Soft Computing (2022) 26:7175–7193
https://doi.org/10.1007/s00500-022-07194-6

DATA ANALYTICS AND MACHINE LEARNING

Estimation of fault probability in medium voltage feeders through calibration techniques in classification models

Enrico De Santis[1] · Francesco Arnò[1] · Antonello Rizzi[1]

Accepted: 24 April 2022 / Published online: 23 June 2022

Abstract
Machine Learning is currently a well-suited approach widely adopted for solving data-driven problems in predictive maintenance. Data-driven approaches can be used as the main building block in risk-based assessment and analysis tools for Transmission and Distribution System Operators in modern Smart Grids. For this purpose, a suitable Decision Support System should be able of providing not only early warnings, such as the detection of faults in real time, but even an accurate probability estimate of outages and failures. In other words, the performance of classification systems, at least in these cases, needs to be assessed even in terms of reliable outputting posterior probabilities, a really important feature that, in general, classifiers very often do not offer. In this paper are compared several state-of-the-art calibration techniques along with a set of simple new proposed techniques, with the aim of calibrating fuzzy scoring values of a custom-made evolutionary-cluster-based hybrid classifier trained on a set of a real-world dataset of faults collected within the power grid that feeds the city of Rome, Italy. Comparison results show that in real-world cases calibration techniques need to be assessed carefully depending on the scores distribution and the proposed techniques are a valid alternative to the ones existing in the technical literature in terms of calibration performance, computational efficiency and flexibility.

Keywords Calibration · Fault recognition · Smart grids · Probability estimation · Evolutionary optimization · Clustering

1 Introduction

Predicting and modeling outages and failures in electrical power grids is of paramount importance, either from the power grid operator and consumer viewpoints. Small- and large-scale faults and disturbances in the grid often cause power outages and thereby affect the system reliability and customer satisfaction, reason why it would be a great achievement for the electricity operator to be aware of which elements of the grid have a high risk of failure in order to deploy a preventive maintenance system.

The conditions related to the physical grid and the environment in which it operates can be detected in real time by placing smart sensors throughout the power grid. Through reliable telecommunication networks, these heterogeneous data can be sent to powerful data-centers for collecting and processing purposes (De Santis et al. 2018b). Modern data-driven techniques, within the Artificial Intelligence field, can exploit this amount of data in order to "x-ray scan" the power grid states, driving a lot of interesting pattern recognition and data science applications. In this context, a very interesting task consists in modeling and recognizing faults in the power grid in order to design a Decision Support System (DSS) that provides decision support for the commanding and dispatching system, for Condition-Based Maintenance (CBM) programs (Raheja et al. 2006) and for providing high-level information to support business strategies, such as in programming and controlling task (De Santis et al. 2018b). As an example, such a DSS, besides the capability of providing real-time early warnings, can be adopted to estimate the failure rate starting from a series of heterogeneous measures collected within the power grid and the surrounding dynamic environment. The underlying model

✉ Enrico De Santis
enrico.desantis@uniroma1.it

Francesco Arnò
arno.1595382@studenti.uniroma1.it

Antonello Rizzi
antonello.rizzi@uniroma1.it

[1] Department of Information Engineering, Electronics, and Telecommunications, "Sapienza" University of Rome, Via Eudossiana 18, 00184 Rome, Italy

can be exploited, from one hand, for gaining insights over the failure phenomena, while from the other, the probability rate can feed an advanced risk assessment tool. For example, it can be used for estimating the risk of power grid equipment, given a suitable series of impacts metrics. Furthermore, once obtained the overall risk associated with the power grid, the system allows driving scenario-based risk identification and analysis.

For this important purpose, in the current paper we deal with a significative extension of our previous works (De Santis et al. 2015a, c, 2017), where it is presented a modeling and recognition system of faults and outages occurring in the real-word power grid managed by Azienda Comunale Energia e Ambiente (ACEA) company in Rome, Italy. The recognition system, known as the OCC_System, has been developed in collaboration with the ACEA personnel within the "ACEA Smart Grids project" (ACEA 2016), as the main core of a larger DSS. It follows a One-class classification paradigm that exploits a cluster-based evolutionary technique in order to learn a model of a specific class of faults, called Localized Faults (LFs). The custom-made system, which works in a supervised fashion, is fed by a historical dataset of power grid states, providing information about endogenous and exogenous factors and is able to learn a model together with a custom-based dissimilarity measure used, in turn, for classifying fault states in real time (De Santis et al. 2018b). On the one hand, the clustered model acts as a gray box for knowledge discovery tasks in line with the explainable AI paradigm (xAI) (Gunning 2017). On the other hand, the output of the recognition system is a Boolean decision together with a score value assigned to a given unseen test pattern, allowing both classifying a power grid state and providing a reliability measure of the decision (score value).

Therefore, the current study starts from the need of extracting from a learned model of fault useful information for programming and control task, such as the estimation of the final risk associated with a set of power grid states. As stated, when dealing with risk assessment and cost benefit analysis for maintenance planning, the availability of reliable probability estimates is of utmost importance. The score values assigned to the test pattern by the recognition system under study are obtained from a suitable fuzzy membership function. In our previous works, these scores were interpreted as a measure of predictions reliability. However, fuzzy values cannot be considered as reliable probability estimates. In fact, a probability measure $P(A)$ defined on a universe U is a mapping function that assigns a number P to each subset of the universe U, and satisfies the so-called Kolmogorov axioms:

1. $P(A) \geq 0$;
2. $P(U) = 1$;
3. for any countably infinite sequence of events $(A_i)_{i\geq 1}$ that are mutually exclusive (i.e., $A_i \cap A_j = \emptyset$):

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i). \quad (1)$$

On the other hand, a fuzzy set F on a universe U is defined by a membership function $\mu_F : U \rightarrow [0, 1]$ and $\mu_F(u)$ is the degree of membership of element u in F. In general, fuzzy memberships represent similarities of objects to imprecisely defined properties, while probabilities convey information about relative frequencies of events. However, the relation between Probability Theory and Fuzzy Logic is a controversial issue in the literature and an in-depth analysis on the argument is provided in Mendel (1995) and Hajek et al. (2013). Due to their nature of fuzzy values, the scores provided by OCC_System do not reflect the empirical probabilities associated with grid states; more specifically these scores are likely to be uncalibrated. Given a classifier which outputs a score $s(x)$ in range [0,1] for each example x, by definition it is said to be *well-calibrated* if $P(c|s(x) = s)$, namely the empirical class membership probability of x, converges to the score $s(x) = s$ as the number of classified patterns tends to infinity (Murphy and Winkler 1977; Zadrozny and Elkan 2002). For instance, a predictive model is well calibrated if among the samples to which it gave a score (or probability) close to, e.g., 0.7 for the membership to the class c, approximately 70% of these samples actually belong to the considered class. In analytical terms, the calibration of a classification system consists in finding a function that maps the scores (or not calibrated probability estimates) into effective probability estimates bounded in range [0, 1] by definition (Martino et al. 2019).

The calibration of predictive models is currently a popular object of study in Machine Learning applications. In particular, calibration is widely used for clinical problems, since probability estimates are very important in order to build reliable diagnostic models. These models can be used to guide clinical decisions in real-world situations concerning the patient's treatment. At a sufficiently abstract level of reasoning, by extension, we can consider a DSS adopted even for CBM tasks as a diagnostic system, where the "patient'' is not an individual but a power grid, following the same rationale adopted in clinical problems.

Among the several calibration techniques described in literature, for our case study, we consider the well-known Platt scaling (Platt 1999), Isotonic Regression (Zadrozny and Elkan 2002; Naeini et al. 2015) and SplineCalib (Lucena 2018). The first two are the most commonly used methods for probability calibration, while SplineCalib is a recently proposed non-parametric calibration method. In addition, in this paper we evaluate three further proposed calibration tech-

niques based on several suitable fitting procedures of the reliability diagram, investigating the effectiveness of these simple techniques in solving real-world calibration problems.

Therefore, the main aim of this paper is to calibrate the OCC_System output (fuzzy) scores by applying several calibration techniques in a post-processing fashion, in order to identify the calibration procedure that provides the best performances in terms of well-suited calibration metrics on several real-world datasets. This study faces the calibration of the OCC_System classifier on three real-world dataset of power grid faults described by heterogeneous data, of which one of them is heavily unbalanced on the Test set, providing misleading results in terms of calibration performances. In this case, after a deep analysis, an ad hoc over-sampling procedure for structured data is designed, allowing to face also unbalanced data. The best calibration method is further adopted to increase the capabilities of the recognition system, which will be in charge of outputting well-calibrated probabilities so that it can be exploited as an important feature within the pipeline used by the distribution utility for estimating the overall network risk, given suitable impact metrics. Part of the calibration techniques are treated in Martino et al. (2019), facing three classical classification problems with the standard SVM algorithm. Specifically, the calibration procedure is experimented on two UCI datasets (Dua and Graff 2019) (ABALONE, ADULT) (Asuncion and Newman 2007) and on a protein contact network datasets for predicting the function of specific proteins (De Santis et al. 2018a). It is noted that the datasets were mainly balanced and no further procedures were needed for evaluating the performance. Moreover, data patterns were standard real-valued n-tuples. Instead, as already stated, in the current paper the calibration procedures are considered as plug-in modules in a real-world custom classification pipeline. This pipeline is used in a decision support system for fault recognition and diagnosis in a MV power grid, where data are heavily structured and the available dataset is heavily unbalanced.

The remainder of this paper is organized as follows: in Sect. 2, we discuss related works, while in Sect. 3, we provide a briefly description of the ACEA power grid and the available dataset of power grid states. In Sect. 4, we review the main functional blocks of the designed recognition system. In Sect. 5, we give an overview of existing calibration techniques and the suitable figures of merit for addressing the goodness of calibration, along with three new procedures to be compared with state-of-the-art approaches. In Sect. 6, we describe the experimental setting and compare the results providing a discussion about the reliability of the considered techniques on the case-study. Finally, in Sect. 7, conclusions are drawn.

2 Related work

In this section, the main adopted techniques for fault prediction and modeling in power grids are reviewed and discussed, along with the use of calibration procedures to improve the reliability of predictive models. Specifically, a series of updated related works on fault detection, classification and localization in power grids through Artificial Intelligence algorithms are discussed in Sect. 2.1, while Sect. 2.2 summarizes some related works about the use of calibration techniques.

2.1 Fault detection and prediction methods for power grids

Several machine learning-based approaches for power grid fault detection and prediction have been proposed in the last two decades, along with the raise of Smart Grid concept, namely an interconnected power system able to provide a huge amount of data about the status of the grid thanks to an Advanced Metering Infrastructures (AMI). The last one consists of at least a reliable telecommunication network with a set of suitable data centers and smart sensors. Some of the existing literature on power grid fault analysis are discussed in the following.

Collecting heterogeneous data about the power grid state together with environmental information where the power grid works can lead to several data mining and pattern recognition problems, such as events classification (Afzal and Pothamsetty 2012) or diagnostic systems for cables and accessories (Rizzi et al. 2009). A Decision Support System for the preventive maintenance of New York city power grid is described in detail in Rudin et al. (2012). The proposed system includes classification algorithms, such as Support Vector Machines (SVM), for ranking the electrical components according to their failure probability, and regression algorithms, such as Classification and Regression Tree (CART), to estimate the mean time between failure (MTBF). Dealing with fault detection and localization problems in power grids, in Jiang et al. (2014) authors propose a system based on a clustering algorithm and a Hidden Markov Model (HMM) that operate on frequency signals acquired by frequency disturbance recorders scattered in the power grid. In Souza Pereira et al. (2018) an evolutionary algorithm for fault localization is proposed, while in Thukaram et al. (2005) the authors present a combined approach SVM–Artificial Neural Network (ANN), where the former detects the type of fault, while the latter estimates the location of the fault. In the field of fault analysis, modeling and recognition, in Tokel et al. (2018) the authors present a system based upon an ANN which processes Phasor Measurement Unit (PMU) data. PMU data are exploited in Bhattacharya and Sinha (2017) too, where the authors propose an innovative

approach based on Long Short Term Memory (LSTM) network. Two fault detection and classification techniques based on the One-Class Quarter-Sphere SVM algorithm are proposed in Shahid et al. (2012). Other approaches may include fuzzy logic (FL) (Das 2006; Sang-Won Min et al. 2004), fuzzy Petri-nets (Luo and Kezunovic 2008; Jing Sun et al. 2004), decision trees (Samantaray 2009). As concern faults diagnosis in power grids, in Zufeng Wang and Pu Zhao (2009) is proposed a SVM-based method to perform the recognition of faults related to high-voltage transmission lines. In Kordestani and Saif (2017) a fusion method, grounded on circuit breakers data, wavelet transform and radial basis function network, is proposed for fault diagnosis.

As concerns a DSS equipped with a fault recognition system, one can find several heterogeneous approaches depending both on the final objective and the available data. For a fault recognition system, it is very important the nature of information acquired that is intrinsic to the physical power grid, extrinsic or both, but also the granularity with which data are acquired through smart sensors. In other words, the architecture of the systems, and specifically the features engineering phase, is related to the objectives of the applications. One may include in the learned model exogenous causes such as weather conditions or endogenous ones, such as the electrical load. In this way, further studies can be undertaken, such as for example the one explained in Guikema et al. (2006), where authors have established a fruitful relationship between environmental features and fault causes.

2.2 Calibration of machine learning models: a review

Calibration techniques play an important role in many machine learning applications.

Calibrating probability estimates is a crucial step for risk analysis tools in many settings (Pleiss et al. 2017). Most of available works in the literature about calibration of predictive models deal with medical predictive analytics with numerous publications focusing on models that estimate patients' risk of a disease or a future health state based on machine learning algorithms (Van Calster et al. 2019). Pereira et al. (2020) a Decision Support System for the treatment of Alzheimer's disease patients based on calibration techniques is presented. Specifically, the authors propose an ensemble-based approach, where outputs from multiple classifiers, such as SVM, Gaussian Naïve Bayes, Neural Networks, are combined with calibration models, such as Platt scaling and Isotonic Regression, and other uncertainty methods [Venn-ABERS predictors (Vovk and Petej 2014; Vovk 2012) and Conformal Predictors (Vovk et al. 2005)]. In order to optimize the quality of predictions, the best pair (classifier—uncertainty method) is chosen for the data under study. Remaining in clinical setting, in Walsh et al. (2017) calibration is used to improve the reliability of a decision making system for the readmission of patients. The authors propose a framework to select the best calibration method, among Platt scaling, Logistic calibration (Steyerberg et al. 2004) and Prevalence Adjustment (Morise et al. 1996), for a risk readmission predictive model, created via a L1-regularized Logistic Regression. In addition, the effect of miscalibration on clinical cost is evaluated.

Dealing with calibration of fault predictive models, in Cremer and Strbac (2019) the authors propose a machine learning-based Dynamic Security Assessment (DSA) for a power grid. An ensemble of classifiers that combines multiple CART is learned and calibrated by using Platt scaling to provide accurate probability estimates. The reported experimental evaluation is performed on a real-world dataset about the French Transmission Grid. In the field of fraud detection in electric power distribution, in Massaferro et al. (2020) the authors describe a machine learning solution to make long- and short-term decisions about customer inspections. Decisions are based on the estimates of posterior fraud probabilities, obtained by means of the calibration of score values output by a suitable classifier. Simulations on real-world datasets, regarding the customers distributed across Montevideo (Uruguay), involve classifiers such as SVM, Random Forest, ANN, and two calibration methods, which are Platt scaling and Isotonic Regression. Furthermore, the best pair is selected for the data under study. It is worth noting that our approach follows a similar pipeline, since we seek to estimate the posterior fault probability given a power grid state described by several suitable features.

Calibration techniques are used for pattern recognition tasks as well. In Gosztolya and Busa-Fekete (2018), calibration methods such as Linear Scaling, Platt scaling and Isotonic Regression are applied to the AdaBoost.MH algorithm (Schapire and Singer 1999) for a speech recognition task. In Blair et al. (2014) calibration is applied as post-processing step for an object detection task. The authors investigate the performance of Platt scaling and Isotonic Regression in converting a score, obtained from a machine learning detector based on an Adaboost classifier or an SVM, to a probability representing confidence in measuring the presence of an object in a given location.

3 The dataset of localized faults

In this paper, we deal with the problem of estimating the probability of faults occurring in Medium Voltage (MV) feeders of the power grid managed by ACEA S.p.A. in Rome, Italy.

The ACEA power grid is a wide and interconnected grid located in the middle of Italy. It is made up of MV backbones of uniform section exerting radially. Each backbone is fed by two distinct Primary Station (PS), and each half-

line is protected against faults through breakers. The ACEA power grid consists of lines (feeders) in which the nominal voltage is 20 kV, and some few *legacy* lines that still work at 8.4 kV. Each MV line supplies a given number of Secondary Stations (SSs) through cables that can be placed on air or underground. The cable' section is variable along the backbone, with the presence of bottlenecks. By several years the power grid is undergoing a modernization process in line with the concept of Smart Grids. In fact, the power grid is equipped with several TLC systems able to transport data related to the main system, collected from smart sensors, as well as environmental variables. Data are collected and sent to the main server for storing and processing tasks (De Santis et al. 2015c). Further details about the ACEA power grid are available in (De Santis et al. 2015b, c).

Within the "ACEA Smart Grids Project", together with ACEA field-experts, a two-class dataset of real power grid states, including standard functioning states (SFs) and LF states, has been built (De Santis et al. 2018c). Specifically, a state of the power grid is described by several features obtained from data related to environmental factors, such as weather condition, temporal data (i.e., when the fault happens), geo-spatial data (i.e., latitude and longitude pairs), physical data related to the state of the power grid and its electric equipment (e.g., measured voltage and currents). As concerns the data type, features belong to categorical (nominal), quantitative (i.e., data belonging to a normed space) and time series (TSs) data. TS data are in form of unevenly spaced sequence of short outages that are automatically registered by the protection systems (Petersen Alarm System) as soon as they occur. A detailed description of the dataset structure, the features and the preprocessing stage are treated in (De Santis et al. 2015b, c, 2017).

A state of the power grid depends upon a lot of different factors. Hence, in order to improve the fault probability estimation, as suggested by a group of field-experts and in line with the real-world case study, it could be convenient to build datasets constituted by a subset of the total features, where each subset is more specific to certain parts of the power grid.

Following these prescriptions, three datasets have been built:

Nodes: the NODES dataset is composed by 486 instances and 19 attributes. It contains data relative to the substations where generators, load or transmission lines interconnect.

Branches: the BRANCHES dataset is composed by 689 instances and 23 attributes. It contains data relative to transmission lines connecting two nodes.

Standard (Std): the STANDARD dataset is composed by 2651 instances and 26 attributes. It contains information about the whole power grid.

We remark that the STANDARD (STD) dataset, built together with the ACEA field experts, is conceived at a more abstract level compared to the NODES and BRANCHES ones. While the last ones are specific for the real-world power grid partition in nodes and branches, in the STANDARD case the power grid as a whole is considered, taking into account the overall set of equipment. The STANDARD dataset is constructed on more features and in the following study has to be treated as a benchmark dataset.

Table 1 reports the list of considered features, including a brief description and the dataset they belong.

4 A brief review of the OCC_System

The designed OCC_System for the recognition of LF patterns in the ACEA power grid is based on a clustering-evolutionary hybrid approach. The clustering approach derives from the assumption that similar states of the power grid have similar chances of generating a LF (De Santis et al. 2018b).

A dataset of target patterns is partitioned in k (disjoint) clusters, where each cluster contains faults having similar features. For clustering computation, the recognition system uses a custom-based dissimilarity measure computed as a weighted Euclidean distance. Given two LF patterns $\mathbf{x}_1$, $\mathbf{x}_2$, the proposed weighted dissimilarity is defined as follows:

$$d(\mathbf{x}_1, \mathbf{x}_2; \mathbf{W}) = \sqrt{(\mathbf{x}_1 \ominus \mathbf{x}_2)^T \mathbf{W}^T \mathbf{W} (\mathbf{x}_1 \ominus \mathbf{x}_2)}, \tag{2}$$

where $\mathbf{W}$ is a diagonal matrix whose elements comes from suitable vector of weights $\mathbf{w}$ and the $\ominus$ operator represents a generic dissimilarity measure, which has to be specified depending on the semantic of data at hand, since the dissimilarity measure is component-wise (De Santis et al. 2018b). A complete description of the various dissimilarity measures, involved in computing the distance between the patterns, is treated in De Santis et al. (2015c).

The architecture of the classifier is grounded on a set of clusters C_i, each one represented by its medoid c_i. The ensemble of clusters defines a decision region through several parameters, such as the radius $\delta(C_i)$, computed as the average distance of cluster members from the medoid, plus a threshold σ. These last, together with the dissimilarity weights $\mathbf{w} = diag(\mathbf{W})$, constitute the parameters of the classification model. Once obtained the model grounded on clusters, for a given test pattern $\mathbf{x}$, the classification procedure need a decision rule in two steps. The first one consists of finding the closest cluster, the second foresees evaluating if the test pattern falls inside or outside the overall faults decision region made up of that cluster. Hence, the overall learning phase involves clustering the Training set composed by LF (target) patterns, employing a standard Genetic Algorithm (GA). The latter is in charge of evolving a family of

Table 1 List of the considered features in power grid dataset records

Feature	Data type	Data set	Description
Day start	Quantitative (integer)	Nodes–Branches–Std	Day in which the LF was detected
Time start	Quantitative (integer)	Nodes–Branches–Std	Time stamp (min) in which the LF was detected
Primary station code	Categorical (string)	Nodes–Branches–Std	Unique backbone identifier
Protection tripped	Categorical (string)	Nodes–Branches–Std	Type of intervention of the protective device
Kind of element	Categorical (string)	Nodes–Branches–Std	Kind of faulty element
Voltage line	Categorical (string)	Nodes–Branches–Std	Nominal voltage of the backbone
Material	Categorical (string)	Nodes–Branches–Std	Constituent material element (CU, AL)
Location element	Categorical (string)	Nodes–Branches–Std	Element positioning (aerial or underground)
#Secondary stations (SS)	Quantitative (integer)	Std	Number of out of service secondary stations due to the LF
Current out of bounds (CoBs)	Quantitative (integer)	Nodes–Branches–Std	The maximum operating current of the backbone is less than or equal to 60% of the threshold "out of bounds", typically established at 90% of capacity
Cable section	Quantitative (real)	Branches–Std	Section of the cable, if applicable
Max. temperature	Quantitative (real)	Nodes–Branches–Std	Maximum registered temperature
Min. temperature	Quantitative (real)	Nodes–Branches–Std	Minimum registered temperature
Delta temperature	Quantitative (real)	Nodes–Branches–Std	Difference between the maximum and minimum temperature
Rain	Quantitative (real)	Nodes–Branches–Std	Millimeters of rain calculated as the average in the 24 h preceding the LF
Interruption (breaker)	Time series (integer)	Nodes–Branches–Std	Outages caused by the opening of the breakers in the primary station
Petersen alarms	Time series (integer)	Nodes–Branches–Std	Alarms detected by the device called "Petersen's coil" due to loss of electrical insulation on the power line
Saving intervention	Time series (integer)	Nodes–Branches–Std	Decisive interventions of Petersen's coil which have prevented the LF
Backbone electric current (BEC)	Quantitative (real)	Branches–Std	Absolute difference between the average current in two non-overlapping windows each one of 12 h registered in the 24 h preceding the LF
Secondary station type	Categorical (string)	Nodes–Std	Type of secondary station
Transformer voltage	Quantitative (integer)	Nodes–Std	Size of the transformer
Number of couplings	Quantitative (integer)	Branches–Std	Number of couplings in a branch
Year	Quantitative (integer)	Branches–Std	Year of installation
Branch length	Quantitative (real)	Branches–Std	Length of the branch
branch positioning type	Categorical (string)	Branches–Std	Position of the branch (air or ground)

Std Standard

cluster-based classifiers by finding the parameter values that minimize a suitable objective function or fitness. The fitness function consists in a convex linear combination of the accuracy of the classification computed on the Validation set that should be maximized, and the value of the thresholds that we seek to minimize. The Validation set is composed by LFs and normal functioning states. Since the classification model is built using only target patterns, while non-target ones are used only in the cross-validation phase, the adopted learning paradigm is the One-Class classification one (Khan and Madden 2010; Pimentel et al. 2014). Due to the high sensibility of the standard k-means algorithm to the random initialization of cluster representatives, the OCC_Systems runs more than one instance of the clustering algorithm with different random initializations. Therefore, during the test phase (and also during validation) a voting procedure for each cluster model is adopted. In this way, it is provided a more robust data-driven model of the power grid faults (De Santis et al. 2018b).

Together with the Boolean classification rules able to check if a state of the power grid is a fault or not, the system is in charge of computing a soft decision value in the real-valued range [0, 1]. For this purpose, we equip each cluster C_i with a suitable membership function, denoted in the following as μ_{C_i}. The membership function allows us to quantify the reliability (or the uncertainty) about the recognition of a test pattern. In our previous works (De Santis et al. 2015b, c, 2018b), the soft decision value was computed by a sigmoidal fuzzy membership function, whose parameters were related to clusters geometry. Instead, in this work, we consider a Gaussian fuzzy membership function. Given a test pattern $\mathbf{x}$, its score value is computed as follows:

$$s(\mathbf{x}) = \mu_{C_i}(d(c_i, \mathbf{x})) = e^{\frac{-(d(c_i,\mathbf{x}))^2}{2t_i^2}}, \quad (3)$$

where c_i is the representative of the cluster C_i. The Gaussian membership function associated with each cluster has zero mean, while the variance t_i^2 is set in order to output a value close to 0.5 for patterns which are placed close to the decision region boundary. Figure 1 depicts this idea by an intuitive illustration; since the function is symmetric around the mean value, only half of the bell shaped curve is depicted in Fig. 1.

The last procedure allows the classifier to assign to $\mathbf{x}$ a score value in the unitary interval that increases as the dissimilarity decreases, depending on the dissimilarity between the test pattern $\mathbf{x}$ and the representative of a given cluster. We underline that the computed score is not a fault probability estimation since it is the output of a fuzzy membership function.

Figure 2 reports a schematic representation of the various subsystems of the proposed OCC_System. Specifically, we have the Clustering subsystem and the GA one. Furthermore, it is shown the test subsystem, where for a test pattern the classifier associates a predicted label (Boolean classification) and a fault score value (soft decision).

Further information about the OCC_System learning model, the data description and preprocessing can be found in our previous works (De Santis et al. 2014, 2015a, b, c).

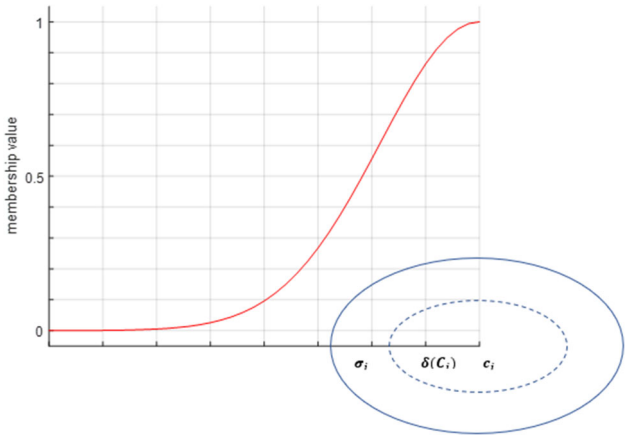


Fig. 1 Gaussian membership function associated with a decision region

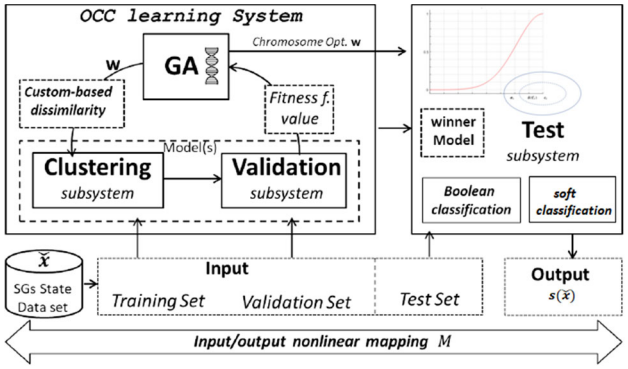


Fig. 2 Schematic of the recognition system able to learn a model of faults providing a reliability decision score s of a fault input pattern $\mathbf{x}$

5 On calibration techniques

5.1 The reliability diagram

In the first instance, the calibration of a classifier can be visualized through a reliability diagram (DeGroot and Fienberg 1983a).

The reliability diagram is a graph where the empirical probabilities $P(c|s(x) = s)$, namely the number of patterns with score s that belongs to class c divided by the total number of patterns with score s, are plotted against the predicted scores/probabilities.

If the classifier is well calibrated, all points fall near the diagonal line as the scores are equal to empirical probabilities. In the context of binary classification, as the problem concerned in this study, the empirical probabilities are computed for positive patterns only, that is, the total number of positive examples with score s divided by the total number of examples with score s (Martino et al. 2019).

In practical applications, the number of possible scores is large compared to the number of available patterns, since they are real-valued scalars, and reliable empirical probabilities cannot be calculated using the procedure described above.[1]

In this case a binning of the score space is needed:

– on the x-axis, the average score value for each bin is considered;

[1] For each score s there will be only one pattern in most cases, consequently $P(c|s(x) = x)$ would be either 0 or 1.

– on the y-axis, we get the ratio between the number of positive patterns lying in that bin and the total number of patterns in the same bin (i.e., the true fraction of positive instances).

It is worth noting that the bin size must be selected carefully, in order to have enough examples in each bin for calculating reliable probability estimates (Zadrozny and Elkan 2002), although in some works, such as Niculescu-Mizil and Caruana (2005), scores are merely divided into 10 equally spaced bins in the range [0, 1] regardless of their distribution. For the considered dataset, a 10-bins discretization procedure does not perform well, so other methods have been considered, such as:

– Scott's normal reference rule (Scott 1979), which evaluates the bin width taking into account the total number of observations (scores for this case) n and their standard deviation σ as:

$$\text{bin width} = \frac{3.5 \cdot \sigma}{n^{1/3}}, \tag{4}$$

– Freedman–Diaconis' choice (Freedman and Diaconis 1981), which computes the bin width as follows:

$$\text{bin width} = \frac{2 \cdot IQR}{n^{1/3}}, \tag{5}$$

where IQR is the interquartile range of scores.

Other interesting binning methods are reported in Martino et al. (2019).

5.2 Current approaches

Given a set of labeled examples, if c is the positive class we can assume that $P(c|x) = 1$ for positive examples and $P(c|x) = 0$ for negative ones. Calibration techniques work similarly to supervised learning methods. In other words, they need a calibration set, made up of predicted scores and actual labels, in order to learn a function (or model) formally defined as:

$$f : s(x) \rightarrow \hat{P}(c|x). \tag{6}$$

In practice, Eq. (6) is a function in charge of mapping scores into probability estimates.

Platt scaling (Platt 1999) is one of the main methods for calibrating scores in a binary classification problem. In order to get calibrated probabilities, Platt proposed to pass output scores through a sigmoid function:

$$\hat{P}(c|x) = \frac{1}{1 + e^{As(x)+B}}, \tag{7}$$

where the parameters A and B are found by minimizing the negative log likelihood of the training data (calibration set). This parametric approach was motivated by Platt showing that the relationship between SVM scores and empirical probabilities $P(c|x)$ can be often fitted well by a sigmoid function. Obviously this technique can be applied for calibrating any type of classifier, not only SVM, and in general, it works well if the reliability diagram of the dataset shows a sigmoidal trend. With regard to the sigmoid parameter estimation, an improved optimization process based on Newton's method is proposed in Lin et al. (2007).

Isotonic regression (Zadrozny and Elkan 2002; Naeini et al. 2015) is a nonparametric approach for model calibration, in which the calibration function is chosen from the class of all isotonic functions. Given a calibration set made up of example labels and their scores $(\mathbf{y}, \mathbf{s})$, the Isotonic Regression applied for a calibration problem consists in finding the non-decreasing function $\hat{m}$ such that:

$$\hat{m} = \arg\min_{z} \sum_{i}^{N} (y_i - z(s_i))^2. \tag{8}$$

Pair-adjacent violators (PAV) (Ayer et al. 1955) are one of the main algorithms in order to compute Isotonic Regression.

PAV works as follows:

1. sort $(\mathbf{y}, \mathbf{s})$ according to $\mathbf{s}$;
2. initialize a vector $\hat{\mathbf{y}} = \mathbf{y}$;
3. while $\exists\, i \;\; s.t.\; \hat{y}_i \leq \hat{y}_{i-1}$;

   set $\hat{y}_i = \hat{y}_{i-1} = \dfrac{y_i + y_{i-1}}{2}$;
4. return $\hat{\mathbf{y}}$.

This procedure makes $\hat{\mathbf{y}}$ contains probability estimates for scores in $\mathbf{s}$. Further, due to the piecewise nature of isotonic regression, $\hat{\mathbf{y}}$ will contain sequences of repeated values, namely the same probability estimate is associated with several scores. Following the procedure described above, Pair-Adjacent Violators returns more samples in the score space where the classifier ranks them incorrectly according their score and less samples where patterns have been ranked properly by the classifier.

Platt scaling and Isotonic Regression are the two main methods for calibrating score in a binary classification problems, but both of them have some limitations:

– Platt scaling gives good results only when the calibration function is well approximated by a sigmoid function (strict parametric approach);
– Isotonic regression works well for many problems due its nonparametric approach with the piecewise constant

approximation, but it tends to overfitting when the calibration set is small (less than 1000 instances).

SplineCalib (Lucena 2018) is a new calibration method, which aims at overcoming limitations of previous models. SplineCalib is a nonparametric approach, and as its name suggests, it uses (cubic) smoothing splines to fit the calibration function, rather than a piecewise constant or sigmoid function. Standard spline regression requires to place K different knots throughout the range of data and fit a polynomial (usually with degree 3 or 4) for each interval: using more knots leads to better fitting of data, but also high risk of overfitting. Smoothing splines (Wahba 1990) performs a regularized regression and use all of the available points as knots. Given a set of predictors and labels $(x_i, y_i), i = 1, ..., N$, the smoothing splines estimate is obtained by finding, among all the twice-differentiable functions $f(x)$, the one that minimizes the following relation:

$$\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int f''(t)\,\mathrm{d}t, \tag{9}$$

In Lucena (2018), instead of minimizing the residual sum of squares, the author adopt a log-likelihood criterion:

$$-\sum_{i=1}^{N}[(y_i \cdot \log f(x_i) + (1 - y_i) \cdot \log(1 - f(x_i))] + \frac{1}{2}\lambda \int f''(t)\,\mathrm{d}t, \tag{10}$$

which is referred as *non parametric logistic regression* (Hastie et al. 2001). The regularization term $\lambda \geq 0$ in Eqs. (9) and (10) is a *smoothing parameter*. In detail, for $\lambda = 0$, the second term from Eqs. (9) and (10) is 0, so f can be any function that interpolates data, no smoothing is tolerated. If $\lambda \to \infty$ no curvature at all can be tolerated, leading to a simple least square line fit.

The most basic version of SplineCalib follows this steps:

1. take K knots randomly[2];
2. use the sampled knots to compute the natural basis expansion matrix $\mathbf{X} \in \mathbb{R}^{n \times k}$ of the values in $\mathbf{s}$. Let be $\{\xi_1, ..., \xi_K\}$ the set of knots, the natural cubic spline basis is defined as:

$$\begin{aligned} &N_1(s) = 1,\ N_2(s) = s, \\ &N_{k+2}(s) = d_k(s) - d_{K-1}(s), \quad \forall k = 1, ..., K-2 \\ &\text{where} \quad d_k(s) = \frac{(s - \xi_k)_+^3 - (s - \xi_k)_+^3}{\xi_K - \xi_k}; \end{aligned} \tag{11}$$

3. fit a l_2-regularized logistic regression model to the set $(\mathbf{X}, \mathbf{y})$ by considering a proper range of value for the regularization term λ and choose the value λ^* which gives the best cross-validated Log-Loss;
4. re-fit the model on the pair $(\mathbf{X}, \mathbf{y})$ using λ^*;
5. output the calibration function $f(s) : [0, 1] \to [0, 1]$ applying in sequence the operation from step 2 and 4, in order to predict probability estimates.

[2] All the available points could be used, yet the Author states that 200 points should be sufficient.

5.3 Fitting-based calibration techniques

In Sect. 5.2, we showed that the reliability diagram of a calibration set is obtained by plotting uncalibrated scores against empirical probabilities. A calibration function, learned by means of one of the three methods described in the previous section, should fit the Reliability Diagram well, because this means that for each uncalibrated score the function returns a probability estimate near to its empirical probability. It follows that a potential calibration function could be computed by finding directly the function which better approximates the reliability diagram's trend. Actually, as we have described in Sect. 5.2, for real problems the reliability diagram is produced after a binning of the score space, so for this new approach, instead of working with "score-label" pairs, we work with "average bin value-fraction of positive patterns in that bin" pairs.

In real cases, as the one shown in Sect. 6, a reliability diagram is unlikely to show a linear trend; therefore, more complex functions shall be used for this fitting. For these purposes, we consider (Martino et al. 2019):

1. polynomial fitting: the relationship between the independent variable and the dependent variable of the reliability diagram is modeled as a 3-degree and 4-degree polynomial;
2. spline fitting: after sampling a suitable set of knots $K = \{k_1, k_2, ..k_t\}$, a natural cubic spline fitting is performed. It returns a piecewise cubic polynomial function f such that:
   - f is a polynomial of 3-degree for each interval $[k_1, k_2], ..., [k_{t-1}, k_t]$;
   - f is linear to the left of the leftmost knot $(-\infty, k_1]$, and to the right of the rightmost knot $[k_t, \infty]$;
   - f, f' and f'' are continuous at the knots $k_1, ..., k_m$.

   Actually we exploits the smoothing spline estimate—described in the previous section for the SplineCalib method—also for this proposed approach in order to avoid the knots selection. We remind that the solution of Eq. (9) is a natural spline.

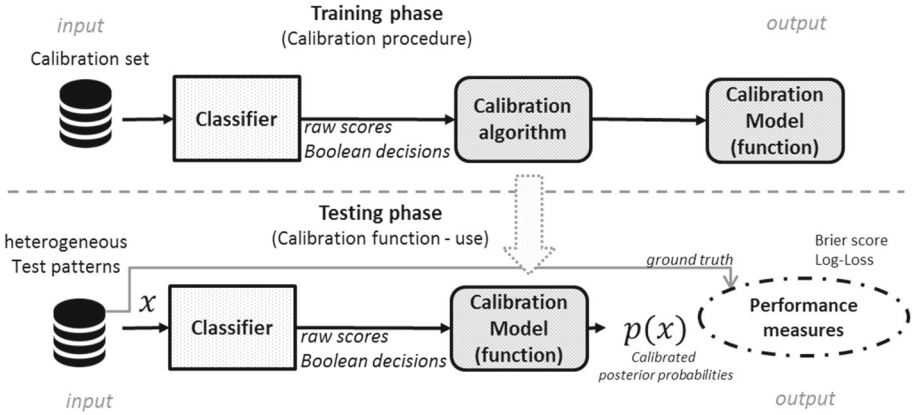


Fig. 3 Generic scheme of a calibration system for classification tasks for both the training phase and the test phase (use)

5.4 Performance metrics

For assessing the accuracy of probability estimates obtained after a calibration procedure, two metrics have been proposed in literature: the Brier Score (DeGroot and Fienberg 1983b; Brier 1950) and the Log-Loss score.

In the context of binary classification, the Brier score is defined as:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(T(y_i = 1|x_i) - P(y_i = 1|x_i))^2\,, \tag{12}$$

where N is the number of samples, $T(y_i = 1|x_i) = 1$ if $y_i = 1$ and $T(y_i = 1|x_i) = 0$ otherwise and $P(y_i = 1|x_i)$ is the probability estimated for pattern x_i to belong to the positive class (label "1"). Since the Brier score is the mean squared error between actual labels ("0" or "1") and predicted probabilities (which must be between 0 and 1), it always takes a value in range [0, 1]. Obviously the lower is the Brier score for a set of predictions, the better the predictions are calibrated.

As regards the Log-Loss (also known as cross-entropy), for a binary classification problem it is defined as follows:

$$LL = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1 - y_i)\log(1 - p_i))], \tag{13}$$

where p_i is the predicted probability and y_i the actual label ("0" or "1"). A perfect calibrated model would have a Log-Loss score of 0. The Log-Loss measures the confidence of predicted probabilities using a logarithmic penalty, namely the index increases rapidly as the estimated probability diverges from the actual label.

5.5 Calibration procedure summary

In general, the calibration procedure consists of learning a suitable function that maps input raw scores (that are the outputs of the generic classifier) with posterior probabilities. Referring to the general scheme of Fig. 3, calibrating a classifier means instantiating a learning procedure as for the classifier system itself. In other words, it is possible to distinguish two phases: (i) the training phase where the calibration model or function is learned on the calibration set, that is a set of patterns extracted from the available dataset, (ii) the test phase where, given a test pattern, is applied the calibration model in order to obtain the (calibrated) posterior probabilities. The test phase coincides with the real use of the calibrated classification system.

Finally, for the next experimental section, we will use specifically this scheme in order to evaluate the overall performance of the OCC_System classifier, adopting the calibration procedures described in the current section.

6 Experimental settings and results

As concerns the experiments, the three ACEA datasets described in Sect. 3 have been divided into Training, Validation and Test sets according to the partition shown in Table 2.

Experiment are conducted with a workstation equipped with an A8-555M CPU@2.10 GHz and 8GB RAM. Simulations are performed with MATLAB R2020a on Windows 10 Home Edition ($\times x64$).

As first task, the OCC_System is trained on the three ACEA datasets, setting the k parameter for the k-means algorithm to 16. We repeated the training procedure five different times for each dataset, by changing the random seed of

Table 2 ACEA dataset splits

Dataset	Training set size	Validation set size	Test set size
Nodes	141	234	111
Branches	215	315	159
Standard	511	1067	1073

Table 3 Average test results on the ACEA datasets

Data	Nodes	Branches	Standard
A	0.96757 ± 0.027	0.93459 ± 0.034	0.99273 ± 0.005
TPR	0.96364 ± 0.038	0.94444 ± 0.021	0.95854 ± 0.010
FPR	0.03146 ± 0.035	0.08627 ± 0.064	0.00444 ± 0.006
S	0.96854 ± 0.035	0.91372 ± 0.064	0.99556 ± 0.006
P	0.89410 ± 0.102	0.95894 ± 0.030	0.95054 ± 0.063
F1	0.92458 ± 0.059	0.95158 ± 0.024	0.95368 ± 0.035
AUC	0.99724 ± 0.002	0.95995 ± 0.019	0.98813 ± 0.008

pseudo-random number generator. Table 3 shows the average performances of the OCC_System. We focus on Accuracy (A), false positive rate (FPR), true positive rate (TPR), Specificity (S), Precision (P), F1 Score (F1), Area Under a Curve (AUC).

Besides the good classification performance reported in Table 3, we are interested here in assessing the reliability of the calibration procedures. Both measures will provide a clear indication of the goodness of the recognition system.

For addressing the calibration performance of three state-of-the-art-methods, namely Platt scaling (PS), Isotonic Regression (IR), SplineCalib (SC) and the three fitting methods reported in Section 5.3, namely 3-degree polynomial (Poly3), 4-degree polynomial (Poly4) and smoothing spline (SplineFit), we consider the score values yield by the Gaussian membership functions of the OCC_System models trained previously. In particular, we have five set of scores for each dataset, since we trained the model five different times. For each dataset, we use the Validation set to create a calibration function; therefore, from here on we refer to it as the calibration set and report its performance on the Test set.

In Fig. 4, we plot the reliability diagrams for calibration and Test set for the three ACEA datasets to asses how well calibrated is the OCC_System. The binning for the NODES and BRANCHES datasets has been performed using the Scott's normal reference rule, described in Sect. 5.1; therefore, the number of points of the reliability diagrams depends upon the distribution of the input scores. Instead, for the STANDARD dataset we have used 10 uniformly spaced bins, since there are enough examples to calculate reliable empirical probability estimates for each bin. In all cases, the trend is way far from the $y = x$ diagonal line, a clear sign that the custom-made classifier is not well calibrated. For example, focusing on the NODES calibration set, we see that the reliability diagram always lies above the diagonal line, namely for all instances of the NODES calibration set the true probability belonging to the fault class is greater than the score value assigned by the classifier.

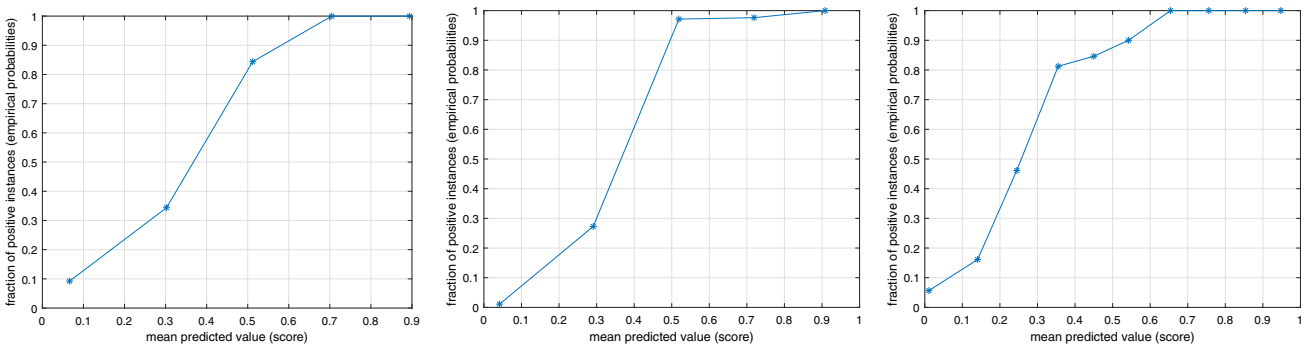
Figure 5 shows the resulting calibration functions overlaid on the reliability diagrams of test sets for the three considered datasets. Conversely, in Fig. 6, we plot the reliability diagrams of Test sets after calibration. For the NODES (a) and BRANCHES (b) datasets, the trend of the reliability diagrams after calibration is close to the $y = x$ line, while for the STANDARD dataset (c), the points of the reliability diagrams fall away from the diagonal line. In particular, it seems that the probabilities obtained by the calibration functions are close to zero for instances with a score lower than 0.5. However, the graphs in Figs. 4, 5 and 6 refer to one of the five sets of scores produced for each dataset.

Tables 4 and 5 report the value of the two figure of merits (Brier score and Log-Loss score, respectively) on both calibration set and Test set as averages over five trials for each of the three datasets. For ease of comparison, Fig. 7 shows through a bar-plot the Brier score (a) and the Log-Loss for each dataset and each calibration method, by considering only the performance on Test set.

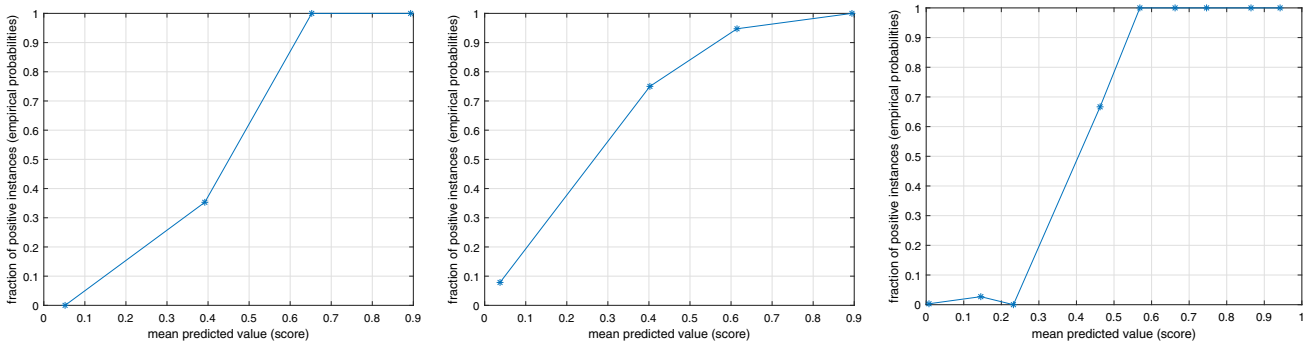
We remark that the Log-Loss index matches the estimated probability with the class label with logarithmic penalty. Hence, for small deviations between y_i (predicted score) and p_i (true probability) the penalty is low, whereas for large deviations the penalty is high. In other words, this figure of merit amplifies the penalty for large deviations.

As concerns the NODES dataset, both the Brier score and the Log-Loss measures exhibit somewhat similar performance, higher than the uncalibrated case, as expected. In particular, in terms of Brier score SplineCalib is the best method, followed by Platt scaling. In terms of Log-Loss, SplineFit reaches the higher values, followed very close by Isotonic Regression and Poly3 techniques. This means that Poly3 well behaves also in the extreme regions of the score. For the BRANCH dataset, regarding the Brier score measure, Platt Scaling and SplinCalib equally outperform other methods, even if the differences are in terms of small percentage values. This is not true for the Log-Loss, as the Platt Scaling and SplineCalib show very low values in comparison with the other techniques. Specifically, Isotonic Regression, Poly3 and SplineFit get worse results than the uncalibrated case. Poly4 takes fourth place in the ranking. As concerns the STANDARD dataset, all the techniques reach small values for the Brier score that are very close to the uncalibrated case. In terms of Log-Loss, the Isotonic Regression obtains the smaller value, followed by SplineCalib. Comparing the results obtained with the calibration set and the Test set, it is clear that all the calibration algorithms seem to lose the generalization capability, as, e.g., SplineCalib on the calibration set well behaves in comparison with the uncalibrated case and performs better than the other techniques. However, a deep investigation—proposed below—shows that for the STANDARD dataset the problem is related to the class unbalance of the Test set.

In general, for at least two datasets (NODES, and BRANCHES) the three alternative methods (Poly3, Poly4, SplineFit) have

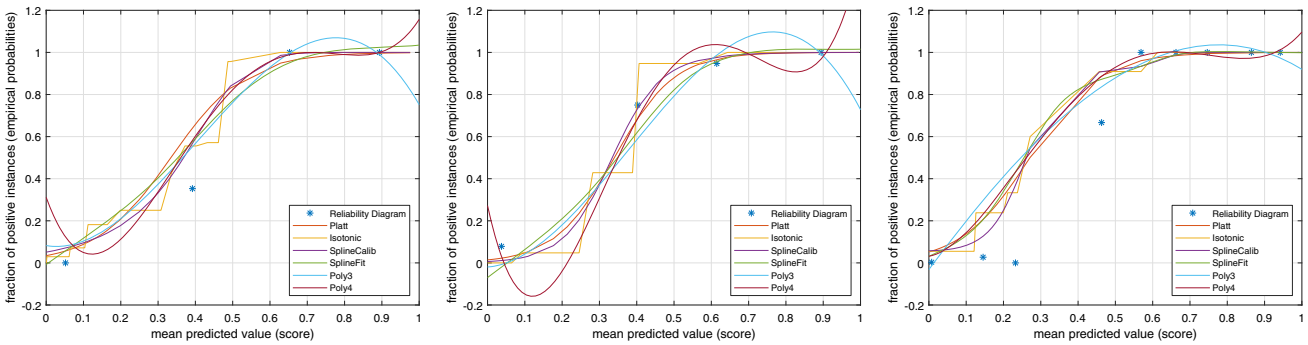


(a) NODES (calibration set) **(b)** BRANCHES (calibration set) **(c)** STD (calibration set)



(d) NODES (Test set) **(e)** BRANCHES (Test set) **(f)** STD (Test set)

Fig. 4 Reliability diagrams



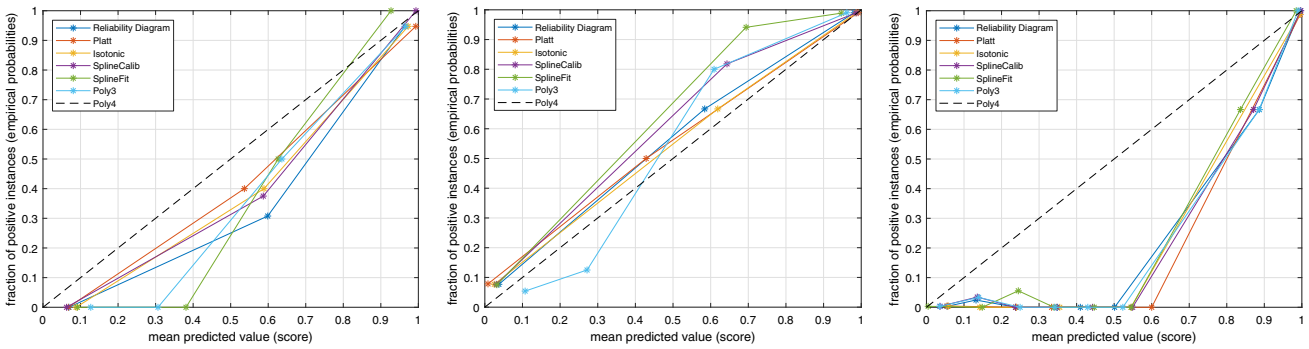
(a) NODES **(b)** BRANCHES **(c)** STD

Fig. 5 Reliability diagrams vs. fitted curves

a Brier score comparable to state-of-the-art-techniques (PS, IR, SC). In terms of Log-Loss, it is not true for Poly3 that loses performance at the borders of the score scale due to the high oscillations caused by the small polynomial degree.

Calibration's results for the STANDARD dataset are certainly not satisfying for our purposes. By looking at the class distribution for this dataset in Table 6, it can be easy to see that the STANDARD Test set is highly imbalanced.

One of the appreciable features of the OCC_System is that it learns a models of faults on the target class and the unbalanced class problem is mitigated. This is not true for the calibration procedure. Specifically, comparing the results obtained for the calibration set and the ones obtained for the Test set (in STANDARD dataset case) in Table 4 or Table 5 it seems that the problem is related to the unbalancing of the Test set. In order to investigate this specific issue, two techniques for balancing the test is further adopted. The first



(a) NODES **(b)** BRANCHES **(c)** STD

Fig. 6 Reliability diagrams after calibration

Table 4 Brier score

Method	Nodes		Branches		Std	
	Calibration set	Test set	Calibration set	Test set	Calibration set	Test set
Uncalibrated	0.12463 ± 0.017	0.03743 ± 0.005	0.05859 ± 0.011	0.08625 ± 0.020	0.05178 ± 0.011	0.00910 ± 0.004
PS	0.05399 ± 0.019	0.02782 ± 0.015	0.02730 ± 0.014	0.05553 ± 0.027	0.02809 ± 0.011	0.00964 ± 0.005
IR	0.04479 ± 0.018	0.02911 ± 0.019	0.02188 ± 0.013	0.05959 ± 0.031	0.02590 ± 0.011	0.00961 ± 0.004
SC	0.02657 ± 0.014	0.02657 ± 0.014	0.05554 ± 0.028	0.05554 ± 0.028	0.00962 ± 0.004	0.00962 ± 0.004
Poly3	0.05980 ± 0.017	0.03362 ± 0.016	0.03257 ± 0.012	0.05959 ± 0.024	0.02960 ± 0.012	0.01100 ± 0.004
Poly4	0.06032 ± 0.018	0.04522 ± 0.013	0.03546 ± 0.011	0.05633 ± 0.025	0.02837 ± 0.011	0.01052 ± 0.004
SplineFit	0.05703 ± 0.018	0.03074 ± 0.016	0.02939 ± 0.014	0.05678 ± 0.025	0.02771 ± 0.011	0.00965 ± 0.004

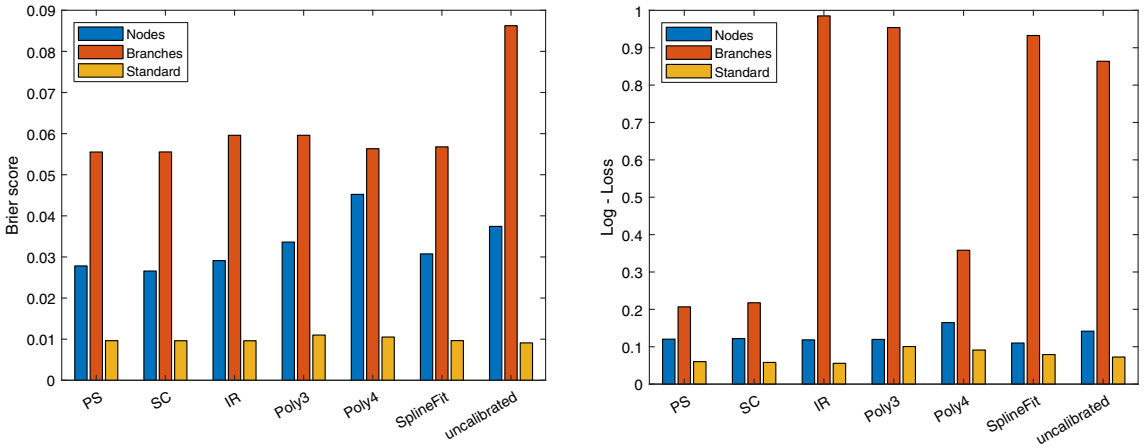
Table 5 Log-loss score

Method	Nodes		Branches		Std	
	Calibration set	Test set	Calibration set	Test set	Calibration	Test set
Uncalibrated	0.45449 ± 0.064	0.14160 ± 0.009	0.21803 ± 0.030	0.86381 ± 0.336	0.32532 ± 0.128	0.07242 ± 0.045
PS	0.19140 ± 0.050	0.12026 ± 0.053	0.10296 ± 0.043	0.20670 ± 0.080	0.10743 ± 0.039	0.06011 ± 0.023
IR	0.19140 ± 0.050	0.11833 ± 0.061	0.10296 ± 0.043	0.98519 ± 0.287	0.10743 ± 0.039	0.05563 ± 0.027
SC	0.17989 ± 0.056	0.12178 ± 0.048	0.09548 ± 0.045	0.21766 ± 0.079	0.10446 ± 0.041	0.05801 ± 0.027
Poly3	0.52169 ± 0.220	0.11959 ± 0.054	0.19125 ± 0.092	0.95396 ± 0.241	0.42450 ± 0.285	0.10052 ± 0.068
Poly4	0.39156 ± 0.252	0.16463 ± 0.049	0.24352 ± 0.111	0.35816 ± 0.144	0.14152 ± 0.038	0.09118 ± 0.087
SplineFit	0.48529 ± 0.192	0.10989 ± 0.052	0.18406 ± 0.110	0.93273 ± 0.232	0.15037 ± 0.053	0.07910 ± 0.068

one is a simple under-sampling of the non-target class, while the other is specifically designed for the problem at hand and consists in the over-sampling of the target class with an ad hoc procedure tailored for heterogeneous and structured data pattern.

In imbalanced cases, misclassification costs tend to be asymmetric, namely incorrectly classifying minority class examples is usually more costly than making mistakes in the other direction. The same concept is applied in estimating class membership probabilities context.

As an example, in Fig. 8 we show the residual errors of probability estimates, produced by Platt scaling, for the instances of the STANDARD Test set. Specifically, each sub-plot displays the absolute differences between the true labels and the corresponding probability estimates, namely $|y_i - P(y_i|x_i)|$. Lower values implies a better calibration, since it means that the probabilities agree with the true labels. Figure 8a shows the histogram for all instances, corresponding to overall calibration. More than 1000 instances lie in the first bin; therefore, it would seem that Platt calibration had worked well. But if we look at the right-most plot, which is the same figure but includes only minority instances, it shows that for few instances the probability estimates highly disagree with the observed labels. These errors have a huge

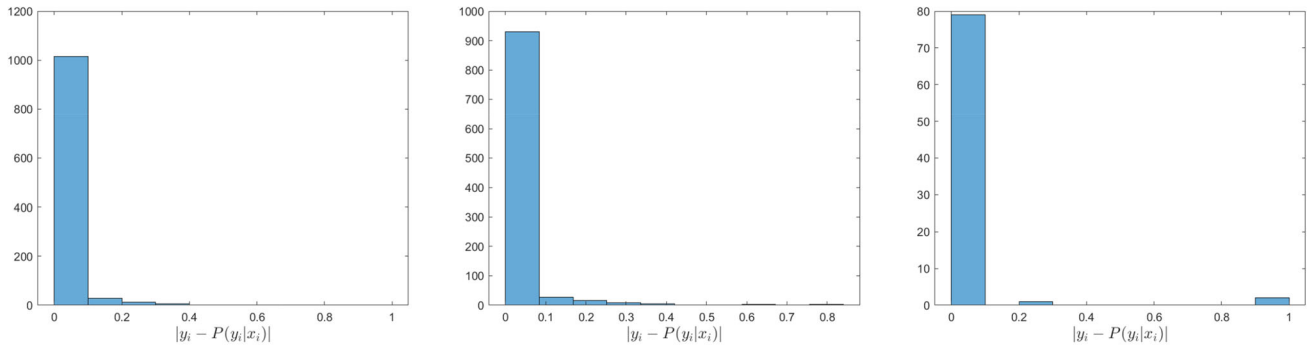


(a) BRIER SCORE **(b)** LOG-LOSS

Fig. 7 Performance of calibration algorithms

Table 6 STANDARD class distribution

Set	LFs patterns (positive class)	SFs patterns (negative class)
Calibration set	569	498
Test set	82	991



(a) Overall **(b)** Negative **(c)** Positive

Fig. 8 Residual error of probability estimates STANDARD test set

impact on the overall score, producing a Brier score for Platt scaling that is higher than the uncalibrated case.

Therefore, the available STANDARD Test set might be not the most appropriate set for evaluating the calibration functions learned from the balanced STANDARD calibration set. As stated above, in order to have a balanced STANDARD Test set, which is more reliable for assessing calibration results, we follow two approach:

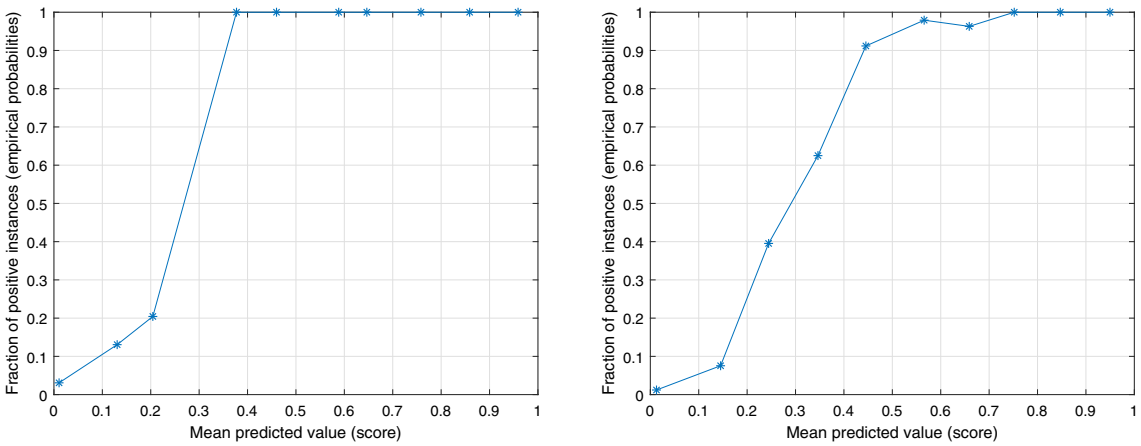
1. under-sampling majority class (SFs patterns);
2. over-sampling minority class (LFs patterns).

As regard the first approach, an under-sampled STANDARD Test set is built by taking 100 SF instances randomly chosen from the original STANDARD Test set and all of the LFs patterns. The new class distribution is 45% (LFs)—55% (SFs).

The over-sampling approach requires a more sophisticated procedure since features belong to different data types, as we described in Sect. 3.

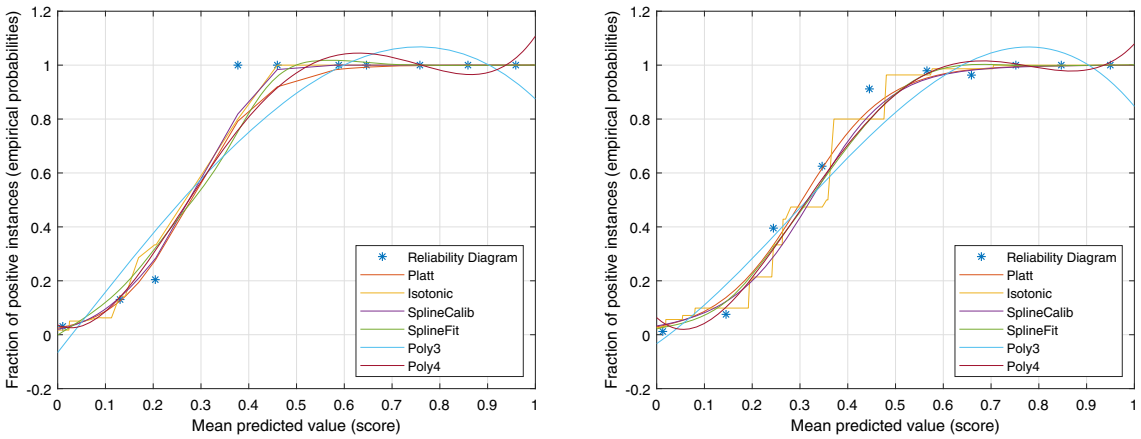
The procedure is the following:

- Split of STANDARD Test set in more subsets, in such way that each of them presents the same values for the categorical features.
- For each subset, a new set of LFs instances is generated, by using a different synthesis technique for each kind of data:



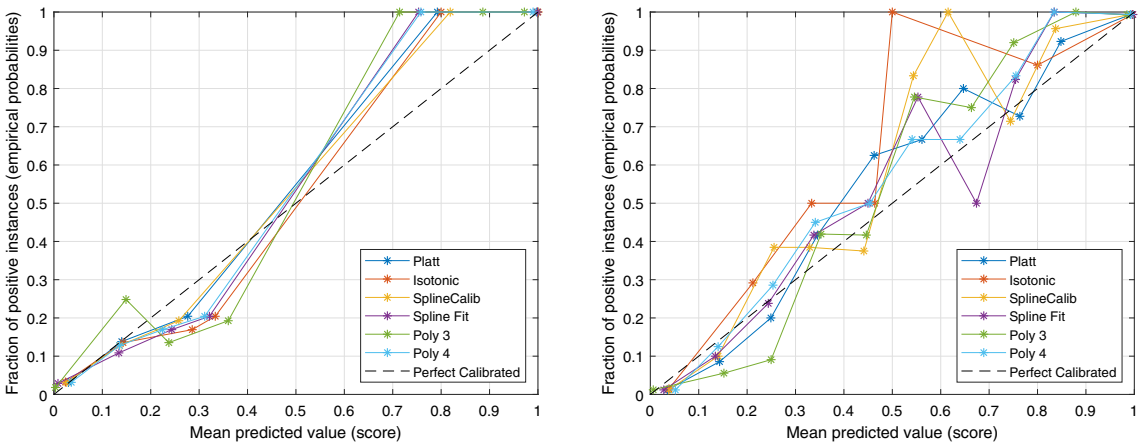
(a) STD (Under-sampled Test set) **(b)** STD (Over-sampled Test set)

Fig. 9 Reliability diagrams



(a) STD (Under-sampled Test set) **(b)** STD (Over-sampled Test set)

Fig. 10 Reliability diagrams vs. fitted curves



(a) STD Under-sampled Test set **(b)** STD Over-sampled Test set

Fig. 11 Reliability diagrams after calibration

Table 7 Brier score for the under-sampled and over-sampled test set (STANDARD dataset)

Method	Dataset std		
	Calibration set	Under-sampled test set	Over-sampled test set
Uncalibrated	0.05178 ± 0.011	0.03745 ± 0.011	0.02255 ± 0.012
PS	0.02809 ± 0.011	0.01490 ± 0.006	0.00850 ± 0.006
IR	0.02590 ± 0.011	0.01458 ± 0.007	0.00868 ± 0.006
SC	0.00962 ± 0.004	0.01412 ± 0.007	0.00856 ± 0.006
Poly3	0.02960 ± 0.012	0.01525 ± 0.007	0.00985 ± 0.006
Poly4	0.02837 ± 0.011	0.01464 ± 0.008	0.00908 ± 0.006
SplineFit	0.02771 ± 0.011	0.01398 ± 0.007	0.00869 ± 0.006

Table 8 Log-Loss score for the under-sampled and over-sampled test set (STANDARD dataset)

Method	Dataset std		
	Calibration set	Under-sampled test set	Over-sampled test set
uncalibrated	0.32532 ± 0.128	0.32817 ± 0.220	0.14809 ± 0.101
PS	0.10743 ± 0.039	0.07773 ± 0.034	0.04353 ± 0.025
IR	0.10743 ± 0.039	0.07459 ± 0.038	0.04748 ± 0.025
SC	0.10446 ± 0.041	0.07496 ± 0.041	0.04322 ± 0.026
Poly3	0.42450 ± 0.285	0.31294 ± 0.184	0.10037 ± 0.103
Poly4	0.14152 ± 0.038	0.11360 ± 0.108	0.06285 ± 0.059
SplineFit	0.15037 ± 0.053	0.13823 ± 0.111	0.06682 ± 0.056

1. Numerical values of the new pattern are generated using the well-known SMOTE technique (Chawla et al. 2002);
2. Categorical values of the new pattern are the same of the correspondent subset obtained grouping categories;
3. Time-series data that are in form of unevenly spaced sequences are obtained by adding Gaussian noise to the Time-series medoid of each subset built through categorical variables in Item 2 above.

– The new LF patterns are then concatenated with the original STANDARD Test set.

We add 910 new LF instances to the STANDARD Test set by applying this procedure. In the over-sampled STANDARD Test set the two classes are totally balanced (50–50%).

The OCC_System model trained for the STANDARD dataset is used for generating the sets of score for the two sampled Test set. Since the calibration set is the same, there is no need to re-train the calibration algorithms, but we only apply the already learned calibration functions to the two new Test sets in order to produce the corresponding probability estimates. In Fig. 9, we show the reliability diagrams for the under-sampled Test set and the over-sampled Test set for the STANDARD dataset. In both cases, the trend does not follow the $y = x$ diagonal line, confirming that the classifier is not well-calibrated for the STANDARD dataset. Figure 10 shows the resulting calibration functions overlaid the reliability diagram for the two Test sets, while in Fig. 11 we plot the reliability diagrams of the Test sets after calibration. For the under-sampled Test set (a) the trend of the reliability diagrams after calibration is close to the $y = x$ line, while for the over-sampled Test set (b) the reliability diagrams are slightly noisy.

For ease of comparison, in Tables 7 and 8, we show the two figure of merits (Brier score and Log-Loss score, respectively) on both the calibration set and the under-sampled and over-sampled Test sets as averages over five trials for STANDARD dataset. The performance on calibration set (reported again for better readability) obliviously is the same of Tables 4 and 5. By considering the performance on the two Test sets, firstly it is possible to see that the calibration algorithms work much better on them compared to the original STANDARD Test set, since they mostly performs better than the uncalibrated case. As concerns the under-sampled Test set, SplineFit reaches the best value for the Brier score, followed by Isotonic Regression. In terms of Log-Loss score, Isotonic Regression and SplineCalib outperform other methods. For the over-sampled scenario, Platt scaling and SplineCalib are the best methods. The former produces the lowest Brier score, while the latter reaches the lowest Log-Loss score, but actually their values are very close for both metrics.

Finally, in Table 9 is reported the performance in terms of execution time for the 7 calibration procedures and for each investigated dataset.

Table 9 Comparison of the execution time for the adopted calibration procedures, for each dataset (measures are in seconds)

Method	Nodes	Branches	Std
Platt	0.05897948	0.01953612	0.03728476
Isotonic	0.20002866	0.01791418	0.09415118
SplineCalib	41.45577378	9.57505456	14.718603
SplineFit	3.85662254	1.2260622	1.6423341
Poly3_fit	1.69367574	1.01460842	1.16662092
Poly4_fit	1.29023974	1.03621046	1.00169828

In general, Platt scaling is the fastest procedure in performing the calibration training at least on our experimented datasets, followed by Isotonic regression while, as expected, SplineClib was the slowest. Poly3_fit and Poly4_fit, although slower than Platt scaling and Isotonic regression, are very far from the SplineCalib execution time for all three datasets.

7 Conclusions

This study is part of a wider project concerning the design and implementation of a modeling and recognition system of faults and outages occurring in the real power grid managed by "Azienda Comunale Energia e Ambiente" (ACEA) company in Rome, Italy. The recognition system, based on a one-class classification approach as the main core of a larger system, has been developed within the "ACEA Smart Grid Project". An important task consists in extracting from the learned fault classification model, called OCC_System, useful information for programming and control procedures, such as the estimation of the financial risk associated with a set of power grid states and network resilience analysis. Reliable fault probability estimates are precious information in order to assess risks and make cost benefit analysis, related to maintenance planning and network expansion. The current paper moves along this direction presenting an approach based on calibration for estimating reliable fault probabilities from classification scores yield by OCC_System. In this study, we reviewed three state-of-the-art posterior calibration methods for calibrating the OCC_System classifier. The three techniques (PS, IR and SC) have been experimented on real-world data, coming from the MV power grid of Rome, against three simple methods (Poly3, Poly4 and SplineFit), which basically perform a plain curve fitting on the reliability diagram. Tests confirm that the three state-of-the-art calibration techniques well behaved for at least two experimented datasets and, especially in terms of Brier scores, results obtained by the proposed techniques are comparable with state of the art competitors. Furthermore, our work shows that calibrating a classifier is a challenging task, especially for custom-based classifiers adopted for predictive maintenance purposes and real-world datasets. Specifically, the performances evaluation can be challenging in unbalanced class problems and a solution is provided for this case. It is noted that the proposed method can meet some limitation with very high unbalanced datasets, despite the balancing procedure. Specifically, when the reliability diagram is populated with few points, or when the points are clustered at the extremes of the unit interval in the reliability diagram. In addition, as regards the polynomial methods presented here, although a smaller degree means having a less complex fitting model, such polynomials are subject to edge oscillations that violate the monotone growth property, especially at the edges of the unitary interval. This problem can be alleviated by increasing the polynomial degree or by using other more exact fitting methods. Further investigations are needed at this point. However, being the fitting methodologies less heavy that other methods, a trade-off between the accuracy and the computational complexity has to be taken into account in applications. Considering the performance for the three investigated datasets, we are confident for the application of the reviewed calibration techniques on the existing on-going faults recognition system. As future-work directions, further investigations will be carried out on the performance of calibration methods in real scenarios and in the big data environment. Moreover, the calibrated posterior probabilities will be adopted for computing the vulnerability and the risk associated with equipment, a fundamental tool for scenario analysis in predictive maintenance. In particular, posterior probabilities will be adopted for estimating simpler and interpretable models of faults (e.g., linear models) usable by non-machine learning experts in line with the explainable AI paradigm. This simpler models could be adopted also as a driver for correlation analysis between fault causes and power grid constitutive parameters.

Acknowledgements The authors wish to thank ACEA Distribuzione S.p.A. for providing the data and for their continuous support during the design and test phases. Special thanks to Ing. Stefano Liotta, Chief Network Operation Division, to Ing. Silvio Alessandroni, Chief Electric Power Distribution, and to Ing. Maurizio Paschero, Chief Remote Control

Funding Open access funding provided by Universitá degli Studi di Roma La Sapienza within the CRUI-CARE Agreement. The authors declare that no funds, grants, and other support were received during the preparation of this manuscript.

Data availability The datasets generated during and/or analyzed during the current study are not publicly available due to property of ACEA S.p.A. but are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

ACEA (2016) The ACEA smart grid pilot project (in Italian). https://ses.jrc.ec.europa.eu/acea-distribuzione-smart-grid-pilot-project

Afzal M, Pothamsetty V (2012) Analytics for distributed smart grid sensing. In: 2012 IEEE PES innovative smart grid technologies (ISGT), pp 1–7

Asuncion A, Newman D (2007) UCI machine learning repository

Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E (1955) An empirical distribution function for sampling with incomplete information. Ann Math Stat 26(4):641–647. http://www.jstor.org/stable/2236377

Bhattacharya B, Sinha A (2017) Intelligent fault analysis in electrical power grids. In: 2017 IEEE 29th international conference on tools with artificial intelligence (ICTAI). https://doi.org/10.1109/ictai.2017.00151

Blair CG, Thompson J, Robertson NM (2014) Introspective classification for pedestrian detection. In: 2014 sensor signal processing for defence (SSPD), pp 1–5

Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78(1):1. https://doi.org/10.1175/1520-0493

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: Synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953

Cremer JL, Strbac G (2019) A machine-learning based probabilistic perspective on dynamic security assessment. arXiv:1912.07477

Das B (2006) Fuzzy logic-based fault-type identification in unbalanced radial power distribution system. IEEE Trans Power Deliv 21(1):278–285

De Santis E, Livi L, Mascioli F, Sadeghian A, Rizzi A (2014) Fault recognition in smart grids by a one-class classification approach. In: Neural networks (IJCNN), 2014 international joint conference on, pp 1949–1956. https://doi.org/10.1109/IJCNN.2014.6889668

De Santis E, Rizzi A, Sadeghian A, Frattale Mascioli F (2015a) A learning intelligent system for fault detection in smart grid by a one-class classification approach. In: Neural networks (IJCNN), 2015 international joint conference on, pp 1–8. https://doi.org/10.1109/IJCNN.2015.7280756

De Santis E, Rizzi A, Sadeghian A, Mascioli F (2015b) A learning intelligent system for fault detection in smart grid by a one-class classification approach. In: 2015 international joint conference on neural networks (IJCNN), pp 1–8. https://doi.org/10.1109/IJCNN.2015.7280756

De Santis ED, Livi L, Sadeghian A, Rizzi A (2015c) Modeling and recognition of smart grid faults by a combined approach of dissimilarity learning and one-class classification. Neurocomputing 170:368–383. https://doi.org/10.1016/j.neucom.2015.05.112

De Santis E, Rizzi A, Sadeghian A (2017) A cluster-based dissimilarity learning approach for localized fault classification in smart grids. Swarm Evolut Comput. https://doi.org/10.1016/j.swevo.2017.10.007

De Santis E, Martino A, Rizzi A, Mascioli FMF (2018a) Dissimilarity space representations and automatic feature selection for protein function prediction. In: 2018 international joint conference on neural networks (IJCNN). IEEE, pp 1–8

De Santis E, Paschero M, Rizzi A, Mascioli FMF (2018b) Evolutionary optimization of an affine model for vulnerability characterization in smart grids. In: 2018 international joint conference on neural networks (IJCNN), pp 1–8. https://doi.org/10.1109/IJCNN.2018.8489749

De Santis E, Rizzi A, Sadeghian A (2018c) A cluster-based dissimilarity learning approach for localized fault classification in smart grids. Swarm Evol Comput 39:267–278

DeGroot MH, Fienberg SE (1983a) The comparison and evaluation of forecasters. J R Stat Soc Ser D (Stat) 32(1/2):12–22. http://www.jstor.org/stable/2987588

DeGroot MH, Fienberg SE (1983b) The comparison and evaluation of forecasters. J R Stat Soc Ser D (Stat) 32(1/2):12–22. http://www.jstor.org/stable/2987588

Dua D, Graff C (2019) UCI machine learning repository. University of California, School of Information and Computer Science. Irvine, CA. http://archive.ics.uci.edu/ml

Freedman DA, Diaconis P (1981) On the histogram as a density estimator: L2 theory. Z Wahrscheinlichkeitstheor Verwa Geb 57:453–476

Gosztolya G, Busa-Fekete R (2018) Calibrating adaboost for phoneme classification. Soft Comput. https://doi.org/10.1007/s00500-018-3577-z

Guikema SD, Davidson RA, Liu H (2006) Statistical models of the effects of tree trimming on power system outages. IEEE Trans Power Deliv 21(3):1549–1557

Gunning D (2017) Explainable artificial intelligence (XAI). Defense Adv Res Proj Agency (DARPA), nd Web 2:2

Hajek P, Godo L, Esteva F (2013) Fuzzy logic and probability. In: Proc of UAI'95

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer Series in Statistics, Springer, New York

Jiang H, Zhang JJ, Gao W, Wu Z (2014) Fault detection, identification, and location in smart grid based on data-driven computational methods. IEEE Trans Smart Grid 5(6):2947–2956

Khan SS, Madden MG (2010) A survey of recent trends in one class classification. In: Coyle L, Freyne J (eds) Artificial intelligence and cognitive science. Springer, Heidelberg, pp 188–197

Kordestani M, Saif M (2017) Data fusion for fault diagnosis in smart grid power systems. In: 2017 IEEE 30th Canadian conference on electrical and computer engineering (CCECE), pp 1–6

Lin HT, Lin CJ, Weng RC (2007) A note on Platt's probabilistic outputs for support vector machines. Mach Learn 68(3):267–276. https://doi.org/10.1007/s10994-007-5018-6

Lucena B (2018) Spline-based probability calibration. arXiv:1809.07751

Luo X, Kezunovic M (2008) Implementing fuzzy reasoning petri-nets for fault section estimation. IEEE Trans Power Deliv 23(2):676–685

Martino A, De Santis E, Baldini L, Rizzi A (2019) Calibration techniques for binary classification problems: a comparative analysis. In: IJCCI, pp 487–495. https://doi.org/10.5220/0008165504870495

Massaferro P, Martino JMD, Fernández A (2020) Fraud detection in electric power distribution: An approach that maximizes the economic return. IEEE Trans Power Syst 35(1):703–710

Mendel JM (1995) Fuzzy logic systems for engineering: a tutorial. Proc IEEE 83(3):345–377. https://doi.org/10.1109/5.364485

Min S-W, Sohn J-M, Park J-K, Kim K-H (2004) Adaptive fault section estimation using matrix representation with fuzzy relations. IEEE Trans Power Syst 19(2):842–848

Morise AP, Diamond GA, Detrano R, Bobbio M, Gunel E (1996) The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. Med Dec Mak 16(2):133–142. https://doi.org/10.1177/0272989X9601600205 (**PMID: 8778531**)

Murphy AH, Winkler RL (1977) Reliability of subjective probability forecasts of precipitation and temperature. Journal of the Royal Statistical Society Series C (Applied Statistics) 26(1):41–47, http://www.jstor.org/stable/2346866

Naeini MP, Cooper GF, Hauskrecht M (2015) Obtaining well calibrated probabilities using Bayesian binning. In: Proceedings of the 29th AAAI conference on artificial intelligence. AAAI Press, AAAI'15, pp 2901–2907. http://dl.acm.org/citation.cfm?id=2888116.2888120

Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on machine learning. ACM, New York, ICML '05, pp 625–63. https://doi.org/10.1145/1102351.1102430

Pereira T, Cardoso S, Guerreiro M, Mendonça A, Madeira SC (2020) Targeting the uncertainty of predictions at patient-level using an ensemble of classifiers coupled with calibration methods, Venn-ABERS, and conformal predictors: a case study in ad. J Biomed Inf 101:103350. https://doi.org/10.1016/j.jbi.2019.103350

Pimentel MAF, Clifton DA, Clifton LA, Tarassenko L (2014) A review of novelty detection. Signal Process 99:215–249

Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in large margin classifier. MIT Press, pp 61–74

Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ (2017) On fairness and calibration. arXiv:1709.02012

Raheja D, Llinas J, Nagi R, Romanowski C (2006) Data fusion/data mining-based architecture for condition-based maintenance. Int J Product Res 44(14):2869–2887. https://doi.org/10.1080/00207540600654509

Rizzi A, Frattale Mascioli FM, Baldini F, Mazzetti C, Bartnikas R (2009) Genetic optimization of a PD diagnostic system for cable accessories. IEEE Trans Power Deliv 24(3):1728–1738

Rudin C, Waltz D, Anderson RN, Boulanger A, Salleb-Aouissi A, Chow M, Dutta H, Gross PN, Huang B, Ierome S, Isaac DF, Kressner A, Passonneau RJ, Radeva A, Wu L (2012) Machine learning for the New York city power grid. IEEE Trans Pattern Anal Mach Intell 34(2):328–345

Samantaray SR (2009) Decision tree-based fault zone identification and fault classification in flexible ac transmissions-based transmission line. IET Gener, Trans Distrib 3(5):425–436

Schapire RE, Singer Y (1999) Improved boosting algorithms using confidence-rated predictions. Mach Learn 37(3):297–336

Scott DW (1979) On optimal and data-based histograms. Biometrika 66(3):605–610. http://www.jstor.org/stable/2335182

Shahid N, Aleem SA, Naqvi IH, Zaffar N (2012) Support vector machine based fault detection classification in smart grids. In: 2012 IEEE Globecom workshops, pp 1526–1531

Souza Pereira D, Almeida C, Kagan N (2018) Fault location in the smart grids context based on an evolutionary algorithm. J Control, Autom Electr Syst. https://doi.org/10.1007/s40313-018-0406-7

Steyerberg E, Borsboom G, van Houwelingen JH, Eijkemans M, Habbema J (2004) Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 23:2567–86. https://doi.org/10.1002/sim.1844

Sun J, Qin S-Y, Song Y-H (2004) Fault diagnosis of electric power systems based on fuzzy petri nets. IEEE Trans Power Syst 19(4):2053–2059

Thukaram D, Khincha HP, Vijaynarasimha HP (2005) Artificial neural network and support vector machine approach for locating faults in radial distribution systems. IEEE Trans Power Deliv 20(2):710–721

Tokel HA, Halaseh RA, Alirezaei G, Mathar R (2018) A new approach for machine learning-based fault detection and classification in power systems. In: 2018 IEEE power energy society innovative smart grid technologies conference (ISGT), pp 1–5

Van Calster B, McLernon D, van Smeden M, Wynants L, Steyerberg E (2019) Calibration: the achilles heel of predictive analytics. BMC Med. https://doi.org/10.1186/s12916-019-1466-7

Vovk V (2012) Venn predictors and isotonic regression. arXiv:1211.0025

Vovk V, Gammerman A, Shafer G (2005) Algorithmic learning in a random world. Springer, Boston, pp 17–51. https://doi.org/10.1007/b106715

Vovk V, Petej I (2014) Venn-abers predictors. In: Proceedings of the 30th conference on uncertainty in artificial intelligence, UAI'14. AUAI Press, Arlington, pp 829–838

Wahba G (1990) Spline models for observational data. Society for Industrial and Applied Mathematics, Philadelphia

Walsh C, Sharman K, Hripcsak G (2017) Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk. J Biomed Inf. https://doi.org/10.1016/j.jbi.2017.10.008

Wang Z, Zhao P (2009) Fault location recognition in transmission lines based on support vector machines. In: 2009 2nd IEEE international conference on computer science and information technology, pp 401–404

Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '02. ACM, New York, pp 694–699. https://doi.org/10.1145/775047.775151