Imprecision in Machine Learning and AI

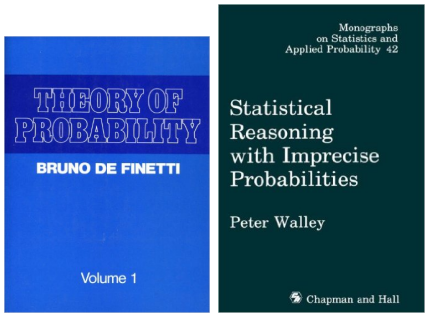
Cassio P. de Campos and Alessandro Antonucci

I. Motivating Imprecise Probabilities

In this note we consider five different relevant problems in AI and machine learning. We argue that possible solutions to such problems might be achieved by replacing the probability distributions in the systems with sets of them. Such a robust approach is based on the so-called imprecise-probabilistic framework. The proposed solutions provide a persuasive justification of the imprecise framework. The problems we consider are:

- proper treatment of missing data,
- reliable classification,
- sensitivity analysis,
- feature selection,
- elicitation of qualitative expert knowledge.

Before reporting a separate discussion for each problem, let us briefly resume the general ideas characterising imprecise-probabilistic methods.



II. Beyond classical probability

Standard approaches to uncertainty modelling assume that the lack of knowledge about the actual state of a quantity is described by probabilities over its possible states (or by densities when coping with continuous variables). Following a subjective (also called *epistemic*) interpretation, these numbers can be regarded as relative strengths (e.g., measured in behavioural terms) for the beliefs that the quantity is in a particular state. Those probabilities might be elicited from expert knowledge or summarise the result of a statistical processing of historical data. Sharp (or, say, *precise*) values are typically used to quantify these probabilities. In many cases there are not compelling reasons for that and a set-valued specification might offer a better, or at least more cautious, description. The seminal work of Peter Walley in line with de Finetti's theory of subjective probability has formalised such a possibility, which can be addressed by replacing standard, precise, distributions with sets of them (e.g., see the figure here below).

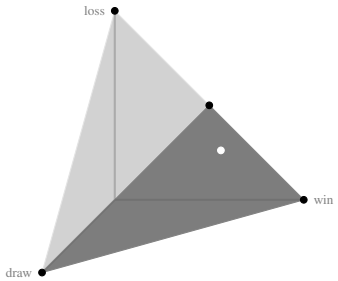


Fig. 1. Geometric view of probabilities for ternary quantities. The possible outcomes of the result of football match is win, draw, or loss. The white point is a precise probability distribution modelling the fact that win is six times more probable than loss. The dark gray area contains all the probability distributions consistent with the fact that win is more probable than draw.

III. Treatment of missing data

Consider a simple medical example. A patient presents symptoms that could be related to lung cancer. A physician can run tests for bronchitis and do X-rays, as well as check for dyspnea. However, (supposedly) he/she can only assess whether the patient is a smoker by asking the patients themselves. Some patients did not answer whether or not they are smokers in the questionnare. As an additional (somehow hidden) information consider that patients have a discount in their insurance because they declared not to be a smoker to the insurance company. Should smoking be ignored? Should it be marginalized out? Should it be treated with (greater) care?

Ignoring missing data is a common practice. Yet, it can be only justified under specific assumptions about the process making the output of an observation/measurement missing. Those assumptions reflect the lack of a selective mechanism taking into account the actual value of the observed quantity. This is clearly not the case in the above medical example: the answer about the smoking habits of the patient is more likely to be missing for smokers (see the table here below). On the other hand, a statistical modelling of the process making the data missing can be hard to assess because the lack (by definition) of complete data about that. The most conservative approach consists therefore in considering all the possible completions of the missing data and learning a different model from each one. This corresponds to an imprecise probabilistic approach, in which a *vacuous* set (i.e., the set of all the possible distributions modelling a condition of near ignorance) is used to describe the incompleteness process. Although possibly leading to less informative results, this approach should be regarded as the most reliable approach to the conservative treatment of missing data.

Alessandro Antonucci is a Senior Researcher at Dalle Molle Institute for AI (IDSIA), Manno-Lugano, Switzerland. He also teaches at the University of Applied Sciences and Arts of Southern Switzerland (SUPSI).
e-mail: alessandro@idsia.ch website: www.idsia.ch
Cassio P. de Campos is a Reader with the Knowledge and Data Engineering Cluster of the Queens University Belfast, Belfast, Northern Ireland, UK.
e-mail: c.decampos@qub.ac.uk website:www.qub.ac.uk/eeecs

PATIENT	ANSWER	TRUTH
1	*smoker*	*smoker*
2	*smoker*	*smoker*
3	*smoker*	*smoker*
4	*smoker*	*smoker*
5	*non-smoker*	*non-smoker*
6	*non-smoker*	*non-smoker*
7	*unanswered*	*smoker*
8	*unanswered*	*smoker*
9	*unanswered*	*smoker*
10	*unanswered*	*non-smoker*

Fig. 2. Results of a questionnaire about the smoking habits of ten patients. The real ratio of smokers is 70%, ignoring the missing answers would give 67%, while a conservative treatment considering all the completions gives a 40-80% range.

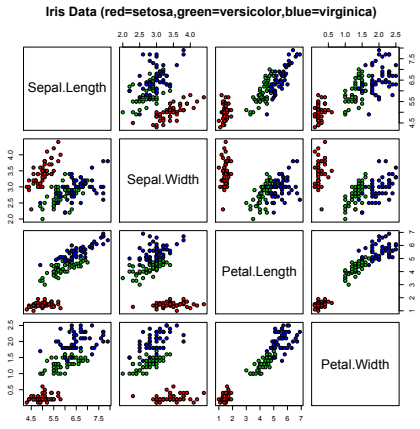


Fig. 3. Two dimensional representation of the Iris dataset with the four features and three classes.

IV. RELIABLE CLASSIFICATION

Consider a standard classification setup with a collection of objects containing some defining features that can possibly be used to identify them. The objects can be categorized into classes, while the class of an object might be unknwown to us. Given a collection of objects of known classes, the problem is to build a model that can guess the class of an object of unknown class. To demonstrate this machine learning task we consider the naive Bayes classifer on the Iris dataset, where the species of an iris flower (setosa, versicolor or virginica) is described by four features: sepal and petal width and length. Can we improve classification accuracy by using a imprecise-probabilistic model, for instance by providing a subset of the classes that certainly contains the correct one? Can we identify hard- and easy-to-classify instances? We have used an imprecise-probabilistic naive Bayes classifier to process the Iris dataset and compared the results with the standard naive Bayes classifier. In our separation of train and test instances, the standard classifier obtains 72% of accuracy in predicting the flower class. The imprecise-probabilistic classifier may return a set of classes for each test instance instead of a single answer. Its set accuracy (whether the true class is within the returned classes) reaches 100% and the accuracy of the standard classifier drops to 60% when only considering the instances where the imprecise-probabilistic classifier has returned more than a single class. Hence, the imprecise-probabilistic classifier is able to identify the hard-to-classify instances from the dataset.

V. SENSITIVITY ANALYSIS

Probabilistic graphical models such as Markov Random Fields are popular tools in AI. Suppose that using a Markov Random Field, we have reached a conclusion about the most probable explanation for the variables in a domain, that is, we have computed the mode of the underlying joint probability distribution. Is this conclusion sensitive to modifications of the model, that is, would the mode be different under some small change in the model's parameters? The most common procedure is to apply local modifications to the model and to check whether the conclusion remains inaltered. An imprecise-probabilistic network, or simply *credal* network, can be efficiently used to verify whether the mode is unique for every joint distribution that is encoded by the network. The result is declared reliable if that is the case. By building an ε-box around the original MRF model (each parameter of the original model is allowed to vary inside such boxes), we obtain a credal network. We increase the value of ε until the limiting moment where all distributions still yield the same mode. Such limiting value of ε is regarded as the robustness of the decision.

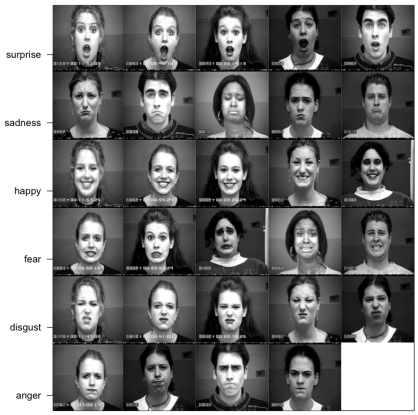


Fig. 4. Examples of posed faces from the Cohn-Kanade dataset.

We have applied the robustness analysis to the problem of detecting facial action units in posed images using the Cohn-Kanade dataset. For each test case, we have used 23 binary variables corresponding to facial action units, which need to be explained (computation of the mode given image observations). The Hamming distance between predicted and true values gives the accuracy of our model for a given test case, and ε is computed as well (as described above). We have found an association between ε and the accuracy of the predictions, as shown in Figure 5 [1].

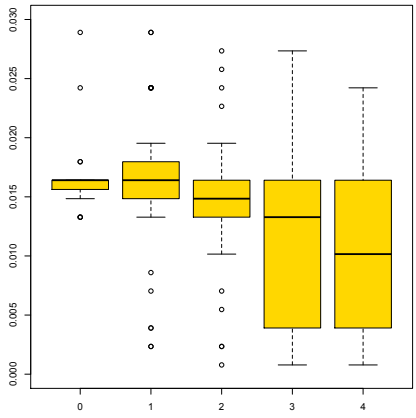


Fig. 5. Relation of Hamming distance (x axis) and robustness ε (y axis) for the test cases from the Cohn-Kanade dataset. Accuracy decreases with the lack of robustness.

VI. Feature selection

We are given a (potentially large) number of covariates and want to identify those which are useful to predict a binary response. An usual procedure is to employ some statistical tests. An example is the Mann-Whitney u-test (aka Wilcoxon rank-sum test) to test whether the probability of a quantity from individuals of one group being greater than that of the other group is greater than half.

Consider the Australian AIDS dataset, where analyses suggested that a difference in survival time existed when discriminating individuals with AIDS by the use (or not) of drugs (for whom a different survival was arguably expected), but also suggested that individuals with AIDS from the Queensland region in Australia have significantly worse survival time than those from the New South Wales region.

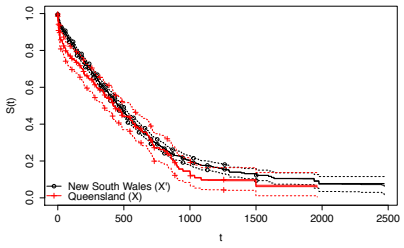


Fig. 6. Survival curves for individuals of two different Australian regions. Standard tests identify a significant difference in the curves, while the imprecise-probabilistic method correctly deems such result as unreliable. x axis represents time in days and y is the survival curve.

Even if the latter conclusion has been questioned in the original studies because such difference was at first not expected, no formal analysis was used to assess the reliability of the result. Using a robust version of the u-test tailored for survival analysis, we have identified such doubtful situation through the use of the imprecise-probabilistic version of the test, which responds an indeterminate outcome in that case, suggesting that further data should be collected for a better decision. Other comparisons which were deemed correct (regarding *drug usage*, *blood* and *haemophilia*) are confirmed by the imprecise-probabilistic method as reliable [2].

VII. Qualitative assessments

Let us go back to the medical example. Assume that you adopt a Bayesian network to implement a knowledge-based expert systems over the relevant quantities (see graph here below). The quantification of the network requires the assessment of the probability for conditional states of each quantity given any possible configuration of the direct predecessors. For lung cancer, we should decide the probability of being sick for patients who are smokers and for those who are not. In a simulated scenario, assume the physician is only able to report the following qualitative statement: *smokers are more likely (than non-smokers) to have lung cancer*. How do we translate such a qualitative statement with sharp probabilistic values?

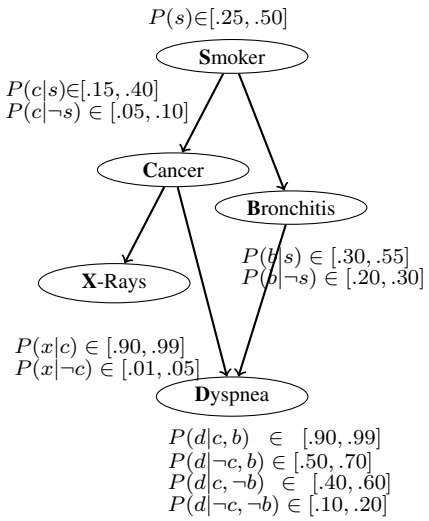


Fig. 7. A simplified version of the Asia diagnostic network (originally proposed as a Bayesian network) quantified by intervals.

Linear constraints over probabilities offers a natural way to express qualitative judgements. This is straightforward for the the above considered comparative judgement, being $P(c|s) \geq P(c|\neg s)$. Verbal-numerical scales can be used to describe any qualitative expert judgements in a similar way. E.g., the judgement *patients with lung cancer are very likely to display positive X-rays* with $.90 \leq P(x|c) \leq .99$. With such a quantification the original Bayesian network becomes an imprecise-probabilistic graphical model called *credal network*, for which a huge number of inference algorithms have been developedv [3].

VIII. Conclusions

We advocated the use of imprecise probability in AI and machine learning. Replacing single probability distributions with sets of them increases realism in the modelling phase, thus leading to more cautious and reliable inferences. These approaches appear especially suited to describe non-ignorable missingness processes, evaluate classifiers reliability and robustness of inferences in graphical models, evaluate relevance of covariates, and properly elicit expert knowledge. Lots of further developments are possible. E.g., a proper description of non-stationarity in dynamic systems [4].

REFERENCES

[1] http://papers.nips.cc/paper/5472-global-sensitivity-analysis-for-map-inference-in-graphical-models
[2] http://dx.doi.org/10.1002/bimj.201500062
[3] http://dx.doi.org/10.1002/9781118763117.ch9
[4] http://dx.doi.org/10.1016/j.neucom.2015.08.095