Building classifiers that cope with small training sets

Erik Quaeghebeur

Supervisor(s): Gert de Cooman, Dirk Aeyels

Abstract— **Classifiers provide an automated way of attaching a class to an object described by one or more attributes. They are constructed using a training set of (manually) pre-classified objects.**

Classical classifiers try to attach a unique class to an object, but are unreliable when constructed with a small training set. This limitation can be overcome by allowing a classifier to attach sets of classes to an object.

As a basis for such a so-called credal classifier, we propose a probabilistic model for the classes and for attributes that are distributed according to an exponential family. This family includes most common distributions, such as the normal, Poisson and multinomial.

Keywords— **Classification, exponential family, imprecise probabilities.**

I. Introduction

This short paper contains two main sections. In the first, we explain classification and the use of classifiers. In the second, we more closely investigate one type of classifier that is based on probabilistic models. At the end of that section we can pinpoint our main contribution to the field.

II. Classification

Classifying is the act of attaching a class c to an object described by a vector of attributes a. Consider, as an illustrative running example, the classification of animals. Given an object with attributes such as a number of legs, hairiness, and weight, we want to attach a class, i.e., say if it is a human, a dog, an octopus, etc.

We are interested in so-called *supervised classification*, where the classes c are pre-defined and form a finite set $\mathscr{C}$. The set of attributes is denoted $\mathscr{A}$. These attributes can range over finite or continuous sets. For example,

$$\mathscr{C} = \{\text{“bee”}, \text{“dog”}, \text{“human”}, \text{“octopus”}\},$$
$$\mathscr{A} = \mathscr{A}_{\text{legs}} \times \mathscr{A}_{\text{hair}} \times \mathscr{A}_{\text{weight}},$$

where

$$\mathscr{A}_{\text{legs}} = \{0, 4, 6, \geq 8\},$$
$$\mathscr{A}_{\text{hair}} = \{\text{none}, \text{light}, \text{heavy}, \text{complete}\},$$
$$\mathscr{A}_{\text{weight}} = \mathbb{R}^+ \quad \text{(in kg)}.$$

We all classify (possibly mental) objects in our daily lives. Sometimes this task becomes tedious and then a classifier, which automates it, is useful. A *classifier* is a function that maps attributes to classes. It is defined using a *training set* of manually pre-classified objects, i.e., couples (c, a). Some examples of pre-classified animals could be

$$\text{“John Normstudent”} \rightarrow (\text{“human”}, (4, \text{heavy}, 75.3)),$$
$$\text{“Lassie”} \rightarrow (\text{“dog”}, (4, \text{complete}, 28.1)).$$

E. Quaeghebeur is a member of the SYSTeMS research group of the department of Electrical Energy, Systems and Automation, Ghent University (UGent), Belgium.

E-mail: Erik.Quaeghebeur@UGent.be.

To build a very simple example classifier we could, for example, first define a norm (distance) on the attribute set $\mathscr{A}$, then locate the center of mass of the learning objects belonging to a certain class, and finally attach the class of the center of mass nearest to the object to classify. A wide range of classifiers exist; our field of interest lies with classifiers that use probabilistic models for the classes and the attributes to give ‘the most likely class’ for an object. These probabilistic models are of course based on the learning set.

Before looking at these probabilistic models in the next section, we comment on two important aspects of building a classifier. *Prior information*, what we know or assume about the classification problem at hand, such as the structure of the attribute set, has a big influence on the design of the classifier. In our simple example classifier, it determines the definition of the norm on the attribute set. The *size of the training set* is an aspect that will greatly influence the output of the classifier. To give an extreme example: if the two pre-classified animals given above form the training set for our simple example classifier, every other animal would be classified as either “dog” or “human”.

Small learning sets also pose problems in less extreme cases. Coping with these problems starts by making a good use of prior information. Zaffalon [1] has shown that using a design that allows the classifier to attach a set of classes to an object (and not only a single class), gives an important conceptual improvement. It is then called a *credal classifier*. Modifying our extreme example, this would mean that “Ella Smallchild”, with attributes $(4, \text{light}, 23.4)$, could get {“dog”, “human”} attached.

III. Probabilistic models

With our simple example classifier, we decided on the class to attach by making a comparison of the distances between the attribute vector a of an object to classify and the different centers of mass corresponding the classes c. Similarly, when using probabilistic models to construct a classifier, this decision will be based on *pairwise comparisons* between classes c' and c''.

To make these pairwise comparisons, we need some building blocks. First of all, we need a function that encodes the utility of attaching a certain class c' to an object with class c. A simple choice for such a *utility function* would be the indicator function $I_{c'}$, for which $I_{c'}(c)$ is 1 if $c = c'$ and 0 otherwise. Using our running example, this means that attaching the class “octopus” to some bee is considered completely useless, though not harmful ($I_{\text{“octopus”}}(\text{“bee”}) = 0$). On the other hand, attaching “bee” would be considered useful ($I_{\text{“bee”}}(\text{“bee”}) = 1$).

The difference $I_{c'} - I_{c''}$ of two utility functions can then be used to encode the utility of switching from attaching one class c'' to attaching another class c'. So if, for a given attribute vector a, it is to be expected, based on the prior information and the data in the training set, that this difference is positive, then c' is the better choice.

Using this type of pairwise comparison, we can then create an order for the classes. The maximal elements of this order will form the set we attach to the object described by a. The possibility that classes are *incomparable* is an important property of this order. So it might happen that "dog" is not better than "human" and vice-versa. If both are better in comparison to other animals, they will form the set of maximal elements.

Previously, we did not mention how we can determine 'what can be expected, given an attribute vector a'. For this, we need a second building block: a *conditional lower expectation* (operator) $\underline{P}(\cdot \mid \mathscr{A})$ on the set $\mathscr{L}(\mathscr{C})$ of bounded real-valued functions on $\mathscr{C}$.

Let us explain what this operator is, starting from the usual expectation (operator) P on $\mathscr{L}(\mathscr{C})$. A utility function is a typical element of $\mathscr{L}(\mathscr{C})$ and so $P(I_{c'})$ is the expected utility of attaching a class c' to an object, based on the prior information and the learning set, but not on any information about the object's attributes.

A conditional expectation $P(\cdot \mid \mathscr{A})$ returns more detailed information: $P(I_{c'} \mid a)$ gives the expected utility of attaching a class c', when we additionally know the object's attribute vector is a. The following conditional expectations would be reasonable examples:

$$P(I_{\text{“octopus”}} \mid (\geq 8, \text{none}, 3.1)) = 0.9,$$
$$P(I_{\text{“octopus”}} \mid (6, \text{heavy}, 35.2\ 10^{-6})) = 0.1,$$
$$P(I_{\text{“bee”}} \mid (6, \text{heavy}, 35.2\ 10^{-6})) = 0.8.$$

Quite often the prior information and the information in the (small) learning set are not enough to determine a unique conditional expectation operator, but a whole set $\mathscr{M}$ of them is compatible with the available information. This so-called *credal set* can be equivalently described by its lower envelope, a conditional lower expectation:

$$\underline{P}(\cdot \mid \mathscr{A}) = \inf\{P(\cdot \mid \mathscr{A}) \in \mathscr{M}\}.$$

Set $a_{\text{“Samson”}} = (4, \text{complete}, 10.5)$. Now if, for example, both $P_1(I_{\text{“dog”}} \mid a_{\text{“Samson”}}) = 0.6$ and $P_2(I_{\text{“dog”}} \mid a_{\text{“Samson”}}) = 0.7$ are compatible with the available information, then

$$\underline{P}(I_{\text{“dog”}} \mid a_{\text{“Samson”}}) = \min_{i \in \{1,2\}} P_i(I_{\text{“dog”}} \mid a_{\text{“Samson”}}) = 0.6.$$

(More information about *imprecise probabilities*, of which lower expectations are an example and our tool for probabilistic modelling, can be found in Walley's excellent book [2].)

Now we can finally give the formal criterion to be used for the pairwise comparison between classes: If, for an object with attribute vector a, the lower expected utility $\underline{P}(I_{c'} - I_{c''} \mid a)$ of switching from class c'' to class c' is strictly positive, then c' should be chosen over c''.

What remains to be done, is show how $\underline{P}(\cdot \mid \mathscr{A})$ can be built. The learning set is seen as a random sample of the set of all objects to classify. Therefore, it contains information on how likely it is to encounter a certain class and how likely it is to get a certain set of attributes for a given class. Note that we don't directly have access to information in the form that we need, i.e., how likely a certain class is for a given attribute vector.

We can, however, get this information indirectly. In a first step, we use prior information and the information in the learning set on how likely a certain class is, to build a lower expectation $\underline{P}$ on $\mathscr{L}(\mathscr{C})$. We call this the *class model*. In our animal example, it would contain information about the relative number of the different animals.

A second, similar step consists of using prior information and the information in the learning set on how likely it is to get a certain set of attributes for a given class, to create a conditional lower expectation $\underline{P}(\cdot \mid \mathscr{C})$ on $\mathscr{L}(\mathscr{A})$. We call this the *attribute model*. This model would, in our example, contain information about the number of legs, etc. for each of the animals.

Out of the attribute and class models we create, in a third step, a joint lower expectation on $\mathscr{L}(\mathscr{C} \times \mathscr{A})$. Finally, conditioning of this joint lower expectation on the attributes gives us the desired lower expectation $\underline{P}(\cdot \mid \mathscr{A})$ on $\mathscr{L}(\mathscr{C})$.

Zaffalon [1] introduced a credal classifier, based on the probabilistic approach described above, for classification problems where all the attributes belong to a finite set (i.e., are distributed according to a multinomial distribution). Our main contribution to this field [3] was to show that this approach can be generalized to problems where the attributes are distributed according to an exponential family distribution (normal distribution, multinomial distribution, Poisson distribution,...). For this, we designed a specific class of attribute models that are especially well suited for holding information about such attributes.

IV. Conclusions

In this short paper we have first introduced classifiers and there pointed out two aspects that are important for their construction, prior information and the size of the learning set. This last aspect led to the introduction of credal classifiers.

Then we took a look at classifiers based on probabilistic models. We showed their building blocks and how they are put together. One of these building blocks is where our main contribution to the field lies: we propose a set of imprecise probabilistic models meant to allow credal classifiers to be built for classification problems with attributes that are distributed according to an exponential family.

Acknowledgments

This paper presents research results of the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The scientific responsibility rests with its author.

Erik Quaeghebeur's research is financed by a Ph.D. grant of the Institute for the Promotion of Innovation through science and technology in Flanders (IWT-Vlaanderen).

References

[1] Marco Zaffalon, "Statistical inference of the naive credal classifier," in *ISIPTA '01: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, G. De Cooman, T. L. Fine, and T. Seidenfeld, Eds., Ithaca, New York, United States, 2001, pp. 384–393.

[2] Peter Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London, 1991.

[3] Erik Quaeghebeur and Gert De Cooman, "Imprecise probability models for inference in exponential families," in *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, F. G. Cozman, R. Nau, and T. Seidenfeld, Eds., Pittsburgh, Pennsylvania, United States, 2005, pp. 287–296.