Lawry, J. (2001). Query Evaluation from Linguistic Prototypes. In Proc. of 10'th IEEE Intl Conf on Fuzzy Systems, Melbourne, Australia. 10.1109/FUZZ.2001.1007240

Query Evaluation from Linguistic Prototypes

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

Abstract

A framework for modelling with words is introduced based on label semantics. It is shown how within this framework that linguistic prototypes, defined as vectors of mass assignments on sets of labels, can be used to evaluate linguistic queries. This provides a flexible knowledge representation framework for data mining and knowledge discovery as well as an environment well suited to information fusion and modelling with words in general.

1 Introduction

The area of automated learning from data is becoming increasingly important in an age of almost continuous data collection. Large companies collect a stream of data relating to the behaviour of their customers which is augmented each time a particular individual uses their services. Such data must be analysed to provide flexible models of customer behaviour that can then be used to aid a wide variety of decision-making processes. In other areas such as medical or engineering systems it may be, for a variety of reasons, difficult or even impossible to formulate analytical models, but where data is available that implicitly describes the behaviour of the system. Here again we need to be able to learn models from the data which are flexible enough to facilitate a wide range of queries needed to gain a comprehensive understanding of the system. Another important feature that is often present in systems of this type is inherent uncertainty and imprecision. In fact it is this property that often means they are not amenable to classical modelling techniques. This uncertainty is not only due to lack of precision or errors in measured features but is present in the model itself since the available features may not be sufficient to provide a complete model of the system. In many application domains, such as medical systems, background knowledge is often available in the form of natural language facts and rules provided by practitioners in the field. Ideally this information should also be incorporated into the model, an observation that necessitates a knowledge representation framework that will allow for the fusion of data derived and linguistic background knowledge.

In this paper we argue that an appropriate paradigm for modelling of the above type is "modelling with words". In other words, we propose that a linguistic based knowledge representation framework should be adopted for data inferred models and that this framework should allow for the transparent handling of uncertainty and imprecision as well as being flexible enough to allow for a wide range of queries, both qualitative and quantitative. The linguistic nature of these models will then permit high level fusion with background knowledge. More specifically, in the sequel an approach to modelling with words based on linguistic prototypes will be outlined. Here a linguistic prototype will represent an amalgam of objects of a certain type or class described in terms of the propensity for certain words to be used to label the attributes of the model for that class. The formal framework used will be label semantics ([3] and [4]). The central idea is that a set of words is selected with a varying level of certainty, from some finite set, the label set, as appropriate labels for a given attribute value. These can then be amalgamated across a set or database of attribute values to form prototypes.

2 Label Semantics

For an attribute (or variable) X into Ω we identify a finite set of words LA with which to label the values of X. Then for a specific value $x \in \Omega$ of X an individual I identifies a subset of LA, denoted D_x^I to stand for the description of x given by I, as the set of words appropriate to label x. If we allow I to vary across a population of individuals V then we obtain a random set D_x from V into the power set of LA where $D_x(I) = D_x^I$ and with an associated probability distribution (or mass assignment) m_{D_x} determined by the underlying distribution on V.

Definition 2.1 (Value Description)
For $x \in \Omega$ the label description of x is a random set from V into LA, denoted D_x^I, with associated distribution m_{D_x} given by $\forall S \subseteq LA \ \ m_{D_x}(S) = \boldsymbol{Pr}\left(\left\{I \in V : D_x^I = S\right\}\right)$

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

Definition 2.2 (Appropriateness Degree)
$\forall x \in \Omega,\ \forall L \in LA \ \mu_L(x) = \sum_{S \subseteq LA : L \in S} m_{D_x}(S)$.

Clearly, then as x varies μ_L defines a fuzzy set on Ω representing the meaning of the word L in terms of the values of X.

Here and in the sequel we make the assumption that for all $x \in \Omega$ m_{D_x} is consonant. This assumption may seem, on first inspection, very strong. However, in the current context consonance simply requires the restriction that individuals in V differ regarding the composition in terms of D_x^I, only in terms of its generality or specificity. This assumption means that m_{D_x} can be completely determined by the values of $\mu_L(x)$ for $L \in LA$ since a consonant mass assignment is completely determined by its fixed point coverage. Specifically, we have that if $\{\mu_L(x) | L \in LA\} = \{y_1, \cdots, y_n\}$ ordered such that $y_i > y_{i+1}$ for $i = 1, \cdots, n-1$then for $S_i = \{L \in LA | \mu_L(x) \geq y_i\}$, $m_{D_x}(S_i) = y_i - y_{i+1}$ for $i = 1, \cdots, n-1$, $m_{D_x}(S_n) = y_n$ and $m_{D_x}(\emptyset) = 1 - y_1$. 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_{D_x}. Rather, for reasoning with label semantics in practice we need only define a set of fuzzy sets μ_L for $L \in LA$ corresponding to the fuzzy definition of each label. Also, given fuzzy set definitions for the labels we can determine a subset of the power set of LA consisting of those appropriate label sets that occur with non-zero probability. For instance, if the fuzzy sets μ_{low} and μ_{high} do not overlap then no subsets of LA containing both *low* and *high* will occur as sets of appropriate labels for any value $x \in \Omega$. More formally, we can define a set of focal elements $F = \{S \subseteq LA | \exists x \in \Omega \; m_{D_x}(S) > 0\}$and restrict attention to this subset of 2^{LA}.

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

Definition 2.3 (Possible Label Sets)

The set of possible label sets identified by a label expression θ is defined recursively as follows:

1. For $L \in LA$ $\lambda(L) = \{S \subseteq F | L \in S\}$
2. For label expressions θ and ϕ $\lambda(\theta \wedge \phi) = \lambda(\theta) \cap \lambda(\phi)$
3. For label expressions θ and ϕ $\lambda(\theta \vee \phi) = \lambda(\theta) \cup \lambda(\phi)$
4. For label expression θ $\lambda(\neg\theta) = \overline{\lambda(\theta)}$
5. For label expressions θ and ϕ $\lambda(\theta \rightarrow \phi) = \lambda(\neg\theta \vee \phi)$

The notion of appropriateness measure given in definition 2.2 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 label expression θ can be applied to x.

Definition 2.4 (Compound Appropriateness Measure)

For θ a label expression and $x \in \Omega$ $\mu_\theta(x) = \sum_{S \in \lambda(\theta)} m_{D_x}(S)$

In the above we have only considered labelling a precise value x of the variable X. However, in many situations we may not have sufficient information to uniquely determine the value of X. Instead we might have some evidence e restricting X, in which case we would want to determine a label description of X conditional on e. Allowing X to vary as well as I naturally generates as random set from $V \times \Omega$ into the power set of LA, denoted D_X, such that $D_X(x, I) = D_x^I$. The mass assignment for D_X can then be determined from the cross product of the distribution on V and the posterior distribution on Ω conditional on e.

Definition 2.5 (Variable Description)

The label description of variable X on the basis of evidence e, denoted D_X, is a random set from $V \times \Omega$ into the power set of LA with associated distribution $m_{D_X}(\bullet|e)$ given by

$\forall S \subseteq LA$ $m_{D_X}(S|e) = \boldsymbol{Pr}\left(\left\{\langle x, I\rangle : D_x^I = S\right\}\middle|e\right)$

$= \sum_{x \in \Omega : \boldsymbol{Pr}(x|e) > 0} \boldsymbol{Pr}(x|e)\, m_{D_x}(S)$ (or $= \int_\Omega p(x|e)\, m_{D_x}(S)\, dx$ in the continuous case)

Given this notion we can extend definitions 2.2 and 2.4 to give a generalised appropriateness measure quantifying the degree to which a label expression θ can be applied to a variable X in light of evidence e.

Definition 2.6 (Generalised Appropriateness Measure)

$\mu_\theta(X|e) = \sum_{S \in \lambda(\theta)} m_{D_X}(S|e) = \sum_{x \in \Omega : \boldsymbol{Pr}(x|e) > 0} \boldsymbol{Pr}(x|e) \mu_\theta(x)$ (or $= \int_\Omega p(x|e)\, \mu_\theta(x)\, dx$in the continuous case)

Notice that this is a strict generalisation of appropriateness measure as given in definition 2.4 since $\mu_\theta(x) = \mu_\theta(X|e)$ where e is the evidence that $X = x$.

3 Linguistic Prototypes

In its most general form we would interpret the term linguistic prototype as referring to any high level description of a type or class of objects within the framework of label semantics. For the scope of this current work, however, we shall introduce a specific type of linguistic prototype and show how they can be applied in application domains where there a clearly defined classes of objects that are of interest. Here we define a linguistic prototype as a vector of mass assignments on words describing the distribution of appropriate labels for various attributes across a particular class or sub-class of objects.

Definition 3.1

Let the model attributes be random variables X_i into Ω_i for $i = 1, \cdots, n$ and let LA_i be the label set for X_i. Then a label prototype for object type T in the context of background information e is a vector $\left\langle m_{D_{X_1}}\left(\bullet \middle| e, T\right), \cdots, m_{D_{X_n}}\left(\bullet \middle| e, T\right)\right\rangle$.

Typically we might expect that e would correspond to a database linking variable values with different classification classes and T would correspond either to a classification class or sub-class identified by some clustering algorithm. For example, let $DB = \left\{\left\langle x_1(k), \cdots, x_n(k), C(k)\right\rangle : k = 1, \cdots, N\right\}$ where k is an index referring to a particular object, $x_i(k)$ is the value of variable X_i for object k, and $C(k)$ is the classification class of k. This is the typical format of a classification problem in machine learning. In this case associating e with the information contained in DB and T with a class c_j we have that $m_{D_{X_i}}\left(S \middle| DB, c_j\right) = \sum_{k:C(k)=c_j} m_{D_{x_i(k)}}(S) \Big/ \left|\left\{k : C(k) = c_j\right\}\right|$. This expression is obtained from definition 2.5 by taking $\boldsymbol{Pr}_i\left(x \middle| DB, c_j\right) = \left|\left\{k : x_i(k) = x, C(k) = c_j\right\}\right| \Big/ \left|\left\{k : C(k) = c_j\right\}\right|$.

In [3] we have described how linguistic prototypes of this kind can be used to estimate classification probabilities which can then be combined in a Naïve Bayes or Semi-Naïve Bayes classifier. In the following section we will outline how such prototypes can be used to evaluate a more general form of linguistic queries.

4 Query Evaluation

The subject of linguistic query evaluation from databases has been widely discussed within the fuzzy reasoning community (see for example [1]). Such evaluation takes a number of forms but often involves evaluating from data, the degree to which some fuzzy quantified proposition holds. In most cases this will mean carrying out some calculation directly on the data. This has the disadvantage that the whole of the data must be stored for the duration of the time period in which the query evaluation system is in use. In the case where DB is large this can be costly and instead we propose to replace DB with a set of summarising models in the form of linguistic prototypes. Queries would then be evaluated on the basis of these prototypes alone without reference to the original data. In this section we shall attempt to illustrate the potential of this approach in the context of examples taken from a benchmark problem; the Pima/diabetes database.

Example 4.1 (Pima/diabetes)

This is a benchmark classification problem taken from the UCI repository [5]. 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*

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. In terms of label semantics we have that $LA_i = \left\{very\ low(vl)_i, low(l)_i, medium(m)_i, high(h)_i, very\ high(vh)_i\right\}$ and $F_i = \left\{\{vl_i\}, \{vl_i, low_i\}, \{l_i\}, \{l_i, m_i\}, \{m_i\}, \{m_i, h_i\}, \{h_i\}, \{h_i, vh_i\}\{vh_i\}\right\}$. The following is the linguistic prototype inferred for the diabetic class.

e=training database DB, T=diabetic

$\langle(\{vl_1\}{:}0.1269, \{vl_1,l_1\}{:}0.1269, \{l_1,m_1\}{:}0.0597, \{m_1\}{:}0.1269, \{m_1,h_1\}{:}0.1318, \{h_1\}{:}0.0945, \{h_1,vh_1\}{:}0.2199, \{vh_1\}{:}0.1134), (\{vl_2\}{:}0.0102, \{vl_2,l_2\}{:}0.02808, \{l_2\}{:}0.0597, \{l_2,m_2\}{:}0.0899, \{m_2\}{:}0.0913, \{m_2,h_2\}{:}0.1507, \{h_2\}{:}0.1278, \{h_2,vh_2\}{:}0.133, \{vh_2\}{:}0.3094), (\{vl_3\}{:}0.0806, \{vl_3,l_3\}{:}0.1105, \{l_3\}{:}0.1015, \{l_3,m_3\}{:}0.1179, \{m_3\}{:}0.0746, \{m_3,h_3\}{:}0.0842, \{h_3\}{:}0.1067, \{h_3,vh_3\}{:}0.223, \{vh_3\}{:}0.1011), (\{vl_6\}{:}0.0219, \{vl_6,l_6\}{:}0.066, \{l_6\}{:}0.0805, \{l_6,m_6\}{:}0.1178, \{m_6\}{:}0.1356, \{m_6,h_6\}{:}0.1675, \{h_6\}{:}0.1326, \{h_6,vh_6\}{:}0.211, \{vh_6\}{:}0.0671), (\{vl_7\}{:}0.0576, \{vl_7,l_7\}{:}0.118, \{l_7\}{:}0.1288, \{l_7,m_7\}{:}0.0656, \{m_7\}{:}0.1207, \{m_7,h_7\}{:}0.0986, \{h_6\}{:}0.1162, \{h_7,vh_7\}{:}0.2189, \{vh_7\}{:}0.0755), (\{vl_8\}{:}0.0522, \{vl_8,l_8\}{:}0.0411, \{l_8\}{:}0.0933, \{l_8,m_8\}{:}0.0746, \{m_8\}{:}0.0945, \{m_8,h_8\}{:}0.1234, \{h_8\}{:}0.1716, \{h_8,vh_8\}{:}0.1682, \{vh_8\}{:}0.1810)\rangle$

Query type I (Single Attribute)

Now consider the query/hypothesis

(Do) *most* diabetic patients have between *medium* and *very high* diastolic blood pressure (?)

This statement is interpreted in label semantics as follows: Let P be a variable representing the unknown value of $\mu_{\theta}\left(X_3 | diab.\right)$ where $\theta = m_3 \vee h_3 \ \vee vh_3$. Also let the label set

for P be QL (to stand for quantifier labels) where, for example, *QL*={*all(a), almost all(aa), most(mst), several(s), few(f), hardly any(ha)*, *none(n)*}. Let the meaning of *most* be characterised by a trapezoidal fuzzy set μ_{mst} = [0.6:0 0.7:1 0.75:1 0.8:0]. Now in order to evaluate the above query we must evaluate the truth of the statement $\mathcal{D}_P \in \lambda(mst)$. To do this we must determine the degree to which *most* is an appropriate label for P on the basis of the evidence *e=DB* (i.e. the database). Now given *DB* we can determine a value p_{DB} for P where $p_{DB} = \mu_{\theta}\left(X_3 | DB, diab.\right) = \sum_{S \in \lambda(\theta)} m_{\mathcal{D}_{X_3}}\left(S | DB, diab.\right)$. In order to evaluate the query we then need only calculate $\mu_{mst}\left(p_{DB}\right)$.Now $\lambda(\theta) = \lambda\left(m_3 \vee h_3 \vee vh_3\right) = \{\{l_3, m_3\}, \{m_3\}, \ \{m_3, h_3\}, \{h_3\} \ \{h_3, vh_3\}, \{vh_3\}\}$ from which we obtain that

$$p_{DB} = m_{\mathcal{D}_{X_3}}\left(\{l_3, m_3\} | DB, diab.\right) + m_{\mathcal{D}_{X_3}}\left(\{m_3\} | DB, diab.\right)$$
$$+ m_{\mathcal{D}_{X_3}}\left(\{m_3, h_3\}, | DB, diab.\right) + m_{\mathcal{D}_{X_3}}\left(\{h_3\}, | DB, diab.\right) +$$
$$m_{\mathcal{D}_{X_3}}\left(\{h_3, vh_3\}, | DB, diab.\right) + m_{\mathcal{D}_{X_3}}\left(\{vh_3\}, | DB, diab.\right)$$
$$= 0.1179 + 0.0746 + 0.0842 + 0.1067 + 0.223 + 0.1179$$
$$= 0.7243$$

Hence, the support for the query is $\mu_{mst}(0.7243)$=1

Query type II (Multiple Attribute)

Consider the query/ hypothesis

> (Do) *several but not most* diabetic patients have between *high* and *very high* diastolic blood pressure and between *medium* but not *low* and *very high* plasma glucose concentration (?)

Let P be a variable representing the unknown value of $\mu_{\theta_2, \theta_3}\left(X_2, X_3 | diab.\right)$ where $\theta_2 = h_2 \vee vh_2$ and $\theta_3 = \left(m_3 \wedge \neg l_3\right) \vee h_3 \vee vh_3$. This is the joint distribution indicating the degree to which θ_2 will be appropriate to label X_2, while at the same time θ_3 will be appropriate to label X_3. Now given *DB* we want to determine a value $\mu_{\theta_2, \theta_3}\left(X_2, X_3 | DB, diab.\right)$ for P. This latter value cannot be determined from the linguistic prototype without further assumptions although both $\mu_{\theta_2}\left(X_2 | DB, diab.\right)$ and $\mu_{\theta_3}\left(X_3 | DB, diab.\right)$ can be evaluated and these provide some bounds on P. More specifically,

$$P \in \left[\boldsymbol{max}\left(0, \mu_{\theta_2}\left(X_2 | DB, diab.\right) + \mu_{\theta_3}\left(X_3 | DB, diab.\right) - 1\right),\right.$$
$$\left.\boldsymbol{min}\left(\mu_{\theta_2}\left(X_2 | DB, diab.\right), \mu_{\theta_3}\left(X_3 | DB, diab.\right)\right)\right]$$

It can easily be seen that $\mu_{\theta_2}\left(X_2 | DB, diab.\right)$ =0.7209. Also $\mu_{\theta_3}\left(X_2 | DB, diab.\right) = \sum_{S \in \lambda(\theta_3)} m_{\mathcal{D}_{X_3}}\left(S | DB, diab.\right)$ where $\lambda(\theta_3) = \{\{m_3\}, \{m_3, h_3\}, \ \{h_3\}, \{h_3, vh_3\}, \{vh_3\}\}$ so that $\mu_{\theta_2}\left(X_2 | DB, diab.\right)$ = 0.0746+ 0.0842 +0.1067+ 0.223+ 0.1011=0.5896 and $P \in [0.3105, 0.5896]$. Therefore, in order to evaluate the query we must compute

$$\mu_{s \wedge \neg mst}\left(P | P \in [0.3105, 0.5896]\right) = \frac{\int_{0.3105}^{0.5896} \mu_{s \wedge \neg mst}(P) dP}{0.5896 - 0.3105}$$

This is based on the assumption that the underlying prior distribution on P is uniform. Now if we assume that μ_s =[0.4:0 0.5:1 0.6:1 0.7:0] then from definition 2.4 we obtain that $\mu_{s \wedge \neg mst}$=[0.4:0 0.5:1 0.6:1 0.65:0]. Therefore,

$$\mu_{s \wedge \neg mst}\left(P | P \in [0.3105, 0.5896]\right) = 3.583\left\{\int_{0.4}^{0.5} (10x - 4) dx + \int_{0.5}^{0.5896} dx\right\}$$

=0.5002

Alternatively we could assume that the random sets $\mathcal{D}_{X_2}$ and $\mathcal{D}_{X_3}$ are conditionally independent given *T*. This gives a framework analogous to Naïve Bayes in machine learning. In this case, $\mu_{\theta_2, \theta_3}\left(X_2, X_3 | diab.\right) = \mu_{\theta_2}\left(X_2 | diab.\right) \times \mu_{\theta_3}\left(X_3 | diab.\right)$ so that we can obtain a value for P, p_{DB}, on the basis of the database given by

$$p_{DB} = \mu_{\theta_2}\left(X_2 | Db, diab.\right) \times \mu_{\theta_3}\left(X_3 | DB, diab.\right) = 0.42504.$$

From this the support for the query is given by $\mu_{s \wedge \neg mst}(0.42504) = 0.2504$.

5 Conclusion

A methodology for modelling with words based on label semantics and linguistic prototypes has been introduced. This approach provides a flexible knowledge representation framework for modelling with words. Specifically, we have shown that linguistic prototypes can be an effective tool for the evaluation of linguistic queries in databases by providing a linguistic summary of the data so that the whole database need no longer be stored.

References

[1] P. Bose, M. Galibourg and G. Hamon (1988), "Fuzzy Quering with SQL: Extensions and Implementation Aspects", *Fuzzy Sets and Systems 28,pp333-349*

[2] M. Delgado, D. Sanchez, M.A. Vila (1999), "A Survey of Methods for Evaluating Quantified Sentences", *Proceedings of EUSFLAT99, pp 279-282*

[3] J. Lawry (2001) "Label Prototypes for Modelling with Words", *Proceedings of NAFIPS2001*

[4] J. Lawry (2001) "Label Semantics: A Formal Framework for Modelling with Words", *Proceedings of ECSQARU 2001.*

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