A Comparison of Venn Machine with Platt's Method in Probabilistic Outputs

Chenzhe Zhou[1], Ilia Nouretdinov[1], Zhiyuan Luo[1], Dmitry Adamskiy[1], Luke Randell[2], Nick Coldham[2], and Alex Gammerman[1]

[1] Computer Learning Research Centre, Royal Holloway, University of London,
Egham, Surrey, TW20 0EX, UK

[2] Veterinary Laboratories Agency, Weybridge, Surrey, KT15 3NB, UK

Abstract. The main aim of this paper is to compare the results of several methods of prediction with confidence. In particular we compare the results of Venn Machine with Platt's Method of estimating confidence. The results are presented and discussed.

Keywords: Venn Machine, Support Vector Machine, Probabilistic Estimation

1 Introduction

There are many machine learning algorithms that allow to make classification and regression estimation. However, many of them suffer from the absence of a confidence measure to assess the risk of error made by an individual prediction.

Sometimes, however, the confidence measure is introduced but very often it is an ad hoc measure. An example of this is a Platt's algorithm developed to estimate confidence for SVM[1]. We recently developed a set of new machine learning algorithms [2,3] that allow not just to make prediction but also to supply this prediction with a measure of confidence. What's more important is that this measure is valid and based on a well-developed algorithmic randomness theory.

The algorithm introduced in this paper is Venn Machine[3], a method that outputs the prediction with an interval of probability that prediction is correct. What follows is an introduction to Venn Machine and Platt's Method, then description of used data and results of experiments.

1.1 Venn Machine

Let us consider a training set consisting of object, x_i, and label, y_i, as pairs: $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$. The possible labels are finite, that is, $y \in \mathbf{Y}$. Our task is to predict the label y_n for the new object x_n and give the estimation of the likelihood that our prediction is correct.

In brief, Venn Machine operates as follows. First, we define a *taxonomy* that can divide all examples into categories. Then, we try all the possible labels of the new object. In each attempt, we can calculate the frequencies of the labels in the category which the new object falls into. The minimum frequency is called the *quality* of this column. At last, we output the assumed label with the highest *quality* among all the columns as our prediction and output the minimum and the maximum frequencies of this column as the interval of the probability that this prediction is correct.

Taxonomy (or, more fully, *Venn taxonomy*) is a function A_n, $n \in \mathbf{N}$ of the space $\mathbf{Z^{(n-1)}} \times \mathbf{Z}$ that divide every example into one of the finite categories τ_i, $\tau_i \in \mathbf{T}$. Then we consider z_i as the pair (x_i, y_i),

$$\tau_i = A_n(\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}, z_i) \tag{1}$$

We assign z_i and z_j to the same category if and only if

$$A_n(\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}, z_i) = A_n(\{z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_n\}, z_j) \tag{2}$$

Here is an example of a simplest *taxonomy* based on 1-nearest neighbour (1NN).

We assign the category of an example the same to the label of its nearest neighbour based on the distance between two objects (e.g. Euclidean distance).

$$A_n(\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}, z_i) = \tau_i = y_j \tag{3}$$

where

$$j = \arg \min_{j=1,\ldots,i-1,i+1,\ldots,n} ||x_i - x_j|| \tag{4}$$

For every attempt (x_n, y), of which the category is τ, let p_y be the empirical probability distribution of the labels in category τ.

$$p_y\{y'\} := \frac{|\{(x^*, y^*) \in \tau : y^* = y'\}|}{|\tau|} \tag{5}$$

this is a probability distribution on $\mathbf{Y}$. The set $P_n := \{p_y : y \in \mathbf{Y}\}$ is the multiprobability predictor consists of K probabilities, where $K = |Y|$.

After all attempts, we get a $K \times K$ matrix P. Let the *best* column with the highest quality, which is the minimum entry of a column, be j_{best}. j_{best} is our prediction and the interval of the probability that the prediction is correct is

$$[\min_{i=1,\ldots,K} P_{i,j_{best}}, \max_{i=1,\ldots,K} P_{i,j_{best}}] \tag{6}$$

1.2 Platt's Method

Standard Support Vector Machines (SVM) [4] only output the value of $sign(f(x_i))$, where f is the decision function. So we can say that SVM is a non-probabilistic binary linear classifier. But in many cases we are more interested in the belief that the label should be +1, that is, the probability $P(y = 1|x)$. Platt introduced a method to estimate posterior probabilities based on the decision function f by fitting a sigmoid for SVM.

$$P(y = 1|f) = \frac{1}{1 + exp(Af + B)} \tag{7}$$

The best parameter A and B are determined by using maximum likelihood estimation from a training set (f_i, y_i). Let us use regularized target probabilities t_i as the new training set (f_i, t_i) defined as:

$$t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2}, \text{ if } y_i = +1 \\ \frac{1}{N_- + 2}, \text{ if } y_i = -1 \end{cases} \tag{8}$$

where N_+ is the number of positive examples, while N_- is the number of negative examples. Then, the parameters A and B are found by minimizing the negative log likelihood of the training data, which is a cross-entropy error function.

$$-\sum_i (t_i \log(p_i) + (1 - t_i)\log(1 - p_i)) \longrightarrow \min_{p_i} \tag{9}$$

where the solution is

$$p_i = \frac{1}{1 + exp(Af_i + B)} \tag{10}$$

With parameters A and B we can calculate the posterior probability that the label should be +1 of every example using (10). But in many cases, probability that the prediction is correct is more useful and easy to compare with Venn Machine. In this binary classification problem, one example with the probability p_i means its label should be +1 with the likelihood of p_i, that is to say, its label should be −1 with the likelihood of $1 - p_i$. So we use the complementary probability when the probability is less than the optimal threshold (in this paper we set it to 0.5 as explained later).

2 Data Sets

The data sets we used in this paper is Salmonella mass spectrometry data provided by VLA[3] and Wisconsin Diagnostic Breast Cancer (WDBC) data from UCI.

The aim of the study of Salmonella data is to discriminate Salmonella vaccine strains from wild type field strains of the same serotype. We analysed the set of 50 vaccine strains (Gallivav vaccine strain) and 43 wild type strains. Both vaccine and wild type strains belong to the same serotype Salmonella enteritidis.

Each strain was represented by three spots; each spot produced 3 spot replicates. Therefore, there are 9 replicates per strain. Pre-processing was applied to each replicate and resulted in representation of each mass spectra as a vector of 25 features corresponding to the intensity of most common peaks. The median was later taken for each feature across replicates of the same strain. In the data set, label +1 corresponds to vaccine strains, label −1 to wild type strains. Table 1 shows some quantitive properties of the data set.

In Figure 1, there is a plot of the class-conditional densities $p(f|y = \pm 1)$ of Salmonella data. The plot shows histograms of the densities of the data set with

[3] Veterinary Laboratories Agency

Table 1. Salmonella Data Set Features

Number of Instances	Number of Attributes	Number of Positive Examples	Number of Negative Examples
93	25	50	43

bins 0.1 wide, derived from Leave-One-Out Cross-Validation. The solid line is $p(f|y=+1)$, while the dot line is $p(f|y=-1)$. What we observed from the plot is that this a linearly non-separable data set.

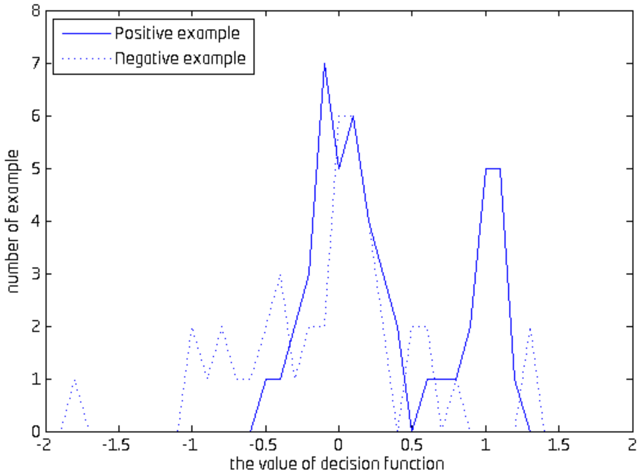

Fig. 1. The histograms for $p(f|y=\pm 1)$ for a linear SVM trained on the Salmonella Data Set.

The second data set is Wisconsin Diagnostic Breast Cancer data. There are ten real-valued features computed for each cell nucleus, resulting in 30 features in the data set. These features are from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. And the diagnosis includes two predicting fields, label +1 corresponding to Benign and label -1 corresponding to Malignant. Data set is linearly separable using all 30 input features. Table 2 shows some quantitive properties of the data set.

Table 2. Wisconsin Diagnostic Breast Cancer Data Set Features

Number of Instances	Number of Attributes	Number of Positive Examples	Number of Negative Examples
569	30	357	212

3 Empirical Result

There are two experiments in this paper to compare the performance of Venn Machine with the SVM+sigmoid combination in Platt's Method.

3.1 Taxonomy Design

The taxonomy used in both experiments is newly designed and it is based on the decision function the same as Platt's Method.

Let the number of categories $K_T = |T|$ and the taxonomy is further referred to as K_T-SVM. Then we train an SVM for the whole data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and calculate the decision values for all examples. We put the examples into the same category if the decision values of them are in the same interval which is generated depending on K_T.

For an instance, if $K_T = 8$, the intervals can be $(-\infty, -1.5]$, $(-1.5, -1.0]$, $(-1.0, -0.5]$, $(-0.5, 0]$, $(0, 0.5]$, $(0.5, 1.0]$, $(1.0, 1.5]$, $(1.5, \infty)$.

3.2 Experiments

The first experiment dealing with Salmonella data set is using a radial basis function (i.e. RBF) kernel in SVM and a Venn Machine with 8-SVM taxonomy since Salmonella data is a linearly non-separable data set . And the second experiment dealing with WDBC data set is using a linear kernel in SVM (i.e. Standard SVM) and a Venn Machine with 6-SVM taxonomy. The Venn Machine can be compared to Platt's Method and the raw SVM for accuracies and estimated probabilities. Assuming equal loss for Type I and Type II errors, the optimal threshold for the Platt's Method is $P(y = 1|f) = 0.5$. And all of the results in this paper are presented using Leave-One-Out Cross-Validation (LOOCV).

Table 3 shows the parameters setting for experiments. The C value is the *cost* for the SVM. And the Underlying Algorithm is the algorithm used in the taxonomy for Venn Machine and the kernel used in SVM. The Kernal Parameter is σ, the parameter of RBF.

Table 4 is the results of experiments. The table lists the accuracies and the probabilistic outputs for raw SVM, Platt's Method, and Venn Machine using both data sets. For Platt's Method, the probabilistic output is the average estimated probability that the prediction is correct. And for Venn Machine, the probabilistic output is the average estimated interval of probability that the prediction is correct.

Table 3. Experimental Parameters

Data Set	Task	C	Underlying Algorithm	Kernal Parameter
Salmonella	SVM	1	RBF	0.05
	Platt's Method	1	RBF	0.05
	Venn Machine	1	8-SVM RBF	0.05
WDBC	SVM	1	Linear	
	Platt's Method	1	Linear	
	Venn Machine	1	6-SVM Linear	

Table 4. Experimental Results

Data Set	Task	Accuracy	Probabilistic Outputs
Salmonella	SVM	81.72%	
	Platt's Method	82.80%	84.77%
	Venn Machine	90.32%	[83.49%, 91.03%]
WDBC	SVM	97.72%	
	Platt's Method	98.07%	96.20%
	Venn Machine	98.24%	[97.22%, 98.27%]

3.3 Results

Table 5 lists some comparisons between two methods. As shown in the table, Venn Machine got better results in both two data sets. For Salmonella data set, Venn Machine got a significant improvement (7.52%) comparing with Platt's Method in accuracy when it used a 8-SVM RBF taxonomy. In the aspect of probabilistic outputs, Venn Machine output an interval of probability with the accuracy included while the probabilistic output of Platt's Method is 1.93% higher than the accuracy. For WDBC data set, Venn Machine increased by 0.52% in accuracy while Platt's Method got 0.35%. In the aspect of probabilistic outputs, Venn Machine output an interval of probability with the accuracy included while the probabilistic output of Platt's Method is 1.87% lower than the accuracy.

Sensitivity and specificity are also calculated and shown in Table 5. For Salmonella Data Set, Venn Machine got a outstanding result in sensitivity, 16.00% better than Platt's Method. It is obvious that Venn Machine got a better ability of identity salmonella vaccine. And for WDBC Data Set, they got approximate results in both sensitivity and specificity. It is hard to tell which method is better, but we can still find Venn Machine has made a slight improvement in both aspects.

Another interest thing we observed is that Platt's Method performs better on linearly separable data set (that is WDBC in this paper) than linearly non-separable data set (that is Salmonella data set), while Venn Machine can achieve good results on both data sets. But it needs conducting experiments on more data sets to prove this.

Table 5. Comparisons Between Two Methods

Data Set	Task	Accuracy	Probabilistic Outputs	Sensitivity	Specificity
Salmonella	Platt's Method	82.80%	84.77%	76.00%	67.44%
	Venn Machine	90.32%	[83.49%, 91.03%]	92.00%	67.44%
WDBC	Platt's Method	98.07%	96.20%	97.52%	99.02%
	Venn Machine	98.24%	[97.22%, 98.27%]	97.53%	99.50%

Table 6 shows several examples in Salmonella data set predicted by Venn Machine and Platt's Method. For each example, the table contains the true label, prediction of Venn Machine and intervals of probability that the prediction is correct, the prediction of Platt's Method and the probabilistic outputs. The table indicates that both methods can be proper or erroneous. For instance, wild type strain 2, 4, 5 and vaccine strain 44 are both wrong for the two methods.

Table 6. Prediction for Individual Examples in Salmonella Data Set

No.	True Label	Prediction of VM	Probabilistic Outputs of VM	Prediction of PM	Probabilistic Outputs of PM
1	−1	−1	[88.89%, 100.00%]	−1	95.65%
2	−1	+1	[60.00%, 63.33%]	+1	77.72%
3	−1	−1	[88.89%, 100.00%]	−1	98.49%
4	−1	+1	[60.00%, 63.33%]	+1	63.98%
5	−1	+1	[60.00%, 63.33%]	+1	56.57%
6	−1	−1	[76.19%, 80.95%]	−1	78.57%
7	−1	−1	[76.19%, 80.95%]	−1	71.69%
...	...	...	...	...	...
44	+1	−1	[80.95%, 85.71%]	−1	71.92%
45	+1	+1	[90.00%, 93.33%]	+1	96.91%
46	+1	+1	[90.00%, 93.33%]	+1	77.96%
47	+1	+1	[56.67%, 60.00%]	−1	61.81%
48	+1	+1	[56.67%, 60.00%]	−1	58.43%
49	+1	+1	[90.00%, 93.33%]	+1	96.15%
50	+1	+1	[90.00%, 93.33%]	+1	94.12%
...	...	...	...	...	...

4 Conclusion

From our experience on these data sets we see the following. The Platt's estimation for the accuracy of prediction can be too optimistic or too pessimistic, while Venn's bounds estimate it more correctly: two-sided estimation is safer than single one. As for the accuracy itself, we see that if Platt's and Venn Machines are based on the same kind of SVM, accuracy of Venn Machine is also a bit better. This may be because Venn Machine do not rely on a fixed transformation of the SVM output, but makes its own transformation for each taxonomy, based on the actual data set.

We applied different probabilistic approaches to the dataset of Salmonella strains. As it can be seen from Figure 1 and Table 6, this data set is hard to separate: there are few errors in the class $+1$, but large part of examples from the class -1 seems to be hardly distinguishable from the class $+1$. This is why in this case we need to have individual assessment of prediction quality: being unable to make a confident prediction on any example, we still can select some of them where our prediction has higher chance to be correct.

The results have been observed on two particular data sets. We plan to conduct experiments on bigger data sets. Another possible direction is to compare Venn Machine and Platt's Method theoretically.

5 Acknowledgements

This work was supported in part by funding from BBSRC for ERASySBio+ Programme: Salmonella Host Interactions PRoject European Consortium (SHIPREC) grant; VLA of DEFRA grant on Development and Application of Machine Learning Algorithms for the Analysis of Complex Veterinary Data Sets; EU FP7 grant O-PTM-Biomarkers (2008–2011); and by grant PLHRO/0506/22 (Development of New Conformal Prediction Methods with Applications in Medical Diagnosis) from the Cyprus Research Promotion Foundation.

References

1. John, C., Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in Large Margin Classifiers, pp.61-74. MIT Press, 1999.
2. A. Gammerman, V. Vovk. Hedging Predictions in Machine Learning (with discussion). The Computer Journal, Vol.50, No.2, pp.151-163, March 2007.
3. V. Vovk, A. Gammerman, G. Shafer. "Algorithmic Learning in a Random World", Springer, 2005.
4. Vladimir Vapnik. "The Nature of Statistical Learning Theory", Springer-Verlag, 1995.