Lawry, J. (2002). A New Calculus for Linguistic Prototypes in Data Analysis. In *Proc. of Soft Methods in Probability, Statistics & Data Analysis* (pp. 116 - 125) https://doi.org/10.1007/978-3-7908-1773-7_9

Early version, also known as pre-print

Link to published version (if available):
[10.1007/978-3-7908-1773-7_9](https://doi.org/10.1007/978-3-7908-1773-7_9)

[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/red/research-policy/pure/user-guides/ebr-terms/>

A New Calculus for Linguistic Prototypes in Data Analysis

Jonathan Lawry¹

Department of Engineering Mathematics, University of Bristol, Bristol, UK,
j.lawry@bris.ac.uk

Abstract. A random set semantics for imprecise concepts is introduced. It is then demonstrated how label descriptions of data sets can be learnt in this framework. These descriptions take the form of linguistic prototypes representing amalgams of elements. The potential of this approach for classification and query evaluation is then investigated.

1 Introduction

The area of automated learning from data is becoming increasingly important in an age of almost continuous data collection. From data we must be able to learn models which are flexible enough to facilitate a wide range of queries and which allow for insight into the underlying nature of the system under consideration. A principal requirement of such models is that they should be transparent and in order to achieve such clarity, ideally the representation framework should be high-level and capture certain aspects of natural language. In this paper we shall focus on modelling the imprecision associated with adjective labels that describe a quantity or the value of a measurement. Typical examples of this type of label occur in expressions of the form ‘the diastolic blood pressure is *high*’ or ‘The sodium concentration is *quite low*’. Specifically, we aim to provide label descriptions of attributes values for sets of similar elements contained in a database. In a sense such descriptions can be viewed as imprecise prototype definitions. These prototypes can then be used for clustering as well as classification and prediction tasks. In the sequel we introduce a random set based calculus [4] for attribute labels with a clear underlying semantics.

2 Label Semantics

For an attribute (or variable) x into a domain of discourse Ω we identify a finite set of words LA with which to label the values of x . Then for a specific value $a \in \Omega$ an individual I identifies a subset of LA , denoted $\mathcal{D}_a^I$ to stand for the description of a given by I , as the set of words with which it is appropriate to label a . Within this framework then, an expression such as ‘the diastolic blood pressure is *high*’, as asserted by I , is interpreted to mean $high \in \mathcal{D}_{bp}^I$ where bp denotes the value of the variable blood pressure. If we allow I to vary across a

population of individuals V then we naturally obtain a random set $\mathcal{D}_x$ from V into the power set of LA where $\mathcal{D}_x(I) = \mathcal{D}_x^I$. A probability distribution (or mass assignment) associated with this random set can be defined and is dependent on the prior distribution over the population V . We can view the random set $\mathcal{D}_x$ as a description of the variable x in terms of the labels in LA .

Definition 1. (*Value Description*) For $x \in \Omega$ the label description of x is a random set from V into the power set of LA , denoted $\mathcal{D}_x$, with associated distribution $m_{\mathcal{D}_x}$, given by

$$\forall S \subseteq LA \quad m_{\mathcal{D}_x}(S) = Pr(\{I \in V : \mathcal{D}_x^I = S\})$$

Another high level measure associated with $m_{\mathcal{D}_x}$ is the following quantification of the degree of appropriateness of a particular word $L \in LA$ as a label of x .

Definition 2. (*Appropriateness Degrees*)

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

Now clearly μ_L is a function from Ω into $[0, 1]$ and therefore can technically be viewed as a fuzzy set. However, we shall use the term ‘appropriateness degree’ partly because it more accurately reflects the underlying semantics and partly to highlight the quite distinct calculus for these functions that will be introduced in the sequel. We now make the additional assumption that value descriptions are consonant random sets (see [2]). In the current context consonance simply requires the restriction that individuals in V differ regarding what labels are appropriate for a value only in terms of generality or specificity. Certainly, given that the meaning of the labels in LA must be sufficiently invariant across V to allow for effective communication then some strong restriction on $\mathcal{D}_x$ should be expected. The consonance restriction could be justified by the idea that all individuals share a common ordering on the appropriateness of labels for a value and that the composition of $\mathcal{D}_x^I$ is consistent with this ordering for each I . The consonance assumption means that $m_{\mathcal{D}_x}$ can be completely determined from the values of $\mu_L(x)$ for $L \in LA$ as follows [2]: If $\{\mu_L(x) : L \in LA\} = \{y_1, \dots, y_n\}$ ordered such that $y_i > y_{i+1}$ for $i = 1, \dots, n-1$ then for $S_i = \{L \in LA : \mu_L(x) \geq y_i\}$,

$$m_{\mathcal{D}_x}(S_i) = y_i - y_{i+1} \text{ for } i = 1, \dots, n-1$$

$$m_{\mathcal{D}_x}(S_n) = y_n \text{ and } m_{\mathcal{D}_x}(\emptyset) = 1 - y_1$$

This has considerable practical advantages since we no longer need to have any knowledge of the underlying population of individuals V in order to determine $m_{\mathcal{D}_x}$. Rather, for reasoning with label semantics in practice we need only define appropriateness degrees μ_L for $L \in LA$ corresponding to the imprecise definition of each label.

For more general linguistic reasoning a mechanism is required for evaluating compound label expressions. For example, we may wish to know whether or not

expressions such as $medium \wedge low$, $medium \vee low$ and $\neg high$ can be applied to a value $x \in \Omega$. In the context of this assertion-based framework we interpret the main logical connectives in the following manner: $L_1 \wedge L_2$ 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 $\neg L$ means that L is not an appropriate label. More generally, if we consider label expressions formed from LA by recursive application of the connectives then an expression θ identifies a set of possible label sets $\lambda(\theta)$ as follows:

Definition 3. *Possible Label Sets*

- For $L \in LA$ $\lambda(L) = \{S \subseteq LA : L \in S\}$
- For label expressions θ and φ $\lambda(\theta \wedge \varphi) = \lambda(\theta) \cap \lambda(\varphi)$
- For label expressions θ and φ $\lambda(\theta \vee \varphi) = \lambda(\theta) \cup \lambda(\varphi)$
- For label expression θ $\lambda(\neg\theta) = \overline{\lambda(\theta)}$

The notion of appropriateness measure given above can now be extended so that it applies to compound label expressions. The intuitive idea here is that $\mu_\theta(x)$ quantifies the degree to which expression θ is appropriate as a description of x .

Definition 4. *(Compound Appropriateness Degrees)* For θ a label expression and $x \in \Omega$ the appropriateness of θ to x is given by:

$$\mu_\theta(x) = \sum_{S \in \lambda(\theta)} m_{\mathcal{D}_x}(S)$$

3 Label Descriptions of Data Sets

Suppose we have a database DB of N elements, associated with each of which are n measurements $x_1, \dots, x_n$ so that $DB = \{\langle x_1(i), \dots, x_n(i) \rangle : i = 1, \dots, n\}$ where $x_j(i)$ denotes the value of x_j for object i . Further, suppose that we select a set of labels LA_j for each attribute x_j for $j = 1, \dots, n$ where each label is defined by an appropriateness measure. The label description of DB is now defined to be a vector of mass assignments as follows:

Definition 5. *(Label Description of DB)* The label description of DB is a vector $\mathcal{L}(DB) = \langle m_1, \dots, m_n \rangle$ where

$$\forall S \subseteq LA \quad m_j(S) = \frac{1}{N} \sum_{i=1}^N m_{\mathcal{D}_{x_j(i)}}(S)$$

Given a label description of DB we now evaluate a joint mass assignment on $2^{LA_1} \times \dots \times 2^{LA_n}$ so that:

$$\forall S_j \in 2^{LA_j}, j = 1, \dots, n \quad m_{DB}(S_1, \dots, S_n) = \prod_{j=1}^n m_j(S_j)$$

Clearly, we are making an independence assumption here and in some cases this may not be appropriate. In order to overcome this problem one approach is to partition DB into a number of disjoint sets $P_1, \dots, P_c$, perhaps according to some standard clustering algorithm, where the elements contained in each partition set are assumed to be sufficiently similar to allow an independence assumption. We can then learn label descriptions $\mathcal{L}(P_k)$ for $k = 1, \dots, c$ and combine them to form an overall mass assignment for DB as follows: Let $\mathcal{L}(P_k) = \langle m_{1,k}, \dots, m_{n,k} \rangle$ then

$$\forall S_j \in 2^{LA_j}, j = 1, \dots, n \quad m_{DB}(S_1, \dots, S_n) = \sum_{k=1}^c \frac{|P_k|}{N} \prod_{j=1}^n m_{j,k}(S_j)$$

Given a joint mass assignment on DB and tuple of label expressions $\theta = \langle \theta_1, \dots, \theta_n \rangle$ where θ_j is an expressions based on labels LA_j , we can now use labels semantic to evaluate the appropriateness of θ for describing DB in the following way.

$$\mu_\theta(DB) = \sum_{S_1 \in \lambda(\theta_1)} \dots \sum_{S_n \in \lambda(\theta_n)} m_{DB}(S_1, \dots, S_n)$$

For many types of data analysis it is useful to be able to estimate the distribution on underlying variables given the information contained in DB . In the current context our knowledge of DB is represented by the mass assignment m_{DB} and hence we need to be able to evaluate a distribution on the base variables $x_1, \dots, x_n$ conditional on m_{DB} . For simplicity, we now assume that all variables are continuous with domains of discourse comprising of bounded closed intervals of the real line. Furthermore, we assume a prior joint distribution $p(x_1, \dots, x_n)$ for the base variables. In the case that p is unknown we will assume it to be the uniform distribution. Furthermore, to simplify the following definition we will at least assume that $x_1, \dots, x_n$ are *a priori* independent so that $p(x_1, \dots, x_n) = \prod_{j=1}^n p_j(x_j)$, p_j being the marginal prior on x_j . The following definition is based on a Bayesian argument the details of which are given in [3].

Definition 6. (*Conditional Density given a Mass Assignment*)

Let x be a variable into Ω with prior distribution $p(x)$, LA be a set of labels for x and m be a posterior mass assignment for the set of appropriate labels of x (i.e. $\mathcal{D}_x$) inferred from some database DB . Then the posterior distribution of x conditional on m is given by:

$$\forall x \in \Omega \quad p(x|m) = p(x) \sum_{S \subseteq LA} \frac{m(S)}{pm(S)} m_{\mathcal{D}_x}(S)$$

where pm is the prior mass assignment generated by the prior distribution p according to

$$pm(S) = \int_{\Omega} m_{\mathcal{D}_x}(S) p(x)$$

This definition is motivated by the following argument based on the theorem of total probability:

$$p(x|m) = \sum_{S \subset LA} p(x|\mathcal{D}_x = S)Pr(\mathcal{D}_x = S) = \sum_{S \subset LA} p(x|\mathcal{D}_x = S)m(S)$$

Also

$$p(x|\mathcal{D}_x = S) = \frac{Pr(\mathcal{D}_x = S|x)p(x)}{Pr(\mathcal{D}_x = S)} = \frac{m_{\mathcal{D}_x}(S)p(x)}{pm(S)}$$

Making the relevant substitutions and then simplifying gives the expression from definition 6. This definition is then extended to the case where the posterior knowledge consists of a set of prototype descriptions of DB .

Definition 7. (*Conditional Densities from Prototype Descriptions*) For m_{DB} generated from label descriptions $\mathcal{L}(P_k), k = 1, \dots, c$

$$\forall \mathbf{x} \in \Omega_1 \times \dots \times \Omega_n \quad p(\mathbf{x}|m_{DB}) = \sum_{k=1}^c \frac{|P_k|}{N} \prod_{j=1}^n p(x_j|m_{j,k})$$

4 Label Models and Linguistic Queries

To illustrate the potential of this framework we shall briefly describe how it can be applied to classification problems. In principle, however, the approach can also be applied to prediction and cluster analysis. Suppose then that the objects of DB can be categorised as belonging to one of the classes $C_1, \dots, C_t$ and let DB_j denote the subset of DB containing only the elements with class C_j . We can now determine m_{DB_j} on the basis of some partition and evaluate $p(\mathbf{x}|m_{DB_j})$ as described above. If we now take $p(\mathbf{x}|m_{DB_j})$ as an approximation for $p(\mathbf{x}|C_j)$ then from Bayes theorem we have $Pr(C_j|\mathbf{x}) \propto p(\mathbf{x}|m_{DB_j})|DB_j|$. Given this estimate for each class probability, classification can be then carried out in the normal way. In the limit case when the partition of DB_j has only one set (i.e. DB_j) then this method corresponds to a version of the well known Naive Bayes algorithm [5]. Also, in the context of classification problems we extend the vector notation for linguistic queries as follows:

$$\langle \theta_1, \dots, \theta_n \rangle : C_j$$

This represents the question: Do elements of class C_j satisfy $\boldsymbol{\theta}$ (i.e. x_1 is θ_1 and x_2 is θ_2 and ... and x_n is θ_n)? The support for this query is given by $Pr(\boldsymbol{\theta}|C_j) = \mu_{\boldsymbol{\theta}}(DB_j)$.

$$\langle \theta_1, \dots, \theta_n \rangle$$

This represents the question: Do elements of DB satisfy $\boldsymbol{\theta}$? The support for this query is given by $Pr(\boldsymbol{\theta}) = \mu_{\boldsymbol{\theta}}(DB) = \sum_{k=1}^t Pr(DB_k)\mu_{\boldsymbol{\theta}}(DB_k)$.

$$C_j : \langle \theta_1, \dots, \theta_n \rangle$$

This represents the question: Do elements satisfying θ belong to class C_j ? The support for this query is given by

$$Pr(C_j|\theta) = \frac{\mu_\theta(DB_j)Pr(DB_j)}{\mu_\theta(DB)}$$

Example 1. The Naive Bayes version of the above algorithm was applied to the UCI repository problem on glass categorisation. The problem has 6 classes and 9 continuous attributes. 5 labels were define for attributes 1-2 and 4-7. Attribute 3 was allocated 4 labels and attributes 8 and 9 were not used as their variance across DB was too low for effective labelling. For all attributes the labels were defined by trapezoidal appropriateness degrees positioned according to a simple percentile method (see figure 1). The database of 214 elements was randomly split into a test and training set of 107 elements each and a classification accuracy of 78.5% on the training set and 72.9% on the test set was obtained. This is comparable with other approaches; for example a feedforward Neural Network with architecture 9-6-6 gives 72% on a smaller test set where the network was trained on 50% of the data, validated on 25% and tested on 25%. The density function for attribute 1 generated from the label description of class 1 according to definition 6 is shown in figure 2.

Fig. 1. Non uniform appropriateness degrees for, from left to right, *very small*, *small*, *medium*, *large* and *very large* for attribute 1 generated using a percentile algorithm.

Fig. 2. Density function for attribute 1 conditional on class 1 generated from the mass assignment for attribute 1 in the label description of class 1.

Now suppose for the attributes with five labels that these correspond to domain specific versions of *very low*(vl), *low*(l), *medium*(m), *high*(h) and *very high* (vh) (see figure 1). Consider the following queries:

Query 1

What is the probability that float processed building window glass (class 1) has a medium to low or high refractive index (att. 1) and a very low or low sodium concentration (att. 2)?

To answer this query we note that $\mathcal{L}(DB_1) = \langle m_{1,1}, \dots, m_{n,1} \rangle$ where

$m_{1,1} = \{vl\} : 0.01373, \{l, vl\} : 0.04342, \{l\} : 0.02804, \{l, m\} : 0.37391, \{m\} : 0.12208, \{m, h\} : 0.08424, \{h\} : 0.10059, \{h, vh\} : 0.17233, \{vh\} : 0.06167.$

In this case the vector representation of the query is given by

$$\theta : C_1 = \langle \theta_1, \theta_2, \top, \dots, \top \rangle : C_1$$

where $\top$ denotes a tautology, $\theta_1 \equiv (medium \wedge low) \vee high$ and $\theta_2 \equiv very\ low \vee low$. For this query we have that

$$\begin{aligned} \mu_{\theta}(DB_1) &= \sum_{S_1 \in \lambda(\theta_1)} \sum_{\lambda \in (\theta_2)} \sum_{S_3 \subseteq LA_3} \dots \sum_{S_n \subseteq LA_n} \prod_{i=1}^n m_{i,1}(S_i) \\ &= \left(\sum_{S_1 \in \lambda(\theta_1)} m_{1,1}(S_1) \right) \times \left(\sum_{S_2 \in \lambda(\theta_2)} m_{2,1}(S_2) \right) \end{aligned}$$

Now from the training database we have

$$\sum_{S \in \lambda(\theta_1)} m_{1,1}(S) =$$

$$m_{1,1}(\{l, m\}) + m_{1,1}(\{m, h\}) + m_{1,1}(\{h\}) + m_{1,1}(\{h, vh\}) = 0.73107$$

Similarly, $\sum_{S \in \lambda(\theta_2)} m_{2,1}(S) = 0.92923$ so that $\mu_{\theta}(DB_1) = 0.73107 \times 0.92923 = 0.67933$ this being the required probability.

Query 2

What is the probability that a glass fragment has a medium to low or high refractive index (att. 1) and a very low or low sodium concentration (att. 2)?

This query has vector representation

$$\langle \theta_1, \theta_2, \top, \dots, \top \rangle$$

and the required probability is given by $\mu_{\theta}(DB)$. To evaluate this we note that for classes $C_1, \dots, C_6$ the probabilities of each class satisfying θ are given by $\mu_{\theta}(DB_1) = 0.67933, \mu_{\theta}(DB_2) = 0.18429, \mu_{\theta}(DB_3) = 0.09036, \mu_{\theta}(DB_4) = 0.42108, \mu_{\theta}(DB_5) = 0, \mu_{\theta}(DB_6) = 0.02173$. Also the number of data elements in $DB_1, \dots, DB_6$ are respectively 35, 38, 8, 7, 4 and 15. From this we can evaluate: $\mu_{\theta}(DB) = \frac{1}{107} (0.67933(35) + 0.18429(38) + 0.09036(8) + 0.42108(7) + 0.02173(15)) = \frac{34.7760}{107} = 0.32501$.

Query 3

What is the probability that a glass fragment with a medium to low or high refractive index (att. 1) and a very low or low sodium concentration (att. 2) is a fragment of float processed building window glass (class 1)?

For this query the vector representation is

$$C_1 : \langle \theta_1, \theta_2, \top, \dots, \top \rangle$$

and the required probability $Pr(C_1|\theta)$ is given by:

$$Pr(C_1|\theta) = \frac{\mu_{\theta}(DB_1)Pr(DB_1)}{\mu_{\theta}(DB)} = \frac{0.67933 \frac{35}{107}}{0.32501} = 0.6837$$

Example 2. In this problem a figure eight shape (see figure 4) was generated according to the parametric equation $x = 2^{-0.5}(\sin 2t - \sin t)$, $y = 2^{-0.5}(\sin 2t + \sin t)$ where $t \in [0, 2\pi]$. Points in $[-1.6, 1.6]^2$ are classified as legal if they lie within the figure and illegal if they lie outside. The training database consisted of a regular grid of points on $[-1.6, 1.6]^2$. Clearly, a Naive Bayes independence assumption is inappropriate for this problem and will lead to significant decomposition error. Instead, c-means [1], was used to partition both the legal and illegal sub-databases into four. A joint mass assignment and joint distribution (figure 3) were then generated for each class, as described in section 3 and based on six labels for each variable. A classification accuracy of 95.4% on the training set and 95.8% on a denser test set of 2116 elements was obtained (see figure 4). This compares with an accuracy for the Naive Bayes model of 85% on the training set and 85.1% on the test set.

Fig. 3. Density function for legal class generated from the label model consisting of four legal prototypes

Fig. 4. Scatter plot showing true positives, false negative and false positive points for the figure eight test set

5 Conclusions

A framework for evaluating label descriptions of a database has been introduced. This involves a new random set based semantics for imprecise concepts. The potential of this approach has been demonstrated by its application to classification problems and to the evaluation of linguistic queries.

References

1. Duda R., Hart P.: Pattern Classification and Scene Analysis Wiley, New York (1973)
2. Goodman I.R., Nguyen H.T.: Uncertainty Models for Knowledge Based Systems North Holland (1985)
3. Lawry, J.: Label Prototypes for Modelling with Words. Proceedings of The North American Fuzzy Information Processing Society 2001 Conference, (2001)
4. Lawry, J.: Label Semantics: A Formal Framework for Modelling with Words. Lecture Notes in Artificial Intelligence, **2143** (ed. S. Benferhat, P. Besnard), (2001) 374-384
5. Lewis, D.D.: Naive Bayes at Forty: The independence Assumption in Information Retrieval. Lecture Notes in Artificial Intelligence, **1398** (1998) 4-15