Honghua Dai · James N.K. Liu
Evgueni Smirnov *Editors*

Reliable Knowledge Discovery

Springer

Reliable Knowledge Discovery

Honghua Dai • James N.K. Liu • Evgueni Smirnov
Editors

Reliable Knowledge Discovery

Springer

Editors
Honghua Dai
Deakin University
Burwood, Victoria, Australia

James N.K. Liu
The Hong Kong Polytechnic University
Hong Kong

Evgueni Smirnov
Maastricht University
Maastricht, The Netherlands

ISBN 978-1-4614-1902-0 e-ISBN 978-1-4614-1903-7
DOI 10.1007/978-1-4614-1903-7
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2012932410

Preface

1 Description

With the rapid development of the data mining and knowledge discovery, a key issue which could significantly affect the real world applications of data mining is the reliability issues of knowledge discovery. It is natural that people will ask if the discovered knowledge is reliable. Why do we trust the discovered knowledge? How much can we trust the discovered knowledge? When it could go wrong. All these questions are very essential to data mining. It is especial crucial to the real world applications.

One of the essential requirements of data mining is validity. This means both the discovery process itself and the discovered knowledge should be valid. Reliability is a necessary but not sufficient condition for validity. Reliability could be viewed as stability, equivalence and consistency in some ways.

This special volume of the book on the reliability issues of Data Mining and Knowledge Discovery will focus on the theory and techniques that can ensure the discovered knowledge is reliable and to identify under which conditions the discovered knowledge is reliable or in which cases the discovery process is robust. In the last 20 years, many data mining algorithms have been developed for the discovery of knowledge from given data bases. However in some cases, the discovery process is not robust or the discovered knowledge is not reliable or even incorrect in certain cases. We could also find that in some cases, the discovered knowledge may not necessary be the real reflection of the data. Why does this happen? What are the major factors that affect the discovery process? How can we make sure that the discovered knowledge is reliable? What are the conditions under which a reliable discovery can be assured? These are some interesting questions to be investigated in this book.

2 Scope and Topics of this Book

The topics of this book covers the following:

- The theories on reliable knowledge discovery
- Reliable knowledge discovery methods
- Reliability measurement criteria of knowledge discovery
- Reliability estimation methods
- General reliability issues on knowledge discovery
- Domain specific reliability issues on knowledge discovery
- The criteria that can be used to assess the reliability of discovered knowledge.
- The conditions under which we can confidently say that the discovered knowledge is reliable.
- The techniques which can improve reliability of knowledge discovery
- Practical approaches that can be used to solve reliability problems of data mining systems.
- The theoretical work on data mining reliability
- The practical approaches which can be used to assess if the discovered knowledge is reliable.
- The analysis of the factors that affect data mining reliability
- How reliability can be assessed
- In which condition, the reliability of the discovered knowledge is assured.

3 The Theme and Related Resources

The main purpose of this book is to encourage the use of Reliable Knowledge Discovery from Databases (RKDD) in critical-domain applications related to society, science, and technology. The book is intended for practitioners, researchers, and advanced-level students. It can be employed primarily as a reference work and it is a good compliment to the excellent book on reliable prediction Algorithmic learning in a random world by Vladimir Vovk, Alex Gammerman, and Glenn Shafer (New York: Springer, 2005). Extra information sources are the proceedings of the workshops Reliability Issues in Knowledge Discovery held in conjunction with the IEEE International Conferences on Data Mining. Other relevant conferences are the Annual ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), the International Conference on Machine Learning (ICML), The pacific-Asia Conference on Knowledge Discovery (PAKDD), and the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD). Many AI-related journals regularly publish work in RKDD. Among others it is worth mentioning the Journal of Data Mining and Knowledge Discovery, the Journal of Machine Learning Research, and the Journal of Intelligent Data Analysis.

4 An Overview of the Book

This book presents the recent advances in the emerging field of **Reliable Knowledge Discovery from Data (RKDD)**. In this filed the knowledge is considered as reliable in the sense that its generalization performance can be set in advance. Hence, RKDD has a potential for a broad spectrum of applications, especially in critical domains like medicine, finance, military etc. The main material presented in the book is based on three consequent workshops Reliability Issues in Knowledge Discovery held in conjunction with the IEEE International Conferences on Data Mining (ICDM) in 2006, 2008, and 2010, respectively. In addition we provided an opportunity to authors to publish the results of their newest research related to RKDD.

This book is organized in seventeen chapters divided into four parts.

Part I includes three chapters on Reliability Estimation.

Chapter 1 provides an overview of typicalness and transductive reliability estimation frameworks. The overview is employed for introducing an approach for accessing reliability of individual classifications called joint confidence machine. Chapter 1 describes an approach that compensates the weaknesses of typicalness-based confidence estimation and transductive reliability estimation by integrating them into a joint confidence machine. It provides better interpretation of the performance of any classifiers. Experimental results performed with different machine learning algorithms in several problem domains show that there is no reduction of discrimination performance and is more suitable for applications with risk-sensitive problems with strict confidence limits.

Chapter 2 introduces new approaches to estimating and correcting individual predictions in the context of stream mining. It investigates the online reliability estimation of individual predictions. It proposes different strategies and explores techniques based on local variance and local bias, of local sensitivity analysis and online bagging of predictors. Comparison results on benchmark data are given to demonstrate the improvement of prediction accuracy.

Chapter 3 deals with the problem of quantifying the reliability in the context of neural networks. It elaborates on new approaches to estimation of confidence and prediction intervals for polynomial neural networks.

Part II includes seven chapters on Reliable Knowledge Discovery Methods.

Chapter 4 investigates outliers in regression targeting robust diagnostic regression. The chapter discusses both robust regression and regression diagnostics, presents several contemporary methods through numerical examples in linear regression.

Chapter 5 presents a conventional view on the definition of reliability; points out the three major categories of factors that affect the reliability of knowledge discovery, examined the impact of model complexity, weak links, varying sample sizes and the ability of different learners to the reliability of graphical model discovery, proposed reliable graph discovery approaches.

Chapter 6 provides a generalization of version spaces for reliable classification implemented using support vector machines.

Chapter 7 presents a unified generative model ONM which characterizes the life cycle of a ticket. The model uses maximum likelihood estimation to capture reliable ticket transfer profiles which can reflect how the information contained in a ticket is used by human experts to make reliable ticket routing decisions.

Chapter 8 applies the methods of aggregation functions for the reliable web based knowledge discovery from network trafic data.

Chapter 9 gives two new versions of SVM for the regression study of features in the problem domain. It provides means for feature selection and weighting based on the correlation analysis to give better and reliable result.

Chapter 10 describes in detail an application of transductive confidence machines for reliable handwriting recognition. It introduces a TCM framework which can enhance classifiers to reduce the computational costs and memory consumption required for updating the non-conformity scores in the offline learning setup of TCMs. Results are found to have outperformed previous methods on both relatively easy data and on difficult test samples.

PART III includes four Chapters on Reliability Analysis.

Chapter 11 addresses the problem of reliable feature selection. It introduces a generic-feature-selection measure together with a new search approach for globally optimal feature-subset selection. It discusses the reliability in the feature-selection process of a real pattern-recognition system, provides formal measurements and allows consistent search for relevant features in order to attain global optimal solution.

Chapter 12 provides three detailed case studies to show how the reliability of an induced classifier can be influenced. The case study results reveal the impact of data-oriented factors to the relaibility of the discovered knowledge.

Chapter 13 analyzes recently-introduced instance-based penalization techniques capable of providing more accurate predictions.

Chapter 14 investigates subsequence frequency measurement and its impact on the reliability of knowledge discovery in single sequences.

PART IV includes three chapters on Reliability Improvement Methods.

Chapter 15 proposed to use the inexact field learning method and parameter optimized one-class classifiers to improving reliability of unbalanced text mining by reducing performance bias.

Chapter 16 proposes a formal description technique for ontology representation and verification using a high level Petri net approach. It provides the capability of detection and identification of potential anomalies in ontology for the improvement of the discovered knolwedge.

Chapter 17 presents an UGDSS framework to provide reliable support for multicriteria decision making in uncertainty problem domain. It gives the system design and architecture.

5 Acknowledgement

We would like to thank many people that made this book possible. We start with the organizers of the workshops held in conjunction with the IEEE International Conferences on Data Mining (ICDM): Shusaku Tsumoto, Francesco Bonchi, Bettina Berendt, Wei Fan and Wynne Hsu. We express our gratitude to the authors whose contributions can be found in the book. Finally, we thank our colleagues from Springer that made the publication process possible in a short period.

Burwood Victoria (Australia), *Honghua Dai*
Hong Kong (China), *James Liu*
Maastricht (The Netherlands), *Evgueni Smirnov*
August 2011

Contents

Part IV Reliability Improvement Methods

Part I
Reliability Estimation

Chapter 1
Transductive Reliability Estimation for Individual Classifications in Machine Learning and Data Mining

Matjaž Kukar

Abstract Machine learning and data mining approaches are nowadays being used in many fields as valuable data analysis tools. However, their serious practical use is affected by the fact, that more often than not, they cannot produce reliable and unbiased assessments of their predictions' quality. In last years, several approaches for estimating reliability or confidence of individual classifiers have emerged, many of them building upon the algorithmic theory of randomness, such as (historically ordered) transduction-based confidence estimation, typicalness-based confidence estimation, and transductive reliability estimation. In the chapter we describe typicalness and transductive reliability estimation frameworks and propose a joint approach that compensates their weaknesses by integrating typicalness-based confidence estimation and transductive reliability estimation into a joint confidence machine. The resulting confidence machine produces confidence values in the statistical sense (e.g., a confidence level of 95% means that in 95% the predicted class is also a true class), as well as provides us with a general principle that is independent of to the particular underlying classifier

1.1 Introduction

Usually machine learning algorithms output only bare predictions (classifications) for the new unclassified examples. While there are ways for almost all machine learning algorithms to at least partially provide quantitative assessment of the particular classification, so far there is no general method to assess the quality (confidence, reliability) of a single classification. We are interested in the assessment of classifier's performance on a *single example* and not in average performance on an

Matjaž Kukar
University of Ljubljana, Faculty of Computer and Information Science,
Tržaška 25, SI-1001 Ljubljana, Slovenia,
e-mail: matjaz.kukar@fri.uni-lj.si

H. Dai et al. (eds.), *Reliable Knowledge Discovery*,
DOI 10.1007/978-1-4614-1903-7_1,

independent dataset. Such assessments are very useful, especially in risk-sensitive applications (medical diagnosis, financial and critical control applications) because there it often matters, how much one can rely upon a given prediction. In such cases an overall quality measure of a classifier (e.g. classification accuracy, mean squared error, ...) with respect to the whole input distribution would not provide the desired value. Another possible use of quality assessment of single classifications is in ensembles of machine learning algorithms for selecting or combining answers from different classifiers [24].

There have been numerous attempts to assign probabilities to machine learning classifiers' (decision trees and rules, Bayesian classifiers, neural networks, nearest neighbour classifiers, ...) in order to interpret their decision as a probability distribution over all possible classes. In fact, we can trivially convert every machine learning classifier's output to a probability distribution by assigning the predicted class the probability 1, and 0 to all other possible classes. The posterior probability of the predicted class can be viewed as a classifier's confidence (reliability) of its prediction. However, such estimations may in general not be good due to inherent applied algorithm's biases.[1]

1.2 Related work

In statistics, estimation for individual predictions is assessed by confidence values and intervals. On the same basis, the reliability estimation was implemented in machine learning methods, where properties of predictive models were utilized to endow predictions with corresponding reliability estimates. Although these approaches are specific for a particular predictive model and cannot be generalized, they provide favorable results to the general approaches. Such reliability estimates were developed for the Support Vector Machines [10, 33] the ridge regression model [28], the multilayer perceptron [27], the ensembles of neural networks [15, 7] and others.

In contrast to the former group of methods, general (model-independent) methods utilize approaches, such as local modeling of prediction error based on input space properties and local learning [2, 11], meta-predicting the leave-one-out error of a single example [39], transductive reasoning [31, 24], and sensitivity analysis [6, 18, 19, 5, 4].

Sensitivity analysis aims at determining how much the variation of input can influence the output of a system. The idea for putting the reliability estimation in the context of the sensitivity analysis framework is, therefore, in observing the changes in model outputs by modifying its inputs. Treating the predictive model as a black box, the sensitivity analysis approach, therefore, indirectly analyzes qualitatively describable aspects of the model, such as generalization ability, bias, resistance to noise, avoidance of overfitting, and so on. The motivation came from the related

[1] An extreme case of inherent bias can be found in a trivial constant classifier that blindly labels any example with a predetermined class with self-proclaimed confidence 1.

fields of data perturbation [9] and co-learning (using unlabeled examples in supervised learning) [3]. Transductive reliability estimation can be viewed as an intersection of these two fields, as it perturbs the training set with as single unlabelled example.

1.2.1 Transduction

Several methods for inducing probabilistic descriptions from training data, figuring the use of density estimation algorithms, are emerging as an alternative to more established approaches for machine learning. Frequently kernel density estimation [43] is used for density estimation of input data using diverse machine learning paradigm such as probabilistic neural networks [37], Bayesian networks and classifiers [17], decision trees [36]. By this approach a chosen paradigm, coupled with kernel density estimation, is used for modelling the probability distribution of input data. Alternatively, stochastically changing class labels in the training dataset is proposed [13] in order to estimate conditionally class probability.

There is some ongoing work for constructing classifiers that divide the data space into reliable and unreliable regions [1]. Such meta-learning approaches have also been used for picking the most reliable prediction from the outputs of an ensemble of classifiers [35].

Meta learning community is partially dealing with predicting the right machine learning algorithm for a particular problem [30] based on performance and characteristics of other, simpler learning algorithms. In our problem of confidence estimation such an approach would result in learning to predict confidence value based on characteristics of single examples.

A lot of work has been done in applications of the transduction methodology [33], in connection with algorithmic theory of randomness. Here, approximations of randomness deficiency for different methods (SVMs, ridge regression) have been constructed in order to estimate confidence of single predictions. The drawback of this approach is that confidence estimations need to be specifically designed for each particular method and cannot be applied to other methods.

Another approach to reliability estimation, similarly based on the transduction principle, has been proposed in [24]. While it is general and independent of the underlying classifier, interpretation of its results isn't always possible in the statistical sense of confidence levels.

A few years ago typicalness has emerged as a complementary approach to transduction [26, 31, 16]. By this approach, a "strangeness" measure of a single example is used to calculate its typicalness, and consequently a confidence in classifier's prediction. The main drawback of this approach is that for each machine learning algorithm it needs an appropriately constructed strangeness measure.

In the chapter we present a further development of the latter two approaches where transductive reliability estimation serves as a generic strangeness measure in the typicalness framework. We compare the experimental results to that of kernel

density estimation and show that the proposed method significantly outperforms it. We also suggest how basic transduction principle can be used to significantly improve results of kernel density estimation so it almost reaches results of transductive typicalness.

The chapter is organized as follows. In Sec. 1.3 we describe the basic ideas of typicalness and transduction, outline the process of their integration, and review kernel density estimation methods used for comparison. In Sec. 1.4 we evaluate how our methodology compares to other approaches in 15 domains with 6 machine learning algorithms. In Sec. 1.5 we present some conclusions and directions for future work.

1.3 Methods and materials

Reliability estimation of a classification ($\tilde{y}$) of a single example (x), given its true class (y) should have the following property:

$$Rel(\tilde{y}\|x) = t \Rightarrow P(\tilde{y} \neq y) \leq 1-t \tag{1.1}$$

If Eq. 1.1 holds, or even better, if it approaches equality, a reliability measure can be treated as a confidence value [26].

The produced confidence values should be valid in the following sense. Given some possible label space $\tilde{Y}$, if an algorithm predicts some set of labels $Y \subseteq \tilde{Y}$ with confidence t for a new example which is truly labelled by $y \in \tilde{Y}$, then we would expect the following to hold over randomization of the training set and the new example:

$$P(y \notin Y) \leq 1-t \tag{1.2}$$

Note that Eq. 1.2 is very general and valid for both classification (Y is predicted set of classes) and regression problems (Y is a predicted interval). As we deal only with single predictions in this chapter, Eq. 1.2 can be simplified to a single predicted class value ($Y = \{\tilde{y}\}$):

$$P(y \neq \tilde{y}) \leq 1-t \tag{1.3}$$

1.3.1 Typicalness

In the typicalness framework [26, 28, 33] we consider a sequence of examples $(z_1,\ldots,z_n) = ((x_1,y_1),\ldots,(x_n,y_n))$, together with a new example x_{n+1} with unknown label $\tilde{y}_{n+1}$, all drawn independently from the same distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is an attribute space and $\mathcal{Y}$ is a label space. Our only assumption is therefore that the training as well as new (unlabelled) examples are independently and identically distributed (*iid* assumption).

We can use the typicalness framework to gain confidence information for each possible labelling for a new example x_{n+1}. We postulate some labels $\tilde{y}_{n+1}$ and for each one we examine how likely (typical) it is that all elements of the extended sequence $((x_1,y_1),\ldots,(x_{n+1},\tilde{y}_{n+1}))$ might have been drawn independently from the same distribution or how typically *iid* the sequence is. The more typical the sequence, the more confident we are in $\tilde{y}_{n+1}$. To measure the typicalness of sequences, we define, for every $n \in \mathbb{N}$, a typicalness function $t : \mathcal{Z}^n \to [0,1]$ which, for any $r \in [0,1]$ has the property

$$P((z_1,\ldots,z_n) \ : \ t(z_1,\ldots,z_n) \leq r) \leq r \tag{1.4}$$

If a typicalness function returns 0.05 for a given sequence, we know that the sequence is unusual because it will be produced at most 5% of the time by any *iid* process. It has been shown [26] that we can construct such functions by considering the "strangeness" of individual examples. If we have some family of functions

$$f : \mathcal{Z}^n \times \{1,2,\ldots,n\} \to \mathbb{R}, \ \ n \in \mathbb{N}\ldots, \tag{1.5}$$

then we can associate a strangeness value

$$\alpha(z_i) = f(\{z_1,\ldots,z_n\};i), i = 1,2,\ldots n \tag{1.6}$$

with each example and define the following typicalness function

$$t((z_1,\ldots,z_n)) = \frac{\#\{\alpha(z_i) : \alpha(z_i) \geq \alpha(z_n)\}}{n} \tag{1.7}$$

We group individual strangeness functions α_i into a family of functions $A_n : n \in \mathbb{N}$, where $A_n : \mathcal{Z}^n \to \mathbb{R}^n$ for all n. This is called an individual strangeness measure if, for any n, any permutation $\pi : \{1,\ldots,n\} \to \{1,\ldots,n\}$, any sequence $(z_1,\ldots,z_n) \in \mathcal{Z}^n$, and any $(\alpha_{\pi(1)},\ldots,\alpha_{\pi(n)}) \in \mathbb{R}^n$) it satisfies the following criterion [26]:

$$(\alpha_1,\ldots,\alpha_n) = A_n(z_1,\ldots,z_n) \Longrightarrow (\alpha_{\pi(1)},\ldots,\alpha_{\pi(n)}) = A_n(z_{\pi(1)},\ldots,z_{\pi(n)}) \tag{1.8}$$

The meaning of this criterion is that the same value should be produced for each individual element in sequence, regardless of the order in which their individual strangeness values are calculated. This is a very important criterion, because it can be proven [26] that the constructed typicalness function (1.7) satisfies the condition from (1.4), provided that the individual strangeness measure satisfies the criterion (1.8).

From a practical point of view it is advisable [26] to use positive strangeness measures, ranging between 0 for most typical examples, and some positive upper bound, (up to $+\infty$), for most untypical examples.

1.3.1.1 Typicalness in machine learning

In the machine learning setup, for calculating the typicalness of a new example $z_{n+1} = (x_{n+1}, \widetilde{y}_{n+1})$ described with attribute values x_{n+1} and labelled with $\widetilde{y}_{n+1}$, given the training set $(z_1, \ldots, z_n)$, Eq. 1.7 changes to

$$t((z_1, \ldots, z_{n+1})) = \frac{\#\{\alpha(z_i) : \alpha(z_i) \geq \alpha(z_{n+1})\}}{n+1} \tag{1.9}$$

Note that on the right-hand side of Eq. 1.9, z_i belongs to the extended sequence, i.e. $z_i \in \{z_1, \ldots, z_{n+1}\}$. For a given machine learning algorithm, first we need to construct an appropriate strangeness measure and modify the algorithm accordingly.[2] Then, for each new unlabelled example x, all possible labels $\widetilde{y} \in Y$ are considered. For each label $\widetilde{y}$ a typicalness of labelled example $t((x,\widetilde{y})) = t((z_1, \ldots, z_n, (x,\widetilde{y})))$ is calculated. Finally, the example is labelled with "most typical" class, that is the one that maximizes $\{t((x,\widetilde{y}))\}$. By Eq. 1.7 the second largest typicalness is an upper bound on the probability that the excluded classifications are correct [31]. Consequently, the confidence is calculated as follows:

$$\text{confidence}((x,\widetilde{y})) = 1 - \text{typicalness of second most typical label.} \tag{1.10}$$

1.3.2 Transductive reliability estimation

Transduction is an inference principle that takes a training sample and aims at estimating the values of a discrete or continuous function only at given unlabelled points of interest from input space, as opposed to the whole input space for induction. In the learning process the unlabelled points are suitably labelled and included into the training sample. The usefulness of unlabelled data has also been advocated in the context of co-training. It has been shown [3] that for every better-than-random classifier its performance can be significantly boosted by utilizing only additional unlabelled data.

It has been suggested [40] that when solving a given problem one should avoid solving a more general problem as an intermediate step. The reasoning behind this principle is that, in order to solve a more general task, resources may be wasted or compromises made which would not have been necessary for solving only the problem at hand (i.e. function estimation only on given points). This common-sense principle reduces a more general problem of inferring a functional dependency on the whole input space (inductive inference) to the problem of estimating the values of a function only at given points (transductive inference).

[2] This is the main problem of the typicalness approach, as the algorithms need do be considerably changed.

1.3.2.1 A formal background

Let $\mathcal{X}$ be a space of attribute descriptions of points (examples) in a training sample (dataset), and $\mathcal{Y}$ a space of labels (continuous or discrete) assigned to each point. Given a probability distribution $\mathcal{P}$, defined on the input space $\mathcal{X} \times \mathcal{Y}$, a training sample

$$S = \{(x_1, y_1), \ldots, (x_l, y_l)\} \tag{1.11}$$

consisting of l points, is drawn *iid* (identically independently distributed) according to $\mathcal{P}$. Additional m data points (working sample)

$$W = \{x_{l+1}, \ldots, x_{l+m}\} \tag{1.12}$$

with unknown labels are drawn in the same manner. The goal of transductive inference is to label all the points from the sample W using a fixed set $\mathcal{H}$ of functions $f : \mathcal{X} \mapsto \mathcal{Y}$ in order to minimize an error functional both in the training sample S and in the working sample W (effectively, in $S \cup W$). In contrast, inductive inference aims at choosing a single function $f \in \mathcal{H}$ that is best suited to the unknown probability distribution $\mathcal{P}$.

At this point there arises a question how to calculate labels of points from a working sample. This can be done by labelling every point from a working sample with every possible label value; however given m working points this leads to a combinatorial explosion yielding n^m possible labellings. For each possible labelling, an induction process on $S \cup W$ is run, and an error functional (error rate) is calculated.

By leveraging the *iid* sampling assumption and transductive inference, one can for each labelling estimate its reliability (also referred to as confidence, a probability that it is correct). If the *iid* assumption holds, the training sample S as well as the joint correctly labelled sample $S \cup W$ should both reflect the same underlying probability distribution $\mathcal{P}$.

If one could measure a degree of similarity between probability distributions $\mathcal{P}(S)$ and $\mathcal{P}(S \cup W)$, this could be used as a measure of reliability of the particular labelling. Unfortunately, this problem in general belongs to the non-computable class [25], so approximation methods have to be used [42, 22].

Evaluation of prediction reliability for single points in data space has many uses. In risk-sensitive applications (medical diagnosis, financial and critical control applications) it often matters, how much one can rely upon a given prediction. In such a case a general reliability measure of a classifier (e.g. classification accuracy, mean, squared error, ...) with respect to the whole input distribution would not provide the desired warranty. Another use of reliability estimations is in combining answers from different predictors, weighed according to their reliability.

1.3.2.2 Why is transduction supposed to work?

There is a strong connection between the transduction principle and the algorithmic (Kolmogorov) complexity. Let the sets S and $S \cup W$ be represented as binary strings

u and v, respectively. Let $l(v)$ be the length of the string v and $C(v)$ its Kolmogorov complexity, both measured in bits. We define the *randomness deficiency* of the string v as following [25, 42]:

$$\delta(v) = l(v) - C(v) \tag{1.13}$$

Randomness deficiency measures how random is the respective binary string and therefore the set it represents. The larger it is, more regular is the string (and the set). If we could calculate the randomness deficiency (but we cannot, since it is not computable), we could do it for all possible labellings of the set $S \cup W$ and select the labelling of W with largest randomness deficiency as the most probable one [42]. That is, we would select the most regular one. We can also construct a universal Martin-Löf's test for randomness [25]:

$$\sum \{P(x|l(x) = n) : \delta(x) \geq m\} \leq 2^{-m} \tag{1.14}$$

That is, for all binary strings of fixed length n, the probability of their randomness deficiency δ being greater than m is less than 2^{-m}. The value $2^{-\delta(x)}$ is therefore a p-value function for our randomness test [42].

Unfortunately, as the definition of randomness deficiency is based on the Kolmogorov complexity, it is not computable. Therefore we need feasible approximations to use this principle in practice. Extensive work has been done by using Support Vector Machines [10, 33, 42], however no general approach exists so far.

1.3.2.3 A machine learning interpretation

In machine learning terms, the sets S and $S \cup W$ are represented by the induced models M_S and $M_{S \cup W}$. The randomness of the sets reflects in the (Kolmogorov) complexity of the respective models. If for the set $S \cup W$ the labelling of W with largest randomness deficiency is selected, it follows from our definition of randomness deficiency (Eq. 1.13) that since the length $l(v)$ is constant, the Kolmogorov complexity $C(M_{S \cup W})$ is minimal. Therefore the model $M_{S \cup W}$ is most similar to the M_S.

This greatly simplifies our view on the problem, namely it suffices to compare the (finite) models M_S and $M_{S \cup W}$. Greater difference between them means that the set $S \cup W$ is more random than the set S and (under the assumption that S is sufficient for learning effective model) that W consist of (at least some) improperly labelled, untypical examples.

Although the problem seems easier now, it is still a computational burden to calculate changes between model descriptions (assuming that they can be efficiently coded; black-box methods are thus out of question). However, there exists another way.

Since transduction is an inference principle that aims at estimating the values of a function only at given points of interest from input space (the set W), we are interested only in model change considering this examples. Therefore we can compare the classifications (or even better, probability distributions) of models M_S and mod-

els $M_{S\cup W}$. Obviously, the labelling of W that would minimally change the model M_S is as given by M_S. We will examine this approach in more detail in the next section.

The transductive reliability estimation process and its theoretical foundations originating from Kolmogorov complexity are described in more detail in [24]. Basically, we have a two-step process, featuring an *inductive step* followed by a *transductive step*.

- An *inductive step* is just like an ordinary inductive learning process in machine learning. A machine learning algorithm is run on the training set, *inducing* a classifier. A selected example is taken from an independent dataset and classified using the induced classifier. An example, labelled with the classified class is temporarily included into the training set.
- A *transductive step* is almost a repetition of an inductive step. A machine learning algorithm is run on the changed training set, *transducing* a classifier. The same example as before is taken from the independent dataset and and classified using the transduced classifier. Both classifications of the same example are compared and their difference (distance) is calculated, thus approximating the randomness deficiency.
- After the reliability is calculated, the example in question is removed from the training set.

In practice the inductive step is performed only once, namely on the original training set. New examples are not permanently included in the training set; this would be improper since the correct class is at this point still unknown. Although retraining for each new example seems to be highly time consuming, it is not such a problem in practice, especially if incremental learners (such as naive Bayesian classifier) are used.

A brief algorithmic sketch is given in Fig. 1.1. An intuitive explanation of transductive reliability estimation is that we disturb a classifier by inserting a new example in a training set. A magnitude of this disturbance is an estimation of classifier's instability (unreliability) in a given region of its problem space.

Since a prerequisite for a machine learning algorithm is to represent its classifications as a probability distribution over all possible classes, we need a method to measure the difference between two probability distributions. The difference measure D should ideally satisfy all requirements for a distance (i.e. nonnegativity, triangle nonequality and symmetry), however in practice nonnegativity suffices. For calculating the difference between probability distributions, a *Kullback-Leibler divergence* is frequently used [12, 38]. Kullback-Leibler divergence, sometimes referred to as a relative entropy or I-divergence, is defined between probability distributions P and Q

$$I(P,Q) = -\sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} \tag{1.16}$$

In our experiments we use a symmetric Kullback-Leibler divergence, or J-divergence, which is defined as follows:

$$J(P,Q) = (I(P,Q) + I(Q,P)) = \sum_{i=1}^{n} (p_i - q_i) \log_2 \frac{p_i}{q_i} \tag{1.17}$$

$J(P,Q)$ is limited to the interval $[0,\infty]$, where $J(P,P) = 0$. Since in this context we require the values to be from the $[0,1]$ interval we normalize it in the spirit of Martin-Löf's test for randomness.

$$J_N(P,Q) = 1 - 2^{-J(P,Q)} \tag{1.18}$$

However, measuring the difference between probability distributions does not always perform well. There are at least a few exceptional classifiers (albeit trivial ones) where the original approach utterly fails.

1.3.2.4 Assessing the classifier's quality: the curse of trivial models

So far we have implicitly assumed that the model used by the classifier is good (at the very least better than random). Unsurprisingly, our approach works very well with random classifiers (probability distributions are randomly calculated) by effectively labelling their classifications as unreliable [22, 23].

Input :	*machinelearningclassifier, atrainingsetandanunlabelledtest example*	(1.15)
Output :	*Estimationoftestexample′sclassificationreliability*	

Inductive step:

- train a classifier from the provided training set
- select an unlabelled test example
- classify this example with an induced classifier
- label this example with a predicted class
- temporarily add the newly labelled example to the training set

Transductive step:

- train a classifier from the extended training set
- select the same unlabelled test example as above
- classify this example with a transduced classifier

Calculate a randomness deficiency approximation as a *normalized difference* $J_N(P,Q)$ between inductive (P) and transductive (Q) classification.

Calculate the reliability of classification as in a universal Martin-Löf's test for randomness 1-*normalized difference*

Fig. 1.1: The algorithm for transductive reliability estimation