Improving Reliable Probabilistic Prediction by Using Additional Knowledge

Ilia Nouretdinov I.R.NOURETDINOV@RHUL.AC.UK
Royal Holloway, University of London, Egham, Surrey, UK

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos

Abstract

Venn Machine is a recently developed machine learning framework for reliable probabilistic prediction of the labels for new examples. This work proposes a way to extend Venn machine to the framework known as Learning Under Privileged Information: some additional features are available for a part of the training set, and are missing for the example being predicted. We make use of this information by making a taxonomy transfer, where taxonomy is the core detail of Venn Machine framework. The transfer is done from the examples with additional information to the examples without additional information.

Keywords: Venn machine, reliable probabilistic prediction, additional information, transfer.

1. Introduction

A general task of supervised machine learning is to predict the label (classifier)for a new object with a feature vector based on the labels of the previous feature vectors.

Venn machine for reliable probabilistic prediction valid under weak (i.i.d. or exchangeability) assumptions was presented in the book of Vovk, Gammerman & Shafer (2005). It is linked to the underlying method by its code parameter called taxonomy. A useful modification of Venn framework called Venn-Abers machine was developed by Vovk & Petej (2012). It allows to create a taxonomy directly from any underlying machine learning algorithm which outputs probabilities or scores.

The practical advantage of Venn machines over standard probabilistic methods was shown by Zhou et al. (2011). Based on an underlying method, Venn framework rearranges probabilistic outputs so that they become valid in weaker assumptions.

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. This is related to the problem of missing values and to Vapnik's Learning Under Privileged Information Paradigm observed by Vapnik & Izmailov (2015). The principal point is that this additional information is not available for the testing example x as well. This reflects a practical situation when this part of information is expensive: one has time to collect it for the training data, but not for the testing examples as they arrive on-line. The works on LUPI paradigms show that additional information may be a useful 'hint' to increase the speed of learning.

In the current work we do not assume that additional information is available for all training examples. In many realistic scenarios the training set can contain example both with and without these additional features.

© 2017 I. Nouretdinov.

Some preliminary modification of conformal framework for this problem was presented by Yang et al. (2013). It considered all hypotheses about the values of additional feature as possible. But there was no attempt to summarise the knowledge obtainable from the additional information.

The work by Vapnik & Izmailov (2015) also contains a new interpretation of Learning Under Privileged Information, called knowledge transfer, when some elements of knowledge are extracted from the feature space extended with privileged information and then transferred to the primary feature space.

In this work we make another kind of transfer applied to the Venn reliable probabilistic framework. As mentioned above, the core detail of Venn machine is a taxonomy, which usually links Venn framework to the underlying (usually probabilistic) method of prediction. Once the taxonomies are assigned to the examples with additional info, we try to transfer them to all the rest examples.

This work is also influenced by Inductive Venn Machine framework by Lambrou et al. (2015). This approach for quickness divides the data set into two parts: a proper training set used as an external background base of defining taxonomies, and the calibration set which examples are considered comparable to testing ones. It appears that the examples with additional information may be put into an analogue of the proper training set.

The plan of the paper is following. In Sections 2 and 3 we recall the notions related to Venn machines, and what exactly is meant by taxonomy transfer. In Section 4 we develop experimental check. Section 5 is the conclusion part.

2. Machine Learning Background

2.1. Venn Machines and Taxonomies

Following Vovk & Petej (2012), we consider examples $z = (x, y)$ that consist of objects $x \in X$ and the labels $y \in Y$, where X is a measurable set (usually a vector space) and Y is the label space. In this work we consider for simplicity only the binary case $Y = \{0, 1\}$.

A Venn taxonomy T is a measurable function that assigns an equivalence relation on a set $\{z_1, \ldots, z_n\}$ of examples, typically dividing it into a relatively small number of categories.

A Venn predictor is completely defined by the taxonomy relation as a parameter. The prediction algorithm inputs a training set $\{z_1, \ldots, z_l\}$ where $z_i \in X \times Y$, and an unlabelled testing example $x \in X$, and outputs a pair (p_0, p_1). It is calculated as follows.

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

where $t_1, \ldots, t_{l+1}$ are numbers of categories to which the examples $z_1, \ldots, z_l, (x, y)$ 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.

The validity of this output as a probabilistic one is shown by Vovk & Petej (2012).

2.2. Defining Taxonomies

The taxonomy is a way to link Venn framework with an underlying machine learning method. We will illustrate this on the example of Nearest Neighbours.

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.

As far as p_i is discrete, its possible values range from 0 to 1 with step $1/k$, it is possible to use p_i directly as the taxonomy value t_i. But it is known that for the effectiveness on Venn Machine, the number of categories (different values of the taxonomy function) should not be too large. Venn-Abers method presented in details by Vovk & Petej (2012) allows to make an automatic regulation of the number of taxonomies. In this version of the framework, t_i is a monotonic function of p_i. It is constructed in such a way that conditional empirical probability of y_i given t_i is strictly increasing. This usually leads to a smaller number of categories than $k + 1$.

3. Algorithms

3.1. Idea of Taxonomy Transfer

Assume now that some additional information h_j is available for some of the examples z_j is addition to their main feature vector x_j and label y_j. How to use it for the prediction of the label y for a new example x, assuming that h is not available for this example as well?

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. Each of examples without additional info is assigned the same category as its first nearest neighbour amongst the examples with additional info.

Strictly saying, putting together the extended and transferred taxonomies violates the definition of Venn Machine. The solution of this problem, hinted by Inductive Venn Prediction by Lambrou et al. (2015), is to consider all the examples with additional info just as an auxiliary set.

Let us now summarize this as a formal plan.

3.2. Algorithm of Taxonomy Transfer

- INPUT: labelled examples $z_1, \ldots, z_l$ without additional information: $z_j = (x_i, y_i)$.
- INPUT: a testing unlabelled example x.
- INPUT: auxiliary examples supplied with additional information are (x'_j, h'_j, y'_j) with their own numeration $j = 1, \ldots, m$.

  Denote extended feature vectors $\tilde{x}_j = (x'_j, h'_j)$.

- INPUT: an underlying method of probabilistic predictions, applicable to the labelled extended feature vectors $z = (\tilde{x}, y)$
- 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.

4. Experimental Validation

For our experiments we generate a special artificial example of data set. Let there be $2N$ examples with two possible labels (0 and 1) and two dimensions of a feature vector. 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. Assume that the task is to predict the label of the example number i ($1 \le i \le N$), using the labels of all the rest examples $(1, \ldots, i-1, i+1, \ldots, 2N)$ and all their available features.

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.

In the first case the application of a machine learning method may be straightforward. We use Venn-Abers version recommended by Vovk & Petej (2012). To get use of additional info, we apply the plan from Section 3.2 as follows.

- Using the second group of N examples which have the additional info, we construct a Venn-Abers taxonomy using both main and additional features. The taxonomy is based on k-Nearest-Neighbours underlying method. After this stage, each of the examples $N+1, \ldots, 2N$ is assigned one of K taxonomies.
- The taxonomy is extended (transferred) to the examples $1, \ldots, N$ according to 1-nearest-neighbour rule.
- Leave-one-out prediction on examples $1, \ldots, N$ is done using Venn machine with this transferred taxonomy.

If the accuracy increases from the case when the additional info was just ignored, we can conclude that the additional information helped the learning.

In addition we may note that although the second alternative is based on more information, it is actually quicker to run, due to a trick similar to one used in Inductive Venn Machines by Lambrou et al. (2015).

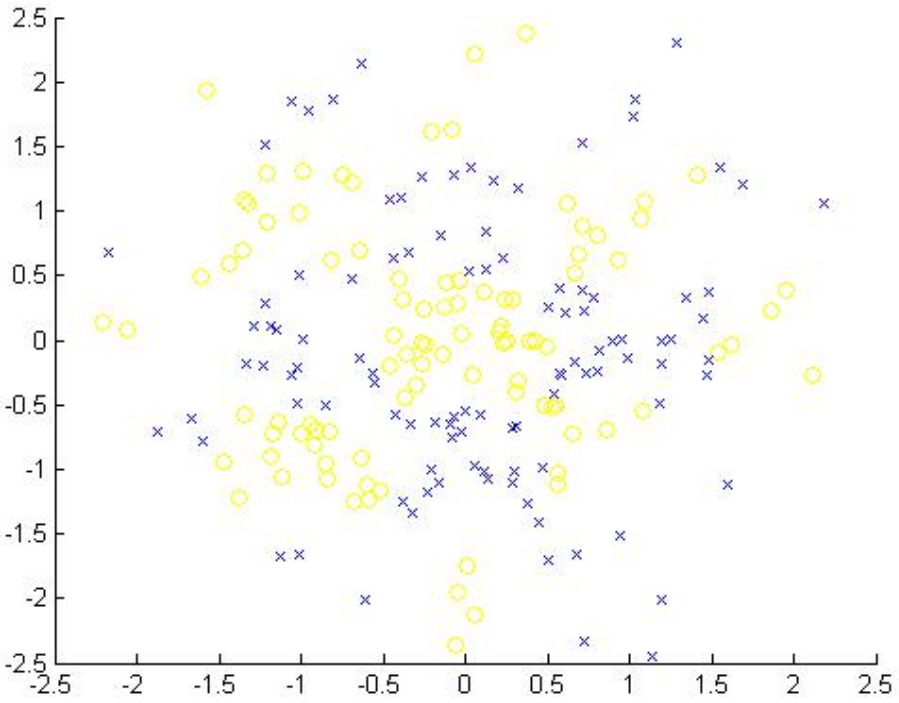

Figure 1: Artificial data example. The first axis represents the main feature x, the second is for the piece h of additional information. Two classes are shown as crosses and circles.

4.1. Data Generation

The data is generated randomly as follows. Each of $2N$ examples (x_i, h_i) is generated by standard two-dimensional normal distribution $N(0, 1)$. The additional feature h_i is hidden (never used for prediction training) for the examples $1, \ldots, N$.

The label is calculated by formula

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

This way we model a case when the separation of two classes is easy with both the features but very hard with only one of them.

A typical result of this generation can be seen on Fig. 4.1.

4.2. Baseline Experimental Setting

For probabilistic prediction we use Venn-Abers algorithm by Vovk & Petej (2012) constructed on the base of k Nearest Neighbours underlying method with $k = 20$.

For the baseline experimental setting, we train on as much examples as possible. They all are assigned equal roles wherever the additional information is available for them or not.

We apply it in leave-one-out mode. For the needs of further comparison we test it only on N first data examples. However, each of them being trained on training set of size $2N-1$.

In this part of the experiment, all the examples are used but the distances for k-NN method are calculated using only the main feature x.

4.3. Extended Taxonomy Generation and Transfer

In proper setting we also start with Venn-Abers machine, but do not make any predictions, the only aim is to assign categories $t_{N+1},\dots,t_{2N}$ to the examples $x_{N+1},\dots,x_{2N}$.

We also use underlying k-Nearest-Neighbours method with $k=20$.

The next step is to create taxonomies $t_1,\dots,t_N$ for the examples $x_1,\dots,x_N$. This is done by simple 1 Nearest Neighbour method: each example x_i is assigned the same taxonomy as its nearest neighbour amongst the examples $x_{N+1},\dots,x_{2N}$.

The number of neighbours is set to 1, because using $k>1$ nearest neighbours would lead to undesirable change of relative sizes of categories.

The distance is calculated based only on the main feature x_i.

The result of this transfer does not depend on the order of the examples $x_1,\dots,x_n$ therefore it is valid as a taxonomy for Venn machine.

Note that the taxonomy as well does not depend on labels $y_1,\dots,y_N$ of the examples $x_1,\dots,x_N$, so there is no need to calculate it more than once.

After the transfer is made, it remains to apply Venn Machine for the examples $x_1,\dots,x_N$ which is done in leave-one-out mode.

4.4. Comparative Evaluation Results

We try different data size. For example, $N=25$ means that there 25 training/testing examples without additional information and 25 extra examples extended by additional info. All the predictions are finally made for the first N examples. When prediction is made without additional information, each prediction is based on the training set of size $2N-1$. If additional information is used, then second N examples are put into the auxiliary set, so the size of training set for a final prediction is $N-1$.

All the results are averaged over 1,000 random generations of the data. Table 1 contains the following results for each size N. First, averaged accuracy of Venn machine applied without and with using privileged information. By accuracy we mean percentage of correctness within the prediction made by the simlified rule: an example's predicted label is 1 if $p_0+p_1>1$ i.e. average of lower and upper bounds is closer to 1 than to 0. All these accuracies are around 0.5.

A more interesting result is presented in the last column, it reflects the number of random data seeds (of 1000) for which the accuracy increases after using additional information. It appeared that for any considered size this chance of improvement is between $55-62\%$. This improvement is significant: the probability to get more than 550 of 1000 improvements by chance is less than 0.0007.

N (examples)	av.accuracy without add.info	av.accuracy with add.info	improvements
25	0.4679	0.5032	551/1000
50	0.4777	0.5014	568/1000
75	0.4863	0.5009	572/1000
100	0.4880	0.4957	583/1000
125	0.4917	0.4969	577/1000
150	0.4912	0.4974	604/1000
175	0.4905	0.5021	606/1000
200	0.4949	0.4983	566/1000
225	0.4953	0.5028	612/1000
250	0.4966	0.4948	586/1000
275	0.4919	0.4945	590/1000
300	0.4912	0.4957	592/1000
325	0.4942	0.4965	595/1000
350	0.4944	0.5000	597/1000
375	0.5005	0.5001	588/1000
400	0.4959	0.5055	623/1000

Table 1: Evaluation

5. Conclusion and Discussion

In this work we have shown a way to combine reliable probabilistic prediction given by Venn machine and using additional information that is partially available for the data used in machine learning.

A promising direction to place the contribution of the additional info appeared to be the stage of taxonomy calculation, which is a sort of heuristic element in the Venn machine framework.

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. There is much space for further work to determine in which real data applications the gain from using the additional information is the largest.

In this work we used well-known Nearest Neighbours underlying algorithm because of its simplicity. In the future it may be interesting to get taxonomies from decision trees and other algorithms.

Another question is how this idea may be reflected for the problem of missing values that complements the privileged information statements: some features are available for the testing example but hidden for a part of the training set. In such case taxonomy transfer would be alternative for two possible 'baselines': ignoring incomplete features and also ignoring incomplete examples.

6. Acknowledgments

This work was supported by Technology Integrated Health Management (TIHM) project awarded to the School of Mathematics and Information Security at Royal Holloway as part of an initiative by NHS England supported by InnovateUK.

It was also supported by European Union grant 671555 ("ExCAPE"), EPSRC grant EP/K033344/1 ("Mining the Network Behaviour of Bots"), Thales grant ("Development of automated methods for detection of anomalous behaviour"), and by the National Natural Science Foundation of China (No.61128003) grant.

We are grateful to the reviewers especially for some recommendations for further research.

References

[1] Lambrou, A., Nouretdinov, I., Papadopoulos, H. Inductive Venn Prediction. Annals of Mathematics and Artificial Intelligence archive Volume 74 Issue 1-2, June 2015, pp.181–201.

[2] 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.

[3] Vovk, V., Gammerman, A., Shafer, G. Algorithmic Learning in a Random World. Springer, 2005

[4] Vovk, V., Petej, I. Venn-Abers predictors. http://alrw.net, Working Paper 7, 2012.

[5] Yang, M., Nouretdinov, I., Luo, Z. Learning by Conformal Predictors with Additional Information The 9th Artificial Intelligence Applications and Innovations Conference (AIAI): 2nd Workshop on Conformal Prediction and its Applications, (IFIP Advances in Information and Communication Technology; vol. 412), 2013, pp. 394–400.

[6] Zhou, C., Nouretdinov, I., Luo, Zh., Adamskiy, D., Coldham, N., Gammerman, A. A Comparison of Venn Machine with Platt's Method in Probabilistic Outputs. 12th INNS EANN-SIG International Conference, EANN 2011 and 7th IFIP WG 12.5 International Conference, Artificial Intelligence Applications and Innovations 2011, Corfu, Greece, September 15-18, 2011, Proceedings , Part II. 2011. p. 483–492.