Estimating confidence values of individual predictions by their typicalness and reliability

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

Abstract. Although machine learning algorithms have been successfully used in many problems, and are emerging as valuable data analysis tools, their serious practical use is affected by the fact that often they cannot produce reliable and unbiased assessments of their predictions' quality. There exist several approaches for estimating reliability or confidence for individual classifications, and many of them build upon the algorithmic theory of randomness, such as transduction-based confidence estimation, typicalness-based confidence estimation, and transductive reliability estimation. Unfortunately, they all have weaknesses: either they are tightly bound with particular learning algorithms, or the interpretation of reliability estimations is not always consistent with statistical confidence levels. In the paper we propose a joint approach that compensates the mentioned weaknesses by integrating typicalness-based confidence estimation and transductive reliability estimation into a joint confidence machine.

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 for the particular classification, so far there exists no general method. Note that we are interested in the assessment of classifier's performance on a *single example* and not in average performance on an independent dataset. Such single assessments are very useful, especially when used in ensembles and risk-sensitive applications [2] since 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) would not provide the desired information. There have been numerous attempts to assign probabilities to machine learning classifiers' in order to interpret their predictions as probability distributions over all 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 algorithm's bias(es) [1].

2 METHODS AND MATERIALS

Given some possible label space $\mathcal{Y}$, if an algorithm predicts some set of labels $Y \subseteq \mathcal{Y}$ with confidence t for a new example which is truly labelled by $y \in \mathcal{Y}$, we expect the confidence to have the following property: $P(y \notin Y) \leq 1-t$. As this paper deals only with single class predictions ($\tilde{y}$) the property can be simplified to $P(y \neq \tilde{y}) \leq 1-t$.

2.1 Typicalness

The typicalness framework [3, 5] can produce nearly precise confidence values estimations for data which is independently and identically distributed (*iid*). To measure the typicalness of sequences of labelled examples, we define, for every $n \in \mathbb{N}$, a function $t : \mathcal{Z}^n \rightarrow [0,1]$ which, for any $r \in [0,1]$ has the property $P((z_1,\ldots,z_n) : t(z_1,\ldots,z_n) \leq r) \leq r$. If such a 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 [3] that we can construct such typicalness functions by considering the strangeness ($\alpha(z_i)$) of individual examples. In machine learning, typicalness of a new example x labelled with $\tilde{y}$ given the training set $(z_1,\ldots,z_n)$ is calculated as

$$t((z_1,\ldots,z_n,x;\tilde{y})) = \frac{\#\{\alpha(z_i) : \alpha(z_i) \geq \alpha(x;\tilde{y})\}}{n+1} \qquad (1)$$

For a given machine learning algorithm, first we need to modify it in order to construct an appropriate strangeness measure. Then, for each new unlabelled example x, all possible labels $\tilde{y} \in Y$ are considered. For each label $\tilde{y}$ a typicalness of labelled example $t(x;\tilde{y}) = t((z_1,\ldots,z_n,x;\tilde{y}))$ is calculated. Finally, the example is labelled with "most typical" class, that is the one that maximizes $\{t(x;\tilde{y})\}$. The second largest typicalness is an upper bound on the probability that the excluded classifications are correct [4]. Consequently, the confidence is calculated as *confidence*$(x;\tilde{y}) = 1 -$ *typicalness of second most typical label*.

2.2 From reliability to confidence: merging typicalness and transduction frameworks

The transductive reliability estimation process and its theoretical foundations originating from algorithmic theory of randomness are described in more detail in [2]. Briefly sketched, an unlabelled example x is predicted a class $\tilde{y}$ and respective class probability distribution P by the given machine learning classifier. The example x is then labelled with the class $\tilde{y}$, the newly labelled example $(x;\tilde{y})$ is temporarily inserted into the training set, and then its class and class probability distribution Q are newly predicted. Reliability ($Rel(x;\tilde{y})$) of the predicted class is calculated as a similarity between the two class probability distributions, and is normalized to the $[0,1]$ interval. While this approach provides a measure that separates correct and incorrect classifications quite well, its numerical values usually cannot be interpreted as confidence levels, and their numerical ranges are very much domain- and algorithm- dependent [1]. This is a good

reason for merging typicalness and transductive reliability estimation frameworks. While transduction gives good reliability estimations, they are often hard to interpret in the statistical sense. On the other hand, the typicalness framework gives clear confidence values, however in order to achieve this a good strangeness measure needs to be constructed. In [3, 4] some ideas on how to construct strangeness measures for different machine learning algorithms are presented. However, we can always use transductive reliability estimation as a strangeness measure. We wish to treat most reliable examples as least strange. Therefore we define a general strangeness measure as follows.

$$\alpha(x_i;\widetilde{y}_i) = 1 - Rel(x_i;\widetilde{y}_i) \quad \in [0,1] \tag{2}$$

It can easily be shown that Eq. 2 satisfies the condition [3] required for strangeness measures.

3 RESULTS

To validate the proposed methodology we perform extensive experiments with 6 different machine learning algorithms on 15 well-known benchmark datasets. All experiments are performed by leave-one-out testing. In this setup, one example is reserved, while learning and preparatory calculations are performed on the rest. Usually, two nested leave-one-out testings are carried out.

The confidence values obtained by typicalness calculation are compared with transductive reliability estimations and kernel density estimation. Confidence values and reliability estimations perform similarly in terms of information gain (that is, their discrimination ability), while confidence values significantly ($p < 0.05$ with two-tailed, paired t-test) outperform reliability estimations in terms of correlation with correctness of classification. From Fig. 1 it is clear that this is because of the shift towards 1 and 0. Comparing confidence values and kernel density estimation shows a slightly different picture. Here, in terms of correlation with correctness as well as for information gain criterion, all differences are significant ($p < 0.01$ with two-tailed, paired t-test).

Figures 1(a) and 1(b) depict how reliability estimations in are transformed to confidence levels. This is a typical example and probably the most important result of our work, as it makes them easily statistically interpretable. On average, the best decision boundary for reliability estimations is 0.74, for confidence it is about 0.45. Also, the mass of correct and incorrect classification has shifted towards 1 and 0, respectively.

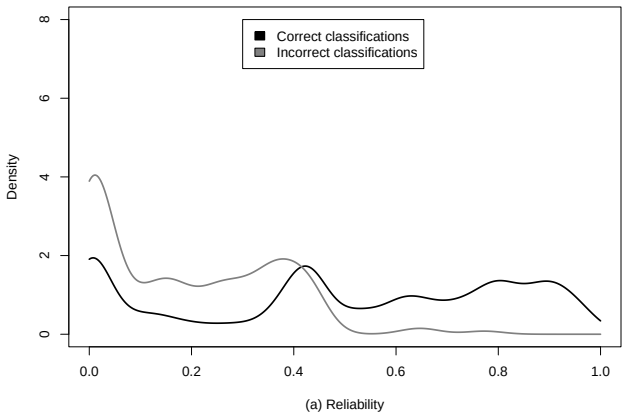
We also compared the results of confidence estimation on KNN (nearest neighbour) algorithm with that of the TCM-NN nearest neighbour confidence machine [4], where a tailor-made strangeness measure for confidence estimation in typicalness framework was constructed. Results of TCM-NN are slightly, though insignificantly better, as could be expected from a proprietary method.

4 DISCUSSION

We propose an approach that compensates the weaknesses of typicalness-based confidence estimation and transductive reliability estimation by integrating them into a joint confidence machine.

The resulting values are true confidence levels, and this makes them much easier to interpret. Contrary to the basic typicalness and transductive confidence estimation, the described approach is not bound to the particular underlying classifier. This is an important improvement since this makes possible to calculate confidence values for almost any classifier, no matter how complex it is.

Experimental results performed with several different machine learning algorithms in several problem domains show that there is no



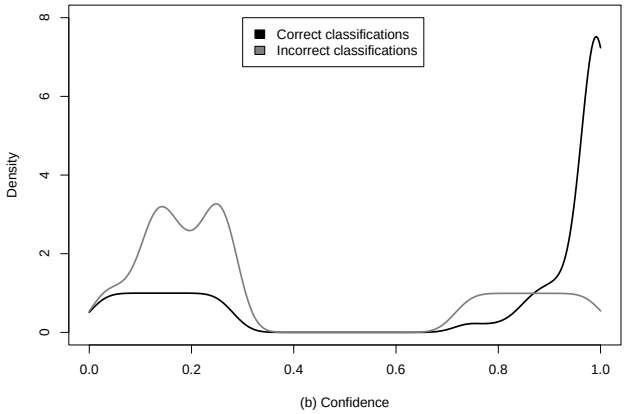


Figure 1. Densities of reliability estimations and confidence levels in Soybean dataset using neural networks.

reduction of discrimination performance with respect to transductive reliability estimation and proprietary approaches. More importantly, statistical interpretability of confidence values makes possible for applications in risk-sensitive problems with strict confidence limits.

The main drawback of our approach is computational complexity, as it needs to perform the leave-one-out testing in advance, and requires temporary re-learning of a classifier for each new example. However, this is not a problem if fast incremental learners (such as naive Bayesian classifier) are used.

ACKNOWLEDGEMENTS

This work was supported by the Slovenian Ministry of Education, Science and Sports.

REFERENCES

[1] M. Kukar, 'Transductive reliability estimation for medical diagnosis', *Artif. intell. med.*, 81–106, (2003).

[2] M. Kukar and I. Kononenko, 'Reliable classifications with Machine Learning', in *Proc. ECML 2002*, ed., T. Elomaa et al., pp. 219–231. Springer-Verlag, Berlin, (2002).

[3] T. Melluish, C. Saunders, I. Nouretdinov, and V. Vovk, 'Comparing the Bayes and typicalness frameworks', in *Proc. ECML 2001*, volume 2167, pp. 350–357, (2001).

[4] K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman, 'Transductive confidence machines for pattern recognition', in *Proc. ECML 2002*, pp. 381–390. Springer, Berlin, (2002).

[5] C. Saunders, A. Gammerman, and V. Vovk., 'Transduction with confidence and credibility', in *Proc. IJCAI'1999*, ed., T. Dean, Stockholm, Sweden, (1999). Morgan Kaufmann, San Francisco, USA.