Lawry, J. (2001). Label Prototypes for Modelling with Words. In Proc. of NAFIPS2001. 10.1109/NAFIPS.2001.943720

Label Prototypes for Modelling with Words

Jonathan Lawry
Dept. Engineering Mathematics,
University of Bristol,
Bristol, BS8 1TR, UK
Email:j.lawry@bris.ac.uk

Abstract
This paper suggests a framework for modelling with words using label prototypes. The underlying methods are based on a random set label semantics together with the voting model interpretation of fuzzy sets. The potential of this methodology will be illustrated by its application to classification problems.

Keywords: random sets, mass assignments, label descriptions, prototypes.

1 Introduction

The phrase "computing with words" was introduced by Zadeh (see [13]) to capture the idea of computation based on natural language terms rather than numerical quantities. Zadeh's formulation was centred on the idea of linguistic variables [12] and with the extension principle motivating many of the proposed inference mechanisms. Unfortunately, many of these are extremely computationally expensive. Furthermore, the focus of much of the research into computing with words has been on the development of inference methods from linguistic information rather than on the learning of linguistic models for (potentially) complex systems. The latter research area can be referred to as modelling with words.

In this paper we outline a new approach to computing and modelling with words. Instead of linguistic variables we introduce the idea of a random set allocating the set of appropriate labels for the value of an underlying variable. The semantics of this approach is based on the voting model interpretation of fuzzy sets (see [4] and [5]). The method described can be used both for inference and for induction of linguistic models. It is this latter application area on which we will focus in the sequel. Specifically, we shall introduce a method for prototype induction where a prototype is defined to be a tuple of mass assignments, each over the sets of appropriate labels for their corresponding variable. In the context of classification problems, a prototype will be learnt for each class and together these can be used for class prediction given new data.

Intuitively, we view prototypes as describing a collection of similar objects all sharing certain properties. For instance, a bank might be interested in prototypes describing customers with certain income and borrowing characteristics. As such, prototypes correspond to descriptions or amalgams of objects rather than individual objects. This is a different view than is taken in the psychology literature (see [8]) where prototypes are viewed as single objects that are in someway descriptive of the whole class.

In the following we describe a random set based framework for modelling with fuzzy labels. This is an alternative to the approach described in [1] and [2] where prototypes are tuples of fuzzy sets on labels.

2 Label Semantics

Suppose we have some domain of discourse Ω that we will assume either to be finite or correspond to a closed interval of the real numbers. The idea of label semantics is that we define some fixed (finite) set of words, *LA*, with which to label the elements of Ω. For instance, the elements of Ω might correspond to the possible values of the height of some individual, say Bill. In this case a statement such as *Bill is tall*, where $tall \in LA$ is taken to mean that *tall* is an appropriate label for the element of Ω corresponding to Bill's height.

The meanings of the words in *LA* are defined by fuzzy sets on Ω, where the membership degree of a value x in label l, denoted μ_l, is taken as quantifying the degree to which l is deemed an appropriate label for x. Conceptually, this can be modelled according to a voting semantics ([4],[5]) as follows: Each voter is asked to provide the subset of words from the finite set LA which are appropriate as labels for the value x. This generates a mass assignment on 2^{LA} giving the distribution of the appropriate label random set, denoted *APL*. The membership degree of x in l is then taken to be the probability that $\{l\} \in APL$. In other words, it is taken to be the proportion of voters who include l in their set of appropriate labels for x. In practice, we would give fuzzy set definitions for the terms in *LA* and use these to find the mass assignment on *APL* corresponding to any element $x \in \Omega$.

Definition 2.1 (Label Descriptions)
The mass assignment on *APL* for a value x is referred to as the (label) description of x and is given by:

$m_{des[x]} = \{l_1, \cdots, l_k\} : \mu_{l_k}(x), \cdots, \{l_1, \cdots, l_i\} : \mu_{l_i}(x) - \mu_{l_{i+1}}(x), \cdots, \{l_1\} : \mu_{l_1}(x) - \mu_{l_2}(x), \quad : 1 - \mu_{l_1}(x)$

where $\{l_1, \cdots, l_k\} = \{l \quad LA \mid \mu_l(x) > 0\}$ and the ordering is such that $\mu_{l_i}(x) \geq \mu_{l_{i+1}}(x)$ for $i = 1, \cdots, k-1$.

In many cases it is desirable that $m_{des[\,]}$ is a normalised random set for all *x*. (i.e. $x \quad \Omega \; m_{des[x]}(\;) = 0$). This holds if *LA* satisfies the following property:

Definition 2.2 (Linguistic Covering)

A set of fuzzy sets $\mu_1, \cdots, \mu_n$ forms a linguistic covering of Ω if and only if $x \quad \Omega \; \max(\mu_1(x), \cdots, \mu_n(x)) = 1$

Hence, we require that $\{\mu_l \mid l \quad LA\}$ forms a linguistic covering of Ω.

We can extend the idea of a label description of a value to obtain a label description of a database of values. This will correspond to a mass assignment representing the probability that a particular subset of *LA* will be the set of appropriate labels for an element of the database.

Definition 2.3

Let $D = \{x(i) \mid i = 1, \cdots, N\}$ be a single variable database then the label description of *D* is a mass assignment on 2^{LA} defined by:

$$S \quad LA \; m_{des[D]}(S) = \frac{1}{N} \sum_{i=1}^{N} m_{des[x(i)]}(S)$$

3 Label Prototypes for Modelling Classification Problems

Consider a standard classification problem where instances or objects from a certain problem domain can be categorised, each as belonging to one of the classes $C_1, \cdots, C_k$. A number of features or attributes of instances can be measures and these are represented by the variables $X_1, \cdots, X_n$. The value of variable X_j for instance *i* is denoted $x_j(i)$. For each variable X_j we define a finite set of labels LA_j.

Now suppose we have a training database $D = \{\langle x_1(i), \cdots, x_n(i) \rangle \mid i = 1, \cdots, N\}$ of *N* instances for which all *n* features have been measured. We are also told to which of the *k* classes each instance belongs. The class of instance *i* is denoted $C(i)$. Now consider the sub-database of instances with class C_j.

$$D_j = \{\langle x_1(i), \cdots, x_n(i) \rangle \mid C(i) = C_j\}$$

Further consider, the projection of this database so that only the values of variable X_r are included.

$$D_{r,j} = \{x_r(i) \mid C(i) = C_j\}$$

Now the label description of class C_j based on the variable X_r can be taken to be the label description of the sub-database $D_{r,j}$, $m_{des[D_{r,j}]}$. Hence, we can view the tuple $\langle m_{des[D_{1,j}]}, \cdots, m_{des[D_{n,j}]} \rangle$ as a decomposed label model of the class. Alternatively, this tuple provides amalgamated information regarding all the data points of class C_j in *D* and can be view as a prototypical description of this class.

Clearly, the above model is decomposed and in situations where significant correlation exists between the variables for some particular class decomposition errors are likely to result. There are a number of possible approaches to this problem.

One idea would be to include compound feature in the model composed of the cross product of highly correlated subsets of $X_1, \cdots, X_n$. For instance, we might include *t* subsets of n_t variables $\{X_{i_1}, \cdots, X_{i_{n_t}}\}$ for $i = 1, \cdots, t$. In this case we have a prototype of the form:

$$\langle m_{des_{1,j}}, \cdots, m_{des_{t,j}} \rangle$$

where $m_{des_{1,j}}$ is the joint mass assignment of variables $X_{i_1}, \cdots, X_{i_{n_t}}$ for class C_j.

An alternative approach is to look for subsets of D_j for which the values of $X_1, \cdots, X_n$ are sufficiently similar to enable a purely decomposed model to be used. For instance, a standard clustering algorithm could be used to partition D_j into *c* sub-databases $D_j^{(1)}, \cdots, D_j^{(c)}$. The mass assignment for the description of these sub-databases can then be found for each variable giving *c* prototypes, $\langle m_{des_{1,r,j}}, \cdots, m_{des_{n,r,j}} \rangle$ for $r = 1, \cdots, c$ where $m_{des_{1,r,j}}$ is the label description of $D_{i,j}^{(r)}$ (see for example [3]).

4 Conditional Distributions from Label Prototypes

In many situations we may only have statistical information regarding *APL* taking the form of a mass assignment m_{des}. In this case it is desirable that we have some means of estimating the distribution of the underlying variable *X*. In other words, we require a means of evaluating a posterior density given the information that *APL* is distributed according to m_{des}.

Definition 4.1 (Posterior Density from Labels)

$$p(x|APL = S) = \frac{m_{des[x]}(S)\,p(x)}{\int_{\Omega} m_{des[x]}(S)\,p(x)p(x)dx}$$

and assuming a uniform prior distribution on Ω gives

$$x \in \Omega \; p(x|APL = S) = \frac{m_{des[x]}(S)}{\int_{\Omega} m_{des[x]}(S)dx}$$

This can then be used to obtain a density on Ω conditional on a mass assignment m_{des}.

$$x \in \Omega \; p(x|m_{des}) = \sum_{S \in LA} m_{des}(S)\,p(x|APL = S)$$

This expression can be rewritten as

$$p(x|m_{des}) = p(x)\sum_{S \in LA} \frac{m_{des}(S)}{m(S)} m_{des[x]}(S)$$

where $S \in LA$ $m(S) = \int_{\Omega} p(x)\, m_{des[x]}(S)\, dx$ and can be viewed as a prior mass assignment on *LA*. Notice that in the case where $S \in LA$ $m_{des}(S) = m(S)$ then it follows that $x \in \Omega$ $p(x|m_{des}) = p(x)$. This is intuitive since if the mass assignment m_{des} provides no new information then we would not expect the conditional density $p(x|m_{des})$ to differ from the prior $p(x)$.

Given the conditional distribution of a mass assignment then we can clearly obtain a point estimate for the underlying variable by taking its expected value according to this distribution. Such distributions can also be used in classification problems as is described in the following section.

5 Using Label Mass Assignments to Estimate Classification Probabilities

We now give the details of two cases where label mass assignments can be used to estimate class probabilities in classification problems. In the first case we make the naïve Bayes assumption (see [4]) that variables are conditionally independent given their associated class. Secondly we consider the situation where it is possible to calculate a joint mass assignment.

Suppose we encounter an instance for which the measured attribute values are given by the vector $\langle x_1, \cdots, x_n \rangle$. Now in order to predict the class to which this instance belongs we need to calculate the probability $\boldsymbol{Pr}(C_j|x_1, \cdots, x_n)$ for $j = 1, \cdots, k$. According to Bayes theorem we have that

$$\boldsymbol{Pr}(C_j|x_1, \cdots, x_n) = \frac{p(x_1, \cdots, x_n|C_j)\,\boldsymbol{Pr}(C_j)}{p(x_1, \cdots, x_n)}$$

The naïve Bayes assumption is that

$$p(x_1, \cdots, x_n|C_j) = \prod_{i=1}^{n} p(x_i|C_j)$$

giving $\boldsymbol{Pr}(C_j|x_1, \cdots, x_n) = \dfrac{\boldsymbol{Pr}(C_j)\prod_{i=1}^{n} p(x_i|C_j)}{p(x_1, \cdots, x_n)}$

We now make the further assumption that the density values $p(x_i|C_j)$ can be estimated from the linguistic prototype(s) for C_j. More specifically, given the prototype $\langle m_{des[D_{1,j}]}, \cdots, m_{des[D_{n,j}]} \rangle$ we take

$p(x_i|C_j) \approx p(x_i|m_{des[D_{i,j}]})$ from which we obtain that

$$\boldsymbol{Pr}(C_j|x_1, \cdots, x_n) = k(x_1, \cdots, x_n)|D_j| \prod_{i=1}^{n} p(x_i|m_{des(D_{i,j})})$$

For the joint mass assignment model we consider only the case where we have two measurable attributes *X* and *Y*. It is trivial to generalise from this to the *n*-dimensional case. For each class C_j we generate a joint mass assignment $m_{des[D_j]}: 2^{LA_1 \times LA_2} \to [0,1]$ such that $S \times R \in LA_1 \times LA_2$

$$m_{des[D_j]}(S,R) = \frac{1}{N}\sum_{i \in D_j} m_{des[x(i)]}(S)\, m_{des[y(i)]}(R)$$

Assuming a uniform prior distribution on $\Omega_1 \times \Omega_2$ then the prior mass assignment on $LA_1 \times LA_2$ is given by

$$m(S,R) = \int_{\Omega_1}\int_{\Omega_2} m_{des[x]}(S)\, m_{des[y]}(R)\, u(x,y)\,dxdy$$

$$= \int_{\Omega_1} m_{des[x]}(S)\, u_1(x)dx \times \int_{\Omega_2} m_{des[y]}(R)\, u_2(y)dy$$

$$= m_1(S)\, m_2(R)$$

Hence the conditional distribution generated by $m_{des[D_j]}$ is given by

$$p(x,y|m_{des[D_j]}) =$$

$$u(x,y)\sum_{S \times R} \frac{m_{des[D_j]}(S,R)}{m_1(S)\,m_2(R)} m_{des[x]}(S)\, m_{des[y]}(R)$$

From this we can obtain an estimate for $\boldsymbol{Pr}(C_j|x,y)$ from Bayes theorem by taking

$$p(x,y|C_j) \approx p(x,y|m_{des[D_j]}).$$

6 Classification using a Decomposed Model

In this section we will describe two decomposable benchmark machine learning problems. We will then show how they can be modelled effectively using the

type of prototypes described above incorporated with a Naïve Bayes classifier.

Example 6.1 (Pima Diabetes Problem)
This is a benchmark classification problem (see [10]) taken from the UCI repository [11]. The problem relates to incidents of diabetes mellitus in the Pima Indian population living near Phoenix Arizona. The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organisation criteria. The database contains details of 768 females all of which are older than 21. This was split into a training and test set each containing 384 instances. There are eight measured attributes: X_1: *number of times pregnant,* X_2: *plasma glucose concentration,* X_3 : *diastolic blood pressure,* X_4: *triceps skin fold thickness,* X_5: *2-hour serum insulin,* X_6: *body mass index,* X_7: *diabetes pedigree function,* X_8: *age*

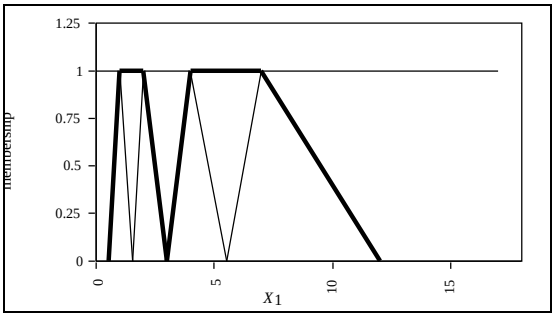

Figure 1 : The linguistic covering for attribute X_1

For each variable a label set was defined for which the associated membership functions were trapezoidal and formed a linguistic covering of the underlying universe. The trapezoids were generated using a simple percentile method to obtain a crisp partition with equal numbers of data points falling within each set and then superimposing trapezoidal membership functions over this partition. A set of five labels was used for each variable although attributes X_4 and X_5 were emitted since their values across the database were not sufficiently distinct for the percentile method to be applied. The linguistic coverings were generated so that, at most, two labels overlapped at anyone time (see figure 1). Hence, if we interpret the five labels as *very small, small, medium, large* and *very large* then the possible focal elements for mass assignments on labels are {*very small*}, {*very small, small*}, {*small*}, {*small, medium*}, {*medium*}, {*medium, large*}, {*large*}, {*large, very large*} and {*very large*}. Also notice from figure 1 that each fuzzy set overlaps exactly half the core region (i.e. region of elements with membership one) of its neighbours. We refer to such fuzzy sets as having a 50% overlap.

For each attribute mass assignments on label sets were learnt for both diabetic and not diabetic as described in section 4 (figure 2). For each mass assignment we can then generate a conditional distribution (figure 3).

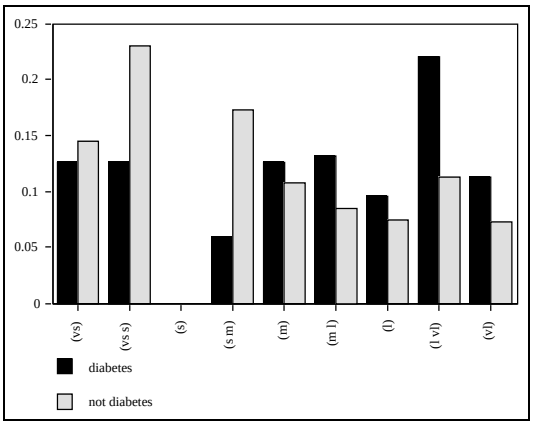

Figure 2: Label mass assignments for diabetes and not diabetes classes for X_1

Using the naïve Bayes approach and estimating class probabilities according to the method discussed in section 5 we obtain an accuracy of 77.34% on the training set and 75.78% on the test set with the following confusion table.

Predicted Class/ True Class	**Diabetes**	**Not Diabetes**
Diabetes	64.2%	35.8%
Not Diabetes	16.5%	83.5%

These results are comparable with more composed algorithms (e.g. decision trees, around 76%, and feed-forward neural networks, around 79% [9])

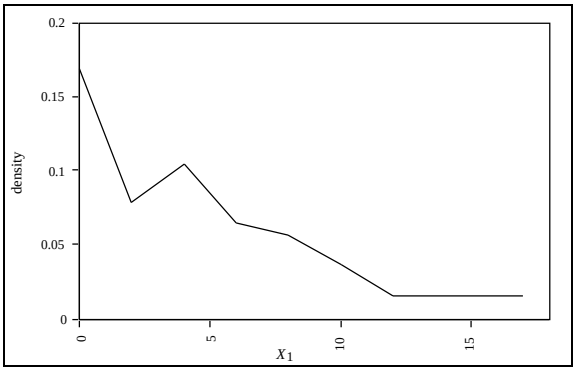

Figure 3: Conditional density for X_1 given the diabetes mass assignment

Example 6.2 (Wisconsin Cancer Problem)
This database originates from a study carried out by Wolberg [8] into cancer diagnosis via linear programming. The data relates to 699 breast tumours and the class variable takes values benign or malignant. This was split into training and test sets containing 322 and 377 elements respectively. The associated variables are: X_1: *clump thickness,* X_2: *uniformity of cell size,* X_3: *uniformity of cell shape,* X_4: *marginal adhesion,* X_5: *single epithelial cell size,* X_6: *bare nuclei,* X_7: *bland chromatin,* X_8: *normal nucleoli,* X_9: *mitoses*

Initially three labels were allocated to each variable where the associated membership functions were uniform trapezoids. The label mass assignments generated for each variable then have the form shown in figure 4.

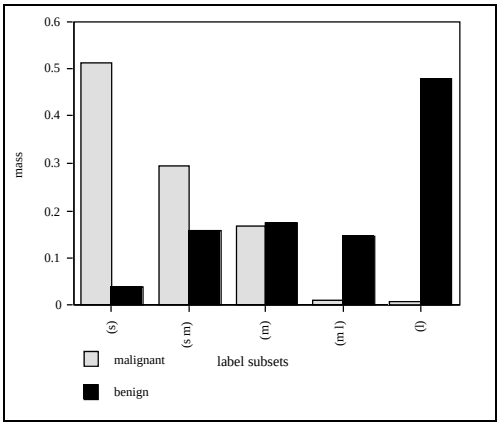

Figure 4: Mass assignments for benign and malignant classes using three labels on X_1

Using these mass assignments to estimate classification probabilities we obtain an accuracy of 97.2% on the training set and 96.55% on the test set.

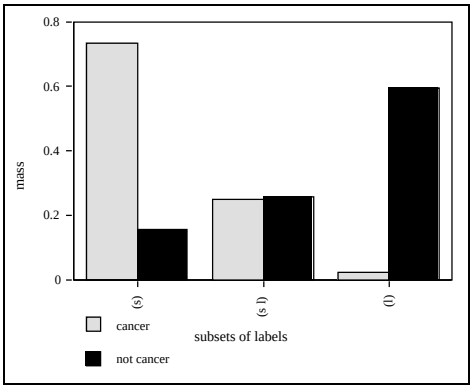

Figure 5 : Mass assignments for malignant and benign using two labels on X_1

We then experimented with lowering the number of labels to two for each attribute giving label mass assignments of the form shown in figure 5. This reduction in the number of labels only brings about a marginal reduction in predictive accuracy to 96.58% on the training set and 95.23% on the test set.

7 Classification using a Joint Mass Assignment Model
In this section we consider an example of composed modelling using joint mass assignments. The problem discussed is a two-dimensional model problem based on a *sin* function.

Example 7.1
In this example a figure eight shape was generated according to the parametric equation $x = 2^{-0.5}\left(\sin 2t - \sin t\right)$ and $y = 2^{-0.5}\left(\sin 2t + \sin t\right)$ where $t \in [0, 2\pi]$ (see figure 6).

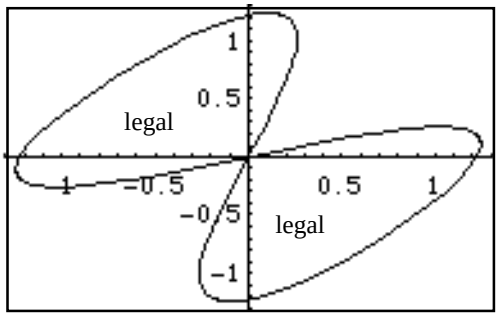

Figure 6: Figure eight classification problem

Points in $[-1.6, 1.6]^2$ are classified as legal if they lie within the figure and illegal if they lie outside. The database consisted of 961 vectors $\langle X, Y \rangle$ generated from a regular grid on $[-1.6, 1.6]^2$.

A linguistic covering of five uniformly spaced trapezoidal fuzzy sets was generated for each attribute. These had an overlap degree of 40%. A joint mass assignment was then learnt for each class. The joint mass assignment obtained for legal is shown in the following table and in histogram form in figure 7.

	{vs}	{s, vs}	{s}	{s, m}	{m}	{m,l}	{l}	{l,vl}	{vl}
{vs}	0	0	0	0.0039	0.017	0.0019	0	0	0
{s,vs}	0	0	0	0.0078	0.036	0.019	0.0063	0	0
{s}	0	0	0	0.005	0.047	0.036	0.041	0.0063	0
{s,m}	0.0039	0.0078	0.0054	0.001	0.026	0.026	0.036	0.019	0.0019
{m}	0.017	0.036	0.047	0.026	0.038	0.026	0.047	0.036	0.017
{m,l}	0.0019	0.019	0.036	0.026	0.026	0.001	0.0055	0.0078	0.0039
{l}	0	0.0063	0.041	0.036	0.047	0.005	0	0	0
{l, vl}	0	0	0.0063	0.019	0.036	0.0078	0	0	0
{vl}	0	0	0	0.0019	0.017	0.0039	0	0	0

Classification based on the posterior distributions from the two joint mass assignments gives a predictive accuracy of 96.25% on the training set and 96.65% on a denser test set of 2116 elements.

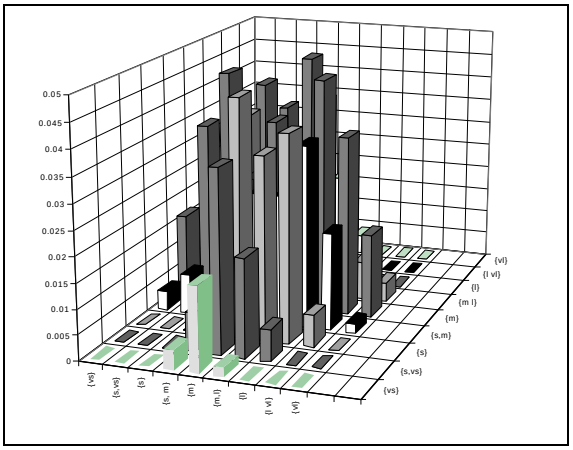

Figure 7: Histogram of the joint mass assignment for legal.

The following scatter plot (figure 8) shows those points correctly classified as legal together with false positives and false negatives.

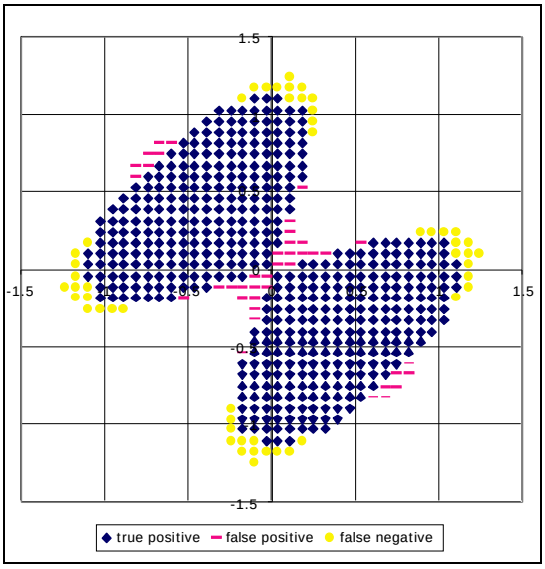

Figure 8: Scatter plot showing classification accuracy for the figure eight problem

5 Conclusion

Prototypes consisting of mass assignments on subsets of labels have been shown to provide a framework for data analysis in classification problems. More specifically, a methodology has been proposed whereby such label mass assignments can be used to estimate class probabilities. The potential of this approach has been demonstrated by its application to a number of benchmark and model problems.

Acknowledgements

This research was partly supported by a grant from the Nuffield Foundation.

References

[1] J.F. Baldwin, J. Lawry, T.P. Martin, "A Mass Assignment Method for Prototype Induction", *International Journal of Intelligent Systems* Vol.14, No. 10, 1999

[2] J.F. Baldwin, J. Lawry, "A c-Fuzzy Means Algorithm for Prototype Induction", *Proceedings of FUZZ-IEEE 2000*, Vol. 1, 2000, pp164-169

[3] C. Borgelt, H. Timm, R. Kruse, "Using Fuzzy Clustering to Improve Naïve Bayes and Probabilistic Networks", *Proceedings of FUZZ-IEEE 2000*, Vol. 1, 2000, pp53-58

[4] B.R. Gaines, "Fuzzy and Probability Uncertainty Logics", *Journal of Information and Control* 38, 1978, pp154-169

[5] J. Lawry, "A Voting Mechanism for Fuzzy Logic", *The International Journal of Approximate Reasoning* 19, 1998, pp315-333.

[6] D.D. Lewis, "Naïve (Bayes) at Forty: The Independence Assumption in Information Retrieval", *Machine Learning ECML-98, LNAI* 1398, 1998, pp4-15

[7] O.L. Mangasarian, W.H. Wolberg, "Cancer Diagnosis via Linear Programming", *SIAM News*, Vol. 23, No. 5, 1990, pp1-18

[8] D.N. Osherson, E.E. Smith, "On the Adequacy of Prototype Theory as a Theory of Concepts", *Cognition* 9, 1981, pp 35-58

[9] B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996

[10] J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, R.S. Johannes, "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus", *Proceedings on the Symposium on Computer Applications and Medical Care*, 1988, pp261-265

[11] UCI Machine Learning repository, http://www.ics.uci.edu/~mlearn/MLRepository.html

[12] L.A. Zadeh, "The concept of a linguistic variable and its applications to approximate reasoning", *Part I: Information Sciences 8*, 1975, pp199-249; *Part II: Information Sciences 8*, 1975, pp301-357; *Part III: Information Sciences* 9, 1976, pp 43-80

[13] L.A. Zadeh, "Fuzzy Logic =Computing with Words", *IEEE Transactions on Fuzzy Systems*, Vol. 4, No. 2, 1996, pp103-111