Extending Bayesian Classifier with Ontological Attributes

Tomasz Łukaszewski

Institute of Computing Science, Poznan University of Technology,
ul. Piotrowo 2, 60-965 Poznan, Poland
tlukaszewski@cs.put.poznan.pl

1 Introduction

The goal of inductive learning classification is to form generalizations from a set of training examples such that the classification accuracy on previously unobserved examples is maximized. Given a specific learning algorithm, it is obvious that its classification accuracy depends on the quality of training data. In learning from examples, *noise* is anything which obscures correlations between attributes and the class [1]. There are many possible solutions to deal with the existence of noise. Data cleaning or detection and elimination of noisy examples constitutes the first approach. Due to the risk of data cleaning, when noisy examples are retained while good examples are removed, efforts have been taken to construct noise tolerant classifiers. Although both these approaches seem very different, they try to somehow 'clean' this noisy training data.

In this paper, we propose an approach to 'admit and utilize' noisy data by enabling to model different *levels of knowledge granularity* both in *training* and *testing* examples. The proposed knowledge representation use hierarchies of sets of attribute values, derived from subsumption hierarchies of concepts from an ontology represented in description logic. The main contributions of the paper are: (i) we propose a novel extension of the naïve Bayesian classifier by hierarchical, ontology based attributes (*ontological attributes*), (ii) we propose an inference scheme that handles ontological attributes.

2 Description-noise and Levels of Knowledge Granularity

There are three major sources of noise: (i) insufficiency of the description for attributes or the class (or both), (ii) corruption of attribute values in the training examples, (iii) erroneous classification of training examples [1]. The second and third source of noise can lead to so-called *attribute-noise* and *class-noise* respectively. Attribute-noise is represented by: (i) erroneous attribute values, (ii) missing or "don't care" attribute values, (iii) incomplete attributes or "don't care" values. The class-noise is represented by: (i) contradictory examples, or (ii) misclassification [2]. However, the first major source of noise, although not easily quantifiable, is important. This insufficiency of the description can lead to

both erroneous attribute values and erroneous classification. Let us call this resulting noise as *description-noise*. Following for example [3] the main reason for description-noise may be in a language used to represent attribute values, which is not expressive enough to model different *levels of knowledge granularity*. In such a case, erroneous or missing attribute values may be introduced by users of a system that are required to provide very specific values, but the level of their knowledge of the domain is too general to precisely describe the observation by the appropriate value of an attribute. Even if the person is an expert of the domain, erroneous or missing attribute values can be observed as a consequence of lack of time, or other resources to make detailed observations (ie. a more complete description). However, if the language enabled modeling different levels of knowledge granularity (very precise or more general descriptions), we would be able to decrease a level of this description-noise.

In order to model different levels of knowledge granularity, each testing and training example would be described by *a set of values* for any attribute. These sets of values should reflect the domain knowledge and could not be constructed arbitrarily. Let us notice, that in some domains, hierarchical or taxonomical relationships between sets of values, represented by so called *concepts*, may be observed and this knowledge could be explored. Such knowledge is currently often available in the form of *ontologies*. The most widely used language to represent ontologies, suitable in particular to model taxonomical knowledge, is *Web Ontology Language (OWL)* [1]. The theoretical counterpart of OWL, from which its semantics is drawn, is constituted by a family of languages called *description logics (DLs)* [4]. A description logic *knowledge base*, KB, is typically divided into *intensional* part (*terminological* one, a $TBox$), and *extensional* part (*assertional* one, an $ABox$).

3 An Ontological Attribute

Given is an attribute A and the set $V = \{V_1, V_2, ..., V_n\}$, where $n > 1$, of nominal values of this attribute. Let us assume that given is a TBox, which specifies domain knowledge relevant to a given classification task. In particular, it expresses a multilevel subsumption ("is-a") *hierarchy of concepts*. Each concept is described by a subset of the set V for every attribute A. Then we can formulate a definition of an *ontological attribute* as follows.

Ontological attribute An ontological attribute $\mathcal{A}$ is defined by a tuple $\langle \mathcal{H}, V \rangle$, where:

- by $\mathcal{H}$ is denoted a multilevel subsumption hierarchy of concepts, derived from a DL knowledge base. This hierarchy of concepts consists of the set of nodes $N^H = \{root, N^C, N^T\}$. This hierarchy defines a *root-node*, denoted by $root$, a set N^C of *complex-nodes* and a set N^T of *terminal-nodes*.

[1] www.w3.org/TR/owl-features/

– by V is denoted a finite set $V = \{V_1, V_2, ..., V_n\}$, where $n > 1$ of nominal values of A.
– each node $N_k \in N^T \cup N^C$ represents a subset of the set V, denoted as $val(N_k)$; the root-node represents the set V

To model actual training examples, an ABox would be used.

3.1 Using Ontological Attributes in the Naïve Bayesian Classifier

In order to apply the proposed ontological attributes in the naïve Bayesian classifier, we further specify the general definition of an ontological attribute given in the former section. Please note, that by making the assumptions presented in the following paragraphs, we will implicitly switch from the usual open world assumption used to reason with a DL knowledge base to produce a concept hierarchy, to the closed world assumption, more appropriate to the case of inference with naïve Bayesian classifier. In particular we will assume that a hierarchy of concepts would represent such *hierarchical partitioning* of the set V of attribute values, such that each concept would correspond to a non-empty subset of V.

Properties of nodes Each complex-node represents a concept from the KB, described by a proper, non-empty subset of V. Each terminal-node represents a concept from the KB, described by a unique value V_i from the set V.

Relations between nodes For a given ontological attribute $\mathcal{A}$, the hierarchy $\mathcal{H}$ is a tree, i.e. each node $N_k \in \{N^C \cup N^T\}$ has exactly one parent, denoted as $pa(N_k)$, such that $val(N_k) \subset val(pa(N_k))$. Moreover, each node $N_k \in \{root \cup N^C\}$ specifies a set $ch(N_k)$ of his children. To model different levels of knowledge granularity, we assume that for each $N_k \in \{root \cup N^C\}$ all his children are pairwise disjoint and this node N_k is a union of his children. Finally, for each node $N_k \in \{root \cup N^C\}$ we define a set $de(N_k)$ of descendants of this node, as a set of its children or children of his descendants.

The role of complex-nodes In the setting of learning with description-noise, each training and testing example can be described in general by a set Z_l of values for each attribute A, where $Z_l \subseteq V$. We can divide training examples into *no-noisy examples* ($|Z_l| = 1$) and *noisy examples* ($|Z_l| > 1$). In order to represent noisy (training and testing) examples, the ontological attribute $\mathcal{A}$ uses complex-nodes. We will call such a hierarchy a *complex-hierarchy*.

Algorithm 1 (Populating a complex-hierarchy). For each ontological attribute $\mathcal{A}$ we proceed as follows:
We associate each training example t described by a set Z_l of values of A and a class label C_j ($t : A = Z_l \wedge C = C_j$) to a node N_k. When $|Z_l| = 1$, Z_l is associated to a terminal-node N_k, such that $Z_l = val(N_k)$. Otherwise, we associate the training example to a complex-node N_k, such that $Z_l \subseteq val(N_k)$, at the *lowest* possible level of the complex-hierarchy.

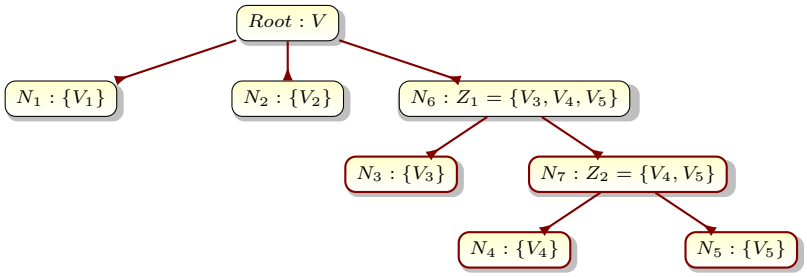

Fig. 1: A complex-hierarchy in setting with description-noise

Example Given is an attribute A such that V=$\{V_1$,V_2,V_3,V_4,$V_5\}$ and given is a class variable C such that it takes values from the set $\{C_1, C_2\}$. Let us assume, that the description-noise is modeled by sets $Z_1 = \{V_3, V_4, V_5\}$ and $Z_2 = \{V_4, V_5\}$. Let us assume a sample scenario in which the single values of the attribute A are determined by conducting three medical tests. The first test is able to partition the set V into the following disjoint subsets: $\{V_1\}$, $\{V_2\}$ and $Z_1 = \{V_3, V_4, V_5\}$. If the result of the first test is Z_1, then in some cases it is conducted a second test, that partitions the set Z_1 into the following disjoint subsets: $\{V_3\}$ and $Z_2 = \{V_4, V_5\}$. Only in critical cases it is conducted the last test, that can partition the set Z_2 into disjoint subsets: $\{V_4\}$ and $\{V_5\}$. Following this domain-knowledge, we have introduced two complex-nodes N_6 and N_7, such that they represent the sets Z_1 and Z_2 respectively. Terminal-nodes N_1, N_2, N_3, N_4, N_5 represent single values from the set V. The root-node represents the set V. The resulting complex-hierarchy is presented in Figure 1.

3.2 Inference with Ontological Attributes

We can approximate the required probability distribution for a noisy *testing* example described by a set $Z_l = val(N_k)$, following principles of the probabilistic theory, by *collecting* frequencies of training examples T, described by sets $Z_m \subseteq Z_l$, as follows:

$$P(Z_l|C_j) = \frac{\sum_{Z_m \subseteq Z_l} |T : A = Z_m \wedge C = C_j|}{|T : C = C_j|} \tag{1}$$

Let us remind, that a set Z_l is assigned to the node N_k, such that $Z_l = val(N_k)$. The key property of an ontological attribute $\mathcal{A}$, is that for the node N_k all its children are *pairwise disjoint*. Since then, all training examples described by sets $Z_m \subseteq Z_l$, are represented by the node N_k or its descendants, and the probability distribution for a noisy *testing* example described by a set Z_l we can define as follows:

$$P(Z_l|C_j) = \frac{|T : A \subseteq val(N_k) \wedge C = C_j| + \sum_{N_d \in de(N_k)} |T : A \subseteq val(N_d) \wedge C = C_j|}{|T : C = C_j|} \tag{2}$$

In this way we are able to *classify a new noisy example using other less noisy and no-noisy training examples*. For example, we can classify a testing example, described by the set Z_1, and associated to the node N_6 using all training examples described by all subsets of the set Z_1. These training examples would be associated to the complex-node N_6 or his descendants.

4 Conclusions

The topic of learning with ontologies is relatively new, and so far there are few approaches in this line of research, for the classification task see for example [5]. The simple use of ontology (Attribute Value Taxonomies) in the naïve Bayesian classifier (AVT-NBL) is presented in [6]. This approach, to the best of our knowledge, is the only one existing approach for learning the naïve Bayesian classifier from noisy (partially specified) data. Both in our approach and in AVT-NBL, noisy (partially specified) data is represented using hierarchical structures and similar aggregation procedures are used. Let us notice, that AVT-NBL requires a *static*, predefined, taxonomy of attribute values. In our approach, the hierarchy of sets of attribute values can be constructed *dynamically* driven by observations and hypotheses to prove. Moreover, our aggregation procedure allows to construct the complex-hierarchy from all possible subsets of attribute values. In this way we would be able to model any noisy training and testing example in order to achieve the highest classification accuracy, that is not possible using an Attribute Value Taxonomy. Due to limitations of the presentation, this generalization is not discussed in the paper. Let us point out, that AVT-NBL uses a propagation procedure, that does not follow principles of the probabilistic theory. Moreover, to the best of our knowldge, AVT-NBL does not classify noisy instances, which is the main goal of our approach.

In the future, we will concentrate on the problem of the optimality of the complex-hierarchy, derived from a knowledge domain of the form of subsumption hierarchies of concepts.

References

1. Hickey, R.J.: Noise Modelling and Evaluating Learning from Examples. Artif. Intell. **82**(1-2) (1996) 157–179
2. Zhu, X., Wu, X.: Class Noise vs. Attribute Noise: A Quantitative Study. Artif. Intell. Rev. **22**(3) (2004) 177–210
3. Clark, P., Niblett, T.: Induction in Noisy Domains. In: EWSL. (1987) 11–30
4. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., eds.: The Description Logic Handbook. Cambridge University Press (2003)
5. d'Amato, C., Fanizzi, N., Esposito, F.: Distance-Based Classification in Owl Ontologies. In Lovrek, I., Howlett, R.J., Jain, L.C., eds.: KES (2). Volume 5178 of Lecture Notes in Computer Science., Springer (2008) 656–661
6. Zhang, J., Honavar, V.: AVT-NBL: An Algorithm for Learning Compact and Accurate Naïve Bayes Classifiers from Attribute Value Taxonomies and Data. In: ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining, Washington, DC, USA, IEEE Computer Society (2004) 289–296