Improving reliable probabilistic prediction by using additional knowledge

Ilia Nouretdinov

Information Security Group
Computer Learning Research Centre
Royal Holloway*, University of London

June 14, 2017

1. Venn machine
2. Additional information and transfer
3. Experimental validation

Venn machine

Let $x \in X = R^k$ be a feature vector, $y \in Y = \{0, 1\}$ be a label.

- Underlying Machine Learning method outputting scores:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l); x \to s$$

  where s is a probabilistic (or scoring) estimate of $y = 1$.

- Venn machine output:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l); x \to (p_0, p_1)$$

  where p_0 and p_1 are lower and upper estimates that $y = 1$.

Based on an underlying method, Venn framework rearranges probabilistic outputs so that they become valid in weaker (i.i.d./exchangeability) assumptions.

Venn Machines and taxonomies

A *Venn taxonomy* T is a function that assigns an equivalence relation on a set of examples (x, y). A *Venn predictor* is completely defined by the taxonomy relation as a parameter.

$$p_y = \frac{card\{i = 1, \ldots, l+1 : t_i = t_{l+1}, y_i = 1\}}{card\{i = 1, \ldots, l+1 : t_i = t_{l+1}\}}$$

where $t_1, \ldots, t_{l+1}$ are numbers of categories to which the examples $(x_1, y_1), \ldots, (x_l, y_l), (x, y)$ belong after applying the Venn taxonomy.

The meaning of p_0 and p_1 is lower and upper estimates of probability that the label of x is 1.

The prediction of y itself is 1 if $p_0 + p_1 > 1$ and 0 otherwise, while p_0 and p_1 reflect its reliability.

k Nearest Neighbours Venn-ABERS taxonomy

Assume that a metric (distance function) d is defined on the space X. For each i, all the examples $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ can be sorted by the distance to this example. Let us select k first examples (neighbours of x_i) and look at the empirical distribution of their labels. Among then there may be at least 0 and at most k examples with label $y_j = 1$, and the average of them p_i can be understood as an estimate of local conditional density of y in the area around x_i.

Venn-Abers method allows to make an automatic regulation of the number of taxonomies. It is constructed as a monotonic function of p_i in such a way that conditional empirical probability of y_i given t_i is increasing.

This leads to a smaller number of categories than $k + 1$.

Additional information

The topic of this work is to get use of *additional information* that is available only for some of the training objects, within Venn framework. In Vapnik's privileged information the principal point is that this additional information is not available for the testing example x as well.

In the current work we also do not assume that additional information is available for *all* training examples.

Idea of taxonomy transfer

It is desirable to use as more features as possible for the initial taxonomy design. Therefore Venn machine is first applied to the subset of example with additional information, and a category is assigned to each of them.

As for the rest of the examples, we calculate transferred taxonomies for them (e.g. by means of 1 Nearest Neighbour.

Strictly saying, putting together the extended and transferred taxonomies violates the definition of Venn Machine. The solution of this problem, inspired by Inductive Venn Prediction to consider all the examples *with* additional info just as an auxiliary set.

Algorithm of taxonomy transfer

- INPUT: labelled examples without additional information: $z_j = (x_i, y_i)$.
- INPUT: a testing unlabelled example x.
- INPUT: auxiliary examples (x'_j, h'_j, y'_j), denote $\tilde{x}_j = (x'_j, h'_j)$.
- INPUT: an underlying method of probabilistic predictions
- Find the corresponding Venn-Abers taxonomy function T_1 corresponding to the underlying method:
  T_1 inputs the examples $(\tilde{x}_1, y'_1), \ldots, (\tilde{x}_m, y'_m)$ and outputs their categories $t'_1, \ldots, t'_m$.
- Define the second taxonomy function T_2 which inputs $(z_1, \ldots, z_l, (x_{l+1}, y))$ and outputs $t_i = t'_j$
  where x'_j is the nearest neighbour of x_i amongst $x'_1, \ldots, x'_m$.
- Run Venn machine with taxonomy function T_2 on $z_1, \ldots, z_l, x$
- OUTPUT probabilistic prediction of x's label y.

Experimental validation

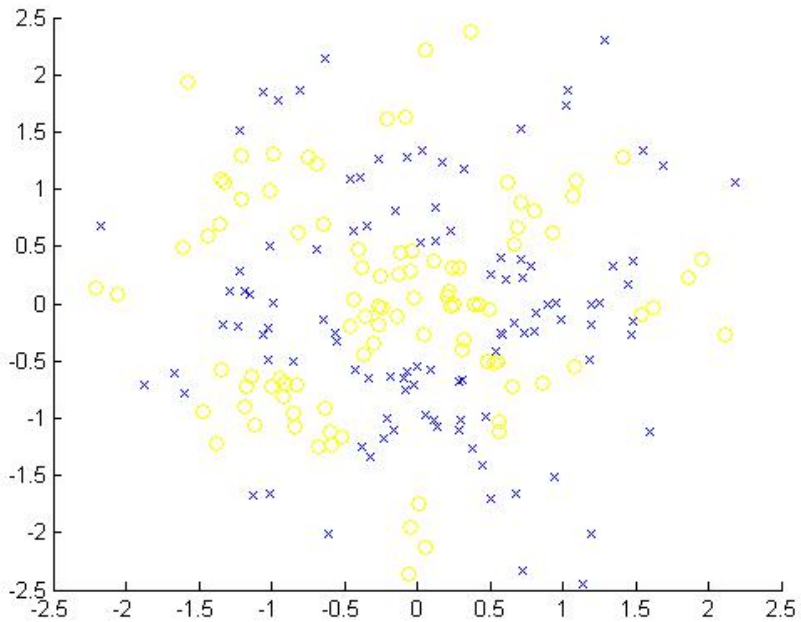
Assume that for N first data examples only the first feature is available, while for N remaining examples there are both features.
We will use the data in the mode of leave-one-out cross-validation.
We compare two principal alternatives:

1. To ignore the privileged information and to use only the features available for all $2N$ examples.
2. To try to get use of all the available features.

Data generation

(x_i, h_i) is generated by standard two-dimensional normal distribution $N(0, 1)$. The label is

$$y_i = \mod\ (round(x_i) + round(h_i), 2)\,.$$



Comparative evaluation results

Leave-one-out for the first N examples based on the training set of size $2N - 1$, averaged over 1,000 random generations.

N (examples)	without add.info	with add.info	improvements
25	0.4679	0.5032	551/1000
50	0.4777	0.5014	568/1000
100	0.4880	0.4957	583/1000
150	0.4912	0.4974	604/1000
200	0.4949	0.4983	566/1000
250	0.4966	0.4948	586/1000
300	0.4912	0.4957	592/1000
350	0.4944	0.5000	597/1000
400	0.4959	0.5055	623/1000

Future work

The experimental part was done on an artificial example, which emulates the case when the additional information is very valuable. However, in real applications it can be put on some scale of importance, between being very useful and being noisy/redundant. Key question: which real data applications will gain from using the additional information?

Another question: how this idea may be reflected for the problem of *missing values*. In such case taxonomy transfer would be alternative for two possible 'baselines': ignoring incomplete features and also ignoring incomplete examples.

Some references

- Vovk, V., Gammerman, A., Shafer, G. Algorithmic Learning in a Random World. Springer, 2005
- Vovk, V., Petej, I. Venn-Abers predictors. `http://alrw.net`, Working Paper 7.
- Vapnik, V., Izmailov, R. Learning with Intelligent Teacher: Similarity Control and Knowledge Transfer. Statistical Learning and Data Sciences. Third International Symposium, SLDS 2015, Egham, UK, April 20-23, 2015, Proceedings, pp.3–32.