1

Introduction

Technological advances in recent decades have made it possible to automate many tasks that previously required significant amounts of manual time, performing regular or repetitive activities. Certainly, computing machines have proven to be a great asset in improving human speed and efficiency as well as in reducing errors in these essentially mechanical tasks. More impressive, however, is the fact that the emergence of computing technologies has also enabled the automation of tasks that require significant understanding of intrinsically human domains that can in no way be qualified as merely mechanical. Although we humans have maintained an edge in performing some of these tasks, e.g., recognizing pictures or delineating boundaries in a given picture, we have been less successful at others, e.g., fraud or computer network attack detection, owing to the sheer volume of data involved and to the presence of nonlinear patterns to be discerned and analyzed simultaneously within these data. Machine learning and data mining, on the other hand, have heralded significant advances, both theoretical and applied, in this direction, thus getting us one step closer to realizing such goals.

Machine learning is embodied by different learning approaches, which are themselves implemented within various frameworks. Examples of some of the most prominent of these learning paradigms include supervised learning, in which the data labels are available and generally discrete; unsupervised learning, in which the data labels are unavailable; semisupervised learning, in which some, generally discrete, data labels are available, but not all; regression, in which the data labels are continuous; and reinforcement learning, in which learning is based on an agent policy optimization in a reward setting. The plethora of solutions that have been proposed within these different paradigms yielded a wide array of learning algorithms. As a result, the field is at an interesting crossroad. On the one hand, it has matured to the point where many impressive and pragmatic data analysis methods have emerged, of course, with their respective strengths

and limitations.[1] On the other hand, it is now overflowing with hundreds of studies trying to improve the basic methods, but only marginally succeeding in doing so (Hand, 2006).[2] This is especially true on the applied front. Just as in any scientific field, the practical utility of any new advance can be accepted only if we can demonstrate beyond reasonable doubt the superiority of the proposed or novel methods over existing ones in the context in which it was designed.

This brings the issue of evaluating the proposed learning algorithms to the fore. Although considerable effort has been made by researchers in both developing novel learning methods and improving the existing models and approaches, these same researchers have not been completely successful at alleviating the users' scepticism with regard to the worth of these new developments. This is due, in big part, to the lack of both depth and focus in what has become a ritualized evaluation method used to compare different approaches. There are many issues involved in the question of designing an evaluation strategy for a learning machine. Furthermore, these issues cover a wide range of concerns pertaining to both the problem and the solution that we wish to consider. For instance, one may ask the following questions: What precise measure is best suited for a quantified assessments of different algorithms' property of interest in a given domain? How can these measures be efficiently computed? Do the data from the domain of interest affect the efficiency of this calculation? How can we be confident about whether the difference in measurement for two or more algorithms denotes a statistically significant difference in their performance? Is this statistical difference practically relevant as well? How can we best use the available data to discover whether such differences exist? And so on. We do not claim that all these issues can be answered in a definitive manner, but we do emphasize the need to understand the issues we are dealing with, along with the various approaches available to tackle them. In particular, we must understand the strengths and limitations of these approaches as well as the proper manner in which they should be applied. Moreover, we also need to understand what these methods offer and how to properly interpret the results of their application. This is very different from the way evaluation has been perceived to date in the machine learning community, where we have been using a routine, de facto, strategy, without much concern about its meaning.

In this book, we try to address these issues, more specifically with regard to the branch of machine learning pertaining to classification algorithms. In particular, we focus on evaluating the performance of classifiers generated by supervised learning algorithms, generally in a binary classification scenario. We wish to emphasize, however, that the overall message of the book and the

[1] These developments have resulted both from empirically studied behaviors and from exploiting the theoretical frameworks developed in other fields, especially mathematics.
[2] Although the worth of a study that results in marginal empirical improvements sometimes lies in the more significant theoretical insights obtained.

insights obtained should be considered in a more general sense toward the study of all learning paradigms and settings. Many of these approaches can indeed be readily exported (with a few suitable modifications) to other scenarios such as unsupervised learning, regression and so on. The issues we consider in the book deal not only with evaluation measures, but also with the related and important issues of obtaining (and understanding) the statistical significance of the observed differences, efficiently computing the evaluation measures in as unbiased a manner as possible, and dealing with the artifacts of the data that affect these quantities. Our aim is to raise an awareness of the proper way to conduct such evaluations and of how important they are to the practical utilization of the advances being made in the field. While developing an understanding of the relevant evaluation strategies, some that are widely used (although sometimes with little understanding) as well as some that are not currently too popular, we also try to address a number of practical criticisms and philosophical concerns that have been raised with regard to their usage and effectiveness and examine the solutions that have been proposed to deal with these concerns.

Our aim is not to suggest a recipe for evaluation to replace the previous de facto one, but to develop an understanding and appreciation of the evaluation strategies, of their strengths, and the underlying caveats. Before we go further and expand our discussion pertaining to the goals of this book by bringing forth the issues with our current practices, we discuss the de facto culture that has pervaded the machine learning community to date.

1.1 The De Facto Culture

For over two decades now, with Kibler and Langley (1988) suggesting the need for a greater emphasis on performance evaluation, the machine learning community has recognized the importance of proper evaluation. Research has been done to both come up with novel ways of evaluating classifiers and to use insights obtained from other fields in doing so. In particular, researchers have probed such fields as mathematics, psychology, and statistics among others. This has resulted in significant advances in our ability to track and compare the performance of different algorithms, although the results and the importance of such evaluation has remained underappreciated by the community as a whole because of one or more reasons that we will soon ponder. More important, however, is the effect of this underappreciation that has resulted in the entrenchment of a *de facto* culture of evaluation. Consider, for example, the following statement extracted from (Witten and Frank, 2005b, p. 144), one of the most widely used textbooks in machine learning and data mining:

> The question of predicting performance based on limited data is an interesting, and still controversial one. We will encounter many different

> techniques, of which one – repeated cross-validation – is gaining ascendance and is probably the evaluation method of choice in most practical limited-data situations.

This, in a sense, prescribes repeated cross-validation as a de facto method for *most practical limited data situations*. And therein lies the problem. Although cross-validation has indeed appeared to be a strong candidate among resampling methods in limited data situations, generalizing its use to most practical situations is pushing our luck a bit too far. Most of the practical data situations warrant looking into broader and deeper issues before zeroing in on an evaluation strategy (or even an error-estimation method such as cross-validation). We will soon look into what these issues are, including those that are generally obvious and those that are not.

The preceding take on choosing an evaluation method makes the implicit statement that cross-validation has been adopted as a standard. This implication is quite important because it molds the mindset of both the researcher and the practitioner as to the fact that a standard recipe for evaluation can be applied without having to consider the full context of that evaluation. This context encompasses many criteria and not simply, as is sometimes believed, the sample size of the application. Other important criteria are the class distribution of the data, the need for parameter selection (also known as model selection), the choice of a relevant and appropriate performance metric, and so on. Witten and Frank (2005b, pp. 144) further state,

> Comparing the performance of different machine learning methods on a given problem is another matter that is not so easy as it sounds: to be sure that apparent differences are not caused by chance effects, statistical tests are needed.

Indeed, statistical tests are needed and are even useful so as to obtain "confidence" in the difference in performance observed over a given domain for two or more algorithms. Generally the machine learning community has settled on merely rejecting the null hypothesis that the apparent differences are caused by chance effects when the t test is applied. In fact, the issue is a lot more involved.

The point is that no single evaluation strategy consisting of a combination of evaluation methods can be prescribed that is appropriate in *all* scenarios. A de facto – or perhaps, more appropriately, a panacea – approach to evaluation, even with minor variations for different cases, is hence neither appropriate nor possible or even advisable. Broader issues need to be taken into account.

Getting back to the issue of our general underappreciation of the importance of evaluation, let us now briefly consider this question: *Why and how has the machine learning community allowed such a de facto or panacea culture to take root?* The answer to this question is multifold. Naturally we can invoke the

argument about the ease of comparing novel results with existing published ones as a major advantage of sticking to a very simple comparison framework. The reasons for doing so can generally be traced to two main sources: (i) the unavailability of other researchers' algorithm implementations, and (ii) the ease of not having to replicate the simulations even when such implementations are available. The first concern has actually encouraged various researchers to come together in calling for the public availability of algorithmic implementations under general public licenses (Sonnenburg et al., 2007). The second concern should not be mistaken for laziness on the part of researchers. After all, there can be no better reward in being able to demonstrate, fair and square – i.e., by letting the creators of the system themselves demonstrate its worth as best as they can – the superiority of one's method to the existing state of the art.

Looking a little bit beyond the issues of availability and simplicity, we believe that there are more complex considerations that underlie the establishment of this culture. Indeed, the implicit adoption of the de facto approach can also be linked to the desire of establishing an "acceptable" scientific practice in the field as a way to validate an algorithm's worth. Unfortunately, we chose to achieve such acceptability by using a number of shortcuts. The problem with this practice is that our comparisons of algorithms' performance, although appearing acceptable, are frequently invalid. Indeed, many times, validity is lost as a result of the violation of the underlying assumptions and constraints of the methods that we use. This can be called the "politically correct" way of doing evaluations. Such considerations are generally, and understandably, never stated as they are implicit.

Digging even deeper, we can discover some of the reasons for this standard adoption. A big part of the problem is attributable to a lack of understanding of the evaluation approaches, their underlying mode of application, and the interpretation of their results. Although advances have been made in finding novel evaluation approaches or their periodic refinements, these advances have not propagated to the mainstream. The result has been the adoption of a "standard" simple evaluation approach comprising various elements that are relatively easily understood (even intuitive) and widely accepted. The downside of this approach is that, even when alternative (and sometimes better-suited) evaluation measures are utilized by researchers, their results are met with scepticism. If we could instill a widespread understanding of the evaluation methodologies in the community, it would be easier to not only better evaluate our classifiers but also to better appreciate the results that were obtained. This can further result in a positive-feedback loop from which we can obtain a better understanding of various learning approaches along with their bottlenecks, leading in turn to better learning algorithms. This, however, is not to say that the researchers adopting alternative, relatively less-utilized elements of evaluation approaches are completely absolved of any responsibility. Instead, these researchers also have the onus of making a convincing case as to why such a strategy is more suitable than

those in current and common use. Moreover, the audience – both the reviewers and the readers – should be open to better modes of evaluation that can yield a better understanding of the learning approaches applied in a given domain, bringing into the light their respective strengths and limitations. To realize this goal, it is indeed imperative that we develop a thorough understanding of such evaluation approaches and promote this in the basic *required* machine learning and data mining courses.

1.2 Motivations for this Book

As just discussed, there is indeed a need to go beyond the de facto evaluation approaches. There are many reasons why this has not happened yet. However, the core reasons can be traced to a relative lack of proper understanding of the procedures. Progress toward realizing the goal of more meaningful classifier evaluation and consequently better understanding of the learning approaches themselves can take place only if both the researchers involved in developing novel learning approaches and the practitioners applying these are better aware of not only the evaluation methods, but also of their strengths and limitations together with their context of application.

There have also been criticisms of specific evaluation methods that were condemned for not yielding the desired results. These criticisms, in fact, arise from unreasonable expectations from the evaluation approaches. It is important to understand what a given evaluation method promises and how the results it obtained should be interpreted. One of the widest criticisms among these has fallen on the statistical significance testing procedure, as we will see later in the book. Although some of these criticisms are genuine, most of them result from a mistaken interpretation. The tests are not definitive, and it is important that both their meaning and the results they produce be interpreted properly. These will not only help us develop a better understanding of the learning algorithms, but they will also lead to a raised awareness in terms of what the tests mean and hence what results should (and can) be expected. A better understanding of the overall evaluation framework would then enable researchers to ask the right questions before adopting the elements of that evaluation framework. Summarizing the goals toward this raised awareness, we need to make sure that both the researchers and practitioners follow these guidelines:

1. To have a better understanding of the entire evaluation process so as to be able to make *informed decisions* about the strategies to be employed.
2. To have *reasonable* expectations from the evaluation methods: For instance, the t test only helps us guard against the claim that one algorithm is better than others when the evidence to support this claim is too weak. It doesn't help us *prove* that one algorithm is better than other in *any* case.

3. To possess a knowledge of the right questions to be asked or addressed before adopting an evaluation framework.

Note that the *de facto* method, even if suitable in many scenarios, is not a panacea. Broader issues need to be taken into account. Such awareness can be brought about only from a better understanding of the approaches themselves. This is precisely what this book is aimed at. The main idea of the book is not to prescribe specific recipes of evaluation strategies, but rather to educate researchers and practitioners alike about the issues to keep in mind when adopting an evaluation approach, to enable them to objectively apply these approaches in their respective settings.

While furthering the community's understanding of the issues surrounding evaluation, we also seek to simplify the application of different evaluation paradigms to various practical problems. In this vein, we provide simple and intuitive implementations of all the methods presented in the book. We developed these by using WEKA and R, two freely available and highly versatile platforms, in the hope of making the discussions in the book easily accessible to and further usable by all.

Before we proceed any further, let us see, with the help of a concrete example, what we mean by the de facto approach to evaluation and what types of issues can arise as a result of its improper application.

1.3 The De Facto Approach

As we discussed in Section 1.1, a de facto evaluation culture has pervaded a big part of experimental verification and comparative evaluation of learning algorithms. The approaches utilized to do so proceed along the following lines, with some minor variations: Select an evaluation metric, the most often used one being accuracy; select a large-enough number of datasets [the number is chosen so as to be able to make a convincing case of apt evaluation and the datasets are generally obtained from a public data repository, the main one being the University of California, Irvine, (UCI) machine learning repository]; select the best parameters for various learning algorithms, a task generally known as model selection but mostly inadvertently interleaved with evaluation; use a k-fold cross-validation technique for error estimation, often stratified 10-fold cross-validation, with or without repetition; apply paired t tests to all pairs of results or to the pairs deemed relevant (e.g., the ones including a possibly new algorithm of interest) to test for statistical significance in the observed performance difference; average the results for an overall estimate of the algorithm's performance or, alternatively, record basic statistics such as win/loss/ties for each algorithm with respect to the others. Let us examine this de facto approach with an illustration.

Table 1.1. *Datasets used in the illustration of the* de facto *evaluation approach*

Datasets	#attr	#ins	#cls
Anneal	39	898	5
Audiology	70	226	24
Balance scale	5	625	3
Breast cancer	10	286	2
Contact lenses	5	24	3
Diabetes	9	768	2
Glass	10	214	6
Hepatitis	20	155	2
Hypothyroid	30	3772	4
Mushroom	23	8124	2
Tic-tac-toe	10	958	2

1.3.1 An Illustration

Consider an experiment that consists of running a set of learning algorithms on a number of domains to compare their generic performances. The algorithms used for this purpose include naive bayes (NB), support vector machines (SVMs), 1-nearest neighbor (1NN), AdaBoost using decision trees (ADA), Bagging (BAG), a C4.5 decision tree (C45), random forest (RF), and Ripper (RIP).

Tables 1.2 and 1.3 illustrate the process just summarized with actual experiments. In particular, Table 1.1 shows the name, dimensionality (#attr), size (#ins), and number of classes (#cls) of each domain considered in the study. Table 1.2 shows the results obtained by use of accuracy, 10-fold stratified cross-validation, and t tests with 95% confidence, and averaging of the results obtained by each classifier on all the domains. In Table 1.2, we also show the results of the t test with each classifier pitted against NB. A "v" next to the result indicates the significance test's success of the concerned classifier against NB, a "*" represents a failure, against NB (i.e. NB wins) and no symbol signals a tie (no statistically significant difference). The results of the t test are summarized at the bottom of the table. Table 1.3 shows the aggregated t-test results obtained by each classifier against each other in terms of wins–ties–losses on each domain. Each classifier was optimized prior to being tested by the running of pairwise t tests on different parameterized versions of the same algorithm on all the domains. The parameters that win the greatest numbers of t tests among all the others, for one single classifier, were selected as the optimal ones.

As can be seen from these tables, results of this kind are difficult to interpret because they vary too much across both domains and classifiers. For example, the SVM seems to be superior to all the other algorithms on the balance scale and it apparently performs worst on breast cancer. Similarly, bagging is apparently

Table 1.2. *Accuracy results of various classifiers on the datasets of Table 1.1*

Dataset	NB	SVM	1NN	ADA(DT)	BAG(REP)	C45	RF	RIP
Anneal	96.43	99.44 v	99.11 v	83.63 *	98.22	98.44 v	99.55 v	98.22 v
Audiology	73.42	81.34	75.22	46.46 *	76.54	77.87	79.15	76.07
Balance scale	72.30	91.51 v	79.03	72.31	82.89 v	76.65	80.97 v	81.60 v
Breast cancer	71.70	66.16	65.74 *	70.28	67.84	75.54	69.99	68.88
Contact lenses	71.67	71.67	63.33	71.67	68.33	81.67	71.67	75.00
Pima diabetes	74.36	77.08	70.17	74.35	74.61	73.83	74.88	75.00
Glass	70.63	62.21	70.50	44.91 *	69.63	66.75	79.87	70.95
Hepatitis	83.21	80.63	80.63	82.54	84.50	83.79	84.58	78.00
Hypothyroid	98.22	93.58 *	91.52 *	93.21 *	99.55 v	99.58 v	99.39 v	99.42
Tic-tac-toe	69.62	99.90 v	81.63 v	72.54 v	92.07 v	85.07 v	93.94 v	97.39 v
Average	78.15	82.35	77.69	71.19	81.42	81.92	83.40	82.05 v
t test		3/6/1	2/6/2	1/5/4	3/7/0	3/7/0	4/6/0	4/6/0

Note: The final row gives the numbers of wins/ties/losses for each algorithm against the NB classifier.
Notes: A "v" indicates the significance test's success in favor of the corresponding classifier against NB while a "∗" indicates this success in favor of NB. No symbol indicates the result between the concerned classifier and NB were not found to be statistically significantly different.

Table 1.3. *Aggregate number of wins/ties/losses of each algorithm against the others over the datasets of Table 1.1*

Algorithm	NB	SVM	1NN	ADA	BAG	C45	RF	RIP
NB		3/6/1	2/6/2	1/5/4	3/7/0	3/7/0	4/6/0	4/6/0
SVM	1/6/3		0/6/4	0/5/5	2/5/3	2/6/2	2/6/2	1/7/2
1NN	2/6/2	4/6/0		0/5/5	2/8/0	2/8/0	3/7/0	2/8/0
ADA	4/5/1	5/5/0	5/5/0		6/4/0	5/5/0	6/4/0	6/4/0
BAG	0/7/3	3/5/2	0/8/2	0/4/6		1/7/2	1/9/0	1/8/1
C45	0/7/3	2/6/2	0/8/2	0/5/5	2/7/1		3/7/0	2/7/1
RF	0/6/4	2/6/2	0/7/3	0/4/6	0/9/1	0/7/3		1/8/1
RIP	0/6/4	2/7/1	0/8/2	0/4/6	1/8/1	1/7/2	1/8/1	

the second best learner on the hepatitis dataset and is average, at best, on breast cancer. As a consequence, the aggregation of these results over domains is not that meaningful either. Several other issues plague this evaluation approach in the current settings. Let us look at some of the main ones.

1.3.2 Issues with the Current Illustration

Statistical Validity – I. First we focus on the sample size of the domains. With regard to the sample size requirement, a rule of thumb suggests a minimum of 30 examples for a paired t test to be valid (see, for instance, Mitchell, 1997).[3] When 10-fold cross-validation experiments on binary datasets are run, this amounts to datasets of at least 300 samples. This assumption is violated in breast cancer and hepatitis. For the multiclass domains, we multiply this requirement by the number of classes and conclude that the assumption is violated in all cases but balance scale and hypothyroid. That is, at the outset, the assumption is violated in 6 out of 11 cases. This, of course, is only a quick rule of thumb that should be complemented by an actual visualization of the data that could help us determine whether the estimates are normally distributed (specific distributional oddities in the data could falsify the quick rule of thumb). In all cases for which the data is too sparse, it may be wiser to use a nonparametric test instead.

Statistical Validity – II. In fact, the dearth of data is only one problem plaguing the validity of the t test. Other issues are problematic as well, e.g., the interdependence between the number of experiments and the significance level of a statistical test. As suggested by Salzberg (1997), because of the large number of experiments run, the significance level of 0.05 used in our t test is not stringent enough: It is possible that, in certain cases, this result was obtained by chance. This is amplified by the fact that the algorithms were tuned on the same datasets

[3] We examine the sample size requirements later in the book.