Reliable Probability Estimates based on Support Vector Machines for Large Multiclass Datasets

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. Motivation and Problem definition.
2. Existing probabilistic methods for SVMs.
3. Extended Multiclass SVM.
4. SVM Venn Predictor.
5. Inductive framework.
6. Experiments and results.
7. Conclusion and future work.
8. Questions.

Problem definition

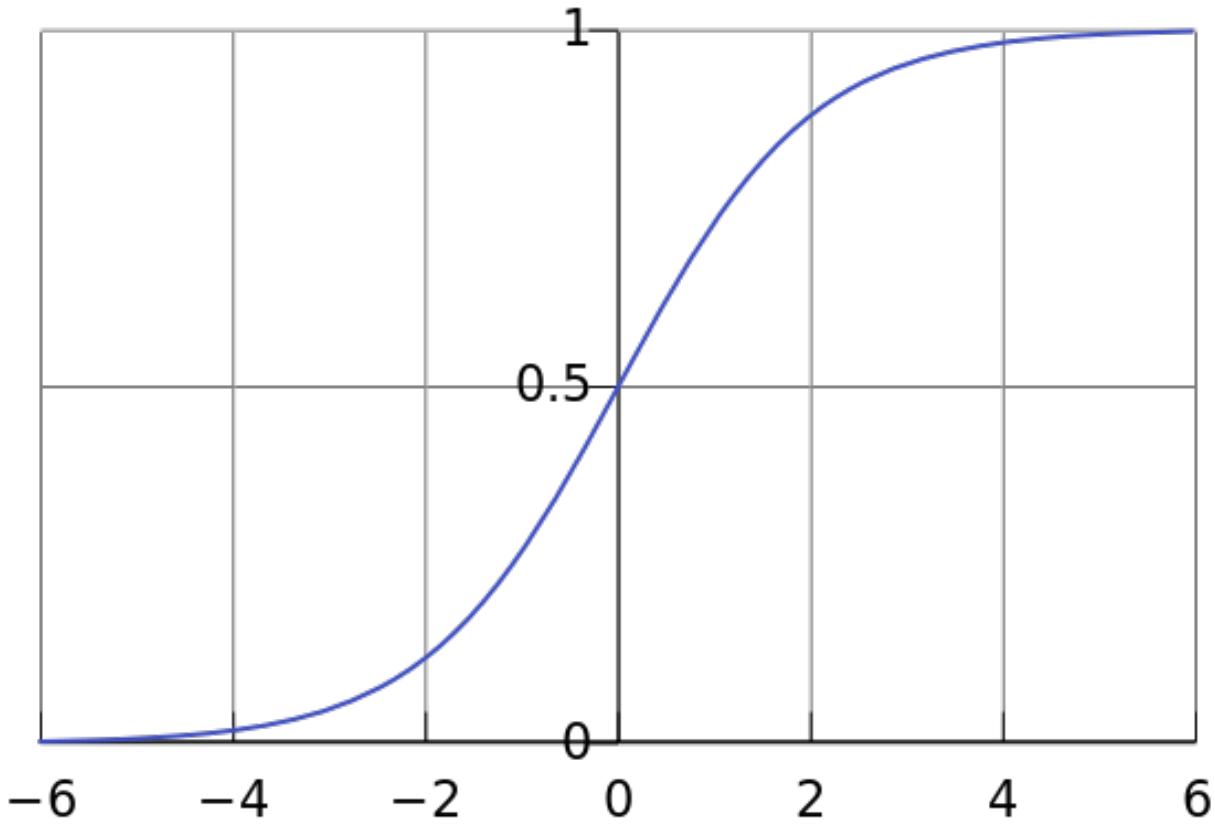
- Our goal is to convert the unthresholded SVM output scores into well-calibrated probability estimates.
- Existing methods that convert SVM scores into probability estimates do not guarantee validity.
- We use Venn Prediction [8] provide validity. Venn predictors output lower and upper bounds of probability estimates that guarantee validity under the i.i.d. assumption.
- Applications that can benefit from this kind of reliability are critical applications such as medical diagnostic systems.

Existing Probabilistic methods for SVMs

- SVM Binning [3]
  - Sort training examples according to their unthresholded SVM output $f(x_i)$ (the distance from the separating hyperplane).
    - Here $y = [-1, 1]$, so negative examples have negative score.
  - Categorize examples according to their SVM score $f(x_i)$ based on equally sized intervals (or bins).
  - Place new example x_i into corresponding bin according to $f(x_i)$.
  - Calculate fraction of positive labels in the bin.

Existing Probabilistic methods for SVMs

- SVM with Isotonic Regression [9]
  - If the scores of the SVM are ranked correctly, we can assume that the probability $P(y = 1|x_i)$ will be strictly increasing as the SVM score increases (isotonic). We can map SVM scores into probability estimates using isotonic regression.



Pair-Adjacent Violators (PAV) algorithm

- The Pair Adjacent Violators (PAV) algorithm learns to map SVM scores into probability estimates:
  1. First, we set $p(x_i) = 1$ if x_i is a positive example, and $p(x_i) = 0$ otherwise.
  2. We rank x_i according to $f(x_i)$. If p is isotonic the function has been learned.
  3. Otherwise, there must be an instance where $p(x_{i-1}) > p(x_i)$. The two examples are called pair-adjacent violators.
  4. The values $p(x_{i-1})$ and $p(x_i)$ are replaced by $\frac{p(x_{i-1})+p(x_i)}{2}$.
  5. Steps 3 and 4 are repeated until all values conform to the isotonic assumption.

Existing Probabilistic methods for SVMs

- Platt's method [7]
  - Trains SVM on the training set and then estimates probabilities based on the decision function f by fitting a sigmoid:

$$P(y = 1|f(x_i)) = \frac{1}{1 + exp(Af(x_i) + B)}. \quad (1)$$

where A and B parameters are determined by using maximum likelihood estimate from the training set.

Multiclass SVM

- The original methods work only for binary classification tasks. Here we apply the one-against-all procedure for multiclass problems.
- We train a binary SVM for each class, and we find probability estimates for each class.
- We merge the probability estimates for each class by normalizing the probabilities to 1. The largest probability is then used to classify the example.

Venn Predictor

- How it works:
  - Examples are divided into categories based on a taxonomy:
    - New example is assigned a possible label and placed in the training set.
    - We train the SVM 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 with the highest quality.
- The probability of this prediction to be correct lies in the interval of the selected label's minimum and maximum value amongst $\{P_1, ..., P_c\}$.

SVM Taxonomy

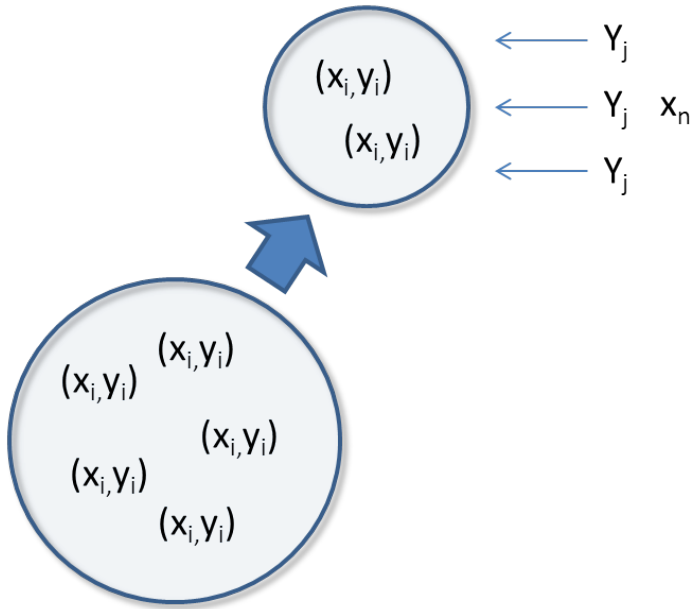
- The simplest taxonomy is to find the largest SVM score of all the binary SVMs:

$$\tau_i = \arg\max_{j=1}^{c} f_j(x_i),$$

where $f_j(x_i)$ is the SVM score of example x_i for class j.

Inductive Framework

- The inductive framework [6, 5] improves time efficiency of transductive methods.
  - In the transductive framework, the test example is being tested for every hypothesis on the training set.
  - The inductive framework, utilises a calibration-set in order to avoid the need to train several times for each test example.



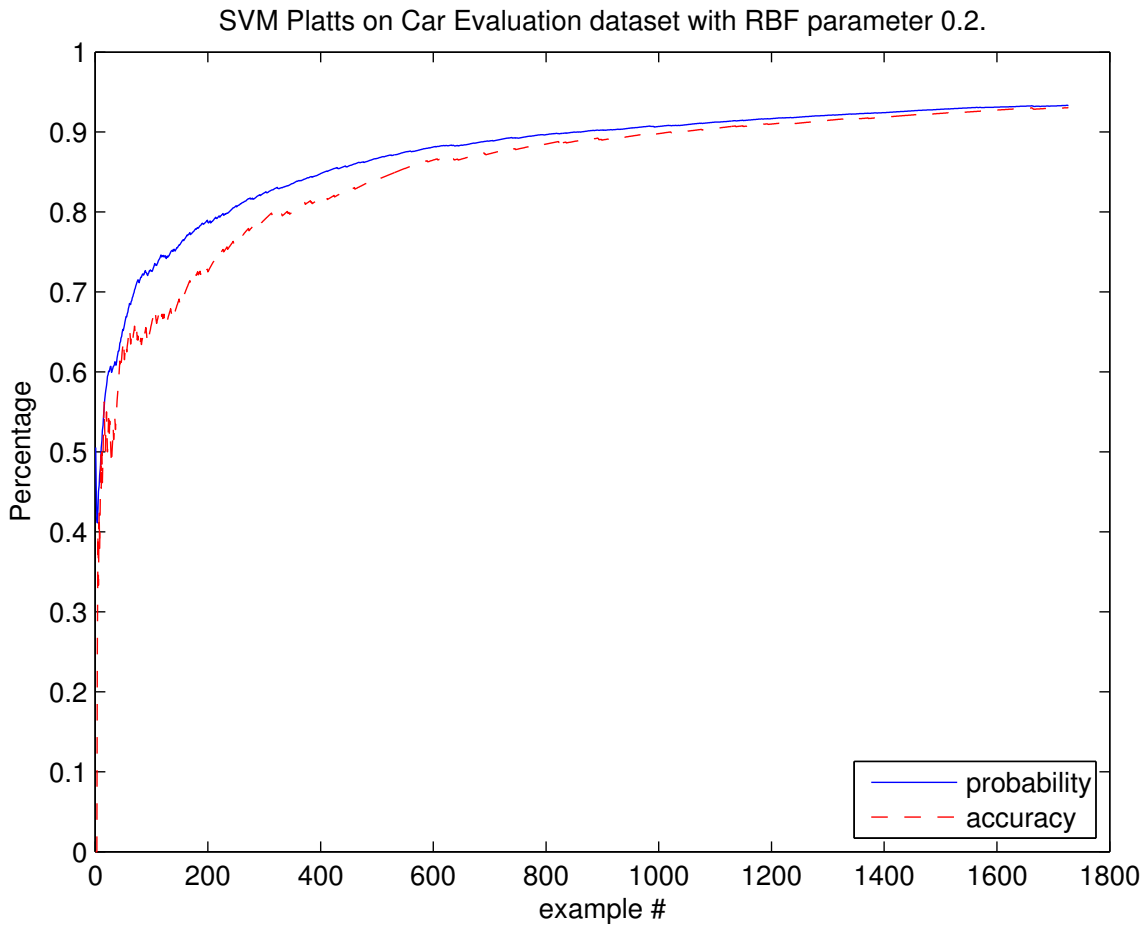
Experiments

- We conduct online experiments:
  - In the online mode, there is no initial training set. Examples are being added to the training set after predictions.
  - We calculate the cumulative average probability and cumulative average accuracy.
  - We have tested two SVM RBF parameters for each method. An optimal value derived from offline tests, and a secondary RBF value which is the optimal value divided by 10.
  - The number of bins for the SVM Binning method is set to 10.
  - For the Inductive Venn Predictor, we were randomly removing 30% of the training set and adding it to the calibration set.

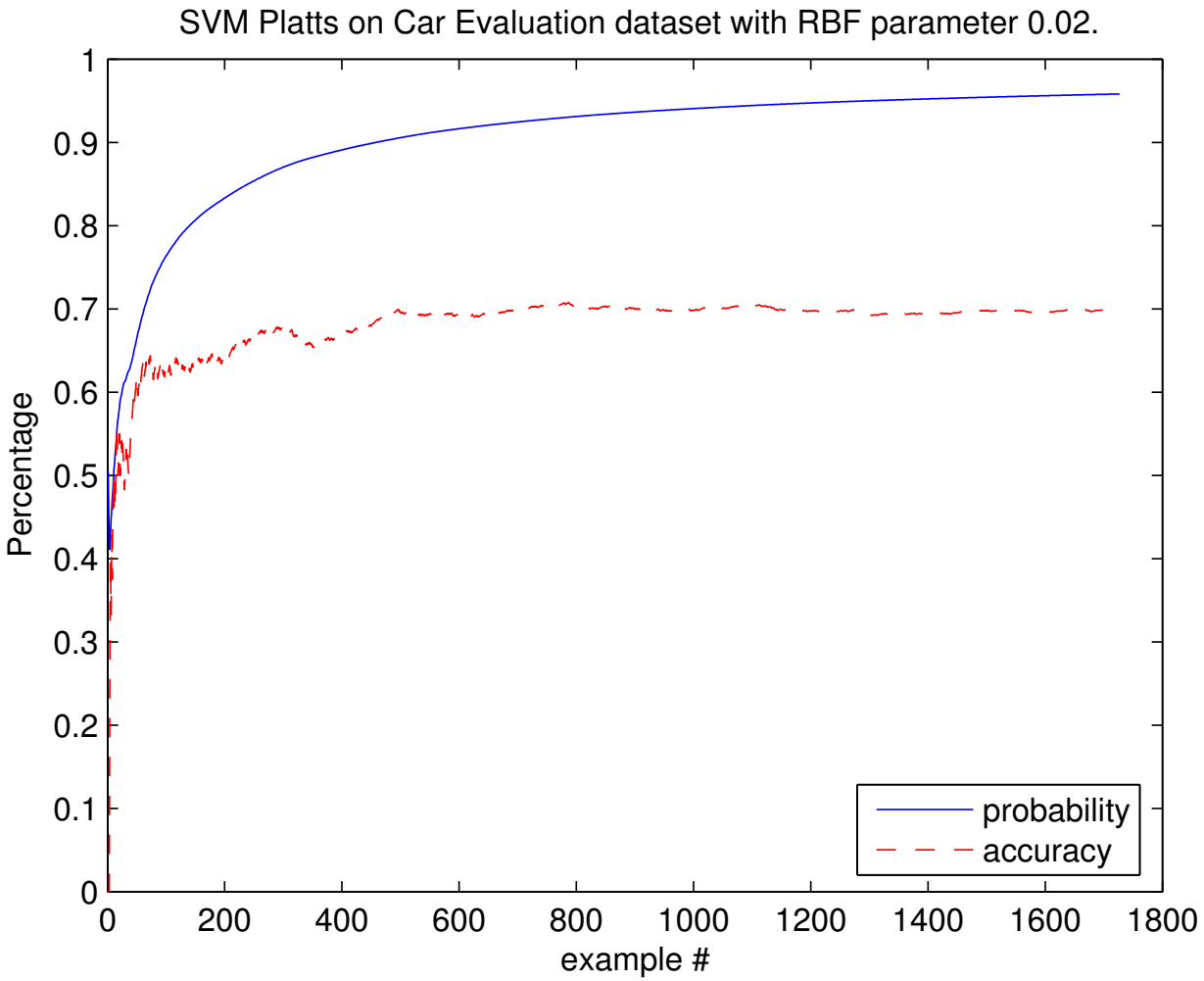
Experiments

- Datasets
  - The Car Evaluation [1] dataset available at UCI ML repository [4]:
    - Contains 1728 instances with 6 features for each instance.
    - The features describe price, technology, and comfort of car.
    - There are 4 classes which describe a car's acceptability.
  - The Red Wine quality [2] dataset available at UCI ML repository [4]:
    - Contains 1599 instances of 11 physiochemical features of re variants of the Portuguese "Vinho Verde".
    - Each instance has a quality score from 1 to 10 (10 classes).

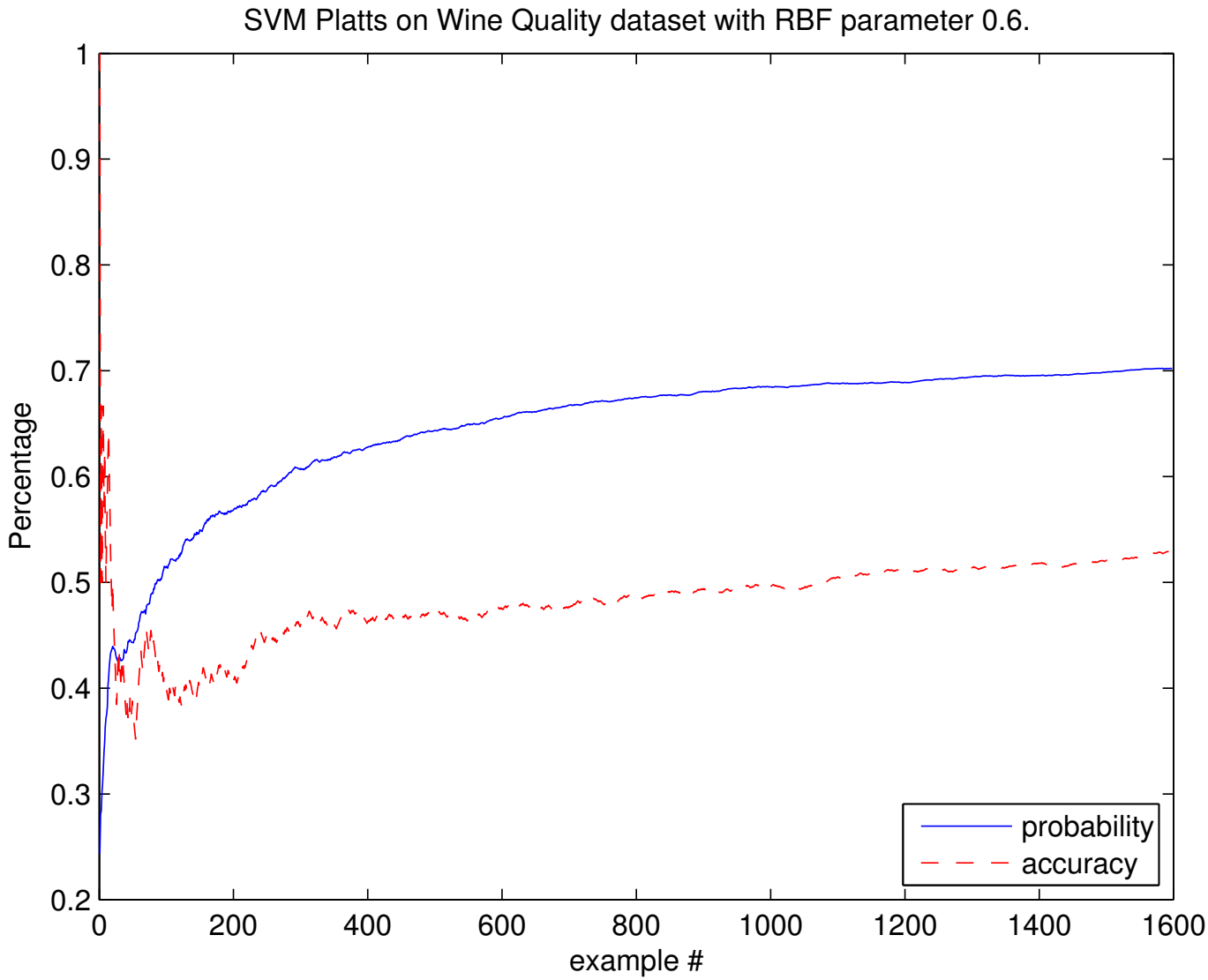
SVM Platt's Results (car evaluation data)



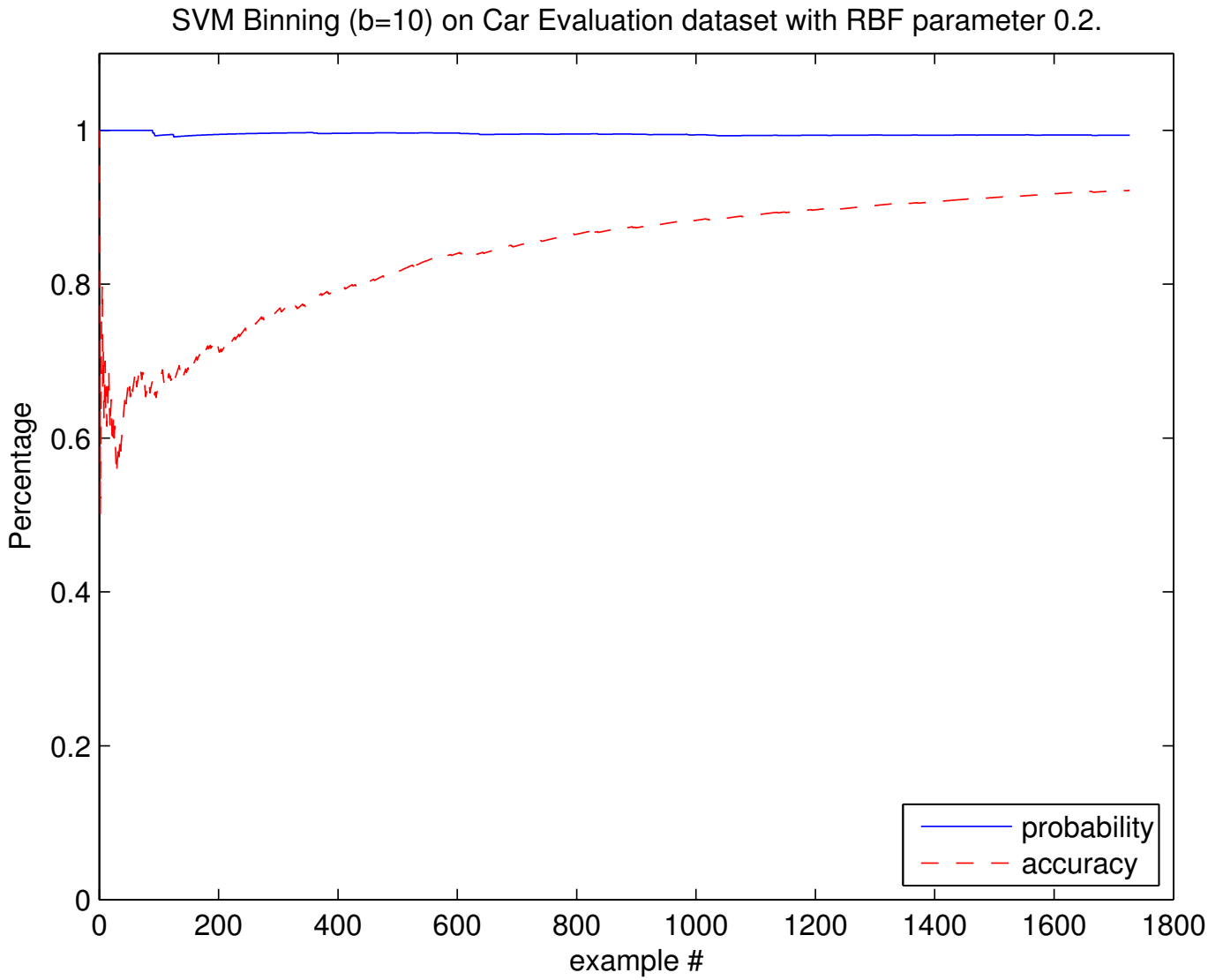
SVM Platt's Results (car evaluation data)



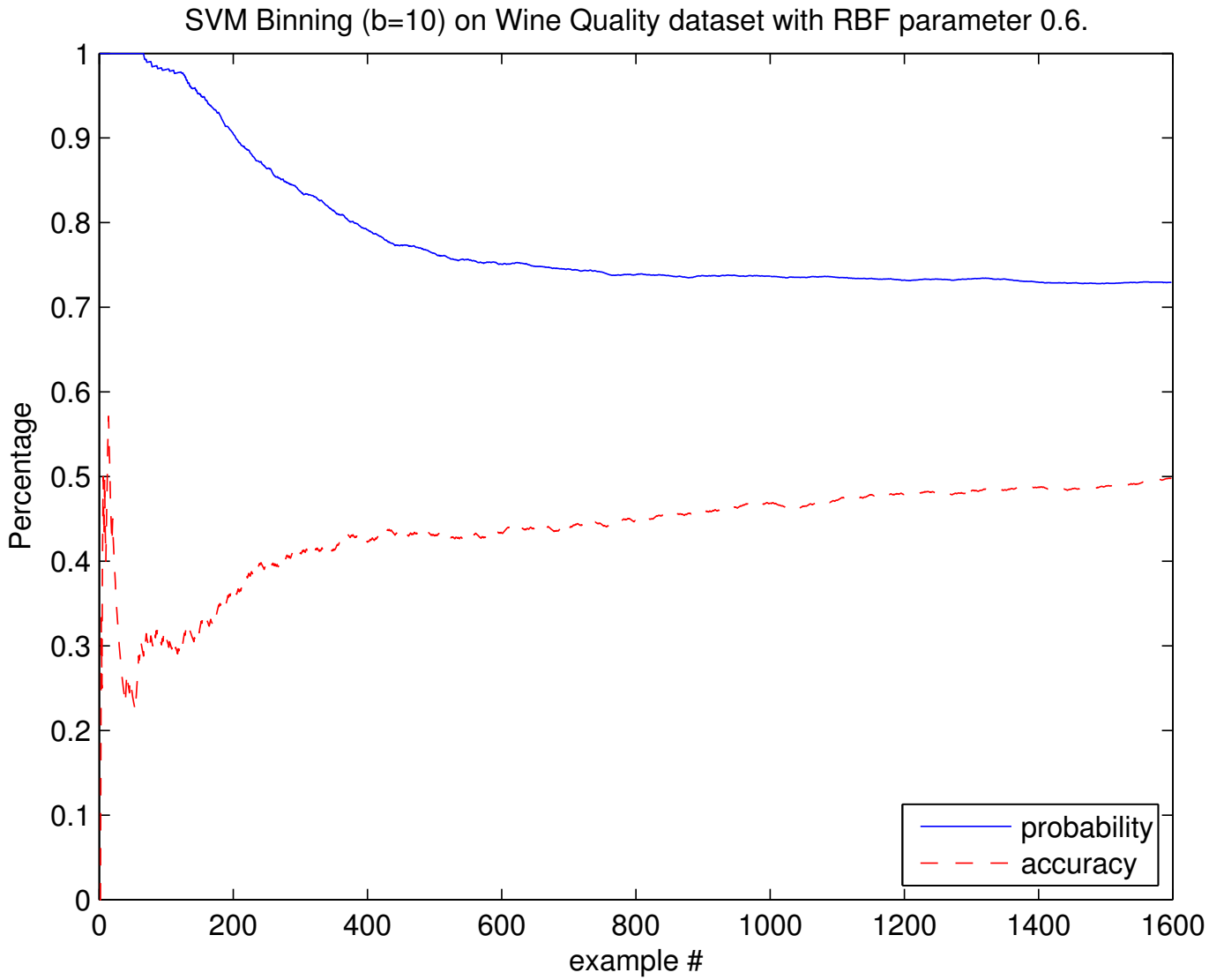
SVM Platt's Results (wine quality data)



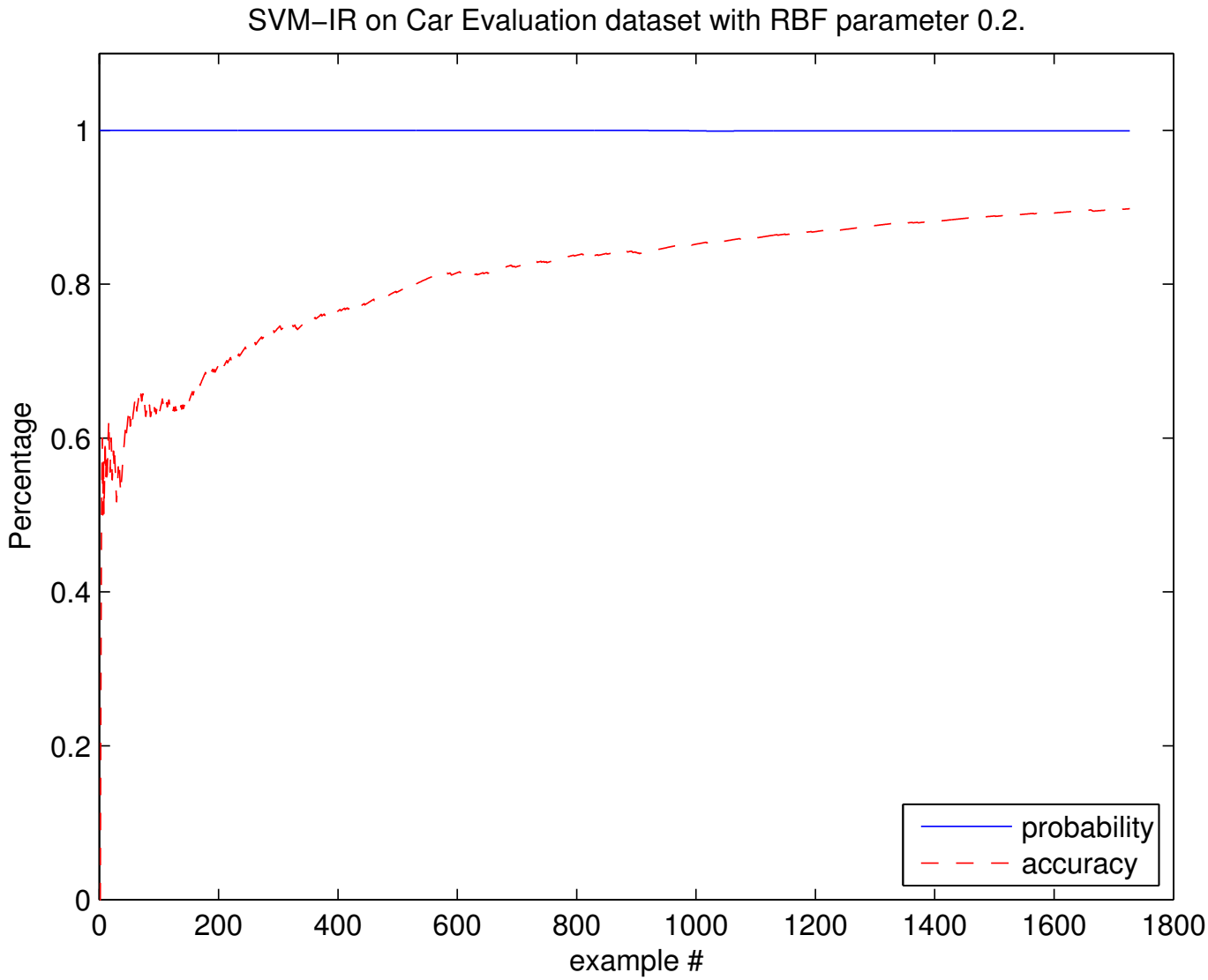
SVM Binning Results (car evaluation data)



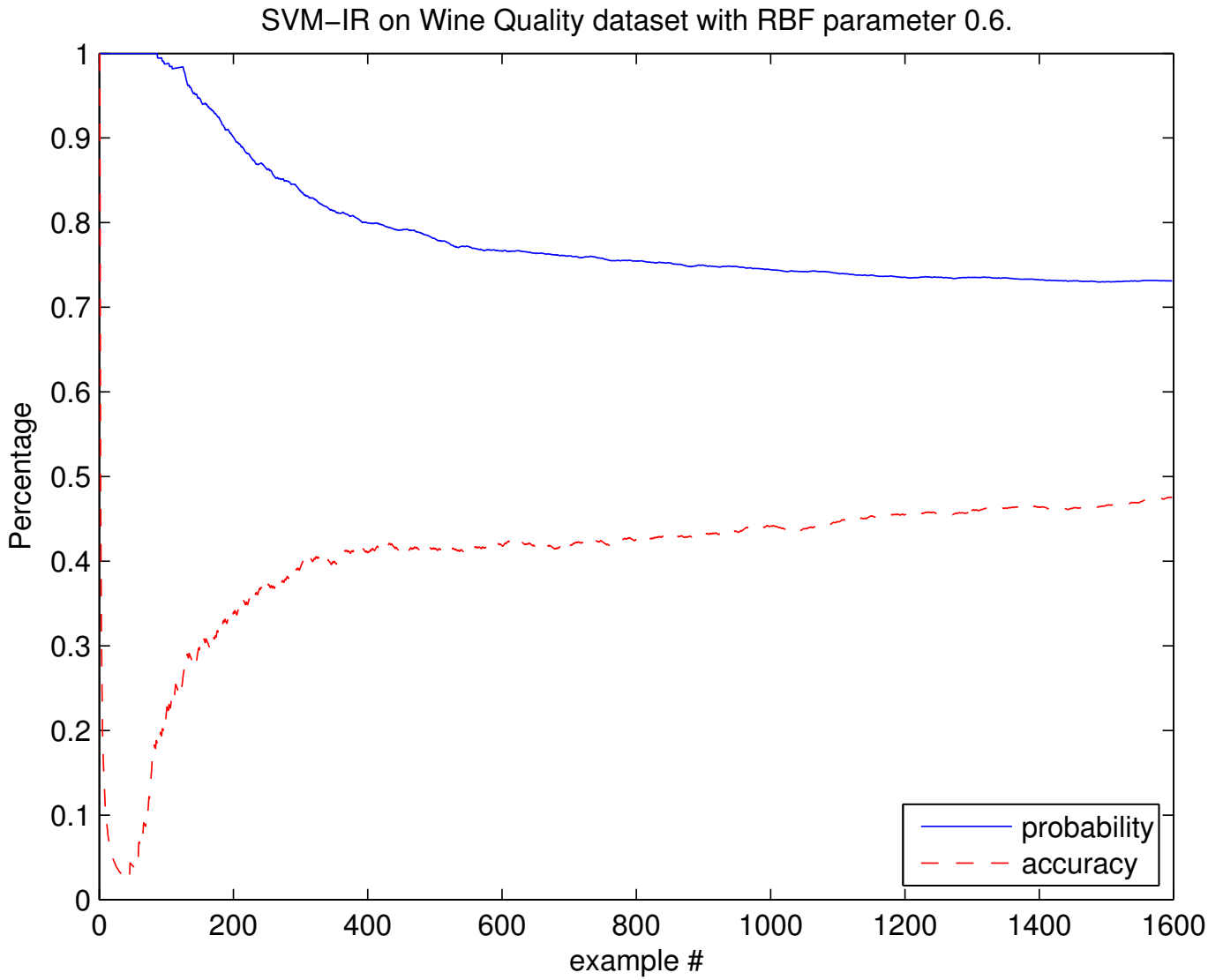
SVM Binning Results (wine quality data)



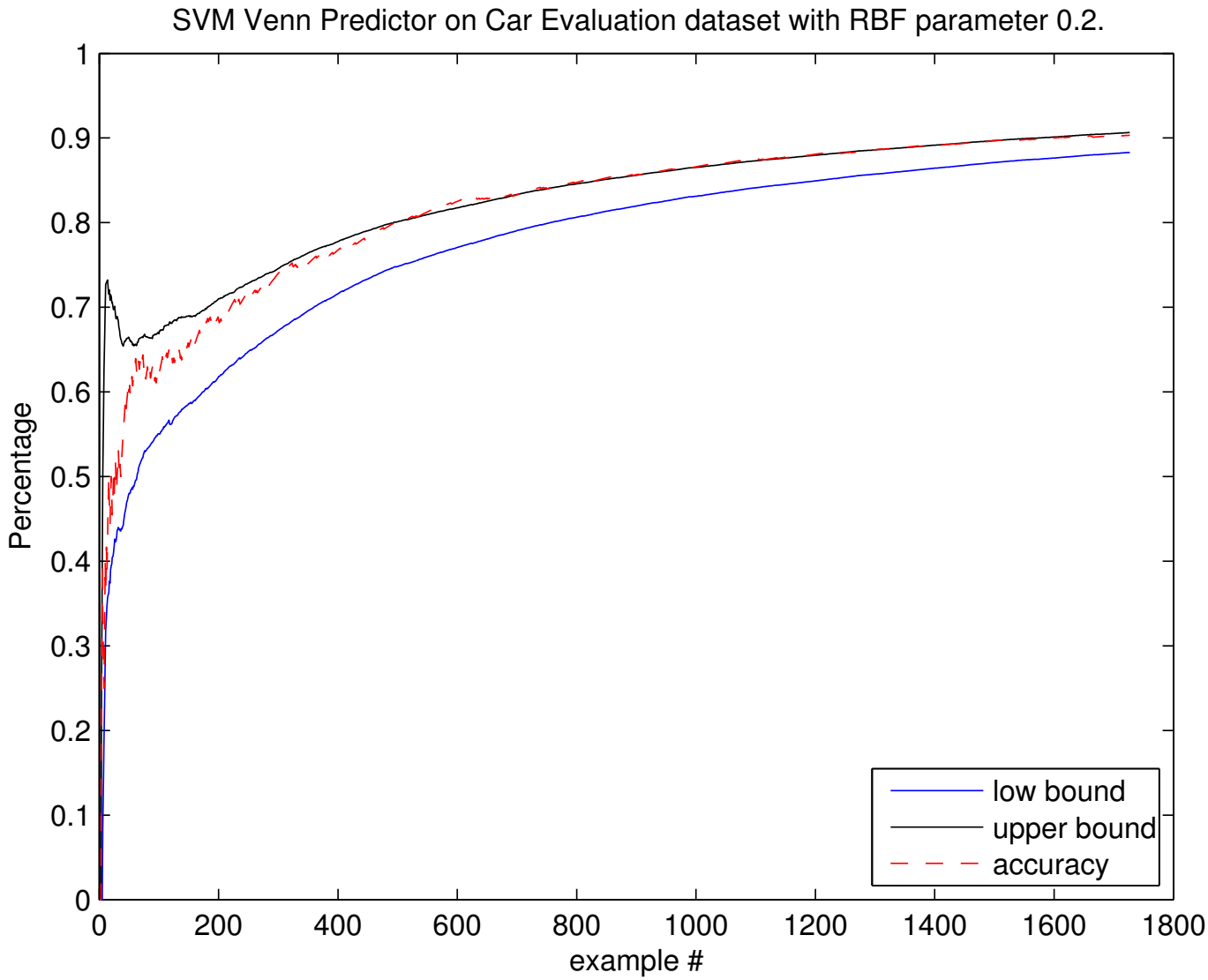
SVM Isotonic Regression Results (car evaluation data)



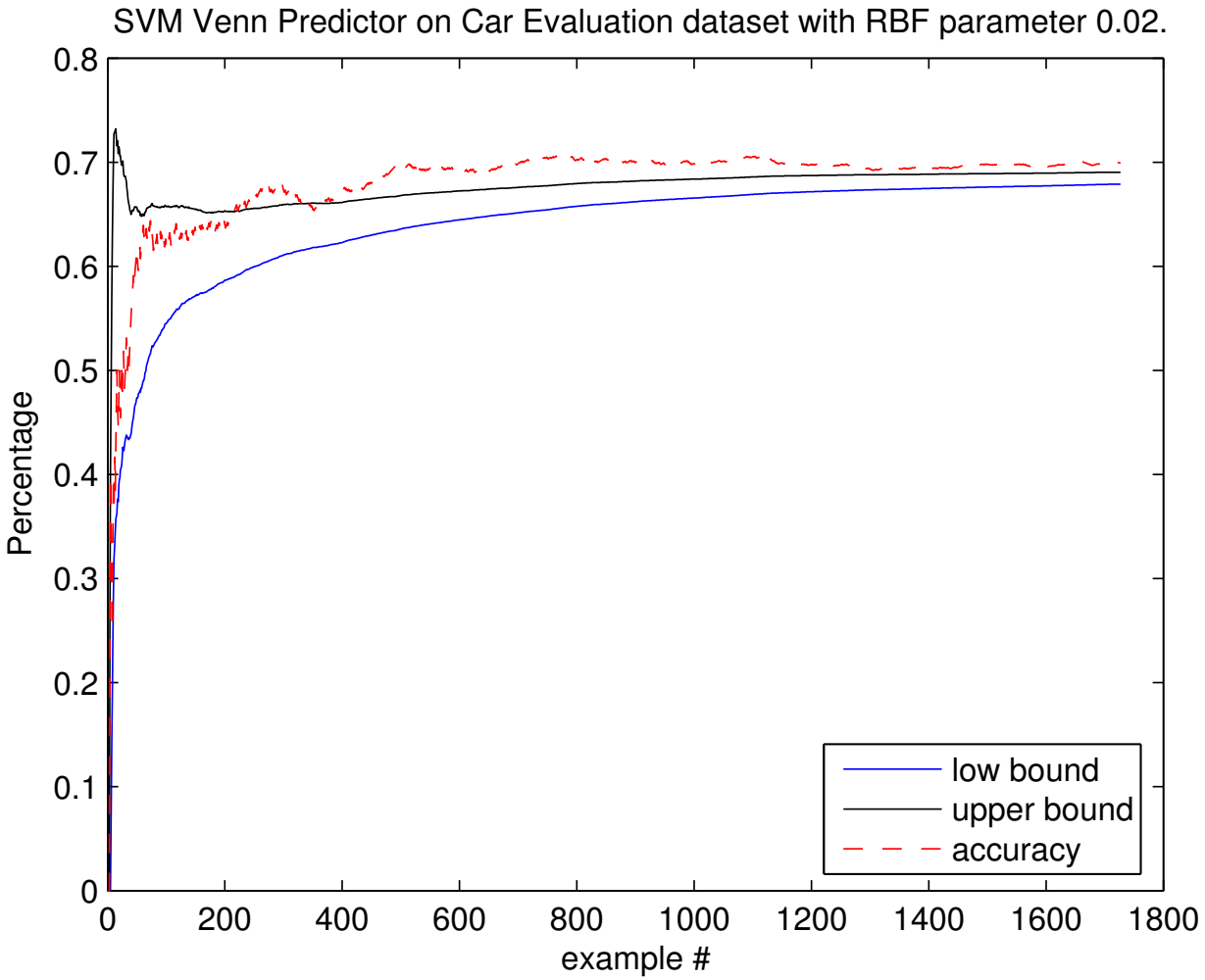
SVM Isotonic Regression Results (wine quality data)



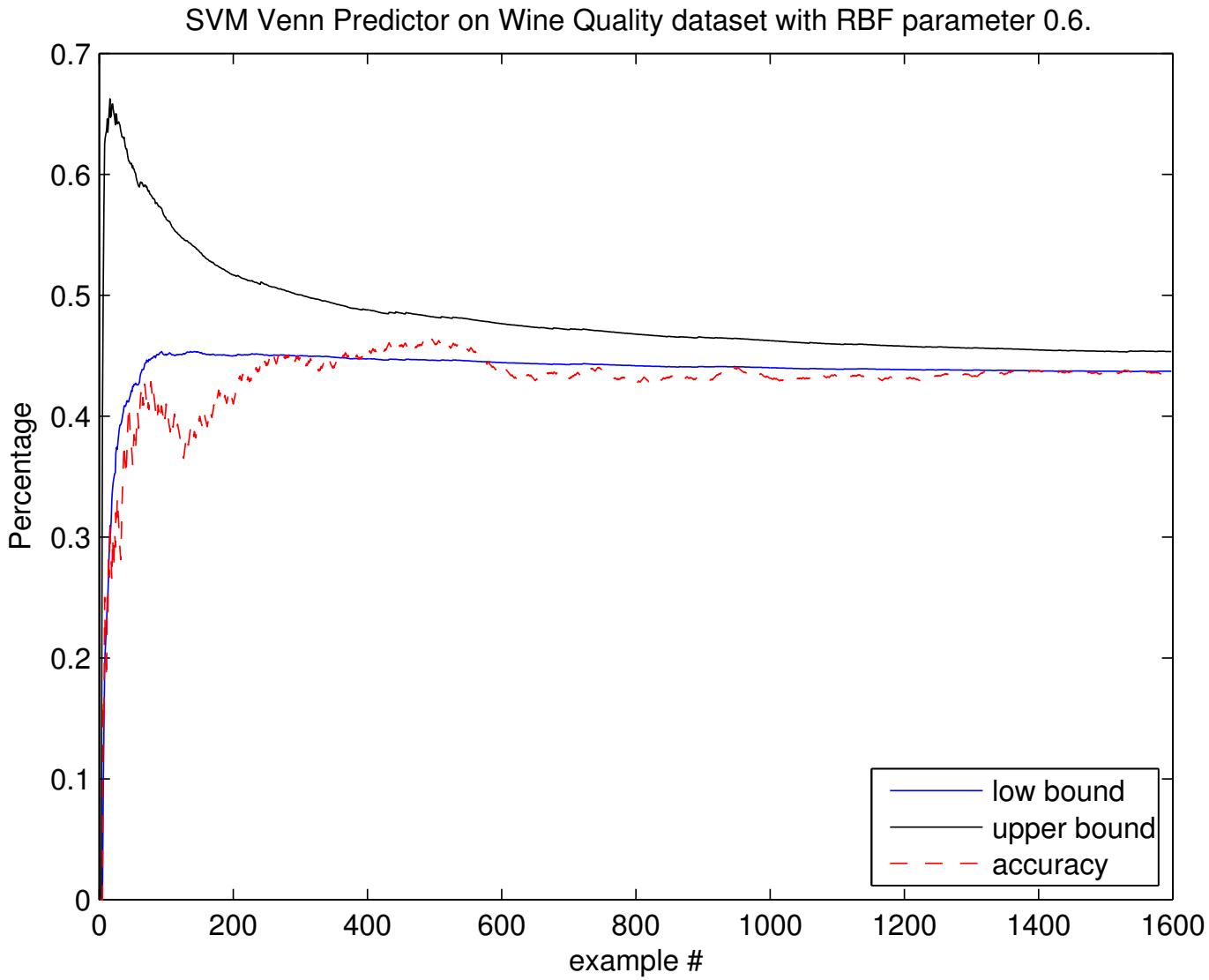
SVM Venn Predictor Results (car evaluation data)



SVM Venn Predictor Results (car evaluation data)



SVM Venn Predictor Results (wine quality data)



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.
- We may need to invastigate the validity of the Inductive Venn Predictor in our online setting.
- More taxonomies will be tested to improve accuracy.

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


Marko Bohanec and Vladislav Rajkovic.
V.: Knowledge acquisition and explanation for multi-attribute decision making.
In *8th International Workshop "Expert Systems and Their Applications"*, 1988.


Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis.
Modeling wine preferences by data mining from physicochemical properties.
Decision Support Systems, 47(4):547–553, November 2009.


Joseph Drish.
Obtaining calibrated probability estimates from support vector machines, 1998.

Bibliography II


A. Frank and A. Asuncion.
UCI machine learning repository, 2010.


Harris Papadopoulos.
Inductive conformal prediction: Theory and application to neural networks.
In Paula Fritzsche, editor, *Tools in Artificial Intelligence*, chapter 18, pages 315–330. I-Tech, Vienna, Austria, 2008.


Harris Papadopoulos, Volodya Vovk, and Alex Gammerman.
Qualified predictions for large data sets in the case of pattern recognition.
In *Proceedings of the 2002 International Conference on Machine Learning and Applications (ICMLA'02)*, pages 159–163. CSREA Press, 2002.

Bibliography III


John C. Platt.
Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.
In *Advances in large margin classifiers*, pages 61–74. MIT Press, 1999.


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


Bianca Zadrozny and Charles Elkan.
Transforming classifier scores into accurate multiclass probability estimates.
In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, pages 694–699. ACM Press, 2002.

Questions

Thank you for your attention.

email: A.Lambrou@frederick.ac.cy