Reliable probability estimates based on Support Vector Machines for large multiclass datasets

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

[1]Frederick Research Center, Nicosia, Cyprus.
[2]Computer Learning Reasearch Centre, Computer Science Department, Royal Holloway, University of London, England.
{A.Lambrou, I.Nouretdinov, A.Gammerman}@cs.rhu.ac.uk
[3]Computer Science and Engineering Department, Frederick University, Cyprus.
H.Papadopoulos@frederick.ac.cy

Abstract. Venn Predictors (VPs) are machine learning algorithms that can provide well calibrated multiprobability outputs for their predictions. The only drawback of Venn Predictors is their computational inefficiency, especially in the case of large datasets. In this work, we propose an Inductive Venn Predictor (IVP) which overcomes the computational inefficiency problem of the original Venn Prediction framework. Each VP is defined by a taxonomy which separates the data into categories. We develop an IVP with a taxonomy derived from a multiclass Support Vector Machine (SVM), and we compare our method with other probabilistic methods for SVMs, namely Platt's method, SVM Binning, and SVM with Isotonic Regression. We show that these methods do not always provide well calibrated outputs, while our IVP will always guarantee this property under the i.i.d. assumption.

Keywords: Support Vector Machine, well calibrated probabilities, multiclass, Inductive Venn Predictor, Machine Learning.

1 Introduction

Support Vector Machines (SVMs) [13] are widely used in the field of Machine Learning for classification or regression analysis. To date, several efforts have been made in order to map the unthresholded SVM outputs into probability estimates. Some of these methods are Platt's method [12], SVM binning [5], and SVM with Isotonic Regression [16]. Nevertheless, there is no guarantee provided that the probability estimates produced by these methods will always be well calibrated. In fact as our experiments show, they can become quite misleading.

In this work, we develop a Venn Predictor (VP) based on the SVM classifier in order to produce probability estimates that are guaranteed to be well calibrated. Venn Prediction is a novel machine learning framework that can be combined with conventional classifiers for producing well calibrated multiprobability predictions under the i.i.d. assumption. In [15], the Venn Prediction framework is described thoroughly and a proof of the validity of its probabilities is given.

In order to overcome the computational inefficiency problem of the original Venn Prediction approach, which renders it not suitable for application to large datasets, we propose an Inductive Venn Predictor (IVP) based on the idea of Inductive Conformal Prediction. As it was shown in many studies, see e.g. [8, 10, 11], Inductive Conformal Predictors are as computationally efficient as the conventional algorithms they are based on. The same is true for the proposed IVP, which is based on SVMs for multiclass tasks. We experiment on two multi-classification datasets, the Car Evaluation [1] and the Wine Quality [2] datasets, which are freely available at the University of California, Irvine (UCI) machine learning repository [6]. We compare our method with Platt's method, SVM Binning, and SVM with Isotonic Regression. We demonstrate that these methods do not always provide well calibrated results, while our method can always guarantee this property under the i.i.d. assumption.

The rest of the paper is structured as follows. In section 2, we outline related work that has been conducted for estimating probabilities. In section 3, we describe the Venn Prediction framework, propose the Inductive version of the framework, and explain the taxonomy we used with multiclass SVM. In section 4, we detail our experimental settings and the obtained results. Finally, in section 5, we give our conclusions and future plans.

2 Related Work

In this section, we provide related work that has been conducted on methods that convert the unthresholded output $f(x_i)$ of the SVM decision rule into a probability estimate. Hereon, $f(x_i)$ will also be referred as the SVM score of the example x_i. We examine Platt's method [12], SVM binning [5], and SVM with Isotonic Regression [16]. Moreover, we describe the approach we have followed for extending the binary SVM into multiclass SVM.

2.1 Platt's method

Platt introduced a method in [12] to estimate posterior probabilities based on the decision function f by fitting a sigmoid:

$$P(Y_j = 1|f(x_i)) = \frac{1}{1 + \exp(Af(x_i) + B)}, \tag{1}$$

where $Y_j \in \{-1, 1\}$. The best parameters A and B are determined so that they minimise the negative log-likelihood of the training data. Platt uses a Levenberg-Marquardt (LM) optimisation algorithm to solve this. As indicated in [12], any method for optimisation can be used. In this work, we use an improved implementation of Platt's method which uses Newton's method with backtracking for optimisation. Further details of this approach are described in [7].

2.2 SVM Binning

The SVM binning method [5] sorts the training examples according to their SVM scores, and then divides them into b equal sized sets, or bins, each having an upper and lower bound. Given a test example x_i, it is placed in a bin according to its SVM score. The corresponding probability $P(Y_j = 1|x_i)$ is the fraction of positive training examples that fall within that bin.

There is no imposed lower or upper bound on SVM scores. Therefore, when using this method it is possible for some scores from the test examples to fall below or above the low and high scores, respectively, of the training examples. If this happens the corresponding probability $P(Y_j = 1|x_i)$ is that of the nearest bin to the score of x_i.

2.3 Isotonic Regression

Isotonic regression has been used in order to map the SVM scores into probability estimates in [16]. An isotonic function has a monotonically increasing trend. If the scores of the SVM are ranked correctly, we can assume that the probability $P(Y_j = 1|x_i)$ will be increasing as the SVM scores increase. Therefore, we can use isotonic regression to map SVM scores into probability estimates. The most common algorithm used for isotonic regression is the Pair-Adjacent-Violators (PAV) algorithm.

The algorithm learns the probability estimate $g(x_i)$ for each ranked example x_i. First, we set $g(x_i) = 1$ if x_i is a positive example, and $g(x_i) = 0$ otherwise. If g is already isotonic the function has been learned. Otherwise, there must be an example where $g(x_{i-1}) > g(x_i)$. The two examples x_{i-1} and x_i are called pair-adjacent violators, because they violate the isotonic assumption. The values of $g(x_{i-1})$ and $g(x_i)$ are then replaced by their average, such that their values no longer violate the isotonic assumption. This process is repeated until an isotonic set of values is obtained. In the end, we have a list of probability estimates together with the adjacent SVM scores of the training examples. When a new example arrives, we assign the mapped probability estimate based on the score that x_i has obtained from the SVM decision rule. Normally, there will be intervals of scores with the same probability estimates. Since there are no imposed boundaries on the SVM scores, the lowest interval begins from $-\infty$ and the highest interval ends at $+\infty$.

2.4 Multiclass SVM

The original SVM works only for binary classification problems. In this work we apply the one-against-all procedure [14] to extend the SVM for multiclass tasks. In one-against-all, we train a binary SVM classifier for each class using as positives the examples that belong to that class, and as negatives all other examples. We then convert the SVM scores of each classifier into probability estimates based on the methods described in the previous subsections, and then we combine the binary probability estimates to obtain multiclass probabilities.

The probabilities are combined by finding the probability $P(Y_j = 1|x_i)$ of each class $j = 1, \ldots, c$ and then by normalizing the probabilities of all classes to 1. The largest probability is then used to classify the example.

3 Venn Prediction

Venn Prediction has been introduced in [15] where the interested reader can find a more detailed description of the framework. Since then VPs have been developed based on k-Nearest Neighbours [4], Nearest Centroid [3] and Neural Networks [9]. Furthermore, a VP based on a binary SVM has been developed in [17], and has been compared with Platt's method in the batch setting.

Typically, we have a training set[1] of the form $\{z_1, \ldots, z_{n-1}\}$, where each $z_i \in Z$ is a pair (x_i, y_i) consisting of the object x_i and its classification y_i. For a new object x_n, we intend to estimate its probability of belonging to each class $Y_j \in \{Y_1, \ldots, Y_c\}$. The Venn Prediction framework assigns each one of the possible classifications Y_j to x_n and divides all examples $\{(x_1, y_1), \ldots, (x_n, Y_j)\}$ into a number of categories based on a *taxonomy*. A taxonomy is a sequence A_n, $n = 1, \ldots, N$ of finite measurable partitions of the space $Z^{(n)} \times Z$, where $Z^{(n)}$ is the set of all multisets of elements of Z of length n. We will write $A_n(\{z_1, \ldots, z_n\}, z_i)$ for the category of the partition A_n that contains $(\{z_1, \ldots, z_n\}, z_i)$. Every taxonomy $A_1, A_2, \ldots, A_N$ defines a different VP. In the next subsection, we define a taxonomy based on the output of the SVM.

After partitioning the examples into categories using a taxonomy, the empirical probability of each classification Y_k in the category τ_{new} that contains (x_n, Y_j) will be

$$p^{Y_j}(Y_k) = \frac{|\{(x^*, y^*) \in \tau_{new} : y^* = Y_k\}|}{|\tau_{new}|}. \tag{2}$$

This is a probability distribution for the class of x_n. So after assigning all possible classifications to x_n we get a set of probability distributions $P_n = \{p^{Y_j} : Y_j \in \{Y_1, \ldots, Y_c\}\}$ that compose the multi-probability prediction of the VP. As proved in [15], these are automatically well calibrated, regardless of the taxonomy used.

The maximum and minimum probabilities obtained for each label Y_k amongst all distributions $\{p^{Y_j} : Y_j \in \{Y_1, \ldots, Y_c\}\}$, define the interval for the probability of the new example belonging to Y_k. We denote these probabilities as $U(Y_k)$ and $L(Y_k)$, respectively. The VP outputs the prediction $\hat{y}_n = Y_{k_{best}}$, where

$$k_{best} = \arg\max_{k=1,\ldots,c} \overline{p(k)}, \tag{3}$$

and $\overline{p(k)}$ is the mean of the probabilities obtained for label Y_k amongst all probability distributions. The probability interval for this prediction is $[L(Y_k), U(Y_k)]$.

[1] The training set is in fact a multiset, as it can contain some examples more than once.

3.1 Inductive Venn Prediction

The transductive nature of the original Venn Prediction framework is computationally inefficient, since it requires training the underlying algorithm for every possible class of each new test example. To address this problem we follow the idea of the Inductive Conformal Prediction, and propose an efficient Inductive Venn Predictor (IVP). Our approach splits the available training examples into two parts, the proper training set and the calibration set. We then use the proper training set to train the underlying algorithm and the calibration set to calculate the set of probability distributions for each new example.

Specifically, on each step of the algorithm in the online mode, we make a Venn Prediction analogue to a step of the Inductive Conformal Prediction ([15], p.98). For each number of available training examples $n-1$, we select $q \leq n-1$ examples to form the training set for the SVM classifier and use the remaining examples as the calibration set. For the taxonomy the training examples $z_1, \ldots, z_q$ are considered as fixed parameters. The original taxonomy function A is transformed to another taxonomy A' such that $A'_{n-q}(\{z_{q+1}, \ldots, z_n\}, z_i) = A_{q+1}(\{z_1, \ldots, z_q\}, z_i)$, for $i = q+1, \ldots, n$. Although slightly different VPs are applied on different steps, we will see that the validity of the outputs is not affected in practice.

3.2 SVM Venn Predictor

We define a taxonomy based on the output of the multiclass SVM. As explained in section 3, the validity of a VP is guaranteed under the i.i.d. assumption, regardless of the taxonomy used. For instance, a taxonomy that puts all examples in one single category would still give a valid predictor. Nevertheless, the performance of each VP is highly affected by the information provided from the categories defined in a taxonomy.

In this work, our taxonomy is simply based on the largest SVM score of the multiclass SVM. Therefore, each example is categorized according to the SVM classification. This taxonomy will give c categories and it is the simplest taxonomy we may define using the output of the SVM. If the SVM is good at classifying examples, then each category should contain sufficient information for the VP to perform well in terms of accuracy.

4 Experiments and results

In order to show the validity of the probability estimates of our method, we conduct experiments in the on-line mode. Initially all examples are test examples and they are added to the training set one by one after a prediction for each one is made. We calculate the cumulative average accuracy of the predictor, and the cumulative average probability. The cumulative average accuracy is calculated as the total accuracy of all tested examples, divided by the total number of tested examples. In the same way we calculate the cumulative average probability. If the methods provide well calibrated probability estimates, the cumulative

average accuracy should be near the cumulative average probability. We test all algorithms described in this paper: Platt's method; SVM Binning; SVM with Isotonic Regression (SVM-IR); and our SVM IVP. In our experiments, we did not try to improve the accuracy of these methods, instead we focused our work on testing the validity of the probability estimates. The underlying SVM algorithm that we have used works with the RBF kernel. We test each algorithm two times, one with a RBF parameter set to an optimal value, and another with a RBF parameter set to the optimal value divided by 10 (we do this in order to test the difference in the results when the predictors do not perform so well). The optimal value for each experiment was chosen based on offline tests (10-fold cross validation) that have been conducted with a standard SVM predictor. The standard SVM predictor was tested with the RBF parameter ranges of [0.1, 1] with steps of 0.1, and [1, 5] with steps of 1. The number of bins for the SVM Binning method was set to $b = 10$. In our experiments with the IVP we have set $q = \lceil 0.7(n-1) \rceil$. In the next two subsections, we describe our results on two multi-classification datasets.

4.1 Car evaluation dataset

The Car Evaluation dataset was derived from hierarchical decision model [1] and is available at [6]. The dataset contains 1728 examples with 6 features for each example. There are 4 classes for this dataset which describe the car acceptability based on features that describe the price, technology, and comfort of a car. In Figure 1, we show the results of the four methods on the Car Evaluation dataset. The best RBF parameter for this dataset is 0.2. For the first three methods we plot the cumulative average probability for the output classifications along with their cumulative accuracy, while for the proposed approach we plot the upper and lower cumulative probability for the output classifications along with their cumulative accuracy. One would expect the curves in each plot to be relatively near if the probabilities produced by the corresponding method were well calibrated. However this is true only for the IVP in both experiments and for Platt's method only with the optimal RBF parameter. When the RBF parameter is 0.2 the accuracy is around 90% for all methods. When we set the RBF parameter to 0.02 the accuracy is reduced to around 70%, while the probability estimates are near 100% for all methods except the IVP. As shown in the last row of the graphs, the IVP probability estimates are automatically lowered to around 68%, which is near the actual accuracy.

To confirm our observations from the graphs we calculated the 2-sided p-values of obtaining a total accuracy with the observed deviation from the expected accuracy given the probabilities produced by each method. In the case of the IVP we used the mean of the upper and lower probabilities as the probability of each prediction being correct. The p-values obtained for the outputs of the Platts's method with the RBF parameter set to 0.2 and the IVP with both parameter values were above 0.15. However, the p-values in all other cases were below 10^{-50}. This shows that the probabilistic outputs produced by the three methods can be far from being well calibrated. Even for Platt's method, a wrong

selection of the RBF parameter leads to misleading outputs. This does not happen with Venn Prediction which produces well calibrated outputs regardless of the underlying algorithm or the taxonomy used.

4.2 Red Wine quality dataset

The Red Wine quality dataset contains 1599 examples of physicochemical features of red variants of the "Vinho Verde" wine [2]. This dataset can be used as a regression or a classification problem. Each example has a quality score from 1 to 10. In this work, we have used the scores as 10 different classes from 1 to 10. This dataset is particularly difficult and requires some pre-proccessing to remove redundant features, or even reduce the number of classes. For instance, some classes have very few or even no examples in the training set. In our experiments, we have intentionally left the dataset to its original state in order to demonstrate the reliability of our probability estimates on difficult problems where the underlying algorithm may not be able to fit the data very well. In Figure 2, we show the online results of the four methods on the Wine quality dataset. The best RBF parameter on this dataset is 0.6. From the results, we can see that Platt's method, SVM-Binning, and SVM-IR did not give reliable probability estimates (due to the difficulty of the task), whereas the IVP has automatically lowered the probability estimates and has given well calibrated results in both cases. The 2-sided p-values for the IVP were above 0.3, whereas for all other methods with both RBF parameters the p-values were below 10^{-50}.

5 Conclusion

In this work, we have examined existing methods that convert SVM scores into probability estimates. We have shown that there is no trust in the probabilities produced by these methods, especially when the algorithm is not well configured or when the dataset is difficult. For the purpose of overcoming this limitation, we have developed an IVP based on SVMs, which guarantees (under the i.i.d. assumption) that the probability estimates will be well calibrated, regardless of the configuration of the algorithm or the difficulty of the task. The proposed IVP overcomes the computational inefficiency problem which renders the original Venn Prediction framework unsuitable in the case of large datasets. Our future aim is to improve the performance of our IVP in terms of accuracy by introducing better taxonomies that can be derived from the unthresholded scores of SVMs. Furthermore, we wish to investigate whether the one-against-all procedure is one of the causes for the non calibrated probability estimates of the existing methods, and we would like to compare our IVP with other multiclass procedures.

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

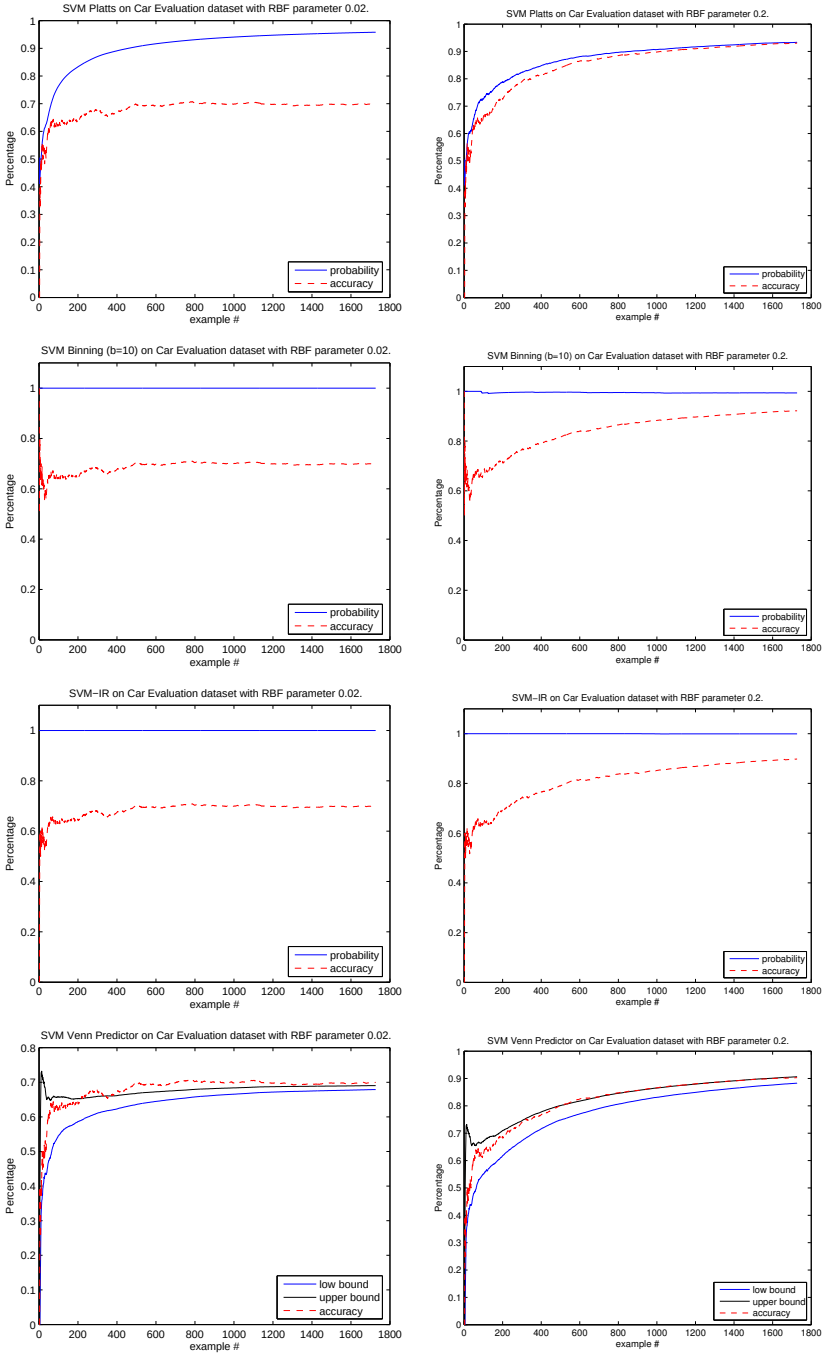

Fig. 1. Online experiments of all four methods on the Car evaluation dataset, RBF parameter is 0.02 on the left column and 0.2 on the right column.

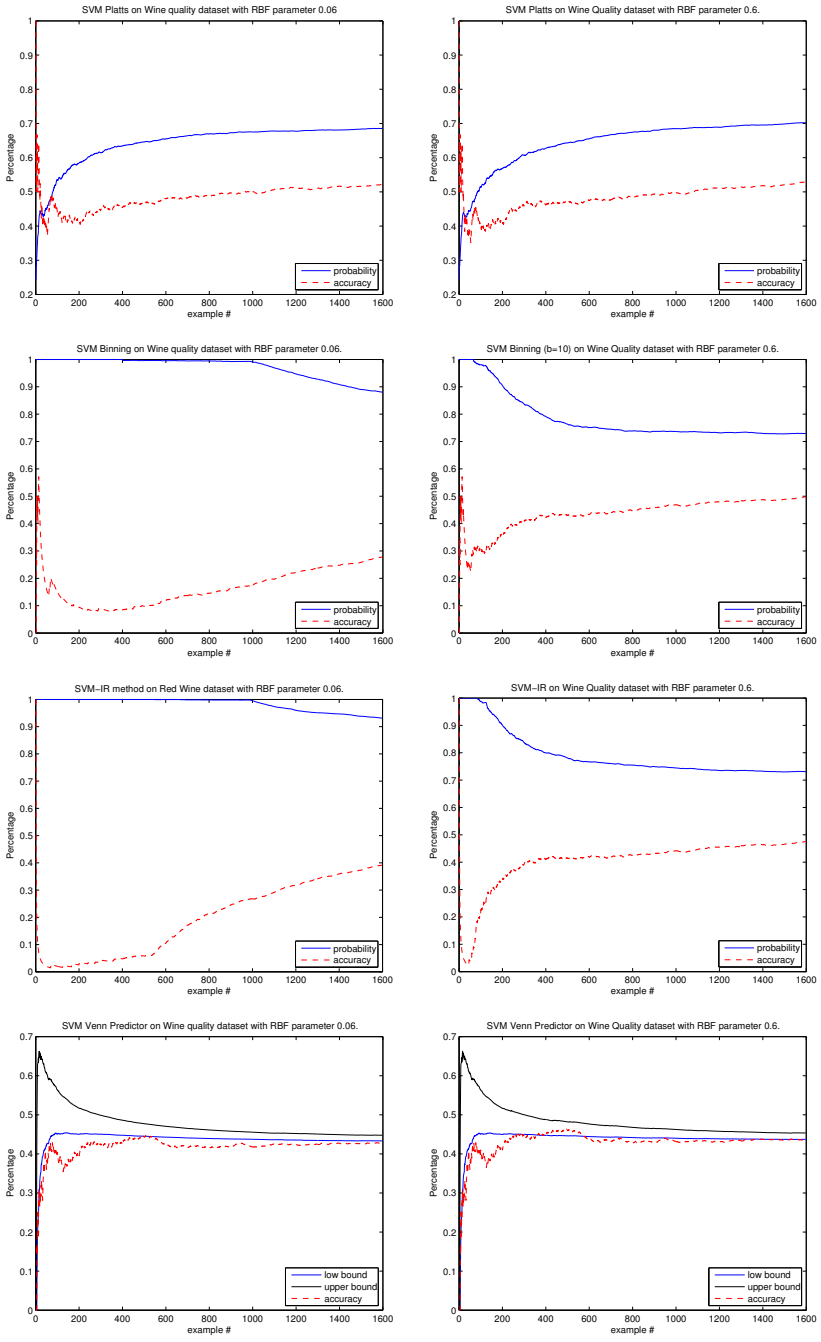

Fig. 2. Online experiments of all four methods on the Wine quality dataset, RBF parameter is 0.06 on the left column and 0.6 on the right column.

References

1. Marko Bohanec and Vladislav Rajkovi. V.: Knowledge acquisition and explanation for multi-attribute decision making. In *8th International Workshop "Expert Systems and Their Applications"*, 1988.
2. 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.
3. Mikhail Dashevskiy and Zhiyuan Luo. Reliable probabilistic classification and its application to internet traffic. In *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, volume 5226 of *LNCS*, pages 380–388. Springer, 2008.
4. Mikhail Dashevskiy and Zhiyuan Luo. Predictions with confidence in applications. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, volume 5632 of *LNCS*, pages 775–786. Springer, 2009.
5. Joseph Drish. Obtaining calibrated probability estimates from Support Vector Machines, 1998.
6. A. Frank and A. Asuncion. UCI machine learning repository, 2010.
7. Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. A note on Platt's probabilistic outputs for Support Vector Machines. *Mach. Learn.*, 68(3):267–276, October 2007.
8. Harris Papadopoulos. Inductive Conformal Prediction: Theory and application to Neural Networks. In Paula Fritzsche, editor, *Tools in Artificial Intelligence*, chapter 18, pages 315–330. InTech, Vienna, Austria, 2008.
9. Harris Papadopoulos. Reliable probabilistic prediction for medical decision support. In *Proceedings of the 7th IFIP International Conference on Artificial Intelligence Appications and Innovations (AIAI 2011)*, volume 364 of *IFIP AICT*, pages 265–274. Springer, 2011.
10. Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive Confidence Machines for Regression. In *Proceedings of the 13th European Conference on Machine Learning (ECML'02)*, volume 2430 of *LNCS*, pages 345–356. Springer, 2002.
11. 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.
12. 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.
13. Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
14. Vladimir N. Vapnik. *Statistical learning theory*. Wiley, 1998.
15. Volodya Vovk, Alexander Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. New York, Springer, 2005.
16. Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.
17. Chenzhe Zhou, Ilia Nouretdinov, Zhiyuan Luo, Dmitry Adamskiy, Luke Randell, Nick Coldham, and Alex Gammerman. A comparison of Venn Machine with Platt's method in probabilistic outputs. In *Proceedings of the 7th IFIP International Conference on Artificial Intelligence Appications and Innovations (AIAI 2011)*, volume 364 of *IFIP AICT*, pages 483–490. Springer, 2011.