be estimated using Bayes theorem:

$$IM_j(S_i) = \frac{\sum_{k \in DB} Pr(F_j|S_i(k)) m_{x_n}(F_j)}{\sum_{k \in DB} Pr(F_j | S_i(k))}$$

where: $Pr(F_j|S_i) = \frac{p(S_i|m_{i,j}) p(F_j)}{p(S_i|m_{i,j}) p(F_j) + p(S_i|m_{i,-j}) (1 - p(F_j))}$

where $m_{i,-j}$ is the mass assignment for group $S_i$ conditional on $F_{LC} - \{F_j\}$

$IM_j(S_i)$ is a measure of importance of the set of variables $S_i$ as discriminators of $F_j$ from the other focal sets. The closer $IM_j(S_i)$ is to 1 the more discriminating the group $S_i$. In this case $\sum_{k \in DB} Pr(F_j|S_i(k)) m_{x_n}(F_j)$ is high relative to $\sum_{k \in DB} Pr(F_j|S_i(k)) (1 - m_{x_n}(F_j))$.

Due to the ‘curse of dimensionality’ (see [2]) careful limits must be set on the maximum number of attributes that can be grouped when running this algorithm. The effect of this can be limited by trading granularity off against dimensionality. The importance measure here is now combined with either a breadth or depth first search to find discriminative groupings. We will now define two methods for measuring whether or not a pair of attributes should be combined.

Definition 7 (Correlation Measure) Let $F_1$ be the focal sets for $S_1$ and $F_2$ the focal sets for
\( S_2 \). Now let \( m_{1,2,j} \) be the joint mass of \( S_1 \cup S_2 \) given the output focal set \( F_j \).

\[
CORR(S_1, S_2) = \sqrt{\frac{1}{|F_1||F_2|} \sum_{R \in F_1} \sum_{T \in F_2} (m_{1,2,j}(R, T) - m_{1,j}(R)m_{2,j}(T))^2}
\]

Here a threshold is used so that the nearer the correlation measure is to 1, the more likely it is that grouping will take place. An alternative to measuring correlation is to trying to maximise the increase in importance of any new grouping formed.

**Definition 8 (Improvement Measure)**

Suppose we have two subsets of attributes \( S_1 \) and \( S_2 \) then the improvement in importance obtained by combining them can be calculated as follows:

\[
IPM_j(S_1, S_2) = \min(IM_j(S_1), IM_j(S_2)) / IM_j(S_1, S_2)
\]

A threshold is once again used so that the closer the improvement measure is to 0 the more likely that the attributes will be combined.

### 6 Performance on a Benchmark Problem

We now give details of the performance of the proposed prediction system. The results obtained from the Fuzzy Bayesian methods are compared here to a \( \varepsilon \)-Support Vector Regression system (\( \varepsilon \)-SVR) [13], implemented in [6] by Gunn [5]. The \( \varepsilon \)-SVR was implemented using a gaussian Radial Basis Function (RBF) kernel with an \( \varepsilon \)-insensitive loss function. We now define a method for evaluating the prediction error, the Mean Square Error (MSE), which is calculated as follows:

\[
MSE = \frac{1}{|DB|} \sum_{i \in DB} (\hat{x}_n(i) - x_n(i))^2
\]

#### 6.1 Surface Based on: \( z = \sin(x \times y) \)

In this example a training set of 529 points were generated describing a surface defined according to the equation \( z = \sin(x \times y) \) where \( x, y \in [0, 3] \), as shown in figure 1:

![Figure 1: Surface defined by the 529 points.](image)
the \( \varepsilon \)-SVR method obtains a marginally better prediction of the surface with an MSE of 0.0011 which is an improvement of 0.0041 compared to that observed using Semi-Naive-Bayes. Though this difference may be reduced by using more fuzzy sets.

6.2 Prediction of Sunspots

This problem is from the time series data library [7] and contains data on J.R. Wolf and Zürich sunspot relative numbers [1] between the years 1700-1979. The data was organized as described in [14], except that form the validation set of 35 examples (1921-1955) was merged into the test set of 24 examples (1956-1979). This is because a validation set is not required in the fuzzy label framework. Hence, a training set of 209 examples (1712-1920) and a test set of 59 examples (1921-1979) were used. The input attributes were \( x_{t-12} \) to \( x_{t-1} \) and the output attribute was \( x_t \). Each attribute had 4 labels defined over the domains using a percentile method obtain the fuzzy partition. The correlation threshold was set to 0.005 and the improvement threshold set to 0.895, with a maximum allowed grouping size of 7 attributes.

Figure 3 give details of the prediction results obtained:

<table>
<thead>
<tr>
<th>Test set results</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \varepsilon )-Support Vector Regression system</td>
<td>418.126</td>
</tr>
<tr>
<td>Best Semi-Naive-Bayes</td>
<td>499.659</td>
</tr>
</tbody>
</table>

Figure 4: Prediction results obtained for the sunspot prediction test set, form applying \( \varepsilon \)-SVR with parameters: \( \sigma = 3, \varepsilon = 0.05, C = 5 \) and best Semi-Naive-Bayes prediction.

We can further compare our results to those given in [14] using the suggested measure of prediction accuracy, Average Relative Variance, which calculated as follows:

\[
ARV(DB) = \frac{1}{\sigma^2} \frac{1}{N} \sum_{k \in DB} (x_k - \hat{x}_k)^2
\]

Figure 5 show the results obtained using our best Semi-Naive-Bayes method (a depth first search with the improvement measure) and the \( \varepsilon \)-SVR are better to those stated by Weigend et al. [14] with Semi-Naive-Bayes performing the best on the 1956-1979 segment of the test set. It should be highlighted that the results of Weigend et al. [14] for the years 1712-1920 and 1921-1955 are though significantly better. This is because these time periods corresponded to the training and validation sets used to train the neural network. The disparity between the results seen for the years 1712-1920 and 1921-1955 and those stated by Weigend et al. over the 1956-1979, suggest over-fitting by the network. However, for a full and fair comparison of the results here we must also allow the validation set to be placed in the training sets as the validation set is used during the training process. The result from allowing this are given in the bottom two rows of figure 5. This shows that, as we would be expected we obtain better prediction results for both Semi-Naive-Bayes and the \( \varepsilon \)-SVR system and the validation set (1921-1955) which now more closely match the results stated by Weigend et al. Also it is possible to see that in this instance there is little different in the prediction obtained between using Semi-Naive-Bayes
and the $\varepsilon$-SVR system. Further we can shown a direct comparison between the predicted results form both the $\varepsilon$-SVR and best Semi-Naïve-Bayes prediction result against the original data, when merging the validation set into the training set, which is shown by figure 6.

<table>
<thead>
<tr>
<th>Average Relative Variance</th>
<th>1712-1920</th>
<th>1921-1955</th>
<th>1956-1979</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single step prediction: (see [14] p 414)</td>
<td>Weight Elimination Net</td>
<td>0.082</td>
<td>0.086</td>
</tr>
<tr>
<td>FCA Model</td>
<td>0.097</td>
<td>0.097</td>
<td>0.28</td>
</tr>
<tr>
<td>Results from merging the validation with test</td>
<td>Best Semi-Naïve-Bayes</td>
<td>0.113</td>
<td>0.204</td>
</tr>
<tr>
<td>F-SVM System</td>
<td>0.155</td>
<td>0.117</td>
<td>0.263</td>
</tr>
<tr>
<td>Results from merging the validation with training</td>
<td>Best Semi-Naïve-Bayes</td>
<td>0.146</td>
<td>0.108</td>
</tr>
<tr>
<td>$\varepsilon$-SVR system</td>
<td>0.127</td>
<td>0.087</td>
<td>0.248</td>
</tr>
</tbody>
</table>

Figure 5: Full comparison of results with those obtained by Weigend et al. [14].

Figure 6: Comparison of prediction result obtained form $\varepsilon$-SVR and best Semi-Naïve-Bayes method.

7 Conclusion

We have introduced a framework for modelling fuzzy labels and shown how this can be applied to the induction of fuzzy models for prediction. In this context input-output relationships are represented by prototypes comprised of vectors of mass assignments. Each of the mass assignments is defined over the label sets describing some subset of the input attributes, where these subset groupings capture the important dependencies in the modelling problem. A number of search strategies are introduced to find important variable groupings based on both measures of correlation and improvement in discrimination. Learnt prototypes can then be used in conjunction with Semi-Naïve-Bayes and a defuzzification method to obtain estimated output values given inputs. In the experiments presented the Fuzzy Bayesian algorithm gives almost identical results to the $\varepsilon$-SVR and neural networks. However, the use of fuzzy labels provides more flexible and transparent models.

The high-level representation of models in terms of fuzzy labels suggests the possibility of extending this research to allow for the evaluation of queries expressed in natural language. This will utilizing the calculus for appropriateness degree proposed in [9]. Query evaluation for prediction problems in still being developed, however, similar research on classification problems is given in [12].

References