# Chapter 1

# Introduction

> *In every phenomenon the beginning remains always the most notable moment.*
>
> Thomas Carlyle

In colloquial terms, learning to classify consists in analyzing a set of objects with different characteristics and of different classes and after that being able to assign a possibly unseen object to one of the classes. The field of Machine Learning has been interested in designing automatic classification algorithms since its inception. The areas of application of such algorithms are immense, ranging from historic Artificial Intelligence (AI in the following) objectives such as “begin able to adapt the behaviour of machines by telling them what is good and what is bad” to far more recent and business oriented objectives such as “targeting direct marketing campaigns”.

Probability theory has always been perceived as a relevant discipline to help in the quest of solving the classification problem and for AI in general. Furthermore, the development in the late eighties and early nineties in the field of Bayesian networks, together with an increased interest from the community, have further increased this belief. In this thesis we show several ways in which the application of probability theory can improve Bayesian network classifiers.

We start this chapter by introducing the objectives of the thesis in section 1.1. After that, we provide a roadmap for the reader in section 1.2 pointing out the key contributions of the thesis. Finally, in section 1.3 we outline a rationale for the contributions.

## 1.1 Objective

This thesis focuses in improving a family of classification algorithms known as *Bayesian network classifiers*. The objective of the thesis is improving Bayesian network classifiers by means of the application of objective Bayesian probability theory techniques.

## 1.2 Roadmap and contributions

In this section we provide an overview of the structure of the thesis which is depicted in Figure 1.1. In that figure, chapters with original contributions are dark gray coloured while introductory and review chapters appear coloured light gray.

The thesis starts with this introductory chapter, where the reader can find the objectives, structure and a rationale for the contributions.

Chapter 2 presents a short introduction to the foundations of probability theory summarizing the results in the first two chapters of “Probability Theory: The logic of science” by E.T. Jaynes. The reason for including this short summary is twofold: on one hand it will help a reader without knowledge about Bayesian statistics understand the philosophy behind the thesis, on the other hand, it gives a deserved visibility to these results. Also in chapter 2 the principle of indifference is introduced. The principle will be used in the contributions of chapters 6 and 8.

In order to ease understanding of the contributions of the thesis, chapter 3 contains an introduction to the problem of classification, a short state of the art of Bayesian network classifiers and a presentation of Bayesian model averaging as a probability technique that is useful for the design of classifiers. The notation and terminology to be used in the thesis are fixed in this chapter. Also in this chapter, the two main classification algorithms that have been improved are introduced: Naive Bayes and Tree Augmented Naive Bayes (TAN).

Discretization techniques help broaden the application of classifiers that cannot deal easily with numerical attributes. In chapter 4 we present a discretization method that can be implemented in parallel providing for a fast and effective method for the discretization of numerical attributes and prove its performance against state of the art discretization methods. The work has been presented at the Third International Conference on Knowledge Discovery and Data Mining.

Chapters 5 and 6 will present two different ways of improving the Naive Bayes classifier. In chapter 5 a variation of the naive Bayes classifier is introduced which is easier to interpret while keeping a reasonable accuracy. This work has been presented at the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery. In chapter 6 Bayesian model averaging and the principle of indifference are applied in order to construct a more accurate Naive Bayes classifier. This work has been presented at The 16th International FLAIRS Conference.

Chapters 7, 8 and 9 show different ways in which Bayesian techniques can

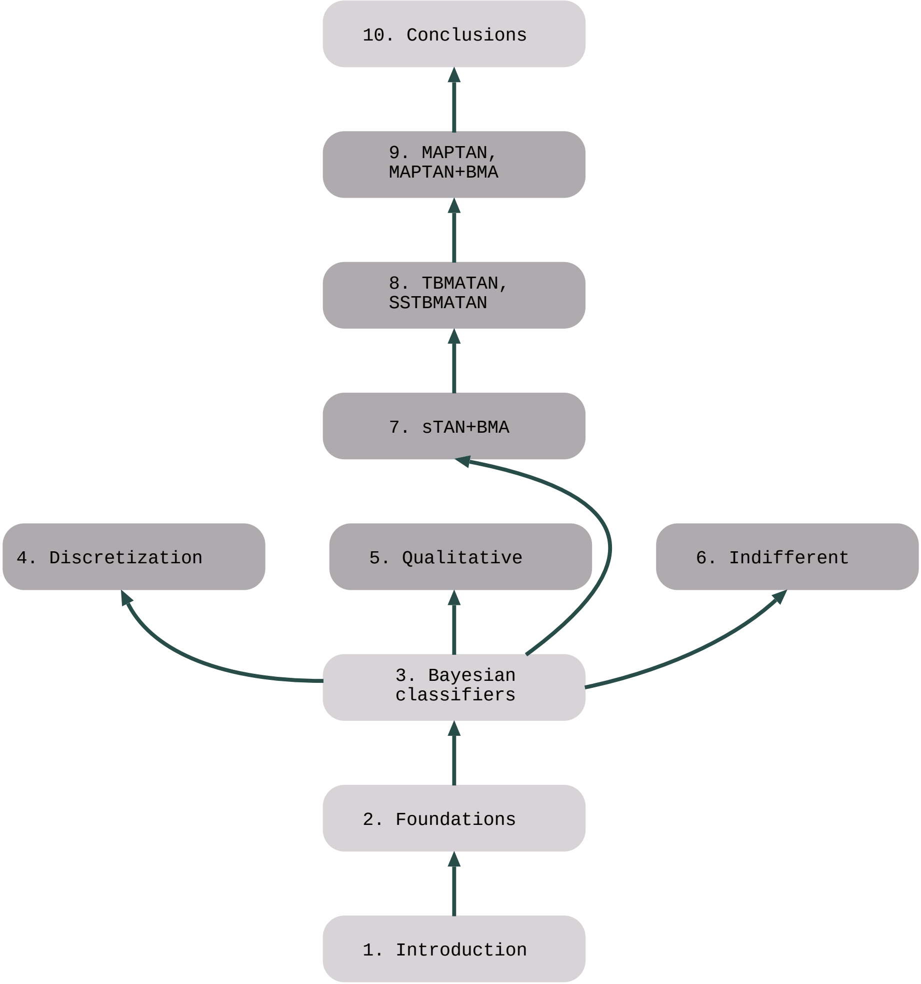

Figure 1.1: Thesis roadmap

improve Tree Augmented Naive Bayes. Chapter 7 shows a hands-on approach to the calculation of the model averaging of TAN by the use of empirical local Bayesian model averaging that results in a classifier (STAN+BMA) that improves both the classification accuracy and the approximation of the class probabilities of the state of the art TAN classifier (Friedman et al., 1997): STAN. This work has been presented at the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chapter 8 presents the most significant result in this thesis, showing that under some conditions, the averaging of models for Tree Augmented Naive Bayes results in an integral that can be calculated in closed form. To do this we introduce decomposable distributions over TANs and show that the expression resulting from the Bayesian model averaging of TAN models can be integrated into closed form if we assume the prior probability distribution to be a decomposable distribution. This result allows the construction of a classifier (TBMATAN) that is most of the cases more accurate than STAN and approximates better the class probabilities. TBMATAN learning time can be very long for large datasets due to computational problems. To fix this problem we introduce an approximation to TBMATAN (SSTBMATAN) which is more efficient. SSTBMATAN has a shorter learning time a longer classification time than STAN. SSTBMATAN is most of the cases more accurate than TBMATAN and approximates better the class probabilities. This work has been presented at the Twentieth International Conference on Machine Learning. Finally, in chapter 9, we show that it is possible to calculate efficiently both the TAN model with maximum a posteriori probability, and the set of $k$ TAN models with maximum a posteriori probability and their relative probability weights. We show that these results allow the construction of two classifiers (MAPTAN and MAPTAN+BMA) which outperform STAN and STAN+BMA respectively in error rate and quality of the predicted probabilities. Furthermore, MAPTAN+BMA learning time complexity is lower than STAN+BMA. In the three chapters, experimental results are given that allow the reader to evaluate the improvements.

## 1.3 Rationale for the contributions

In this section we propose a simplified comparison of classifiers from the perspective of a user and motivate the contributions of this thesis in this simplified framework. The purpose of the section is easing understanding of what the thesis proposes, why we think it can be useful and what have been the main characteristics we have taken into consideration while trying to improve Bayesian classifiers. The reader should not take this framework as a proposal for future use but as a simple tool for understanding what has been done.

### 1.3.1 Characteristics of a classifier

Classifiers can be compared by different characteristics. In order to ease the understanding of the contributions of the thesis from the point of view of a user of machine learning or data mining algorithms, at the end of some chapters the

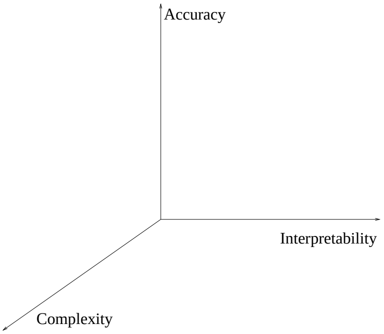

Figure 1.2: Three axis in which a classifier can be placed

different classifiers introduced are plot in a three dimensional space where the three axis are determined by accuracy, interpretability and time complexity (see Figure 1.2). Evidently, these are not the only characteristics we could define, but in our opinion these are the most significant when evaluating a classifier. For a more extended list of classifier characteristics from the perspective of a user, the reader can refer to section 10.7 of (Hastie et al., 2001).

Accuracy can be defined as how well a classifier performs the task that is its main objective, that is, classifying. There are different measures of classification accuracy. The error rate computes the percentage of instances that are misclassified. Some more sophisticated measures do also take into account the probability that the classifier has assigned to the class. This way, those measures punish much more misclassifications where the probability of the class was very high than those were the classifier itself knew that it was confused and assigned a not so high probability to the most probable class. In the last few years, the evaluation of accuracy of classifiers based on the Receiver Operating Characteristics is gaining momentum (Provost et al., 1998; Fawcett, 2003). Independently of how it is measured, increasing classification accuracy is, obviously, an objective when proposing new classifiers.

Interpretability stands for how easily humans can understand the decisions taken by a classifier. Many times, machine learning algorithms, and also classifiers, are used as supporting devices for a human being who is in charge of taking a decision in a concrete situation. In these cases, being able to interpret, understand and explain to interested third parties the rationale behind the decision will influence heavily whether we can use a classifier or not. Increasing the interpretability of classifiers is hence another objective when proposing new classifiers.

Time complexity is the amount of time needed for a classifier to perform its task. Given a dataset, classifiers usually split their work in two steps, a learning step where the dataset is processed and the information from the dataset summarized in some way, and a classification step where new unclassified instances are observed and the classifier guesses the class in which they should be classified. This split generates two different complexity measures for a classifier. We will call learning time complexity the time spent on the first task, and classification time complexity the time spent to classify a new instance once the learning step is over. Decreasing both learning time complexity and classification time complexity is yet another objective when proposing new classifiers.

### 1.3.2 Why do we need new classifiers?

If we accept the simplified view proposed by this three axis, when proposing a new classifier we can be in one of two cases. In the first case, we can move in a positive direction in one or more of the axis while keeping the others fixed. This is the case, for example, when the accuracy of a classifier is increased without modifying its interpretability or time complexity, which is the case for the improvement of Naive Bayes presented provided by INDIFFERENTNB (see chapter 6) as well as for the improvement for STAN provided by MAPTAN (see chapter 9). This kind of "moves" are clear improvements of previous results and are usually expected to substitute the previous version of the classifier. In the second case, we can move in a positive direction in some axis while losing in some other. This is the case, for example, when we create a classifier with a better accuracy but at the cost of increasing the time complexity and decreasing the interpretability, as is the case for the improvements for STAN suggested in chapters 7 and 8. This is also the case when we create a classifier with increased interpretability but higher time complexity, as is the case for the improvement presented for Naive Bayes in chapter 5. This kind of moves provide an increase in the "palette" of tools for the classification user. In the face of a new problem, the user can evaluate which are the constraints in that concrete case for accuracy, interpretability and complexity and choose the classifier best suited for the task. For example, imagine that our task is to create a system to be used for disease self-treatment and that the system has to suggest the user the correct treatment for his disease. Assuming the user is not knowledgeable in the area, he will have to rely on the system decision without trying to interpret the decision, because he should be knowledgeable in medicine to do that. He will not be willing to wait for days to get an answer, but he can wait for one or two hours. Accuracy is probably the most important characteristic for a classifier for this task. If we slightly change the task and assume that the classifier will be used by a doctor in his office, this heavily increases the interpretability and time complexity needs (he will definitely not be willing to make the patient wait for the computer to answer back) and lowers the accuracy needs, because probably the doctor will use the system as a double check for his own decisions and in case of disagreeing he will be able to explore the patient more carefully and generate the adequate treatment.

Due to the increasing computational power, that allows higher complexity calculations, increasing accuracy is specially significant. That is why we consider specially valuable the results in chapter 8, which provide, to the best of our knowledge, the most accurate classifier with a learning time complexity linear on the number of observations of the dataset.

We understand that the new classifiers proposed in this thesis provide reasonable trade-offs and cover areas of the described three dimensional space that where uncovered before. Hence we think that they deserve to be selected by any user for some domains and that they could end up being a valuable tool in an users "palette".