Venn predictors and isotonic regression

Vladimir Vovk
v.vovk@rhul.ac.uk
http://vovk.net

November 2, 2012

Abstract

This note introduces Venn–Abers predictors, a new class of Venn predictors based on the idea of isotonic regression. As all Venn predictors, Venn–Abers predictors are well calibrated under the exchangeability assumption.

1 Introduction

This note is prompted by [2], which demonstrates that the probability forecasting procedure introduced by Zadrozny and Elkan in [4] (an adaptation of the isotonic regression procedure of [1]) can be poorly calibrated, whereas Venn predictors ([3], Chapter 6) are always well calibrated in their experiments and, moreover, are guaranteed to be well calibrated under the exchangeability assumption. This note shows that a simple modification of Zadrozny and Elkan's procedure is also a Venn predictor and so overcomes the problem of potential poor calibration. (The modified procedure, however, is a multiprobability predictor.)

2 Venn–Abers predictors

We consider *examples* $z = (x, y)$ consisting of two components: an *object* $x \in \mathbf{X}$ and its *label* $y \in \mathbf{Y}$. In this note we are only interested in the binary case and for concreteness set $\mathbf{Y} := \{0, 1\}$. We will use the notation $\Lbag a_1, \dots, a_n \Rbag$ for bags (in other words, multisets); the cardinality of the set $\{a_1, \dots, a_n\}$ might well be smaller than n (because of the removal of all duplicates in the bag). As usual, a "training set" is a bag of examples rather than a set. We say that a function f is *increasing* if its domain is an ordered set and $t_1 \le t_2 \Rightarrow f(t_1) \le f(t_2)$.

Many machine-learning algorithms for classification are in fact *scoring algorithms*: when trained on a training set of examples and fed with a test object x, they output a *prediction score* $s(x)$; we will call $s : \mathbf{X} \to \mathbb{R}$ the *scoring function* for that training set. The actual classification algorithm is obtained by fixing a threshold c and predicting the label of x to be 1 if and only if $s(x) \geq c$ (or if and only if $s(x) \geq c$). Alternatively, one could apply an increasing function g to $s(x)$ in an attempt to "calibrate" the scores, so that $g(s(x))$ can be used as the predicted probability that the label of x is 1.

Fix a scoring algorithm and let $\Lbag z_1, \dots, z_l \Rbag$ be a training set of examples $z_i = (x_i, y_i)$, $i = 1, \dots, l$. The most direct application [4] of the method of isotonic regression [1] to the problem of score calibration is as follows. Train the scoring algorithm on the training set and compute the score $s(x_i)$ for each training example (x_i, y_i), where s is the scoring function for $\Lbag z_1, \dots, z_l \Rbag$. Let g be the increasing function on the set $\{s(x_1), \dots, s(x_l)\}$ that maximizes the likelihood

$$\prod_{i=1}^{l} p_i, \qquad \text{where} \quad p_i := \begin{cases} g(s(x_i)) & \text{if } y_i = 1 \\ 1 - g(s(x_i)) & \text{if } y_i = 0. \end{cases} \tag{1}$$

Such a function g is indeed unique ([1], Corollary 2.1) and can be easily found using the "pair-adjacent violators algorithm" (PAVA, described in detail in the summary of [1] and, in a special case, in [4]; see also the proof of Lemma 1 below). We will say that g is the *isotonic calibrator* for $\Lbag (s(x_1), y_1), \dots, (s(x_l), y_l) \Rbag$. To predict the label of a test object x, the direct procedure finds the closest $s(x_i)$ to $s(x)$ and outputs $g(s(x_i))$ as its prediction (we do not go into details such as breaking the ties or the possibility of interpolation).

The direct procedure is prone to overfitting as the same examples $z_1, \dots, z_l$ are used both for training the scoring algorithm and for calibration without taking any precautions. The *Venn–Abers predictor* is the multiprobability predictor that is defined as follows. Try the two different classifications, 0 and 1, for the test object x. Let s_0 be the scoring function for $\Lbag z_1, \dots, z_l, (x, 0) \Rbag$, s_1 be the scoring function for $\Lbag z_1, \dots, z_l, (x, 1) \Rbag$, g_0 be the isotonic calibrator for $\Lbag (s_0(x_1), y_1), \dots, (s_0(x_l), y_l), (s_0(x), 0) \Rbag$, and g_1 be the isotonic calibrator for $\Lbag (s_1(x_1), y_1), \dots, (s_1(x_l), y_l), (s_1(x), 1) \Rbag$. The multiprediction output by the Venn–Abers predictor is $\{p_0, p_1\}$, where $p_0 := g_0(s_0(x))$ and $p_1 := g_1(s_1(x))$. (And we can expect p_0 and p_1 to be close to each other unless the direct procedure overfits grossly.)

In general, Venn–Abers predictors are computationally inefficient, especially if we would like to apply them to a large number of test examples and the same training set. More computationally efficient *pre-trained Venn–Abers predictors* are defined as follows. The training set $\Lbag z_1, \dots, z_l \Rbag$ is split into two parts: the *proper training set* $\Lbag z_1, \dots, z_m \Rbag$ of size $m < l$ and the *calibration set* $\Lbag z_{m+1}, \dots, z_l \Rbag$ of size $l - m$. Let $s : \mathbf{X} \to \mathbb{R}$ be the scoring function for $\Lbag z_1, \dots, z_m \Rbag$, g_0 be the isotonic calibrator for $\Lbag (s(x_{m+1}), y_{m+1}), \dots, (s(x_l), y_l), (s(x), 0) \Rbag$, and g_1 be the isotonic calibrator for $\Lbag (s(x_{m+1}), y_{m+1}), \dots, (s(x_l), y_l), (s(x), 1) \Rbag$. The multiprobability prediction output by the pre-trained Venn–Abers predictor is $\{p_0, p_1\}$, where $p_0 := g_0(s(x))$ and $p_1 := g_1(s(x))$. (This definition is in the spirit of inductive conformal predictors [3], Section 4.1, but we avoid using the term "inductive Venn–Abers predictors" since our pre-trained Venn–Abers predictors are not inductive Venn predictors the sense of [2], Section 3.1.)

Venn predictors are defined as in [3], Chapter 6, except that a probability distribution P on the set $\{0,1\}$ is now represented by the number $P(\{1\}) \in [0,1]$.

Proposition 1. *Venn–Abers predictors are Venn predictors. Pre-trained Venn–Abers predictors are Venn predictors when considered as functions of* $(z_{m+1},\dots,z_l)$.

Proof. Fix a Venn–Abers predictor. The corresponding taxonomy is defined as follows: assign $(\lbag z_1,\dots,z_n \rbag,(x,y))$ and $(\lbag z'_1,\dots,z'_{n'} \rbag,(x',y'))$ to the same cell if and only if $g(s(x)) = g'(s'(x'))$, where s is the scoring function for $\lbag z_1,\dots,z_n,(x,y) \rbag$, s' is the scoring function for $\lbag z'_1,\dots,z'_{n'},(x',y') \rbag$, g is the isotonic calibrator for $\lbag (s(x_1),y_1),\dots,(s(x_n),y_n),(s(x),y) \rbag$, and g' is the isotonic calibrator for $\lbag (s'(x'_1),y'_1),\dots,(s'(x'_{n'}),y'_{n'}),(s'(x'),y') \rbag$. Lemma 1 below shows that the Venn predictor corresponding to this taxonomy gives predictions identical to those given by the original Venn–Abers predictor. This proves the first statement of the proposition.

The second statement follows from the fact that for a fixed bag $\lbag z_1,\dots,z_m \rbag$ the pre-trained Venn–Abers predictor is the Venn–Abers predictor corresponding to a scoring function $s_0 = s_1 = s$ that does not depend on the data $\lbag z_{m+1},\dots,z_l \rbag$ at all. □

Lemma 1. *Let g be the isotonic calibrator for $\lbag (t_1,y_1),\dots,(t_n,y_n) \rbag$, where $t_i \in \mathbb{R}$ and $y_i \in \{0,1\}$, $i=1,\dots,n$. Any $p \in \{g(t_1),\dots,g(t_n)\}$ is equal to the arithmetic mean of the labels y_i of the t_i, $i=1,\dots,n$, satisfying $g(t_i)=p$.*

Proof. The statement of the lemma immediately follows from the definition of the PAVA ([1], summary), which we will reproduce here. Arrange the numbers t_i in the strictly increasing order $t_{(1)} < \dots < t_{(k)}$, where $k \le n$ is the number of distinct elements among t_i. We would like to find the increasing function g on the set $\{t_{(1)},\dots,t_{(k)}\} = \{t_1,\dots,t_n\}$ maximizing the likelihood (defined by (1) with t_i in place of $s(x_i)$ and n in place of l). The procedure is recursive. At each step the set $\{t_{(1)},\dots,t_{(k)}\}$ is partitioned into a number of disjoint cells consisting of adjacent elements of the set; to each cell is assigned a ratio a/N (formally, a pair of integers, with $a \ge 0$ and $N > 0$); the function g defined at this step (perhaps to be redefined at the following steps) is constant on each cell. For $j=1,\dots,k$, let a_j be the number of i such that $y_i = 1$ and $t_i = t_{(j)}$, and let N_j be the number of i such that $t_i = t_{(j)}$. Start from the partition of $\{t_{(1)},\dots,t_{(k)}\}$ into one-element cells, assign the ratio a_j/N_j to $\{t_{(j)}\}$, and set

$$g(t_{(j)}) := \frac{a_j}{N_j} \tag{2}$$

(in the notation used in this proof, a/N is a pair of integers whereas $\frac{a}{N}$ is a rational number, the result of the division). If the function g is increasing, we are done. If not, there is a pair C_1, C_2 of adjacent cells ("violators") such that C_1 is to the left of C_2 and $g(C_1) > g(C_2)$ (where $g(C)$ stands for the common value of $g(t_{(j)})$ for $t_{(j)} \in C$); in this case redefine the partition by merging C_1

and C_2 into one cell C, assigning the ratio $(a_1 + a_2)/(N_1 + N_2)$ to C, where a_1/N_1 and a_2/N_2 are the ratios assigned to C_1 and C_2, respectively, and setting

$$g(t_{(j)}) := \frac{N_1}{N_1 + N_2} g(C_1) + \frac{N_2}{N_1 + N_2} g(C_2) = \frac{a_1 + a_2}{N_1 + N_2} \tag{3}$$

for all $t_{(j)} \in C$. Repeat the process until g becomes constant (the number of cells decreases by 1 at each iteration, so the process will terminate in at most k steps). The final function g is the one that maximizes the likelihood. The statement of the lemma follows from this recursive definition: it is true by definition for the initial function (2) and remains true when g is redefined by (3). □

3 Conclusion

This note has introduced a new class of Venn predictors thereby extending the domain of applicability of the method.

Acknowledgments

Partially supported by the Cyprus Research Promotion Foundation (research contract TPE/ORIZO/0609(BIE)/24).

References

[1] Miriam Ayer, H. Daniel Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 26:641–647, 1955.

[2] Antonis Lambrou, Harris Papadopoulos, Ilia Nouretdinov, and Alex Gammerman. Reliable probability estimates based on support vector machines for large multiclass datasets. In Lazaros Iliadis, Ilias Maglogiannis, Harris Papadopoulos, Kostas Karatzas, and Spyros Sioutas, editors, *Proceedings of the AIAI 2012 Workshop on Conformal Prediction and its Applications*, volume 382 of *IFIP Advances in Information and Communication Technology*, pages 182–191, Berlin, 2012. Springer.

[3] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[4] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699. ACM Press, 2002.