Chapter 19
Learning Verifiable Ensembles for Classification Problems with High Safety Requirements

Sebastian Nusser
Otto-von-Guericke-University, Germany

Clemens Otte
Siemens AG, Germany

Werner Hauptmann
Siemens AG, Germany

Rudolf Kruse
Otto-von-Guericke-University, Germany

ABSTRACT

This chapter describes a machine learning approach for classification problems in safety-related domains. The proposed method is based on ensembles of low-dimensional submodels. The usage of low-dimensional submodels enables the domain experts to understand the mechanisms of the learned solution. Due to the limited dimensionality of the submodels each individual model can be visualized and can thus be interpreted and validated according to the domain knowledge. The ensemble of all submodels overcomes the limited predictive performance of each single submodel while the overall solution remains interpretable and verifiable. By different examples from real-world applications the authors will show that their classification approach is applicable to a wide range of classification problems in the field of safety-related applications - ranging from decision support systems over plant monitoring and diagnosis systems to control tasks with very high safety requirements.

INTRODUCTION

Machine learning methods are successfully applied in a wide range of applications – for instance – object recognition in computer vision, search engines, or stock market analysis. But in the field of safety-related applications such methods are regarded with suspicion by the domain experts because the learned models are often hard to verify, may tend to overfitting, and the exact inter- and extrapolation behavior is often unclear. In this chapter, a machine learning method is proposed that (1) is capable of

DOI: 10.4018/978-1-61520-757-2.ch019

tackling classification problems in application domains with high safety requirements and (2) satisfies the domain experts' demand for verification of the correctness of the learned solution for the given application problem.

A safety-related system is a system whose malfunction or failure can lead to serious consequences – for instance environmental harm, loss or severe damage of equipment, harm or serious injury of people, or even death. Examples of safety-related application domains are: aerospace engineering, automotive industry, medical systems, and process automation. The increasing complexity of such safety-related systems and the growth of the number of requirements and customer requests raise the interest in applying machine learning methods within this domain. For instance, the domain knowledge is often imperfect and, thus, purely analytical solutions cannot be provided by the domain experts. In addition, the data-driven generation of classification models offers a reduction of development time and costs. The classification performance can be improved by advanced classification models. Unfortunately, in the field of safety-related application domains, it is often not possible to rectify a wrong decision. For effectively applying data-driven classification methods within this domain, it is crucial to provide strong evidence that the learned solution is valid within the complete input space and correctly satisfies all given functional specifications. It must be guaranteed that the interpolation and extrapolation behavior of the solution is always correct. Therefore, it is imperative to provide an interpretable solution that can be validated according to the given domain knowledge. Figure 1 illustrates the issue of an unexpected change of a classifier's decision within a region of the input space where no data is available. Such unintended behavior can be regularly discovered by visualization for the two-dimensional case – but for a high-dimensional model this is infeasible. Often, statistical risk estimation methods are not satisfactorily in real-world applications because the observed data is scarce and, therefore, vast regions of the high-dimensional input space do not contain any data.

This contribution is motivated by a real-world application within the field of automotive safety electronics. It concerns the deployment of restraint systems (for instance belt pretensioners and airbags) and serves as an example for control systems with high safety requirements. The malfunction might be fatal and it is impossible to rectify a wrong deployment decision since an airbag system can only be triggered once. Depending on the severity of a crash different restraint systems must be triggered: for instance, the belt pretensioners, the front airbag stage 1 (airbag is inflated to 70%) or stage 2 (airbag is inflated to 100%), the knee airbags, the side airbags (front or rear), or the curtain airbags. Furthermore, the airbag must be triggered within a certain time interval in order to ensure the best passenger protection – a front crash, for instance, must be triggered within 15 ms to 30 ms. The postponed deployment of an airbag can lead to severe injuries of the car occupants.

For each new car platform the control logic of the restraint systems has to be developed nearly from scratch since altered mechanical components, different sensor placements, or new functional requirements of the car platform can dramatically influence the signal characteristics and, for this reason, a solution of a previous platform will not be applicable anymore. Until now, most of the calibration work is done manually by domain experts. For each crash type many different sensor combinations are evaluated and a control logic based on those combinations is developed. This manual calibration is time and cost intensive. Due to cost pressure in the market, the increasing complexity of today's restraint systems must be handled with limited resources. Thus, there is a growing interest in automatically learning the control logic from crash test data in order to reduce