Towards Robust Performance Guarantees for Models Learned from High-Dimensional Data

Rui Henriques and Sara C. Madeira

Abstract Models learned from high-dimensional spaces, where the high number of features can exceed the number of observations, are susceptible to overfit since the selection of subspaces of interest for the learning task is prone to occur by chance. In these spaces, the performance of models is commonly highly variable and dependent on the target error estimators, data regularities and model properties. High-variable performance is a common problem in the analysis of omics data, healthcare data, collaborative filtering data, and datasets composed by features extracted from unstructured data or mapped from multi-dimensional databases. In these contexts, assessing the statistical significance of the performance guarantees of models learned from these high-dimensional spaces is critical to validate and weight the increasingly available scientific statements derived from the behavior of these models. Therefore, this chapter surveys the challenges and opportunities of evaluating models learned from big data settings from the less-studied angle of big dimensionality. In particular, we propose a methodology to bound and compare the performance of multiple models. First, a set of prominent challenges is synthesized. Second, a set of principles is proposed to answer the identified challenges. These principles provide a roadmap with decisions to: *i)* select adequate statistical tests, loss functions and sampling schema, *ii)* infer performance guarantees from multiple settings, including varying data regularities and learning parameterizations, and *iii)* guarantee its applicability for different types of models, including classification and descriptive models. To our knowledge, this work is the first attempt to provide a robust and flexible assessment of distinct types of models sensitive to both the dimensionality and size of data. Empirical evidence supports the relevance of these principles as they offer a coherent setting to bound and compare the performance of models learned in high-dimensional spaces, and to study and refine the behavior of these models.

Key words: high-dimensional data, performance guarantees, statistical significance of learning models, error estimators, classification, biclustering

KDBIO, INESC-ID, Instituto Superior Técnico, Universidade de Lisboa
{rmch,sara.madeira}@ist.utl.pt

1 Introduction

High-dimensional data has been increasingly adopted to derive implications from the analysis of biomedical data, social networks or multi-dimensional databases. In high-dimensional spaces, it is critical to guarantee that the learned relations are statistically significant, that is, they are not learned by chance. This is particularly important when these relations are learned from subspaces of the original space and when the number of observations is not substantially larger than the number of features. Examples of data where the number of observations does not significantly exceed the number of features include collaborative filtering data, omics data (such as gene expression data, structural genomic variations and biological networks), clinical data (such as data integrated from health records, functional magnetic resonances and physiological signals), and random fields (Amaratunga, Cabrera, and Shkedy 2014). In order to bound or compare the performance of models composed by multiple relations, the impact of learning in these high-dimensional spaces on the statistical assessment of these models needs to be properly considered.

Despite the large number of efforts to study the effects of dimensionality and data size (number of instances) on the performance of learning models (Kanal and Chandrasekaran 1971; Jain and Chandrasekaran 1982; Raudys and Jain 1991; Adcock 1997; Vapnik 1998; Mukherjee et al. 2003; Hua et al. 2005; Dobbin and Simon 2007; Way et al. 2010; Guo et al. 2010), an integrative view of their potentialities and limitations is still lacking. In this chapter, we identify a set of major requirements to assess the performance guarantees of models learned from high-dimensional spaces and survey critical principles for their adequate satisfaction. These principles can also be applied to affect the learning methods and to estimate the minimum sample size that guarantees the inference of statistical significant relations.

Some of the most prominent challenges for this task are the following. First, assessing the performance of models based on simulated surfaces and on fitted learning curves often fail to provide robust statistical guarantees. Typically under these settings, the significance of the estimations is tested against loose models learned from permuted data and the performance guarantees are not affected by the variability of the observed errors (Mukherjee et al. 2003; Way et al. 2010). Second, many of the existing assessments assume independence among features (Dobbin and Simon 2005; Hua et al. 2005). This assumption does not hold for datasets in high-dimensional spaces where the values of few features can discriminate classes by chance. This is the reason why learning methods that rely on subsets of the original features, such as rule-based classifiers, have higher variance on the observed errors, degrading the target performance bounds. Third, error estimators are often inadequate since the common loss functions for modeling the error are inappropriate and the impact of test sample size is poorly studied leading to the collection error estimates without statistical significance (Beleites et al. 2013). Fourth, assessment methods from synthetic data commonly rely on simplistic data distributions, such as multivariate Gaussian class-conditional distributions (Dobbin and Simon 2007). However, features in real-world data (biomedical features such as proteins, metabolites, genes, physiological features, etc.) exhibit highly skewed mixtures of distributions (Guo et al. 2010). Finally, existing methods are hardly extensible towards more flexible settings, such as the performance evaluations of descriptive models (focus on a single class) and of classification models in the presence of multiple and unbalanced classes.

In this context, it is critical to define principles that are able to address these drawbacks. In this chapter, we rely on existing contributions and on additional empirical evidence to derive these structural principles. Additionally, their integration through a new methodology is discussed. Understandably, even in the presence of datasets with identical sample size and dimensionality, the performance is highly dependent on data regularities and learning setting as they affect the underlying significance and composition of the learned relations. Thus, the proposed methodology is intended to be able to establish both data-independent and data-dependent assessments. Additionally, it is suitable for distinct learning tasks in datasets with either single or multiple classes. Illustrative tasks include classification of tumor samples, prediction of healthcare needs, biclustering of genes, proteomic mass spectral classification, chemosensitivity prediction, survival analysis, or putative class discovery using clustering.

The proposed assessment methodology offers three new critical contributions to the big data community:

- integration of statistical principles to provide a solid foundation for the definition of robust estimators of the true performance of models learned in high-dimensional spaces, including adequate loss functions, sampling schema (or parametric estimators), statistical tests and strategies to adjust performance guarantees in the presence of high variance and bias of performance;
- inference of general performance guarantees for models tested over multiple high-dimensional datasets with varying regularities;
- applicability for different types of models, including classification models with class-imbalance, regression models, and local and global descriptive models.

This chapter is organized as follows. Below, we provide the background required for the definition and comprehension of the target task – assessing models learned from high-dimensional spaces. *Section 2* surveys research streams with important contributions for this task, covering their major challenges. *Section 3* introduces a set of key principles derived from existing contributions to address the identified challenges. These are then coherently integrated within a simplistic assessment methodology. *Section 4* discusses the relevance of these principles based on experimental results and existing literature. Finally, concluding remarks and future research directions are synthesized.

1.1 Problem Definition

Consider a dataset described by n pairs (x_i, y_i) from (X, Y), where $x_i \in \mathbb{R}^m$ and Y is either described by a set of labels $y_i \in \Sigma$ or numeric values $y_i \in \mathbb{R}$. A space described by $n \in \mathbb{N}$ observations and $m \in \mathbb{N}$ features is here referred as a (n, m)-space, $X^{n,m} \subseteq X$.

Assuming that data is characterized by a set of underlying stochastic regularities, $P_{X|Y}$, a learning task aims to infer a model M from a (n, m)-space such that the error over $P_{X|Y}$ is minimized.

The M model is a composition of relations (or abstractions) from the underlying stochastic regularities. Two major types of models can be considered.

First, *supervised models*, including classification models ($M : X \to Y$, where Y=Σ is a set of categoric values) and regression models ($M : X \to Y$, with Y=$\mathbb{R}$), focus on the discriminative aspects of the conditional regularities $P_{X|Y}$ and their error is assessed recurring to loss functions (Toussaint 1974). Loss functions are typically based on accuracy, area under *roc*-curve or sensitivity metrics for classification models, and on the normalized or root mean squared errors for regression models. In supervised settings, there are two major types of learning paradigms with impact on the assessment of performance: *i)* learning a relation from all features, including multivariate learners based on discriminant functions (Ness and Simpson 1976), and *ii)* learning a composition of relations inferred from specific subspaces $X^{q,p} \subseteq X^{n,m}$ of interest (e.g. rule-based learners such as decision trees and Bayesian networks). For the latter case, capturing the statistical impact of feature selection is critical since small subspaces are highly prone to be discriminative by chance (Iswandy and Koenig 2006).

To further clarify the impact of dimensionality when assessing the performance of these models, consider a subset of original features, $X^{n,p} \subseteq X^{n,m}$, and a specific class or real interval, $y \in Y$. Assuming that these discriminative models can be decomposed in mapping functions of the type $M : X^{n,p} \to y$, comparing or bounding the performance of these models needs to consider the fact that the $(n$,$p)$-space is not selected aleatory. Instead, this subspace is selected as a consequence of an improved discriminatory power. In high-dimensional spaces, it is highly probable that a small subset of the original features is able to discriminate a class by chance. When the statistical assessment is based on error estimates, there is a resulting high-variability of values across estimates that needs to be considered. When the statistical assessment is derived from the properties of the model, the effect of mapping the original $(n$,$m)$-space into a $(n$,$p)$-space needs to be consider.

Second, *descriptive models* ($|Y|$=1) either globally or locally approximate P_X regularities. The error is here measured either recurring to merit functions or match scores when there is knowledge regarding the underlying regularities. In particular, a local descriptive model is a composition of learned relations from subspaces of features J=$X^{n,p} \subseteq X^{n,m}$, samples I=$X^{q,m} \subseteq X^{n,m}$, or both (I, J). Thus, local models define a set of k (bi)clusters such that each (bi)cluster (I_k, J_k) satisfies specific criteria of homogeneity. Similarly to supervised models, it is important to guarantee a robust collection and assessment of error estimates or, alternatively, that the selection of the

(q_k,p_k)-space of each (bi)cluster (where q_k=$|I_k|$ and p_k=$|J_k|$) is statistical significant, that is, the observed homogeneity levels for these subspaces do not occur by chance.

Consider that the asymptotic probability of misclassification of a particular model M is given by ε_{true}, and a non-biased estimator of the observed error in a (n,m)-space is given by $\theta(\varepsilon_{true})$. The problem of computing the *performance guarantees* for a specific model M in a (n,m)-space can either be given by its performance bounds or by the verification of its ability to perform better than other models. The task of computing the $(\varepsilon_{min},\varepsilon_{max})$ *performance bounds* for a M model in a (n,m)-space can be defined as:

$$[\varepsilon_{min},\varepsilon_{max}] : P(\varepsilon_{min} < \theta(\varepsilon_{true}) < \varepsilon_{max} \mid n,m,M,P_{X|Y}) = 1-\delta, \qquad (1)$$

where the performance bounds are intervals of confidence tested with 1-δ statistical power.

The task of *comparing a set of models* $\{M_1,..,M_l\}$ in a (n,m)-space can be defined as the discovery of significant differences in performance between groups of models while controlling the family-wise error, the probability of making one or more false comparisons among all the $l \times l$ comparisons.

Defining an adequate estimator of the true error $\theta(\varepsilon_{true})$ for a target $(n,m,M,P_{X|Y})$ setting is, thus, the central role of these assessments.

In literature, similar attempts have been made for testing the minimum number of observations, by comparing the estimated error for n observations with the true error, $min_n : P(\theta_n(\varepsilon_{true})<\varepsilon_{true} \mid m,M,P_{X|Y}) >$ 1-δ rejected at α, or by allowing relaxation factors $\theta_n(\varepsilon_{true})<(1+\gamma)\varepsilon_{true}$ when the observed error does not rapidly converge to ε_{true}, $\lim_{n\to\infty}\theta_n(\varepsilon_{true}) \neq \varepsilon_{true}$. In this context, the ε_{true} can be theoretically derived from assumptions regarding the regularity $P_{X|Y}$ or experimentally approximated using the asymptotic behavior of learning curves estimated from data.

To illustrate the relevance of target performance bounding and comparison tasks, let us consider the following model: a linear hyperplane $M(x)$ in $\mathbb{R}^m$ defined by a vector w and point b to either separate two classes, $sign(w \cdot x + b)$, predict a real-value, $w \cdot x + b$, or globally describe the observations, $X \sim w \cdot x + b$. In contexts where the number of features exceeds the number of observations ($m > n$), these models are not able to generalize (perfect overfit towards data). As illustrated in Fig.1, a linear hyperplane in $\mathbb{R}^m$ can perfectly model up to $m+1$ observations, either as classifier $X \rightarrow \{\pm 1\}$, as regression $X \rightarrow \mathbb{R}$ or as descriptor of X. Thus, a simple assessment of the errors of these models using the same training data would lead to $\theta(\varepsilon_{true})$=0 without variance across estimates ε_i and, consequently, to ε_{min}=ε_{max}=0, which may not be true in the presence of an additional number of observations. Also, the performance of these models using new testing observations tends to be high-variable. These observations should be considered when selecting the assessment procedure, including the true error estimator $\theta(\varepsilon_{true})$, the statistical tests and the assumptions underlying data and the learning method.

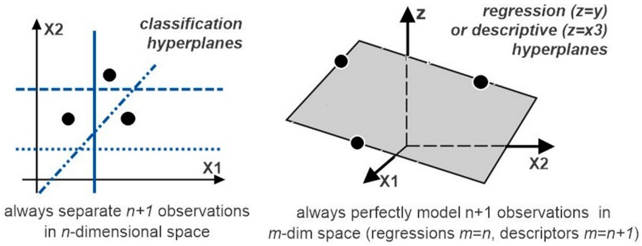
X2
X1
classification hyperplanes
always separate n+1 observations in n-dimensional space
z
X1
X2
regression (z=y) or descriptive (z=x3) hyperplanes
always perfectly model n+1 observations in m-dim space (regressions m=n, descriptors m=n+1)

Fig. 1: Linear hyperplanes cannot generalize when dimensionality is larger than the number of observations (data size), $m \geq n+1$.

2 Related Work

Classic statistical methods to *bound the performance* of models as a function of the data size include power calculations based on frequentist and Bayesian methods (Adcock 1997), deviation bounds (Guyon et al. 1998), asymptotic estimates of the true error ε_{true} (Raudys and Jain 1991; Niyogi and Girosi 1996), among others (Jain and Chandrasekaran 1982). Here, the impact of the data size in the observed errors is essentially dependent on the entropy associated with the target (n,m)-space. When the goal is the *comparison of multiple models*, Wilcoxon signed ranks test (two models) and the Friedman test with the corresponding post-hoc tests (more than two models) are still state-of-the-art methods to derive comparisons either from error estimates or from the performance distributions given by classic statistical methods (Demšar 2006; García and Herrera 2009).

To generalize the assessment of performance guarantees for an unknown sample size n, learning curves (Mukherjee et al. 2003; Figueroa et al. 2012), theoretical analysis (Vapnik 1998; Apolloni and Gentile 1998) and simulation studies (Hua et al. 2005; Way et al. 2010) have been proposed. A critical problem with these latter approaches is that they either ignore the role of dimensionality in the statistical assessment or the impact of learning from subsets of overall features.

We have grouped these existing efforts according to six major streams of research: *1)* classic statistics, *2)* risk minimization theory, *3)* learning curves, *4)* simulation studies, *5)* mutivariate model's analysis, and *6)* data-driven analysis. Existing approaches have their roots on, at least, one of these research streams. These streams of research assess the performance significance of a single learning model as a function of the available data size, which is a key factor when learning from high-dimensional spaces. Understandably, comparing multiple models is a matter of defining robust statistical tests from the assessed performance per model.

First, *classic statistics* cover a wide-range of methods. They are either centered on power calculations (Adcock 1997) or on the asymptotic estimates of ε_{true} by using approximation theory, information theory and statistical mechanics (Raudys and Jain 1991; Opper et al. 1990; Niyogi and Girosi 1996). Power calculations provide a critical view on the model errors (performance) by controlling both sample size n and statistical power 1-γ, $P(\theta_n(\varepsilon_{true}) < \varepsilon_{true})$=1-$\gamma$, where $\theta_n(\varepsilon_{true})$ can either rely on a frequentist view, from counts to estimate the discriminative/descriptive ability of subsets of features, or on a Bayesian view, more prone to deal with smaller and noisy data (Adcock 1997).

Second, *theoretical analysis of empirical risk minimization* (Vapnik 1982; Apolloni and Gentile 1998). To understand the concept of risk minimization, consider two distinct models: one simplistic model achieving good generalization but with high observed error, and a model able to minimize the observed error but overfitted to the available data. As illustrated in Fig.2, this analysis aims to minimize the risk by finding an optimal trade-off between the model capacity (or complexity term) and the observed error. Core contributions from this research stream comes from Vapnik-Chervonenkis (VC) theory (Vapnik 1998), where the sample size and the dimensionality is related through the VC-dimension (h), a measure of the model capacity that defines the minimum number of observations required to generalize the learning in a m-dimensional space. As we illustrated in Fig.1, linear hyperplanes have $h = m+1$. The VC-dimension can be theoretically or experimentally estimated for different models and used to compare the performance of models and approximate lower-bounds. Although under this stream the target overfitting problem is addressed, the resulting assessment tends to be conservative.

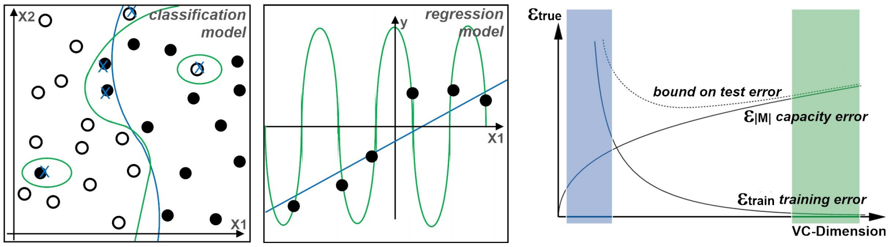

Fig. 2: Capacity and training error impact on true error estimation for classification and regression models

Third, *learning curves* use the observed performance of a model over a given dataset to fit inverse power-law functions that can extrapolate performance bounds as a function of the sample size or dimensionality (Mukherjee et al. 2003; Boonyanunta and Zeephongsekul 2004). An extension that weights estimations according to their confidence has been applied for medical data (Figueroa et al. 2012). However, the estimation of learning curves in high-dimensional spaces requires large data (n>m), which are not always available, and does not consider the variability across error estimates.

Fourth, *simulation studies* infer performance guarantees by studying the impact of multiple parameters on the learning performance (Hua et al. 2005; Way et al. 2010; Guo et al. 2010). This is commonly accomplished through the adoption of a large number of synthetic datasets with varying properties. Statistical assessment and inference over the collected results can be absent. This is a typical case when the simulation study simply aims to assess major variations of performance across settings.

Fifth, true performance estimators can be derived from a *direct analysis of the learning models* (Ness and Simpson 1976; El-Sheikh and Wacker 1980; Raudys and Jain 1991; Raudys 1997). This stream of research is mainly driven by the assessment of multivariate models that preserve the dimensionality of space (whether described by the original m features or by a subset of the original features after feature selection) when specific regularities underlying data are assumed. Illustrative models include classifiers based on discriminant functions, such as Euclidean, Fisher, Quadratic or Multinomial. Unlike learning models based on tests over subsets of features selected from the original high-dimensional space, multivariate learners consider the values of all features. Despite the large attention given by the multivariate analysis community, these models only represent a small subset of overall learning models.

Finally, model-independent size decisions derived from data regularities are reviewed and extended by Dobbin and Simon (2005; 2007). *Data-driven formulas* are defined from a set of approximations and assumptions based on dimensionality, class prevalence, standardized fold change, and on the modeling of non-trivial sources of errors. Although dimensionality is used to affect both the testing significance levels and the minimum number of features (i.e., the impact of selecting subspaces is considered), the formulas are independent from the selected models, forbidding their extension for comparisons or the computation of performance bounds.

These six research streams are closely related and can be mapped through concepts of information theory. In fact, an initial attempt to bridge contributions from statistical physics, approximation theory, multivariate analysis and VC theory within a Bayesian framework was proposed by Haussler, Kearns, and Schapire (1991).

2.1 Challenges and Contributions

Although each of the introduced research streams offer unique perspectives to solve the target task, they suffer from drawbacks as they were originally developed with a different goal – either minimum data size estimation or performance assessments in spaces where $n \gg m$. These drawbacks are either related with the underlying approximations, with the assessment of the impact of selecting subspaces (often related with a non-adequate analysis of the variance of the observed errors) or with the poor extensibility of existing approaches towards distinct types of models or flexible data settings. Table 1 details these drawbacks according to three major categories that define the ability to: *A)* rely on robust statistical assessments, *B)* deliver performance guarantees from multiple flexible data settings, and *C)* extend the target assessment towards descriptive models, unbalanced data, and multi-parameter settings. The latter two categories trigger the additional challenge of inferring per-

Category	Problem Description
A. Statistical Robustness	1. Non-robust estimators of the true performance of models. First, the probability of selecting informative features by chance is higher in high-dimensional spaces, leading to an heightened variability of error estimates and, in some cases, turning inviable the inference of performance guarantees. Second, when the number of features exceeds the number of observations, errors are prone to systemic biases. The simple use of mean and deviation metrics from error estimates to compare and bound the performance is insufficient in these spaces; 2. Inappropriate sampling scheme for the collection of error estimates in high-dimensional spaces (Beleites et al. 2013). Assessing the variance of estimations within and across folds, and the impact of the number of folds and test sample size is critical to tune the level of conservatism of performance guarantees; 3. Inadequate loss functions to characterize the observed error. Examples of loss functions of interest that are commonly ignored include sensitivity for unbalanced classification settings (often preferred against accuracy) or functions that provide a decomposition of errors; 4. Inadequate underlying density functions to test the significance of error estimates. Significance is typically assessed against very loose null settings (Mukherjee et al. 2003), and rarely assessed over more meaningful settings. Additionally, many of the proposed estimators are biased (Hua et al. 2005); 5. Others: approximated and asymptotic error estimators derived from multivariate model analysis (Raudys and Jain 1991) are only applicable for a specific subset of learning models; model-independent methods, such as formulae-based methods for minimum size estimation (Dobbin and Simon 2005), are non-extensible to compare models or bound performance; performance guarantees provided by a theoretical analysis of the learning properties, such as in VC-theory (Vapnik 1982), tend to be very conservative; dependency of large datasets to collect feasible estimates (Mukherjee et al. 2003);
B. Data Flexibility	1. Performance guarantees are commonly only assessed in the context of a specific dataset (e.g. classic statistics, learning curves), and, therefore, the implied performance observations cannot be generalized; 2. Performance comparisons and bounds are computed without assessing the regularities underlying the inputted data (Guo et al. 2010). These regularities provide a context to understand the learning challenges of the task and, thus, providing a frame to assess the significance of the scientific implications; 3. Contrasting with data size, dimensionality is rarely considered a variable to compare and bound models' performance (Jain and Chandrasekaran 1982). Note that dimensionality m and performance $\theta(\varepsilon_{true})$ are co-dependent variables as it is well-demonstrated by the VC theory (Vapnik 1998); 4. Independence among features is assumed in some statistical assessments. However, most of biomedical features (such as molecular units) and extracted features from collaborative data are functionally correlated; 5. Non-realistic synthetic data settings. Generated data should follow properties of real data, which is characterized by mixtures of distributions with local dependencies, skewed features and varying levels of noise; 6. The impact of modeling additional sources of variability, such as pooling, dye-swap samples and technical replicates for biomedical settings, is commonly disregarded (Dobbin and Simon 2005);
C. Extensibility	1. Inadequate statistical assessment of models learned from datasets with heightened unbalance among classes and non-trivial conditional distributions $P_{X\|y_i}$; 2. Weaker guidance for computing bounds for multi-class models ($\|\Sigma\|> 2$); 3. Existing methods are not extensible to assess the performance bounds of descriptive models, including (single-class) global and local descriptive models; 4. Lack of criteria to establish performance guarantees from settings where the impact of numerous parameters is studied (Hua et al. 2005; Way et al. 2010);

Table 1: Common challenges when defining performance guarantees of models learned from high-dimensional data

Approach	Major Problems (non-exhaustive observations)
Bayesian & Frequentist Estimations	Originally proposed for the estimation of minimum data size and, thus, not prepared to deliver performance guarantees; Applied in the context of a single dataset; Impact of feature selection is not assessed; No support as-is for descriptive tasks and hard data settings;
Theoretical Methods	Delivery of worst-case performance guarantees; Learning aspects need to be carefully modeled (complexity); Guarantees are typically independent from data regularities (only the size and dimensionality of the space are considered); No support as-is for descriptive tasks and hard data settings;
Learning Curves	Unfeasible for small datasets or high-dimensional spaces where $m>n$; Dimensionality and the variability of errors does not explicitly affect the curves; Guarantees suitable for a single input dataset; No support as-is for descriptive tasks and hard data settings;
Simulation Studies	Driven by error minimization and not by the statistical significance of performance; Data often rely on simplistic conditional regularities (optimistic data settings); Poor guidance to derive decisions from results;
Multivariate Analysis	Limited to multivariate models from discriminant functions; Different models require different parametric analyzes; Data often rely on simplistic conditional regularities; No support as-is for descriptive tasks and hard data settings; Approximations can lead to loose bounds;
Data-driven Formula	Not able to deliver performance guarantees (model-independent); Estimations only robust for specific data settings; Independence among features is assumed; Suitable for a single inputted dataset; Unfeasible for small samples;

Table 2: Limitations of existing approaches according to the introduced challenges

formance guarantees from multiple settings where data regularities and model parameters are varied.

Since each one of the introduced streams of research were developed with a specific goal and under a large set of assumptions, it is natural that their limitations fall into several of the identified categories in Table 1. In Table 2, we identify the major problems of these streams of research when answering the target tasks.

Although the surveyed challenges are lengthy in number, many of them can be answered recurring to contributions provided in literature. In Table 3, we describe illustrative sets of contributions that can be used to satisfy the requirements derived from the identified limitations. This analysis triggers the need to isolate sets of principles per challenge and to see whether it is possible to integrate these dispersed contributions for the development of more robust comparisons and bound estimation procedures.

Requirements	Contributions
Guarantees from High-Variable Performance (A.1)	Statistical tests to bound and compare performance sensitive to error distributions and loss functions (Martin and Hirschberg 1996; Qin and Hotilovac 2008; Demšar 2006); VC theory and discriminant-analysis (Vapnik 1982; Raudys and Jain 1991); Unbiasedness principles from feature selection (Singhi and Liu 2006; Iswandy and Koenig 2006);
Bias Effect (A.1)	Bias-Variance decomposition of the error (Domingos 2000);
Adequate Sampling Schema (A.2)	Criteria for sampling decisions (Dougherty et al. 2010; Toussaint 1974); Test-train splitting impact (Beleites et al. 2013; Raudys and Jain 1991);
Expressive Loss Functions (A.3)	Error views in machine learning (Glick 1978; Lissack and Fu 1976; Patrikainen and Meila 2006);
Feasibility (A.4)	Significance of estimates against baseline settings (Adcock 1997; Mukherjee et al. 2003);
Flexible Data Settings (B.1/4/5)	Simulations with hard data assumptions: mixtures of distributions, local dependencies and noise (Way et al. 2010; Hua et al. 2005; Guo et al. 2010; Madeira and Oliveira 2004);
Retrieval of Data Regularities (B.2)	Data regularities to contextualize assessment (Dobbin and Simon 2007; Raudys and Jain 1991);
Dimensionality Effect (B.3)	Extrapolate guarantees by sub-sampling features (Mukherjee et al. 2003; Guo et al. 2010);
Advanced Data Properties (B.6)	Modeling of additional sources of variability (Dobbin and Simon 2005);
Unbalanced/Difficult Data (C.1)	Guarantees from unbalanced data and adequate loss functions (Guo et al. 2010; Beleites et al. 2013);
Multi-class Tasks (C.2)	Integration of class-centric performance bounds (Beleites et al. 2013);
Descriptive Models (C.3)	Adequate loss functions and collection of error estimates for global and (bi)clustering models (Madeira and Oliveira 2004; Hand 1986);
Guidance Criteria (C.4)	Weighted optimization methods for robust and compact multi-parameter analysis (Deng 2007);

Table 3: Contributions with potential to satisfy the target set of requirements

3 Principles to Bound and Compare the Performance of Models

The solution space is proposed according to the target tasks of bounding or comparing the performance of a model M learned from high-dimensional spaces. The adequate definition of estimators of the true error is the central point of focus. An illustrative simplistic estimator of the performance of classification model can be described by a collection of observed errors obtained under a k-fold cross-validation, with its expected value being their average:

$E[\theta(\varepsilon_{true})] \approx \frac{1}{k}\Sigma_{i=1}^{k}(\varepsilon_i \mid M, n, m, P_{X|Y})$, where ε_i is the observed error for the i^{th} fold. When the number of observations is not significantly large, the errors can be collected under a leave-one-out scheme, where k=n and the ε_i is, thus, simply given by a loss function L applied over a single testing instance (x_i, y_i): $L(M(x_i)$=$\hat{y}_i, y_i)$.

In the presence of a estimator for the true error, finding performance bounds can rely on non-biased estimators from the collected error estimates, such as the mean and q-percentiles to provide a bar-envelope around the mean estimator (e.g. $q \in \{20\%, 80\%\}$). However, such strategy does not robustly consider the variability of the observed errors. A simple and more robust alternative is to derive the confidence intervals for the expected true performance based on the distribution underlying the observed error estimates.

Although this estimator considers the variability across estimates, it still may not reflect the true performance bounds of the model due to poor sampling and loss function choices. Additionally, when the number of features exceeds the number of observations, the collected errors can be prone to systemic biases and even statistically inviable for inferring performance guarantees. These observations need to be carefully considered to shape the statistical assessment.

The definition of good estimators is also critical for comparing models, as these comparisons can rely on their underlying error distributions. For this goal, either the traditional t-Student, McNemar and Wilcoxon tests can be adopted to compare pairs of classifiers, and Friedman tests with the corresponding post-hoc tests (Demšar 2006) or less conservative tests[1] (García and Herrera 2009) can be adopted for either comparing distinct models, models learned from multiple datasets or models with different parameterizations.

Motivated by the surveyed contributions to tackle the limitations of existing approaches, this section derives a set of principles for a robust assessment of the performance guarantees of models learned from high-dimensional spaces. First, these principles are incrementally provided according to the introduced major sets of challenges. Second, we show that these principles can be consistently and coherently combined within a simplistic assessment methodology.

3.1 Robust Statistical Assessment

Variability of Performance Estimates. Increasing the dimensionality m for a fixed number of observations n introduces variability in the performance of the learned model that must be incorporated in the estimation of performance bounds for a specific sample size. A simplistic principle is to compute the confidence intervals from error estimates $\{\varepsilon_1, .., \varepsilon_k\}$ obtained from k train-test partitions by fitting an underlying distribution (e.g. Gaussian) that is able to model their variance.

However, this strategy two major problems. First, it assumes that the variability is well-measured for each error estimate. This is commonly not true as each error

[1] Friedman tests rely on pairwise Nemenyi tests that are conservative and, therefore, may not reveal a significant number of differences among models (García and Herrera 2009)

estimate results from averaging a loss function across testing instances within a partitioning fold, which smooths and hides the true variability. Second, when the variance across estimates is substantially high, the resulting bounds and comparisons between models are not meaningful. Thus, four additional strategies derived from existing research are proposed: one for robust assessments for models that preserve the original dimensionality, another for correcting performance guarantees for models that rely on subspaces of the original space, a third strategy to reduce variability in $m \gg n$ settings, and a final strategy for obtaining more conservative guarantees.

First, the discriminant properties of multivariate models learned over the original space can be used to approximate the observed error for a particular setting $\theta_n(\varepsilon_{true} \mid m, M, P_{X|Y})$ and the asymptotic estimate of the true error $\lim_{n\rightarrow\infty} \theta_n(\varepsilon_{true} \mid m, M, P_{X|Y})$ (Ness and Simpson 1976). An analysis on the deviations of the observed error from the true error as a function of data size n, dimensionality m and discriminant functions M was initially provided by Raudys and Jain (1991) and extended by more recent approaches (Bühlmann and Geer 2011; Cai and Shen 2010).

Second, the unbiasedness principle from feature selection methods can be adopted to affect the significance of performance guarantees. Learning models M that rely on decisions over subsets of features either implicitly or explicitly use a form of feature selection driven by core metrics, such as Mahalanobis, Bhattacharyya, Patrick-Fisher, Matusita, divergence, mutual Shannon information, and entropy. In this context, statistical tests can be made to guarantee that the value of a given metric per feature is sufficiently better than a random distribution of values when considering the original dimensionality (Singhi and Liu 2006; Iswandy and Koenig 2006). These tests return a p-value that can be used to weight the probability of the selected set of features being selected by chance over the (n, m)-space and, consequently, to affect the performance bounds and the confidence of comparisons of the target models. Singhi and Liu (2006) formalize selection bias, analyze its statistical properties and how they impact performance bounds.

Third, when error estimates are collected, different methods have been proposed for controlling the observed variability across estimates (Raeder, Hoens, and Chawla 2010; Jain et al. 2003), ranging from general principles related with sampling schema and density functions to more specific statistical tests for a correct assessment of the true variability in specific biomedical settings where, for instance, replicates are considered. These options are revised in detail in the next subsections.

Fourth, conservative bounds for a given dimensionality can be retrieved from the VC-dimension (capacity) of a target model (Vapnik 1982; Blumer et al. 1989). These bounds can be used to guide model comparison. The VC-dimension can be obtained either theoretically or experimentally (Vayatis and Azencott 1999). A common experimental estimation option for the VC-dimension is to study the maximum deviation of error rates among independently labeled datasets. An illustrative lower-bound for the estimator of the true performance of a M model composed by h mapping functions (number of decisions from the values of m features) is: $\theta_n(\varepsilon_{true}) \geq \frac{1}{n}(log\frac{1}{\delta} + logh)$ (Apolloni and Gentile 1998), where δ is the statistical power[2]. In high-dimensional spaces, h tends to be larger, which can degrade performance bounds if the number of instances is small. For more complex models, such as Bayesian learners or decision trees, the VC-dimension can be adopted using assumptions that lead to less conservative bounds[3] (Apolloni and Gentile 1998). Still bounds tend to be loose as they are obtained using a data-independent analysis and rely on a substantial number of approximations.

Bias associated with High-Dimensional Spaces. In (n,m)-spaces where $n < m$, the observed error associated with a particular model can be further decomposed in bias and variance components to understand the major cause of the variability across error estimates. While variance is determined by the ability to generalize a model from the available observations (see Fig.2), the bias is mainly driven by the complexity of the learning task from the available observations. High levels of bias are often found when the collection of instances is selected from a specific stratum, common in high-dimensional data derived from social networks, or affected by specific experimental or pre-processing techniques, common in biomedical data. For this reason, the bias-variance decomposition of error provides useful frame to study the error performance of a classification or regression model, as it is well demonstrated by its effectiveness across multiple applications (Domingos 2000). To this end, multiple metrics and sampling schemes have been developed for estimating bias and variance from data, including the widely employed holdout approach of Kohavi and Wolpert (Kohavi and Wolpert 1996).

Sampling Schema. When the estimator of the true performance estimator is not derived from the analysis of the parameters of the learned model, it needs to rely on samples from the original dataset to collect estimates. Sampling schema are defined by two major variables: sampling criteria and train-test size decisions. Error estimations in high-dimensional data strongly depend on the adopted resampling method (Way et al. 2010). Many principles for the selection of sampling methods have been proposed (Molinaro, Simon, and Pfeiffer 2005; Dougherty et al. 2010; Toussaint 1974). Cross-validation methods and alternative bootstrap methods (e.g. randomized bootstrap, 0.632 estimator, mc-estimator, complex bootstrap) have been compared and assessed for a large number of contexts. Unlike cross-validation, bootstrap was shown to be pessimistically biased with respect to the number of training samples. Still, studies show that bootstrap becomes more accurate than its peers for space with very large observed errors as often observed in high-dimensional spaces where $m > n$ (Dougherty et al. 2010). Resubstitution methods are optimistically biased and should be avoided. We consider both the use of k-folds cross-validation and bootstrap to be acceptable. In particular, the number of folds, k, can be adjusted based on the minimum number of estimates for a statistical robust assessment of confidence intervals. This implies a preference for a large number of folds in high-dimensional spaces with either high-variable performance or $n \ll m$.

[2] Inferred from the probability $P(\varepsilon_{true} \mid M,m,n)$ to be consistent across the n observations.

[3] The number and length of subsets of features can be used to affect the performance guarantees. For instance, a lower-bound on the performance of decision lists relying on tests with at most p features chosen from a m-dimensional space and d-depth is $\theta(\varepsilon_{true}) \geq \frac{1}{n}(log\frac{1}{\delta} + \Theta(p^d log_2 p^d))$.

An additional problem when assessing performance guarantees in (n,m)-spaces where $n < m$, is to guarantee that the number of test instances per fold offers a reliable error estimate since the observed errors within a specific fold are also subjected to systematic (bias) and random (variance) uncertainty. Two options can be adopted to minimize this problem. First option is to find the best train-test split. Raudys and Jain (1991) propose a loss function to find a reasonable size of the test sample based on the train sample size and on the estimate of the asymptotic error, which essentially depends on the dimensionality of the dataset and on the properties of the learned model M. A second option is to model the testing sample size independently from the number of training instances. This guarantees a robust performance assessment of the model, but the required number of testing instances can jeopardize the sample size and, thus, compromise the learning task. Error assessments are usually described as Bernoulli process: n_{test} instances are tested, t successes (or failures) are observed and the true performance for a specific fold can be estimated, $\hat{p}$=t/n_{test}, as well as its variance $p(1$-$p)/n_{test}$. The estimation of n_{test} can rely on confidence intervals for the true probability p under a pre-specified precision[4] (Beleites et al. 2013) or from the expected levels of type I and II errors using the statistical tests described by Fleiss (1981).

Loss Functions. Different loss functions capture different performance views, which can result in radically different observed errors, $\{\varepsilon_1,...\varepsilon_k\}$. Three major views can be distinguish to compute each one of these errors for a particular fold from these loss functions. First, error counting, the commonly adopted view, is the relative number of incorrectly classified/predicted/described testing instances. Second, smooth modification of error counting (Glick 1978) uses distance intervals, and it is applicable for classification models with probabilistic outputs (correctly classified instances can contribute to the error) and for regression models. Finally, posterior probability estimate (Lissack and Fu 1976) is often adequate in the presence of the class-conditional distributions. These two latter metrics provide a critical complementary view for models that deliver probabilistic outputs. Additionally, their variance is more realistic than the simple error counting. The problem with smooth modification is its dependence on the error distance function, while posterior probabilities tend to be biased for small datasets.

Although error counting (and the two additional views) are commonly parameterized with an accuracy-based loss function (incorrectly classified instances), other metrics can be adopted to turn the analysis more expressive or to be extensible towards regression models and descriptive models. For settings where the use of confusion matrices is of importance due to the difficulty of the task for some classes/ranges of values, the observed errors can be further decomposed according to type-I and type-II errors.

[4] For some biomedical experiments (Beleites et al. 2013), 75-100 test samples are commonly necessary to achieve reasonable validation and 140 test samples (confidence interval widths 0.1) are necessary for an expected sensitivity of 90%. When this number is considerably higher than the number of available observations, there is the need to post-calibrate the test-train sizes according to the strategies depicted for the first option.

Model	Performance views
Classification model	accuracy (percentage of samples correctly classified); area under receiver operating characteristics curve (AUC); critical complementary performance views can be derived from (multi-class) confusion matrices, including sensitivity, specificity and the F-Measure;
Regression model	simple, average normalized or relative root mean squared error; to draw comparisons with literature results, we suggest the use of the normalized root mean squared error (NRMSE) and the symmetric mean absolute percentage of error (SMAPE);
Descriptive Local model (presence of hidden bics.)	entropy, F-measure and match score clustering metrics (Assent et al. 2007; Sequeira and Zaki 2005); F-measure can be further decomposed in terms of recall (coverage of found samples by a hidden cluster) and precision (absence of samples present in other hidden clusters); match scores (Prelić et al. 2006) assess the similarity of solutions based on the Jaccard index; Hochreiter et al. (2010) introduced a consensus score by computing similarities between all pairs of biclusters; biclustering metrics can be delivered by the application of a clustering metric on both dimensions or by the relative non-intersecting area (RNAI) (Bozdağ, Kumar, and Catalyurek 2010; Patrikainen and Meila 2006);
Descriptive Local model (absence of hidden bics.)	merit functions can be adopted as long as they are not biased towards the merit criteria used within the approaches under comparison (mean squared residue introduced by Cheng and Church (2000) or the Pearson's correlation coefficent; domain-specific evaluations can be adopted by computing statistical enrichment p-values (Madeira and Oliveira 2004);
Descriptive Global model	merit functions to test the fit in the absence of knowledge regarding the regularities; equality tests between multivariate distributions; similarity functions between the observed and approximated distributions;

Table 4: Performance views to estimate the true error of discriminative and descriptive models

A synthesis of the most common performance metrics per type of model is provided in Table 4. A detailed analysis of these metrics is provided in Section 3.3 related with extensibility principles. In particular, in this section we explain how to derive error estimates from descriptive settings.

The use of complementary loss functions for the original task (Eq.1) is easily supported by computing performance guarantees multiple times, each time using a different loss function to obtain the error estimates.

Feasibility of Estimates. As previously prompted, different estimators of the true error can be defined to find confidence intervals or significant differences associated with the performance of a specific model M. For this goal, we covered how to derive estimators from the parametric analysis of the learned models or from error estimates gathered under a specific sampling scheme and loss function. Nevertheless, the performance guarantees defined by these estimators are only valid if they are able to perform better than a null (random) model under a reasonable statistical significance level. An analysis of the significance of these estimators indicates whether we can estimate the performance guarantees of a model or, otherwise, we would need a larger number of observations for the target dimensionality.

A simplistic validation option is to show the significant superiority of M against permutations made on the original dataset (Mukherjee et al. 2003). A possible permutation procedure is to construct for each of the k folds, t samples where the classes (discriminative models) or domain values (descriptive models) are randomly permuted. From the errors computed for each permutation, different density functions can be developed, such as:

$$P_{n,m}(x) = \frac{1}{kt}\Sigma_{i=1}^{k}\Sigma_{j=1}^{t}\theta(x - \varepsilon_{i,j,n,m}), \tag{2}$$

where $\theta(z) = 1$ if $z \geq 0$ and 0 otherwise. The significance of the model is $P_{n,m}(x)$, the percentage of random permutations with observed error smaller than x, where x can be fixed using a estimator of the true error for the target model M. The average estimator, $\varepsilon_{n,m} = \frac{1}{k}\Sigma_{i=1}^{k}(\varepsilon_i \mid n,m)$, or the θ^{th} percentile of the sequence $\{e_1,...,e_k\}$ can be used as an estimate of the true error. Both the average and θ^{th} percentile of error estimates are unbiased estimators. Different percentiles can be used to define error bar envelopes for the true error.

There are two major problems with this approach. First, variability of the observed errors does not affect the significance levels. To account for the variability of error estimates across the $k{\times}t$ permutations, more robust statistical tests can be used, such as one-tailed t-test with ($k{\times}t$)-1 degrees of freedom to test the unilateral superiority of the target model. Second, the significance of the learned relations of a model M is assessed against permuted data, which is a very loose setting. Instead, the same model should be assessed against data generated with similar global regularities in order to guarantee that the observed superiority does not simply result from an overfitting towards the available observations. Similarly, stastical t-tests are suitable options for this scenario.

When this analysis reveals that error estimates cannot be collected with statistical significance due to data size constraints, two additional strategies can be applied. A first strategy it to adopt complementary datasets by either: *1)* relying on identical real data with more samples (note, however, that distinct datasets can lead to quite different performance guarantees (ibid.)), or by *2)* approximating the regularities of the original dataset and to generated larger synthetic data using the retrieved distributions. A second strategy is to relax the significance levels for the inference of less conservative performance guarantees. In this case, results should be provided as indicative and exploratory.

3.2 Data Flexibility

Deriving performance guarantees from a single dataset is of limited interest. Even in the context of a specific domain, the assessment of models from multiple datasets with varying regularities of interest provides a more complete and generalize frame to validate their performance. However, in the absence of other principles, the adoption of multiple datasets leads to multiple, and potentially contradicting, performance guarantees. Principles for the generalization of performance bounds and comparisons[5] retrieved from distinct datasets are proposed in *Section 3.4*.

When real datasets are adopted, their regularities should be retrieved for a more informative context of the outputted performance guarantees. For this goal, distribution tests (with parameters estimated from the observed data) to discover global regularities, biclustering approaches to identify (and smooth) meaningful local correlations, and model reduction transformations to detect (and remove) redundancies (Hocking 2005) can be adopted. When the target real datasets are sufficiently large, size and dimensionality can be varied to approximate learning curves or to simply deliver performance bounds and comparisons for multiple (n,m)-spaces. Since performance bounds and comparisons for the same (n,m)-space can vary with the type of data[6], it is advisable to only combine estimates from datasets that share similar conditional regularities $P_{X|Y}$.

In simulation studies, synthetic datasets should be generated using realistic regularities. Common distribution assumptions include either single or multiple multivariate Gaussian distributions (Way et al. 2010; Guo et al. 2010; Hua et al. 2005; El-Sheikh and Wacker 1980), respectively, for descriptive ($M(X)$) or discriminative models ($M : X \to Y$). In classification settings, it is common to assume unequal means and equal covariance matrices ($X_i \mid y_1 \sim Gaussian(\mu_1, \sigma^2)$, $X_j \mid y_2 \sim Gaussian(\mu_2, \sigma^2)$, where $\mu_1 \neq \mu_2$). The covariance-matrix can be experimentally varied or estimated from real biomedical datasets. In (Way et al. 2010), unequal covariance matrices that differ by a scaling factor are considered. While a few datasets after proper normalization have a reasonable fit, the majority of biomedical datasets cannot be described by such simplistic assumption. In these cases, the use of mixtures, such as the mixture of the target distribution with Boolean feature spaces (Kohavi and John 1997), is also critical to assess non-linear capabilities of the target models. Hua et al. (2005) proposes a hard bimodal model, where the conditional distribution for class y_1 is a Gaussian centered at μ_0=(0,...,0) and the conditional distribution for class y_2 is a mixture of equiprobable Gaussians centered at $\mu_{1,0}$=(1,...,1) and $\mu_{1,1}$=(-1,...,-1). In Guo et al. (2010) study, the complexity of Gaussian conditional distributions was tested by fixing μ_0=0 and by varying μ_1 from 0.5 to 0 in steps of 0.05 for $\sigma_0^2 = \sigma_1^2 = 0.2$. Additionally, one experimental setting generated data according to a mixture of Uniform $U(\mu + 3\sigma, \mu + 6.7\sigma)$ and Gaussian $N(\mu, \sigma^2)$ distributions.

Despite these flexible data assumptions, some datasets have features exhibiting highly skewed distributions. This is a common case with molecular data (particularly from human tissues). The Guo et al. study introduces varying levels of signal-to-noise in the dataset, which resulted in a critical decrease of the observed statistical power for the computed bounds (ibid.). Additionally, only a subset of overall fea-

[5] The comparison of performance of models can be directly learned from multiple datasets using the introduced Friedman framework based on Nemenyi tests (Demšar 2006).

[6] Distinct datasets with identical (n,m)-spaces can have significantly different learning complexities (Mukherjee et al. 2003).

tures was generated according class-conditional distributions in order to simulate the commonly observed compact set of discriminative biomarker features.

The majority of real-world data settings is also characterized by functionally correlated features and, therefore, planting different forms of dependencies among the m target features is of critical importance to infer performance guarantees. Hua et al. (2005) proposes the use of different covariance-matrices by dividing the overall features into correlated subsets with varying number of features ($p \in \{1, 5, 10, 30\}$), and by considering different correlation coefficients ($\rho \in \{0.125, 0.25, 0.5\}$). The increase in correlation among features, either by decreasing g or increasing ρ, increases the Bayes error for a fixed dimensionality. Guo et al. (2010) incorporates a correlation factor just for a small portion of the original features. Other studies offer additional conditional distributions tested using unequal covariance matrices (Way et al. 2010). Finally, biclusters can be planted in data to capture flexible functional relations among subsets of features and observations. Such local dependencies are commonly observed in biomedical data (Madeira and Oliveira 2004).

Additional sources of variability can be present, including technical biases from the collected sample of instances or replicates, pooling and dye-swaps in biological data. This knowledge can be used to shape the estimators of the true error or to further generate new synthetic data settings. Dobbin and Simon work (Dobbin and Simon 2005; Dobbin and Simon 2007) explore how such additional sources of variability impact the observed errors. The variability added by these factors is estimated from the available data. These factors are modeled for both discriminative (multi-class) and descriptive (single-class) settings where the number of independent observations is often small. Formulas are defined for each setting by minimizing the difference between the asymptotic and observed error, $(\lim_{n\to\infty} \varepsilon_{true|n}) - \varepsilon_{true|n}$, where $\varepsilon_{true|n}$ depends on these sources of variability. Although this work provides hints on how to address advanced data aspects with impact on the estimation of the true error, the proposed formulas provide loose bounds and have been only deduced in the the scope of biological data under the independence assumption among features. The variation of statistical power using ANOVA methods has been also proposed to assess these effects on the performance of models (Surendiran and Vadivel 2011).

Synthesizing, flexible data assumptions allow the definition of more general, complete and robust performance guarantees. Beyond varying the size n and dimensionality m, we can isolate six major principles. First, assessing models learned from real and synthetic datasets with disclosed regularities provide complementary views for robust and framed performance guarantees. Second, when adopting multivariate Gaussian distributions to generate data, one should adopt varying distances between their means, use covariance-matrices characterized by varying number of features and correlation factors, and rely on mixtures to test non-linear learning properties. Non-Gaussian distributions can be complementary considered. Third, varying degrees of noise should be planted by, for instance, selecting a percentage of features with skewed values. Fourth, impact of selecting a subset of overall features with more discriminative potential (e.g. lower variances) should be assessed. Fifth, other properties can be explored, such as the planting of local regularities with different properties to assess the performance guarantees of descriptive models and the creation of imbalance between classes to assess classification models. Finally, additional sources variability related with the specificities of the domains of interest can be simulated for context-dependent estimations of performance guarantees.

3.3 Extensibility

Performance Guarantees from Unbalanced Data Settings. Imbalance in the representativity of classes (classification models), range of values (regression models) and among feature distributions affect the performance of models and, consequently, the resulting performance guarantees. In many high-dimensional contexts, such as biomedical labeled data, case and control classes tend to be significantly unbalanced (access to rare conditions or diseases is scarce). In these contexts, it is important to compute performance guarantees in (n,m)-spaces from unbalanced real data or from synthetic data with varying degrees of imbalance. Under such analysis, we can frame the performance guarantees of a specific model M with more rigor. Similarly, for multi-class tasks, performance guarantees can be derived from real datasets and/or synthetic datasets (generated with a varying number and imbalance among the classes) to frame the true performance of a target model M.

Additionally, an adequate selection of loss functions to compute the observed errors is required for these settings. Assuming the presence of c classes, one strategy is to estimate performance bounds c times, where each time the bounds are driven by a loss function based on the sensitivity of that particular class. The overall upper and lower bounds across the c estimations can be outputted. Such illustrative method is critical to guarantee the robustness assessment of the performance of classification models for each class.

Performance Guarantees of Descriptive Models. The introduced assessment principles to derive performance guarantees of discriminative models are applicable to descriptive models under a small set of assumptions. Local and global descriptive models can be easily adopted when considering one of the loss functions proposed in Table 4. The evaluation of local descriptive models can either be made in the presence or absence of hidden (or planted) (bi)clusters, H. Similarly, global descriptive models that return a mixture of distributions that approximate the population from which the sample was retrieved, $X \sim \pi$, can be evaluated in the presence and absence of the underlying true regularities.

However, both descriptive and global models cannot rely on traditional sampling schema to collect error estimates. Therefore, in order to have multiple error estimates for a particular (n,m)-space, which is required for a robust statistical assessment, these estimates should be computed from:

- alternative subsamples of a particular dataset (testing instances are discarded);
- multiple synthetic datasets with fixed number of observations n and features m generated under similar regularities.

3.4 Inferring Performance Guarantees from Multiple Settings

In previous sections, we have been proposing alternative estimators of the true performance, and the use of datasets with varying regularities. Additionally, the performance of learning methods can significantly vary depending on their parameterizations. Some of the variables that can be subject to variation include: data size, data dimensionality, loss function, sampling scheme, model parameters, distributions underlying data, discriminative and skewed subsets of features, local correlations, degree of noise, among others. Understandably, the multiplicity of views related with different estimators, parameters and datasets results in a large number of performance bounds and comparison-relations that can hamper the assessment of a target model. Thus, inferring more general performance guarantees is critical and valid for studies that either derive specific performance guarantees from collections of error estimates or from the direct analysis of the learned models.

Guiding criteria needs to be considered to frame the performance guarantees of a particular model M based on the combinatorial explosion of hyper-surfaces that assess performance guarantees from these parameters. When *comparing models*, simple statistics and hierarchical presentation of the inferred relations can be available. An illustrative example is the delivery of the most significant pairs of values that capture the percentage of settings where a particular model had a superior and inferior performance against another model.

When *bounding performance*, a simple strategy is to use the minimum and maximum values over similar settings to define conservative lower and upper bounds. More robustly, error estimates can be gathered for the definition of more general confidence intervals. Other criteria based on weighted functions can be used to frame the bounds from estimates gathered from multiple estimations (Deng 2007). In order to avoid very distinct levels of difficulty across settings that penalized the inferred performance bounds, either a default parameterization can be made for all the variables and only one variable be tested at a time or distinct settings can be clustered leading to a compact set of performance bounds.

3.5 Integrating the Proposed Principles

The retrieved principles can be consistently and coherently combined according to a simple methodology to enhance the assessment of the performance guarantees of models learned from high-dimensional spaces. First, the decisions related with the definition of the estimators, including the selection of adequate loss functions and sampling scheme and the tests of the feasibility of error estimates, provide a structural basis to bound and compare the performance of models.

Second, to avoid biased performance guarantees towards a single dataset, we propose the estimation of these bounds against synthetic datasets with varying properties. In this context, we can easily evaluate the impact of assuming varying regularities $X|Y$, planting feature dependencies, dealing with different sources of variability, and of creating imbalance for discriminative models. Since the result of varying a large number of parameters can result in large number of estimations, the identified strategies to deal with the inference of performance guarantees from multiple settings should be adopted in order to collapse these estimations into a compact frame of performance guarantees.

Third, in the presence of a model that is able to preserve the original space (e.g. support vector machines, global descriptors, discriminant multivariate models), the impact of dimensionality in the performance guarantees is present by default, and it can be further understood by varying the number of features. For models that rely on subsets of overall features, as the variability of the error estimates may not reflect the true performance, performance guarantees should be adjusted through the unbiasedness principle of feature selection or conservative estimations should be considered recurring to VC-theory.

Finally, for both of these models, the estimator of the true performance should be further decomposed to account for the both the bias and variance underlying error estimates. When performance is highly-variable (loose performance guarantees), this decomposition offers an informative context to understand how the model is able to deal with the risk of overfitting associated with high-dimensional spaces.

4 Results and Discussion

In this section we experimentally assess the relevance of the proposed methodology. First, we compare alternative estimators and provide initial evidence for the need to consider the proposed principles when assessing performance over high-dimensional datasets when $n<m$. Second, we bound and compare the performance of classification models learned over datasets with varying properties. Finally, we show the importance of adopting alternative loss functions for unbalanced multi-class and single-class (descriptive) models.

For these experiments, we rely on both real and synthetic data. Two distinct groups of real-world datasets were used: high-dimensional datasets with small number of instances ($n<m$) and high-dimensional datasets with a large number of instances. For the first group we adopted microarrays for tumor classification collected from BIGS repository[7]: *colon* cancer data (m=2000, n=62, 2 labels), *lymphoma* data (m=4026, n=96, 9 labels), and *leukemia* data (m=7129, n=72, 2 labels). For the second group we selected a random population from the healthcare heritage prize database[8] (m=478, n=20000) which integrates claims across hospitals, pharmacies and laboratories. The original relational scheme was denormalized by mapping each patient as an instance with features extracted from the collected claims (400 attributes), the monthly laboratory tests and taken drugs (72 attributes), and the patient profile (6 attributes). We selected the tasks of classifying the need for upcom-

[7] http://www.upo.es/eps/bigs/datasets.html

[8] http://www.heritagehealthprize.com/c/hhp/data (under a granted permission)

ing interventions (2 labels) and the level of drug prescription ({*low,moderate,high*} labels), considered to be critical tasks for care prevention and drug management.

Two groups of synthetic datasets were generated: multi-label datasets for discriminative models and unlabeled datasets for descriptive models. The labeled datasets were obtained by varying the following parameters: the ratio and the size of the number of observations and features, the number of classes and their imbalance, the conditional distributions (mixture of Gaussians and Poissons per class), the amount of planted noise, the percentage of skewed features, and the area of planted local dependencies. The adopted parameterizations are illustrated in Table 5. To study the properties of local descriptive models, synthetic datasets with varying number and shape of planted biclusters were generated. These settings, described in Table 6, were carefully chosen in order to follow the properties of molecular data (Serin and Vingron 2011; Okada, Fujibuchi, and Horton 2007). In particular, we varied the size of these matrices up to m=4000 and n=400, maintaining the proportion between rows and columns commonly observed in gene expression data.

Features	$m \in \{500, 1000, 2000, 5000\}$
Observations	$n \in \{100, 200, 500, 1000, 10000\}$
Number of Classes	$c \in \{2, 3, 5\}$
Distributions (illustrative)	(c=3) {N(1,σ), N(0,σ), N(-1,σ)} with $\sigma \in \{3, 5\}$ (easy setting) (c=3) {N(u_1,σ),N(0,σ),N(u_3,σ)} with u_1∈{-1,2}, u_2∈{-2,1}} (c=3) mixtures of N(u_i,σ) and P(λ_i) where λ_1=4, λ_2=5, λ_3=6
Noise (% of values' range)	{0%,5%,10%,20%,40%}
Skewed Features	{0%,30%,60%,90%}
Degree of Imbalance (%)	{0%,40%,60%,80%}

Table 5: Parameters for the generation of the labeled synthetic datasets

Features×Observations (♯m×♯n)	100×30	500×60	1000×100	2000×200	4000×400
Nr. of hidden biclusters	3	5	10	15	20
Nr. columns in biclusters	[5,7]	[6,8]	[6,10]	[6,14]	[6,20]
Nr. rows in biclusters	[10,20]	[15,30]	[20,40]	[40,70]	[60,100]
Area of biclusters	9.0%	2.6%	2.4%	2.1%	1.3%

Table 6: Properties of the generated set of unlabeled synthetic datasets

The software implementing the methodology that combines the introduced principles was codified in Java (JVM version 1.6.0-24). The selected supervised learners were adopted from WEKA. The following experiments were computed using an Intel Core i3 1.80GHz with 6GB of RAM.

Challenges. An initial assessment of the performance of two simplistic classification models learned from real high-dimensional datasets is given in Fig.3. The performance bounds[9] from real datasets where $m>n$ confirm the high-variability of performance associated with the learning in these spaces. In particular, the difference between the upper and lower bounds is over 30% for cross-validation options with 10 folds and n folds (leave-one-out). Generally, leave-one-out sampling scheme has higher variability than 10-fold cross-validation. Although leave-one-out is able to learn from more observations (decreasing the variability of performance), the true variability of 10-fold cross-validation is masked by averaging errors per fold. The smooth effect of cross-validation sampling supports the need to increase the levels of significance to derive more realistic performance bounds. Additionally, the use of bootstrap schema with resampling methods to increase the number of instances seems to optimistically bias the true performance of the models. Contrasting with these datasets, models learned from the heritage data setting, where $n \gg m$, have a more stable performance across folds. This leads to a higher number and significance of the superiority comparisons among classification models collected from Friedman tests.

Bounding performance using VC inference or specific percentiles of error estimates introduces undesirable bias. In fact, under similar experimental settings, the VC bounds were very pessimistic (>10 percentage points of difference), while the use of the 0.15 and 0.85 percentiles (to respectively define lower and upper bounds) led to more optimistic bounds against the bounds provided in Fig.3. Although changing percentiles easily allows to tune the target level of conservatism, they do not capture the variability of the error estimates.

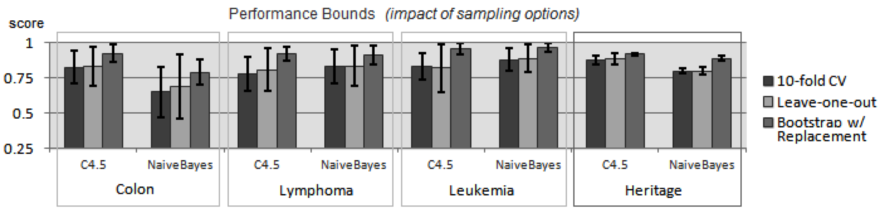

Fig. 3: Performance guarantees from real datasets with varying $\frac{n}{m}$ degree for two classifiers tested under different sampling options.

A set of views on the significance of the learned relations from real and synthetic high-dimensional datasets is respectively provided in Table 7 and Fig.4. Different methods were adopted to compute the significance (p-value) associated with a collection of error estimates. These methods basically compute a p-value by comparing the collected error estimates against estimates provided by loose settings where:

[9] Confidence intervals of a mean estimator from the sample of error estimates assumed to be normally distributed with the same expectation mean, a standard error $\frac{\sigma}{\sqrt{n}}$ and significance α=0.05.

1) the target model is learned from permuted data, *2)* a *null* classifier[10] is learned from the original data, and *3)* the target model is learned from *null* data (preservation of global conditional regularities). We also considered the setting proposed by Mukherjee et al. (2003) (Eq.2). Comparisons are given by one-tailed t-tests. For this analysis, we compared the significance of the learned C4.5, Naive Bayes and support vector machines (SVM) models for real datasets and averaged their values for synthetic datasets. A major observation can be retrieved: p-values are not highly significant ($\ll$1%) when $n<m$, meaning that the performance of the learned models is not significantly better than very loose learners. Again, this observation underlines the importance of carefully framing assessments of models learned from high-dimensional spaces. Additionally, different significance views can result in quite different p-values, which stresses the need to choose an appropriate robust basis to validate the collected estimates. Comparison against null data is the most conservative, while the counts performed under Eq.2 (permutations density function) are not sensitive to distances among error mismatches and easily lead to biased results.

	Colon			Leukemia			Heritage		
	C4.5	NBayes	SVM	C4.5	NBayes	SVM	C4.5	NBayes	SVM
Comparison Against Permutated Data	1.5%	41.3%	1.2%	0.6%	0.1%	0.2%	∼0%	∼0%	∼0%
Comparison Against Null Model	1.1%	32.2%	1.2%	0.1%	0.1%	0.1%	∼0%	∼0%	∼0%
Comparison Against Null Dataset	15.2%	60.3%	9.3%	9.7%	12.0%	7.2%	1.3%	3.8%	1.7%
Permutations Density Function (Eq.2)	14.0%	36.0%	8.4%	8.4%	1.2%	0.8%	0.0%	0.4%	0.0%

Table 7: Significance of the collected error estimates of models learned from real datasets using improvement p-values. p-values are computed by comparing the target models vs. a baseline classification models, and error estimates collected from the original dataset vs. a permuted dataset or null dataset (where basic regularities are preserved).

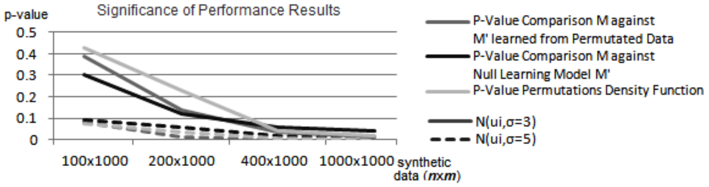

Fig. 4: Significance views on the error estimates collected by classification models from $m>n$ synthetic datasets under easy N(u_i,σ=3) and moderate N(u_i,σ=5) settings against loose baseline settings.

[10] A classifier that defines the average conditional values per feature during the training phase and the mode of feature-based classes during the testing phase was considered.

To further understand the root of the variability associated with the performance of models learned from high-dimensional datasets, Fig.5 provides its decomposition in two components: bias and variance. Bias provides a view on how the expected error deviates across folds for the target dataset. Variance provides a view on how the model behavior differs across distinct training folds. We can observe that the bias component is higher than the variance component, which is partly explained by the natural biased incurred from samples in $n<m$ high-dimensional spaces. The disclosed variance is associated with the natural overfitting of the models in these spaces. Interestingly, we observe that the higher $\frac{m}{n}$ ratio is, the higher the bias/variance ratio. The sum of these components decrease for an increased number of observations, n, and it also depends on the nature of the conditional distributions of the dataset, as it is shown by the adoption of synthetic datasets with conditional Gaussian distributions with small-to-large overlapping areas under the density curve. The focus on each one of these components for the inference of novel performance guarantees is critical to study the impact of the capacity error and training error associated with the learned model (see Fig.2).

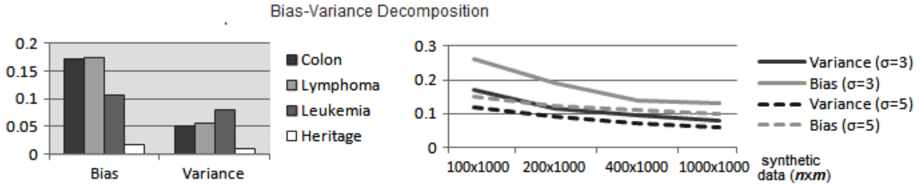

Fig. 5: Decomposition of the performance variability from real and synthetic data (see Table 5) using C4.5: understanding the model capacity (*variance* component) and the model error (*bias* component).

Unbalanced Multi-class Data. The importance of selecting adequate performance views to retrieve realistic guarantees is shown in Fig.6. This is still an underestimated problem that needs to be addressed for both: *1)* balanced datasets where the class-conditional distributions differs in complexity (see the sensitivity associated with the classes from Colon and Leukemia datasets in Fig.6a), and *2)* for unbalanced datasets, where the representativity of each class can hamper the learning task even if the complexity of the class-conditional distributions is similar (see the sensitivity of the classes from datasets with different degrees of imbalance Fig.6b). In this analysis, we adopted sensitivity as a simplistic motivational metric, however many other loss functions hold intrinsic properties of interest to derive particular implications from the performance of the target models. Table 4 synthesizes some of the most common performance views. The chosen view not only impacts the expected true error, but the variability of the error as it is well-demonstrated in Fig.6a. This impacts both the inferred bounds and the number of significant comparisons.

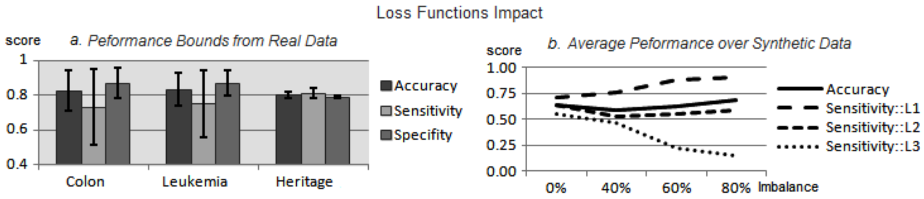

Fig. 6: Impact of adopting alternative loss functions on the: *a*) performance variability of real datasets, and *b*) true performance of synthetic datasets (n=200 and m=500) with varying degrees of imbalance among classes.

Performance Guarantees from Flexible Data Settings. To understand how performance guarantees varies across different data settings for a specific model, we computed C4.5 performance bounds from synthetic datasets with varying degree of planted noise and skewed features. Inferring performance guarantees across settings is important to derive more general implications on the performance of models. This analysis is provided in Fig.7. Generalizing performance bounds from datasets with different learning complexity may result in very loose bounds and, therefore, should be avoided. In fact, planting noise and skewing features not only increases the expected error but also its variance. Still, some generalizations are possible when the differences between collections of error estimates is not high. In these cases, collections of error estimates can be joint for the computation of new confidence intervals (as the ones provided in Fig.7). When the goal is to compare sets of models, superiority relations can be tested for each setting under relaxed significance levels, and outputted if the same relation appears across all settings. In our experimental study we were only able to retrieve a small set of superiority relations between C4.5 and Naive Bayes using the Friedman-test under loose levels of significance (10%).

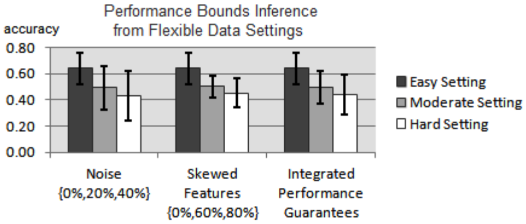

Fig. 7: Inference of performance guarantees in a (n=200,m=500)-space with varying degree of planted noise (as a percentage of domain values) and skewed features (as a percentage of total features).

Fig.8 assesses the impact of adopting different conditional distributions for the inference of general performance guarantees for C4.5. Understandably, the expected error increases when the overlapping area between conditional distributions is higher or when a particular class is described by a mixture of distributions. Combining such hard settings with more easy settings gives rise to loose performance bounds and to a residual number of significant superiority relations between models. Still, this assessment is required to validate and weight the increasing number data-independent implications of performance from the recent studies.

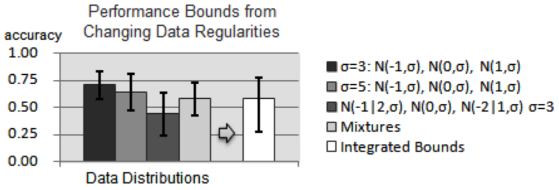

Fig. 8: Inference of performance guarantees from (n=200,m=500)-spaces with different regularities described in Table 5.

Descriptive Models. The previous principles are extensible towards descriptive models under an adequate loss function and sampling method to collect estimates. This means that the introduced significance views, decomposition of the error and inference of guarantees from flexible data settings become applicable to different types of models, such as (bi)clustering models and global descriptive models. Fig.9 illustrates the performance bounds of BicPAM biclustering model[11] using three distinct loss functions computed from estimates collected from datasets generated with identical size, dimensionality and underlying regularities (according to Table 6). The target loss functions are the traditional match scores (Prelić et al. 2006), which assess the similarity of the discovered biclusters B and planted biclusters H based on the Jaccard index[12], and the Fabia consensus[13] (Hochreiter et al. 2010). The observed differences on the mean and variability of performance per loss function are enough to deliver distinct Friedman-test results when comparing multiple descriptive models. Therefore, the retrieved implications should be clearly contextualized as pertaining to a specific loss function, sampling scheme, data setting and significance threshold.

[11] http://web.ist.utl.pt/ rmch/software/bicpam/

[12] $MS(B,H)$ defines the extent to what found biclusters match with hidden biclusters, while $MS(H,B)$ reflects how well hidden biclusters are recovered: $\mathbf{MS}(B,H) = \frac{1}{|B|}\Sigma_{(I_1,J_1)\in B} max_{(I_2,J_2)\in H} \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|}$

[13] Let S_1 and S_2 be, respectively, the larger and smaller set of biclusters from $\{B,H\}$, and MP be the pairs $B \leftrightarrow H$ assigned using the Munkres method based on overlapping areas (Munkres 1957): $\mathbf{FC}(B,H) = \frac{1}{|S_1|}\Sigma_{((I_1,J_1)\in S_1,(I_2,J_2)\in S_2)\in MP} \frac{|I_1 \cap I_2| \times |J_1 \cap J_2|}{|I_1| \times |J_1| + |I_2| \times |J_2| - |I_1 \cap I_2| \times |J_1 \cap J_2|}$

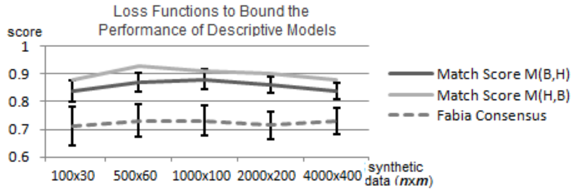

Fig. 9: Performance assessment over biclustering models (BicPAM) using distinct loss functions – Fabia consensus, and match scores $M(B,H)$ and $M(H,B)$ – and a collection of error estimates from 20 data instances per data setting.

Final Discussion. In this chapter, we synthesized critical principles to bound and compare the performance of models learned from high-dimensional datasets. First, we surveyed and provide empirical evidence for the challenges related with this task for (n,m)-spaces where $n<m$. This task is critical as implications are derived from studies where the differences in performance of classification models learned over these spaces against permuted and *null* spaces is not significant. Also, the width between the estimated confidence intervals of performance is considerably high in these spaces, leading to the absence of significant results from Friedman comparisons.

Second, motivated by these challenges, we have shown the importance of adopting robust statistical principles to test the feasibility of the collected estimates. Different tests for computing significance levels have been proposed, each one providing different levels of conservatism, which can be used to validate and weight the increasing number of implications derived from the performance of models in high-dimensional spaces.

Third, understanding the source of variability of the performance of the models is critical in these spaces as this variability can be either related with the overfitting aspect of the models or with the learning complexity associated with the dataset. The variability of performance can, thus, be further decomposed in variance and bias. While the variance captures the differences on the behavior of the model across samples from the target population, which is indicative of the model capacity (see Fig.2), the bias captures the learning error associated within the available samples. These components disclose the why behind the inferred performance guarantees and, thus, are critical to understand and refine the model behavior.

Fourth, we compared alternative ways of bounding and comparing performance, including different sampling schema, loss functions and statistical tests. In particular, we used initial empirical evidence to show how different estimators can bias the true error or smooth its variability.

An alternative to the inference of performance guarantees from estimates is to approximate the true performance from the properties of the learned models. For this latter line of research two strategies can be followed. A first strategy is to retrieve guarantees from the learned parameters from multivariate models that preserve the original dimensionality (Ness and Simpson 1976; El-Sheikh and Wacker 1980). A second strategy is to understand the discriminative significance of the selected local subspaces from the original space when a form of feature-set selection is adopted during the learning process (Singhi and Liu 2006; Iswandy and Koenig 2006).

Fifth, the impact of varying data regularities on the performance guarantees was also assessed, including spaces with varying degrees of the $\frac{n}{m}$ ratio, (conditional) distributions, noise, imbalance among classes (when considering classification models), and uninformative features. In particular, we observed that inferring general bounds and comparisons from flexible data settings is possible, but tends to originate very loose guarantees when mixing data settings with very distinct learning complexities. In those cases, a feasible trade-off would be to simply group data settings according to the distributions of each collection of error estimates.

Finally, we have shown the applicability of these principles for additional types of models, such as descriptive models.

5 Conclusion

Motivated by the challenges of learning from high-dimensional data, this chapter established a solid foundation on how to assess the performance guarantees given by different types of learners in high-dimensional spaces. The definition of adequate estimators of the true performance in these spaces, where the learning is associated with high variance and bias from error estimates, is critical. We surveyed a set of approaches that provide distinct principles on how to bound and compare the performance of models as a function of the data size. A taxonomy to understand their major challenges was proposed. These challenges mainly result from their underlying assumptions and task goals. Existing approaches often fail to provide a robust performance guarantees, are not easily extensible to support unbalanced data settings or assess non-discriminative models (such as local and global descriptive models), and are not able to infer guarantees from multiple data settings with varying properties, such as locally correlated features, noise, and underlying complex distributions.

In this chapter, a set of principles is proposed to answer the identified challenges. They offer a solid foundation to select adequate estimators (either from data sampling or direct model analysis), loss functions, and statistical tests sensitive to the pecularities of the performance of models in high-dimensional spaces. Additionally, these principles provide critical strategies for the generalization of performance guarantees from flexible data settings where the underlying global and local regularities can vary. Finally, we briefly show that these principles can be integrated within a single methodology. This methodology offers a robust, flexible and complete frame to bound and compare the performance of models learned over high-dimensional datasets. In fact, it provides critical guidelines to assess the performance of upcoming learners proposed for high-dimensional settings or, complementary, to determine the appropriate data size and dimensionality required to support decisions related with experimental, collection or annotation costs.

Experimental results support the relevance of these principles. We provided empirical evidence for the importance of computing adequate significance views to adjust the statistical power when bounding and comparing the performance of models, of selecting adequate error estimators, of inferring guarantees from flexible data settings, and of decomposing the error to gain further insights on the source of its variability. Additionally, we have experimentally shown the extensibility of these decisions for descriptive models under adequate performance views.

This work opens a new door for understanding, bounding and comparing the performance of models in high-dimensional spaces. First, we expect the application of the proposed methodology to study the performance guarantees of new learners, parameterizations and feature selection methods. Additionally, these guarantees can be used to weight and validate the increasing number of implications derived from the application of these models over high-dimensional data. Finally, we expect the extension of this assessment towards models learned from structured spaces, such as high-dimensional time sequences.

Acknowledgments

This work was partially supported by FCT under the projects PTDC/EIA-EIA/111239/2009 (Neuroclinomics) and PEst-OE/ EEI/LA0021/2011, and under the PhD grant SFRH/BD/75924/2011.

Software Availability

The generated synthetic datasets and the software implementing the proposed statistical tests are available in http://web.ist.utl.pt/rmch/software/bsize/.

References

Adcock, C. J. (1997). “Sample size determination: a review”. In: *J. of the Royal Statistical Society: Series D (The Statistician)* 46.2, pp. 261–283.

Amaratunga, D., J. Cabrera, and Z. Shkedy (2014). *Exploration and Analysis of DNA Microarray and Other High-Dimensional Data*. Wiley Series in Probability and Statistics. Wiley.

Apolloni, B. and C. Gentile (1998). “Sample size lower bounds in PAC learning by algorithmic complexity theory”. In: *Theoretical Computer Science* 209.1.2, pp. 141 –162.

Assent, I. et al. (2007). “DUSC: Dimensionality Unbiased Subspace Clustering”. In: *ICDM*.

Beleites, C. et al. (2013). “Sample size planning for classification models”. In: *Analytica Chimica Acta* 760.0, pp. 25 –33.

Blumer, A. et al. (1989). “Learnability and the Vapnik-Chervonenkis dimension”. In: *J. ACM* 36.4, pp. 929–965.

Boonyanunta, N. and P. Zeephongsekul (2004). “Predicting the Relationship Between the Size of Training Sample and the Predictive Power of Classifiers”. In: *Knowledge-Based Intelligent Information and Engineering Systems*. Vol. 3215. LNCS. Springer Berlin Heidelberg, pp. 529–535.

Bozdağ, D., A. S. Kumar, and U. V. Catalyurek (2010). “Comparative analysis of biclustering algorithms”. In: *BCB*. Niagara Falls, New York: ACM, pp. 265–274.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer.

Cai, T. and X. Shen (2010). *High-Dimensional Data Analysis (Frontiers of Statistics)*. World Scientific.

Cheng, Y. and G. M. Church (2000). “Biclustering of Expression Data”. In: *Intelligent Systems for Molecular Biology*. AAAI Press, pp. 93–103.

Demšar, J. (2006). “Statistical Comparisons of Classifiers over Multiple Data Sets”. In: *J. Machine Learning Res.* 7, pp. 1–30.

Deng, G. (2007). *Simulation-based optimization*. University of Wisconsin–Madison.

Dobbin, K. and R. Simon (2005). “Sample size determination in microarray experiments for class comparison and prognostic classification”. In: *Biostatistics* 6.1, pp. 27+.

Dobbin, K. K. and R. M. Simon (2007). “Sample size planning for developing classifiers using high-dimensional DNA microarray data.” In: *Biostatistics* 8.1, pp. 101–117.

Domingos, P. (2000). “A Unified Bias-Variance Decomposition and its Applications”. In: *IC on Machine Learning*. Morgan Kaufmann, pp. 231–238.

Dougherty, E. R. et al. (2010). “Performance of Error Estimators for Classification”. In: *Current Bioinformatics* 5.1, pp. 53–67.

El-Sheikh, T. S. and A. G. Wacker (1980). “Effect of dimensionality and estimation on the performance of gaussian classifiers.” In: *Pattern Recognition* 12.3, pp. 115–126.

Figueroa, R. L. et al. (2012). “Predicting sample size required for classification performance”. In: *BMC Med. Inf. & Decision Making* 12, p. 8.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. Wiley P. in Applied Statistics. Wiley.

García, S. and F. Herrera (2009). “An Extension on ”Statistical Comparisons of Classifiers over Multiple Data Sets” for all Pairwise Comparisons”. In: *Journal of Machine Learning Research* 9, pp. 2677–2694.

Glick, N. (1978). “Additive estimators for probabilities of correct classification”. In: *Pattern Recognition* 10.3, pp. 211 –222.

Guo, Y. et al. (2010). “Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms”. English. In: *BMC Bioinformatics* 11.1, pp. 1–19.

Guyon, I. et al. (1998). “What Size Test Set Gives Good Error Rate Estimates?” In: *IEEE Trans. Pattern Anal. Mach. Intell.* 20.1, pp. 52–64.

Hand, D. J. (1986). “Recent advances in error rate estimation”. In: *Pattern Recogn. Lett.* 4.5, pp. 335–346.

Haussler, D., M. Kearns, and R. Schapire (1991). “Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension”. In: *Proceedings of the fourth annual workshop on Computational learning theory*. COLT ’91. Santa Cruz, California, USA: Morgan Kaufmann Publishers Inc., pp. 61–74.

Hochreiter, S. et al. (2010). “FABIA: factor analysis for bicluster acquisition”. In: *Bioinformatics* 26.12, pp. 1520–1527.

Hocking, R. (2005). *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. Wiley Series in Probability and Statistics. Wiley, p. 81.

Hua, J. et al. (2005). “Optimal number of features as a function of sample size for various classification rules”. In: *Bioinformatics* 21.8, pp. 1509–1515.

Iswandy, K. and A. Koenig (2006). “Towards Effective Unbiased Automated Feature Selection”. In: *Hybrid Intelligent Systems*, pp. 29–29.

Jain, A. and B. Chandrasekaran (1982). “Dimensionality and Sample Size Considerations”. In: *Pattern Recognition in Practice*. Ed. by P. Krishnaiah and L. Kanal, pp. 835–855.

Jain, N. et al. (2003). “Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays.” In: *Bioinformatics* 19.15, pp. 1945–1951.

Kanal, L. and B. Chandrasekaran (1971). “On dimensionality and sample size in statistical pattern classification”. In: *Pattern Recognition* 3.3, pp. 225 –234.

Kohavi, R. and G. H. John (1997). “Wrappers for feature subset selection”. In: *Artif. Intell.* 97.1-2, pp. 273–324.

Kohavi, R. and D. H. Wolpert (1996). “Bias Plus Variance Decomposition for Zero-One Loss Functions”. In: *Machine Learning*. Morgan Kaufmann Publishers, pp. 275–283.

Lissack, T. and K.-S. Fu (1976). “Error estimation in pattern recognition via L-distance between posterior density functions”. In: *IEEE Transactions on Information Theory* 22.1, pp. 34–45.

Madeira, S. C. and A. L. Oliveira (2004). “Biclustering Algorithms for Biological Data Analysis: A Survey”. In: *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 1.1, pp. 24–45.

Martin, J. K. and D. S. Hirschberg (1996). *Small Sample Statistics for Classification Error Rates II: Confidence Intervals and Significance Tests*. Tech. rep. DICS.

Molinaro, A. M., R. Simon, and R. M. Pfeiffer (2005). “Prediction error estimation: a comparison of resampling methods”. In: *Bioinformatics* 21.15, pp. 3301–3307. eprint: http://bioinformatics.oxfordjournals.org/content/21/15/3301.full.pdf+html.

Mukherjee, S. et al. (2003). “Estimating dataset size requirements for classifying DNA Microarray data”. In: *Journal of Computational Biology* 10, pp. 119–142.

Munkres, J. (1957). “Algorithms for the Assignment and Transportation Problems”. In: *Society for Ind. and Applied Math.* 5.1, pp. 32–38.

Ness, J. W. van and C. Simpson (1976). “On the Effects of Dimension in Discriminant Analysis”. In: *Technometrics* 18.2, pp. 175–187.

Niyogi, P. and F. Girosi (1996). “On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions”. In: *Neural Comput.* 8.4, pp. 819–842.

Okada, Y., W. Fujibuchi, and P. Horton (2007). “A biclustering method for gene expression module discovery using closed itemset enumeration algorithm”. In: *IPSJ Transactions on Bioinformatics* 48.SIG5, pp. 39–48.

Opper, M et al. (1990). “On the ability of the optimal perceptron to generalise”. In: *Journal of Physics A: Mathematical and General* 23.11, p. L581.

Patrikainen, A. and M. Meila (2006). “Comparing Subspace Clusterings”. In: *IEEE TKDE* 18.7, pp. 902–916.

Prelić, A. et al. (2006). “A systematic comparison and evaluation of biclustering methods for gene expression data”. In: *Bioinf.* 22.9, pp. 1122–1129.

Qin, G. and L. Hotilovac (2008). “Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test”. In: *Stat. Methods Med. Res.* 17.2, pp. 207–221.

Raeder, T., T. R. Hoens, and N. V. Chawla (2010). “Consequences of Variability in Classifier Performance Estimates”. In: *ICDM*, pp. 421–430.

Raudys, S. (1997). “On Dimensionality, Sample Size, and Classification Error of Nonparametric Linear Classification Algorithms”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 19.6, pp. 667–671.

Raudys, S. J. and A. K. Jain (1991). “Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 13.3, pp. 252–264.

Sequeira, K. and M. Zaki (2005). “SCHISM: a new approach to interesting subspace mining”. In: *Int. J. Bus. Intell. Data Min.* 1.2, pp. 137–160.

Serin, A. and M. Vingron (2011). “DeBi: Discovering Differentially Expressed Biclusters using a Frequent Itemset Approach”. English. In: *Algorithms for Molecular Biology* 6 (1), pp. 1–12.

Singhi, S. K. and H. Liu (2006). “Feature subset selection bias for classification learning”. In: *IC on Machine Learning*. Pittsburgh, Pennsylvania: ACM, pp. 849–856.

Surendiran, B. and A. Vadivel (2011). “Feature Selection using Stepwise ANOVA Discriminant Analysis for Mammogram Mass Classification”. In: *IJ on Signal Image Proc.* 2.1, p. 4.

Toussaint, G. (1974). “Bibliography on estimation of misclassification”. In: *IEEE Transactions on Information Theory* 20.4, pp. 472–479.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Vayatis, N. and R. Azencott (1999). “Distribution-Dependent Vapnik-Chervonenkis Bounds”. English. In: *Computational Learning Theory*. Ed. by P. Fischer and H. Simon. Vol. 1572. LNCS. Springer Berlin Heidelberg, pp. 230–240.

Way, T. et al. (2010). “Effect of finite sample size on feature selection and classification: A simulation study”. In: *Medical Physics* 37.2, pp. 907–920.