Learning Qualitative Relations from Categorical Data

Jure Žabkar and **Martin Možina** and **Ivan Bratko** and **Janez Demšar**
University of Ljubljana, Faculty of Computer and Information Science
Tržaška 25, SI-1000 Ljubljana, Slovenia,
email: jure.zabkar@fri.uni-lj.si

Abstract

We address the problem of learning qualitative relations in categorical domains. We propose an algorithm that observes the change of probability of a target class w.r.t. the change in the values of the selected attribute for each learning example. We generalize the notion of a partial derivative by defining the probabilistic discrete qualitative partial derivative (PDQ PD). PDQ PD is a qualitative relation between the target class c and a discrete attribute, given as a sequence of attribute values a_i in the order of $P(c|a_i)$ in a local neighbourhood of the reference point. In a two stage learning process we first compute PDQ PD for all examples in the training data, and then generalize over the entire data set using a machine learning algorithm. The induced model explains the influence of the attribute's values on the target class in different subspaces of the attribute space.

1 Introduction

Qualitative modelling deals with representations of numerical quantities, mostly in physical domains. A branch of qualitative modelling considers learning qualitative models from numerical data (Bratko and Šuc 2003; Žabkar, Bratko, and Demšar 2007; Žabkar et al. 2009). It is a technique similar to regression modelling, except that instead of numerical predictions it predicts the direction of change (increase, decrease, no change) of the target variable. Such models are preferred over regression models when exact numerical models are difficult to obtain or difficult to use due to unreliable measurements, and, in particular, when the task is to predict the effect that a change in input variables will have on the observed variable. For instance, a qualitative model may describe the conditions under which a decrease of interest rates will stimulate investments and how will this affect the unemployment rate. While exact numerical models for this problem are elusive, qualitative relations between these variables may be easier to describe.

In this paper we extend this type of reasoning to categorical domains. We introduce an algorithm Qube for calculating qualitative preferences, a type of qualitative models describing the influence of a categorical (e.g. nominal) variable on a target class probability. Instead of dealing with direction of change, we rank attribute values so that the probability of the target class increases. By ranking, we drop all numerical information (probabilities). Finally, as we generalize over the entire data set, the induced qualitative model describes the conditions under which certain attribute values have greater/lower influence on the target class probability.

For example, consider a dataset describing passengers' experiences with different flight companies. Let the attributes describe each passenger (*gender, age, country, education* etc.) and the flight (*provider, distance, type of food, amount of food* etc.) taken by the passenger. The class variable describes passenger's overall satisfaction (*satisfied, not satisfied*). By setting target class to *satisfied*, Qube can calculate, for example, food preferences for each passenger taking a certain flight. These preferences express qualitative relation between the attribute *type of food* and the target class *satisfied*. Qualitative preferences define the order of attribute values but ignore the magnitudes, for example, *John*'s food preferences when flying with *BA* are:

$$\text{drink-only} \prec \text{vegetarian} \prec \text{sweet} \prec \text{non-vegetarian}.$$

We introduce *Probabilistic discrete qualitative models* (PDQ models) which predict how a change in the attribute values affects the probability of the target class, other attributes being equal.

Our approach intersects with the field called *preference learning* which gained much attention in AI and especially machine learning in recent years. The typical task of preference learning is to induce models for prediction of preferences, based on learning data which includes descriptions of examples and their preferences. We address the problem of learning preference models from data containing implicit preference information. Our algorithm Qube learns explicit preferences which can then be generalized to the preference model.

2 Methods

We will first provide a definition of probabilistic discrete qualitative partial derivatives based on conditional probabilities of the target class given the attribute values. The computation of these probabilities from data requires selecting proper subsets of examples, which we will describe next. Finally, we show how to combine the computed probabilities into partial derivatives and use them to induce qualitative models.

2.1 Probabilistic discrete qualitative partial derivative

Derivative of a function $f(x)$ at a certain point x_0, $f'(x_0)$ tells us, informally, the change of the function value corresponding to a certain (small) change of the value the function's argument, *e.g.*

$$f(x_0 + \Delta x) - f(x_0) \approx f'(x_0)\Delta x. \tag{1}$$

For functions of multiple arguments, *e.g.* $f(x_1, x_2, \ldots, x_n)$, we compute derivatives by each argument x_i separately and denote them by $\partial f/\partial x_i$.

Qualitative derivatives are similar to ordinary derivatives except that they give only the direction of change, that is, whether the function will increase or decrease when its argument increases. Qualitative derivative of a function is positive (negative, zero) if the continuous derivative is positive (negative, zero),

$$\frac{\partial_Q f}{\partial_Q x} = \text{sgn}\frac{\partial f}{\partial x}. \tag{2}$$

Now consider a multivariate distribution which assigns a probability y to each element of Cartesian product of $\mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n$. In machine learning, $(a_1, a_2, \ldots, a_n) \in \mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n$ can be values of discrete attributes $A_1, A_2, \ldots A_n$ describing an example (we will call this example a *reference example*), and y is the probability that such an example belongs to some target class c,

$$y = p(c|a_1, a_2, \ldots, a_n). \tag{3}$$

Recall that qualitative derivative of f w.r.t. x_i computed at $x_1, x_2, \ldots$ tells whether a certain change of x_i will *i*ncrease or decrease the function value if other arguments remain constant. Let the *probabilistic discrete qualitative partial derivative* at $(a_1, \ldots, a_n)$ with respect to A_i tell whether a change of value of the attribute A_i from a_i to a'_i will *increase or decrease the probability* of the target class:

$$\frac{\partial_Q f}{\partial_Q A_i : a_i \to a'_i}(a_1, \ldots, a_n) =$$

$$\begin{cases} +, & p(c|a_1, \ldots, a_i, \ldots, a_n) < p(c|a_1, \ldots, a'_i, \ldots, a_n) \\ \circ, & p(c|a_1, \ldots, a_i, \ldots, a_n) = p(c|a_1, \ldots, a'_i, \ldots, a_n) \\ -, & p(c|a_1, \ldots, a_i, \ldots, a_n) > p(c|a_1, \ldots, a'_i, \ldots, a_n). \end{cases} \tag{4}$$

Let us define a partial order on set $\mathcal{A}_i$, with respect to fixed values of a_j for all $j \neq i$

$$a_i \prec a'_i \Leftrightarrow$$
$$\Leftrightarrow p(c|a_1, \ldots, a_i, \ldots, a_n) \leq p(c|a_1, \ldots, a'_i, \ldots, a_n) \tag{5}$$

This allows us to rewrite (4) as

$$\frac{\partial_Q f}{\partial_Q A_i : a_i \to a'_i}(a_1, \ldots, a_n) = \begin{cases} +, & a_i \prec a'_i \\ \circ, & a_i = a'_i \\ -, & a_i \succ a'_i. \end{cases} \tag{6}$$

The derivative $\partial_Q f/\partial_Q A_i : a_i \to a'_i$ for any pair a_i and a'_i can thus be described by a total ordering of attribute values $\mathcal{A}_i$.

2.2 Computation of conditional probabilities

To compute the derivative we have to estimate the conditional probability $p(c|a_1, \ldots, a_n)$ at a single point $(a_1, a_2, \ldots, a_n)$ from data sample. This probability cannot be estimated directly, for instance by relative frequencies, since there may be only a few or even no examples in the data which match the condition part. We also cannot use a naive Bayesian method since it would reduce the PDQ PD to comparison of $p(c|a_1)$ and $p(c|a'_1)$, cancelling out all the terms corresponding to the values of other attributes. This is easily explained: the naive Bayesian assumption of conditional independence of attributes given the class implies that the derivative $\partial_Q f/\partial_Q A_i : a_i \to a'_i$ is constant on the entire attribute space.

The problem requires a semi-naive Bayesian approach. We will replace the condition in $p(c|a_1, \ldots, a_n)$ with a relaxed condition $\mathcal{D} \subseteq \{a_1, a_2, \ldots\}$, where $\mathcal{D}$ will include only the attribute values which are conditionally dependent on a_i given the class. Ignoring the other, conditionally independent values does not change the computed derivative (see the proof in Appendix).

To construct the set of conditions $\mathcal{D}$ we will use a greedy approach: we start with an empty set $\mathcal{D}$ and iteratively add the most dependent value. Let e_j represent an event that attribute A_j on a certain example has a value a_j. Let event v represent values of attribute A_i on an example ($v \in \mathcal{A}_i$). We need to test the hypothesis that e_j and v are conditionally independent given the class and a set of existing conditions $\mathcal{D}$ (that is, knowing the class of an example and knowing that the example satisfies $\mathcal{D}$, the probability distribution for values of A_i is independent of whether the value of j-th attribute equals a_j or not). In each step of the algorithm we test the dependence between v and e_j, find e_j which most strongly violates the independence and add the corresponding a_j to $\mathcal{D}$.

The independence is tested using the standard χ^2 test. Separate tables with $2 \times |\mathcal{A}_i|$ cells are constructed for class c and its complement. We compute the expected absolute frequencies for the first table as $n(c, \mathcal{D})p(a_j|c, \mathcal{D})p(v|c, \mathcal{D})$ and $n(c, \mathcal{D})(1 - p(a_j|c, \mathcal{D}))p(v|c, \mathcal{D})$, where $v \in \mathcal{A}_i$ and $n(c, \mathcal{D})$ is the number of examples in class c which satisfy the conditions $\mathcal{D}$. Frequencies for the complement of c are computed analogously. The sum of χ^2 statistics for both tables is distributed according to χ^2 distribution with $2(|\mathcal{A}_i| - 1)$ degrees of freedom.

For each a_j we compute its corresponding p-value and select the one with the lowest value. We stop the selection procedure when the lowest p-value is above the specified threshold or when the number of examples matching the conditions $\mathcal{D}$ falls below the given minimum. This is needed to ensure the reliability of χ^2 statistics and of estimated conditional probabilities. Our use of p-value does not require adjustments for multiple hypotheses testing since the p-value is used only as a stopping criteria and not to claim the significance of the alternative hypothesis.

After choosing a set of conditions $\mathcal{D}$, we compute $p(c|v, \mathcal{D})$ for all $v \in \mathcal{A}_i$ using relative frequency, Laplace estimate or m-estimate (Cestnik 1990) on examples match-

ing $\mathcal{D}$.

Note that the χ^2 test is performed over all values of A_i and not only a_i and a_i'. This lets us use the same set of conditions $\mathcal{D}$ for all derivatives with respect to A_i at a certain reference example and ensures that probabilities $p(c|a_1, \ldots, a_i, \ldots, a_n)$ for all $a_i \in \mathcal{A}_i$ are comparable and thus useful for defining a total ordering of a_i.

For another interpretation of this procedure, consider the definition of partial derivative of continuous functions:

$$\frac{\partial f}{\partial x_1}(x_1, \ldots, x_n) = \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h} \tag{7}$$

When computing the derivative with respect to x_1, we subtract the function value in two points where all arguments except x_1 are the same. If the function value at point $(x_1, \ldots, x_n)$ does not depend on, say, x_2, we could use different value of x_2 in the two terms. Omitting the values from (4) is similar to that: the χ^2 test is used to determine whether ignoring the difference between a_j and other values of the attribute A_j will affect the computation or not.

The proposed greedy procedure may seem naive, yet it works well in practice. We must also keep in mind that the selection of attributes needs to be fast since we have to run it for each point in which we compute the derivative, which rules out any advanced search for sets of dependent values.

2.3 Computation of derivatives

The total order of $\mathcal{A}_i$ is determined by the order of the corresponding probabilities as defined in (4). To handle noisy data we will however treat two probabilities (and the corresponding values of A_i) as equal if they differ for less than a user-provided threshold.

For this we use hierarchical clustering of values with average linkage (Sokal and Michener 1958) using the difference of probabilities as distances. The clustering is stopped when the distance between the closest clusters is greater than 0.2.

For example, let A_i be a five-valued attribute with values v_1 to v_5. Probabilities $p(c|v_i, \mathcal{D})$ for these values equal 0.1, 0.2, 0.3, 0.5 and 0.6, respectively. Let the merging threshold be 0.2. We recognize v_1 and v_2 as equivalent and assign them the average probability of $(0.1 + 0.2)/2 = 0.15$. Next we merge v_4 and v_5, the average probability is $0.5 + 0.6 = 0.55$. Finally, we merge v_1 and v_2 with v_3; the average probability is $(0.1 + 0.2 + 0.3)/3 = 0.2$. We then stop since the difference between $p = 0.2$ (v_1 to v_3) and $p = 0.55$ (v_4 to v_5) exceeds the threshold of 0.2. The resulting total ordering of A_i is $v_1 = v_2 = v_3 \prec v_4 = v_5$.

2.4 Induction of qualitative models

To induce a qualitative model with respect to a certain attribute A_i, we first compute the PDQ PD for the entire learning set: for each example, we compute the set of dependent values $\mathcal{D}$ and find the total ordering of attribute values $\mathcal{A}_i$ as explained in the previous two sections. We replace the original class labels with partial derivatives (that is, the total ordering) and induce a model for predicting the ordering. In principle, any learning algorithm can be used for this task.

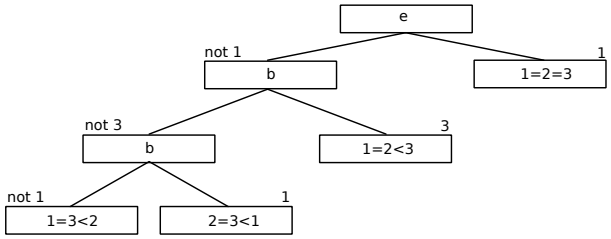


Figure 1: PDQ model for derivative of y w.r.t. attribute a for Monks 1.

3 Experiments

We present the experiments on three UCI (Asuncion and Newman 2007) data sets: Monks 1, Monks 3 and Titanic. The Monks data sets are interesting because the underlying concept is known and can be used to verify the induced models. Titanic is a representative of a real-world data set on which we can check the interpretability of the models.

Conditional probabilities are estimated using the m-estimate (Cestnik 1990) with $m = 2$. We use a threshold of 0.2 for merging attribute values. Our reimplementation of the C4.5 algorithm (Zupan, Leban, and Demšar 2004) is used to obtain qualitative models describing the relation between the target class and each attribute separately.

Monks 1 Monks data sets consists of 6 discrete attributes: a $(1, 2, 3)$, b $(1, 2, 3)$, c $(1, 2)$, d $(1, 2, 3)$, e $(1, 2, 3, 4)$, f $(1, 2)$ and a binary class y $(1, 0)$. For the first data set, the target concept is defined by $y := (a = b) \vee (e = 1)$. We selected target class $y = 1$.

Computation of partial derivatives with respect to all attributes and induction of the corresponding trees took 7.5 seconds on a laptop computer.

Derivative with respect to c was $1 = 2$ on the entire data set, meaning that 1 and 2 have the same influence on the target class. This is in agreement with the target concept since the attribute c is irrelevant. Similarly, derivatives by d and f were $1 = 2 = 3$ and $1 = 2$, respectively.

Figure 1 shows the model for $\partial_Q y / \partial_Q a$; the model for $\partial_Q y / \partial_Q b$ is equivalent. The tree states that when $e = 1$, the values of a (and b) do not matter; this is true since y is already 1 when $e = 1$, disregarding values of other attributes. For $e \neq 1$, the probability of $y = 1$ is greater for the value of a (b) which is equal to the value of b (a). For instance, if $b = 3$, then class 1 has a higher probability of when $a = 3$ then when $a = 1$ or $a = 2$. There is no difference between the latter two values, so $1 = 2 \prec 3$.

We have, for sake of clarity, transformed the model for $\partial_Q y / \partial_Q e$ into a set of rules.

$\partial_Q y / \partial_Q e$:
- IF $b = 1 \ \wedge \ a = 1$ THEN $1 = 2 = 3 = 4$
- IF $b = 2 \ \wedge \ a = 2$ THEN $1 = 2 = 3 = 4$
- IF $b = 3 \ \wedge \ a = 3$ THEN $1 = 2 = 3 = 4$
- IF $b = 1 \ \wedge \ a \neq 1$ THEN $2 = 3 = 4 \prec 1$
- IF $b = 2 \ \wedge \ a \neq 2$ THEN $2 = 3 = 4 \prec 1$
- IF $b = 3 \ \wedge \ a \neq 3$ THEN $2 = 3 = 4 \prec 1$

The rules describe basically the same patterns as the models for a and b but now from the perspective of the attribute e: if $a = b$ (the first three rules), the value of e is irrelevant. If $a \neq b$, $e = 1$ yields a higher probability of the target class than the other three values. For values 2, 3 and 4 the probability is equal.

Monks 3 In Monks 3 the class y is defined as $y := (e = 3 \wedge d = 1) \vee (e \neq 4 \wedge b \neq 3)$ and some noise is added. Again, we select target class $y = 1$. Computation of derivatives and induction of trees took 7.6 seconds. We converted the trees into rules for sake of readability.

As in Monks 1, our algorithm correctly recognizes the irrelevant attributes (a, c and f). The model for $\partial_Q y/\partial_Q b$ is consistent with the definition of the underlying concept.

$\partial_Q y/\partial_Q b$:
IF $e = 4$ THEN $1 = 2 = 3$
ELSE IF $e = 3 \wedge d = 1$ THEN $1 = 2 = 3$
ELSE $3 \prec 1 = 2$

If $e = 4$, the influence of b's values is equal since an example with $e = 4$ has $y = 0$, disregarding the value of b. If $(e = 3 \wedge d = 1)$ the value of b is again irrelevant as this condition already gives $y = 1$. Otherwise, $b = 1$ and $b = 2$ have much greater chances of achieving $y = 1$ than $b = 3$.

The model for $\partial_Q y/\partial_Q d$ is less perfect.

$\partial_Q y/\partial_Q d$:
IF $e = 3 \wedge b = 3$ THEN $2 = 3 \prec 1$
IF $e \neq 3$ THEN $1 = 2 = 3$
IF $e = 3 \wedge b \neq 3 \wedge a = 2 \wedge c = 1$ THEN $3 \prec 1 = 2$
ELSE $1 = 2 = 3$

The rules correctly state that $d = 1$ yields higher probability for $y = 1$ than the other values of d, if b and e both equal 3. It is also true that the values of d are all of the same influence if $e \neq 3$. The third rule is not consistent with the target concept and can be an artefact of noise in the data.

The rule for $\partial_Q y/\partial_Q e$ again correctly implies that e equal to 1, 2 or 3 gives a higher probability for $y = 1$ than $e = 4$. Examples with $e = 3$ yield a higher probability than other values of e when $b = 3$ and $d = 1$. If $b = 3$ and $d \neq 1$, the value of e does not matter.

$\partial_Q y/\partial_Q e$:
IF $b \neq 3$ THEN $4 \prec 1 = 2 = 3$
IF $b = 3 \ \wedge \ d = 1$ THEN $1 = 2 = 4 \prec 3$
IF $b = 3 \ \wedge \ d \neq 1$ THEN $1 = 2 = 3 = 4$

Titanic Titanic data set consists of 2201 examples described by three attributes, *status (first, second, third, crew)*, *age (child, adult)* and *sex (male, female)*, and a class *survived (yes, no)*. We selected target class *survived* = *yes*. The time needed for computation of derivatives and induction of trees was 52 seconds due to a large number examples for which we computed the partial derivatives.

For the probability of survival with respect to the status (Figure 2a) we can observe that the chances of survival are equal for all adult males. For others, the probability of survival is worst in the third class and equal for others, except for girls where the survival in the second class was higher than in the first.

Figure 2b shows relations from perspective of age. Among the crew there were no children. For the third class, the probability is similar for children and adults. The only group where the age makes a difference are males in the first and second class.

A similar story is told in Figure 2c. The gender is not important for children, while among adults females fared much better than males.

4 Case study: Bacterial infections in geriatric population

Compared to younger population, people over 65 years of age usually react to a disease in a different way. Many symptoms may not even present or they are masked by others which makes it a very difficult task for a medical doctor to diagnose a condition, to decide a proper treatment or to estimate the patient's risk of death. Many patients that present with infection have associated chronic diseases such as diabetes, heart, kidney, lung or liver disease which makes the treatment even more complicated. Despite great progress in treating infectious diseases they remain one of the major causes of death in geriatric population. Some differences in the course of illness can be observed compared to younger patients.

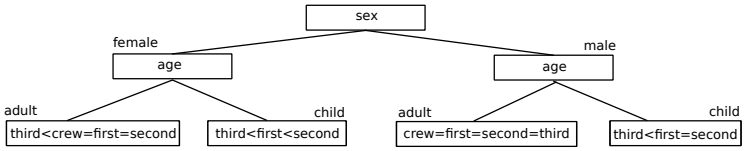
Greater risk of severe bacterial infection is due to the patient being immunocompromised (Ben-Yehuda and Weksler 1992), immobile, nursing home resident or co-morbidity. In elderly, the infections often present with atypical signs, such as the absence of fever (Marco et al. 1995; Castle et al. 1991; Gleckman and Hibert 1982; Mellors et al. 1987), the absence of cough at pneumonia and weakness or changed mental status (Rockwood 1989). These usually cause a delay in making a right diagnosis. A proper and efficient antimicrobial treatment is often given too late, and the risk of fatal outcome is increased (Fontanarosa et al. 1992; Pfitzenmeyer et al. 1995).

4.1 Description of data

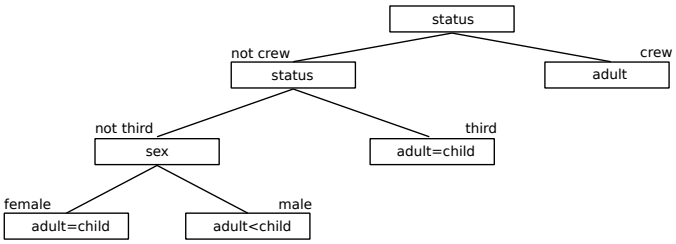
The prospective study was carried out at University Medical Centre Ljubljana, Department of Infectious Diseases from June 1st, 2004 to June 1st, 2005. It included the patients having C-reactive protein in serum (CRP) $60mg/l$ or more, which indicates bacterial etiology of infection. The patients were observed for 30 days from their first visit or death due to observed infection. The study included 602 patients of age 65 and above. Data contains 32 attributes and a binary class *DIED (Yes, No)* having the following distribution: *DIED=Yes*: $77/602 = 12.8\%$ and *DIED=No*: $525/602 = 87.2\%$.

4.2 Learning qualitative relations in medical domain

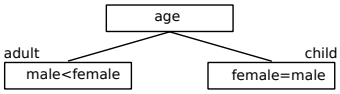
We selected target class *DIED=Yes*. We observed the relations between the target class and individual attributes. In other words, how does a change in the value of the selected attribute influence the probability of the target class. We computed the ranking of attribute's values for each learning example using algorithm Qube. We induced qualitative models using the C4.5 algorithm. Due to the space restriction, we only report a small subset (Table 1) of 32 models.



(a) $\partial_Q survived/\partial_Q status$



(b) $\partial_Q survived/\partial_Q age$



(c) $\partial_Q survived/\partial_Q sex$

Figure 2: PDQ models for Titanic data.

The obtained models were evaluated by two infectologists. The majority (25/32) of the models made perfect sense. These models are clear, accurate and simple. For example, the model for *GENDER* says that both genders have the same influence on the target class while the model for *COMORBIDITY* states that having more than one associated disease significantly raises the probability of death compared to having zero or one associated disease. The model for body temperature (*TEMP*) states that in the case of changed conciousness lower body temperature implies worst outcome than higher body temperature. The medical explanation is that older people with severe bacterial infections (that cause the conciousness to change) can not develop the temperature resistance to the infection due to extremely weak general condition. The same also happens with unchanged conciousness but having *CLINICALINF = OTHER* which in our data usually means severe sepsis. Concious patients with defined infections usually develop the temperature resistance - higher body temperature implies worse outcome. It is also known that lower values of blood pressure (*RR*) and saturation (*SAT*) suggest worse outcome with the patients with bacterial infections.

Seven models were only partially explained, meaning that further medical research should be carried out to affirm or reject the total correctness. Further work is also possible in machine learning, for example by choosing different cut-off values while converting numerical attributes to categorical, or even by making this process automatic. Current cut-off values were suggested by medical doctors based on their experience.

5 Related work

Research of qualitative reasoning in AI has been mostly concerned with qualitative physics (de Kleer and Brown 1984; Kuipers 1986; 1994; Forbus 1997; 1984). In these works the model was provided by an expert and then used in qualitative simulations. There are only a few algorithms for automated induction of such models (Klenk and Forbus 2009) and even these are limited to learning from numerical data (Bratko and Šuc 2003; Žabkar, Bratko, and Demšar 2007). There are, to the best of our knowledge, no algorithms for learning qualitative models from categorical data.

An important part of our method deals with relaxing the strong independence assumption of naive Bayesian approach. There exist a number of methods for this purpose, yet none fits our context. Kononenko introduced *semi-naive Bayes* (Kononenko 1991) and Langley and Sage (Langley and Sage 1994) proposed *Selective Bayesian Classiffier*, a variant of the naive method that uses only a subset of the attributes in making predictions. Since their algorithm is only searching for subset of attributes that yields highest classification accuracy, it can not reveal attribute dependencies. Friedman and Goldszmidt introduced *tree augmented naive Bayes (TAN)* (Friedman and Goldszmidt 1996) which

attribute	qualitative model
GENDER	M=F
AGE	IMMUNITY = Yes: B=C $\prec$ A IMMUNITY = No: A=B=C
COMORBIDITY	ZERO=ONE $\prec$ MORE
IMMUNITY	No $\prec$ Yes
TEMP	CHNGCONC = Yes: [>37.8] $\prec$ [$\leq$ 37.8] CHNGCONC = No: CLINICALINF = OTHER: [>37.8] $\prec$ [$\leq$ 37.8] CLINICALINF $\neq$ OTHER: [$\leq$ 37.8] $\prec$ [>37.8]
FRQBREATH	[(10.00, 20.00] = >20.00] $\prec$ [$\leq$10.00]
SAT	>90.00 $\prec$ $\leq$90.00
RR	>90.00 $\prec$ $\leq$90.00
CHNGCONC	No $\prec$ Yes

Table 1: The qualitative models describing the relations between individual attributes and the target class.

allows for attributes having another attribute as a parent in the bayesian network representation.

A lazy algorithm, named Locally Weighted Naive Bayes (LWNB) is proposed in (Frank, Hall, and Pfahringer 2003). LWNB relaxes the independence assumption by learning local models at prediction time. The models are learned on weighted set of training instances in the neighbourhood of the test instance. In LWNB, the test example neighbourhood is chosen using the k-nearest neighbours algorithm. A step further is the Lazy Bayesian Rules (LBR) algorithm (Zheng, Webb, and Ting 1999). LBR search of the local neighbourhood is not based on a global metric. Instead, for each test example, LBR uses a greedy search to generate a Bayesian rule with an antecedent that matches the test example. The basic difference between these approaches and ours is that these methods are concerned with optimizing the accuracy of predictions and not with estimations of the chosen attribute's influence on the target class probability.

As we have mentioned in the introduction, our work intersects with preference learning (Fürnkranz and Hüllermeier 2005; Brafman 2008). Preference learning considers the problem of learning given learning examples as well as the preferences (Boutilier et al. 2003; Chu and Ghahramani 2005; Brochu, de Freitas, and Ghosh 2007). Our approach starts earlier and calculates the preferences for each learning example from data. While one could continue by using standard preference learning approaches (Chu and Ghahramani 2005; Brochu, de Freitas, and Ghosh 2007), we use simple machine learning algorithms and treat preferences as values of a new class variable. Theoretically, it is possible that the number of class values exceeds a reasonable amount but it can be practically very well controlled by setting the threshold parameter (for joining the values) and the size of the neighbourhood of the reference example (a kind of smoothing the data).

6 Conclusion

We have presented, to our knowledge, the first machine learning method for induction of qualitative models from categorical data. The method has a solid theoretical background and we have also shown a few simple examples on which it performed excellently. Additionally, we presented a real case study which shows that our algorithm is robust regarding noise and produces simple yet accurate and comprehensible models.

Appendix

We need to prove that the ordering of probabilities $p(c|a_1,\ldots,a_i,\ldots,a_n)$ does not change if we omit from the conditional part the values which are conditionally independent from a_i given the class c. Let us first redefine the PDQ PD using conditional log odds ratios.

$$\frac{\partial_Q f}{\partial_Q A_i : a_i \to a_i'}(a_1,\ldots,a_n) = \operatorname{sgn} \ln \frac{p(c|a_1,\ldots,a_i,\ldots,a_n)/p(\bar{c}|a_1,\ldots,a_i,\ldots,a_n)}{p(c|a_1,\ldots,a_i',\ldots,a_n)/p(\bar{c}|a_1,\ldots,a_i',\ldots,a_n)}, \quad (8)$$

where $\bar{c}$ is the complement of the target class c. It is easy to see that (8) is equivalent to (4).

Let us without loss of generality assume that values a_1 to a_k, $k < i$ are conditionally independent of values a_{k+1} to a_n, given the class. Applying Bayesian rule, using the independence assumption, cancelling the identical terms and reapplying the Bayesian rule turns (8) into

$$\frac{\partial_Q f}{\partial_Q A_i : a_i \to a_i'}(a_1,\ldots,a_n) = \operatorname{sgn} \ln \frac{p(c|a_{k+1},\ldots,a_i,\ldots,a_n)/p(\bar{c}|a_{k+1},\ldots,a_i,\ldots,a_n)}{p(c|a_{k+1},\ldots,a_i',\ldots,a_n)/p(\bar{c}|a_{k+1},\ldots,a_i',\ldots,a_n)}. \quad (9)$$

This is equivalent to (4) without values a_1 to a_k. Therefore, $p(c|a_1,\ldots,a_i,\ldots,a_n) \leq p(c|a_1,\ldots,a_i',\ldots,a_n) \iff p(c|a_{k+1},\ldots,a_i,\ldots,a_n) \leq p(c|a_{k+1},\ldots,a_i',\ldots,a_n)$.

References

Asuncion, A., and Newman, D. J. 2007. UCI machine learning repository.

Ben-Yehuda, A., and Weksler, M. 1992. Host resistance and the immune system. *Clin Geriatr Med* 8(4):701–11.

Boutilier, C.; Brafman, R. I.; Hoos, H. H.; and Poole, D. 2003. Cp-nets: A tool for representing and reasoning with

conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research* 21:2004.

Brafman, R. I. 2008. Preferences, planning and control. In *KR*, 2–5.

Bratko, I., and Šuc, D. 2003. Learning qualitative models. *AI Magazine* 24(4):107–119.

Brochu, E.; de Freitas, N.; and Ghosh, A. 2007. Active preference learning with discrete choice data. In *Advances in Neural Information Processing Systems*.

Castle, S.; Norman, D.; Yeh, M.; Miller, D.; and Yoshikawa, T. 1991. Fever response in elderly nursing home residents: are the older truly colder? *J Am Geriatr Soc* 39(9):853–7.

Cestnik, B. 1990. Estimating probabilities: A crucial task in machine learning. In *ECAI*, 147–149.

Chu, W., and Ghahramani, Z. 2005. Preference learning with gaussian processes. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, 137–144. New York, NY, USA: ACM.

de Kleer, J., and Brown, J. 1984. A qualitative physics based on confluences. *Artificial Intelligence* 24:7–83.

Fontanarosa, P. B.; Kaeberlein, F. J.; Gerson, L. W.; and Thomson, R. B. 1992. Difficulty in predicting bacteremia in elderly emergency patients. *Annals of Emergency Medicine* 21(7):842–8.

Forbus, K. 1984. Qualitative process theory. *Artificial Intelligence* 24:85–168.

Forbus, K. 1997. *Qualitative reasoning*. CRC Press.

Frank, E.; Hall, M.; and Pfahringer, B. 2003. Locally weighted naive Bayes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2003)*.

Friedman, N., and Goldszmidt, M. 1996. Building classifiers using Bayesian networks. In *Proceedings of the thirteenth national conference on artificial intelligence*, 1277–1284. AAAI Press.

Fürnkranz, J., and Hüllermeier, E. 2005. Preference learning. *KI* 19(1).

Gleckman, R., and Hibert, D. 1982. Afebrile bacteremia. a phenomenon in geriatric patients. *JAMA* 248(12):1478–81.

Klenk, M., and Forbus, K. 2009. Analogical model formulation for transfer learning in AP physics. *Artificial Intelligence* 173(18):1615–1638.

Kononenko, I. 1991. Semi-naive bayesian classifier. In *EWSL*, 206–219.

Kuipers, B. 1986. Qualitative simulation. *Artificial Intelligence* 29:289–338.

Kuipers, B. 1994. *Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge*. MIT Press, Massachusetts.

Langley, P., and Sage, S. 1994. Induction of selective Bayesian classifiers. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 399–406. Morgan Kaufmann.

Marco, C.; Schoenfeld, C.; Hansen, K.; Hexter, D.; Stearns, D.; and Kelen, G. 1995. Fever in geriatric emergency patients: clinical features associated with serious illness. *Ann Emerg Med* 26(1):18–24.

Mellors, J. W.; Horwitz, R. I.; Harvey, M. R.; and Horwitz, S. M. 1987. A simple index to identify occult bacterial infection in adults with acute unexplained fever. *Arch Intern Med* 147(4):666–71.

Pfitzenmeyer, P.; Decrey, H.; Auckenthaler, R.; and Michel, J. 1995. Predicting bacteremia in older patients. *J Am Geriatr Soc* 43(3):230–5.

Rockwood, K. 1989. Acute confusion in elderly medical patients. *J Am Geriatr Soc* 37(2):150–4.

Sokal, R. R., and Michener, C. D. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* 28:1409–1438.

Žabkar, J.; Možina, M.; Bratko, I.; and Demšar, J. 2009. Discovering monotone relations with padé. In *ECML 2009 workshop: Learning Monotone Models From Data*.

Žabkar, J.; Bratko, I.; and Demšar, J. 2007. Learning qualitative models through partial derivatives by Padé. In *Proc. of the 21th International Workshop on Qualitative Reasoning*.

Zheng, Z.; Webb, G. I.; and Ting, K. M. 1999. Lazy Bayesian rules: a lazy semi-naive Bayesian learning technique competitive to boosting decision trees. In *Proc. 16th International Conf. on Machine Learning*, 493–502. Morgan Kaufmann, San Francisco, CA.

Zupan, B.; Leban, G.; and Demšar, J. 2004. Orange: Widgets and visual programming, a white paper.