Calibrated Probabilistic Predictions for Biomedical Applications

Antonis Lambrou[1,2], Harris Papadopoulos[1,3], and Alexander Gammerman[2]

[1]Frederick Research Center, Nicosia, Cyprus

[2]Computer Learning Research Centre
Royal Holloway University of London

[3]Frederick University
Computer and Engineering Department Nicosia Cyprus

Outline

1. Problem definition.
2. Venn Prediction with SMO.
3. Experiments and results.
4. Conclusion and future work.
5. Questions.

Problem definition

- In biomedical applications, Machine Learning classification methods are used widely. However most of these methods are not probabilistic.
- We propose the use of Venn Prediction for producing well-calibrated probabilistic predictions for biomedical problems.
- Venn Predictors (VPs) can provide reliable probability estimates based on the only assumption that the data are independently and identically distributed (i.i.d.).
- Existing probabilistic methods do not guarantee validity. VPs are valid [5] probabilistic predictors.

Venn Prediction

- Examples are divided into categories based on a taxonomy:
  - New example is assigned a possible label and placed in the training set.
    - We train an underlying algorithm and categorize all examples.
    - We calculate the distribution of labels P_j in the category of the new example.
    - We repeat this process for every possible label $j \in \{1, ..., c\}$ for c number of classes.

- In the end, we have a set of label distributions $\{P_1, ..., P_c\}$.
- We calculate the quality amongst all distributions as the average of each label and we pick the label $\hat{y}_n = Y_{k_{best}}$ with the highest average.
- The probability of this prediction to be correct lies in the interval of the selected label's minimum $L(Y_k) = \min_{j=1}^{c} p^{Y_j}(Y_k)$ and maximum value $U(Y_k) = \max_{j=1}^{c} p^{Y_j}(Y_k)$ amongst $\{P_1, ..., P_c\}$.

SMO Venn Predictor

- In this work, we use the Sequential Minimal Optimisation (SMO) algorithm [4] to categorise examples based on a SMO taxonomy.
- SMO iteratively solves the optimisation problem which arises during the training phase of Support Vector Machines (SVMs) by breaking up the quadratic programming problem into smaller problems with 2 Lagrange multipliers each.
- Here, multi-class problems are solved using pair-wise classification (1-vs-1) [2].
- The simplest taxonomy is to find the class that gives the maximum score.

SMO Venn Predictor

- We have developed four variations of the SMO algorithm:
  - SMO classifier.
  - SMO with Logistic Regression (SMOL).
  - SMO with Feature Selection (SMO-FS). We use Correlation Based Feature Selection (CBFS).
  - SMO with Logistic Regression and FS (SMOL-FS).
- We have developed four variations of VPs:
  - VENN-SMO
  - VENN-SMOL
  - VENN-SMO-FS
  - VENN-SMOL-FS

Experiments

- Offline experiments
  - 10-fold cross validation.
  - We show the mean accuracy, and additionally for probabilistic methods the Brier Score (BS):

$$BS = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{c} (f(x_{ij}) - o_{ij})^2, \quad (1)$$

where $f(x_{ij})$ is the probabilistic output of the algorithm for example x_i and class j, and o_{ij} is set to 1 if example x_i belongs to class j, and 0 otherwise.

Experiments

- Online experiments
  - For the online experiments we have selected the SMOL-FS and VENN-SMO-FS for comparison.
  - No initial training set. A test instance x_t is predicted by the algorithm with a probability estimate. Then (x_t, y_t) is added to the training set which grows every time.

Experiments

- In the results of the SMOL-FS algorithm, we graph the Cumulative Mean Probability (CMP) and the Cumulative Mean Accuracy (CMA) curves:

$$CMP(t) = \frac{1}{t} \sum_{i=1}^{t} \max_{j=1}^{c} f(x_i), \tag{2}$$

$$CMA(t) = \frac{1}{t} \sum_{i=1}^{t} Acc_i, \tag{3}$$

where t is the number of test examples that have been added to the training set, and $Acc_i = 1$ when the prediction for example x_i is correct and 0 otherwise.

- For the VENN-SMO-FS algorithm we graph the Cumulative Mean Lower Probability (CMLP), the Cumulative Mean Upper Probability (CMUP), the Cumulative Mean Central Probability (CMCP), and the CMA curves:

$$CMLP(t) = \frac{1}{t}\sum_{i=1}^{t} L_i(Y_{k_{best}}), \tag{4}$$

$$CMUP(t) = \frac{1}{t}\sum_{i=1}^{t} U_i(Y_{k_{best}}), \tag{5}$$

$$CMCP(t) = \frac{1}{t}\sum_{i=1}^{t} \frac{U_i(Y_{k_{best}}) + L_i(Y_{k_{best}})}{2}. \tag{6}$$

Experiments

- Datasets
  - The Ecoli dataset contains 336 Ecoli proteins that are classified into 8 cellular localization sites [3].
  - The Dermatology dataset contains 366 instances of patients with erythemato-squamous diseases [1]. The instances are classified into 6 diseases.

Experiments

- Ecoli dataset offlne results

Method	RBF	Acc.	BS	Mean Prob.
SMO	9.5	78.87%	-	-
SMO-FS	9.5	78.87%	-	-
SMOL	9.5	77.38%	34.38%	72.01%
SMOL-FS	9.5	77.68%	35.04%	72.08%

Method	RBF	Acc.	BS	Prob. Interval
VENN-SMO	**9.5**	**86.90%**	**22.14%**	**81.08% – 90.59%**
VENN-SMO-FS	9.5	86.90%	22.29%	81.04% – 90.49%
VENN-SMOL	9.5	85.42%	33.88%	22.19% – 91.05%
VENN-SMOL-FS	9.5	85.12%	33.88%	22.19% – 91.05%

Experiments

- Ecoli dataset online results

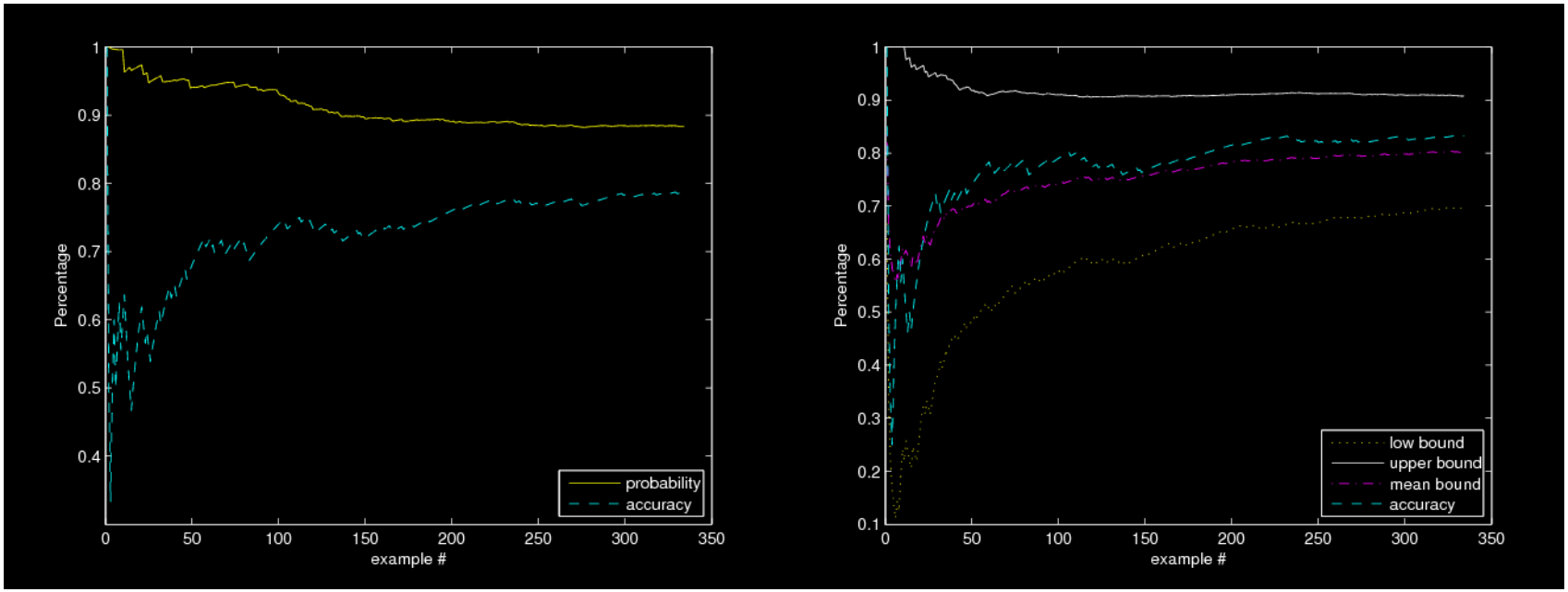

Figure: SMOL-FS (left) and VENN-SMO-FS (right) results.

Experiments

- Dermatology dataset offlne results

Method	RBF	Acc.	BS	Mean Prob.
SMO	0.02	93.30%	-	-
SMO-FS	0.02	96.93%	-	-
SMOL	0.02	90.50%	14.78%	89.26%
SMOL-FS	0.02	96.65%	5.75%	95.15%

Method	RBF	Acc.	BS	Prob. Interval
VENN-SMO	0.02	93.30%	11.39%	78.75% – 97.82%
VENN-SMO-FS	**0.02**	**96.93%**	**5.23%**	**93.35% – 98.37%**
VENN-SMOL	0.02	90.78%	18.33%	50.44% – 96.71%
VENN-SMOL-FS	0.02	96.09%	7.98%	79.46% – 98.37%

Experiments

- Dermatology dataset online results

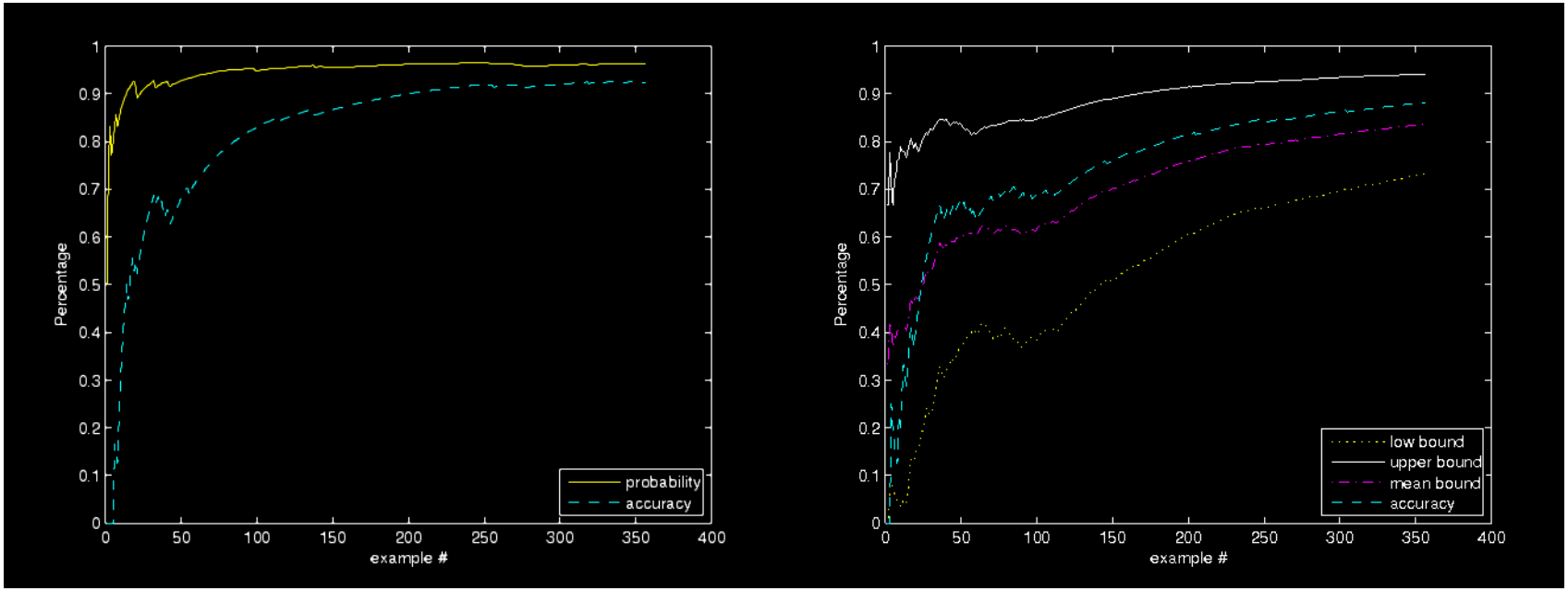

Figure: SMOL-FS (left) and VENN-SMO-FS (right) results.

Conclusion and future work

- Existing methods do not guarantee validity of the probability estimates.
- Venn predictors guarantee validity under the i.i.d. assumption and well-calibrated results are provided.
- More taxonomies will be tested to improve accuracy.
- More biomedical applications will be applied.

Acknowledgement

- This work was co-funded by the European Regional Development Fund and the Cyprus Government through the Cyprus Research Promotion Foundation "DESMI 2009-2010", research contract TPE/ORIZO/0609(BIE)/24 ("Development of New Venn Prediction Methods for Osteoporosis Risk Assessment").

Bibliography I


H. Altay Guvenir, Gulsen Demiroz, and Nilsel Ilter.
Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals.
Artificial Intelligence in Medicine, 13:147–165, 1998.


Trevor Hastie and Robert Tibshirani.
Classification by pairwise coupling.
In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1998.


Paul Horton and Kenta Nakai.
A probabilistic classification system for predicting the cellular localization sites of proteins.
In *In Proceeding of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 109–115, 1996.

Bibliography II

J. Platt.
Fast training of support vector machines using sequential minimal optimization.
In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

Volodya Vovk, Alexander Gammerman, and G. Shafer.
Algorithmic Learning in a Random World.
New York, Springer, 2005.

Questions

Thank you for your attention.

email: A.Lambrou@cs.rhul.ac.uk