Multiprobabilistic prediction in early medical diagnoses

Ilia Nouretdinov[1], Dmitry Devetyarov[1], Volodya Vovk[1], Brian Burford[1], Stephane Camuzeaux[2], Aleksandra Gentry-Maharaj[2], Ali Tiss[3], Celia Smith[3], Zhiyuan Luo[1], Alexey Chervonenkis[1], Rachel Hallett, Mike Waterfield[2], Rainer Cramer[3], John F. Timms[2], Ian Jacobs[2], Usha Menon[2] and Alex Gammerman[1]

[1]Computer Learning Research Centre, Royal Holloway, University of London, London, UK

[2]EGA Institute for Women's Health, University College London, London, UK

[3] BioCentre and Department of Chemistry, University of Reading, West Berkshire, UK

Keywords: Confident prediction, Probabilistic prediction, Risk, Diagnostic

Acknowledgments

This work was supported by EraSysBio+ grant funds from the European Union, BBSRC and BMBF "Living with uninvited guests: comparing plant and animal responses to endocytic invasions" to the Salmonella Host Interactions Project European Consortium; MRC grant G0802594 (Application of conformal predictors to FMRI research); MRC grant G0301107 (Proteomic analysis of the human serum proteome); Veterinary Laboratories Agency (VLA) of Department for Environment, Food and Rural Affairs (Defra) on Machine learning algorithms for analysis of large veterinary datasets; a grant from The National Natural Science Foundation of China (No. 61128003); and by grant "Development of New Venn Prediction Methods for Osteoporosis Risk Assessment" from the Cyprus Research Promotion Foundation.

Abstract

This paper describes the methodology of providing multiprobability predictions for proteomic mass spectrometry data. The methodology is based on a newly developed machine learning framework called Venn machines. Is allows to output a valid probability interval. The methodology is designed for mass spectrometry data. For demonstrative purposes, we applied this methodology to MALDI-TOF data sets in order to predict the diagnosis of heart disease and early diagnoses of ovarian cancer and breast cancer. The experiments showed that probability intervals are narrow, that is, the output of the multiprobability predictor is similar to a single probability distribution. In addition, probability intervals produced for heart disease and ovarian cancer data were more accurate than the output of corresponding probability predictor. When Venn machines were forced to make point predictions, the accuracy of such predictions is for the most data better than the accuracy of the underlying algorithm that outputs single probability distribution of a label. Application of this methodology to MALDI-TOF data sets empirically demonstrates the validity. The accuracy of the proposed method on ovarian cancer data rises from 66.7%11 months in advance of the moment of diagnosis to up to 90.2 % at the moment of diagnosis. The same approach has been applied to heart disease data without time dependency, although the achieved accuracy was not as high (up to 69.9 %). The methodology allowed us to confirm mass spectrometry peaks previously identified as carrying statistically significant information for discrimination between controls and cases.

1 Introduction

Prediction of heart disease (HD), breast cancer (BC) and ovarian cancer (OC) is a critical task. For some of these diseases (e.g., OC) it is especially crucial in their early stages, when the disease has no clinical symptoms. Mass spectrometry techniques are widely deployed in these problems.

When predicting diagnosis based on proteomics data, very often, the classical machine learning approach is to predict the diagnosis without any measure of how strongly we can believe in this prediction. In this work we describe the methodology of hedging predictions made in proteomics mass spectrometry data analysis. For example, in medicine it can be more useful to predict a probability of a disease (disease risk), rather than simply the diagnosis. There is a range of methods that can output probability distribution of a new label (see [1] for a review). However, these methods are usually based on strong statistical assumptions about example distribution. Hence, if the assumed statistical model is not correct, predicted probabilities will not be correct either.

Other known hedged prediction methods include statistical learning theory [7], Bayesian learning [10] and hold-out estimates (by the use of splitting all available examples into a training set and a test set). In statistical learning theory a simple prediction is produced for each object based on the training set of examples and there is a theoretical guarantee that these predictions get more accurate with greater probability when the training set becomes bigger: they are probably approximately correct. This implies that the probability of error will not exceed the threshold which can be calculated unless an event of the preset probability has happened. These theoretical bounds of the probability of error are often useless in practice: the upper bound of the probability error is usually greater that one. Exceptions are easy problems with a large number of training examples.

The main drawback of Bayesian learning is again that it depends on the statistical assumptions used in the model. This approach allows us to automatically output intervals for regression problems because the true probability distribution is known. When the chosen probability distribution is the real one, which we can be sure about only for artificially generated data, Bayesian learning outputs valid intervals. However, if this probability distribution is incorrect, the intervals produced make more errors than expected. This is explored in Section 10.3 of [3].

We suggest hedging predictions by producing a set of probability distributions through the use of Venn machines (also known as Venn predictors). The advantage of Venn machines is their validity regardless of the example distribution: the only assumption made is a simple i.i.d. assumption.
The framework of Venn machines was introduced in [3] and represents a new generation of prediction algorithms. These methods have a range of advantages over the known techniques. Firstly, the prediction which is made is always tailored to the object; as a result, we output a probability interval for each patient's diagnosis. Secondly, the only statistical assumption which is used is the exchangeability assumption which can be satisfied when the data set is randomised. Finally, Venn machines (and related methods such as conformal predictors) are much more confident and accurate. This comparison is rather informal due to the different nature of hedging predictions, however, one can easily see the improvement from the practical point of view.

Venn machines represent a framework that can generate a range of different algorithms. Practically any known machine learning algorithm can be used as an underlying algorithm in this framework and thus result in a new algorithm of prediction with confidence. However, regardless of the underlying algorithm Venn machines output valid results.

Strict definitions of Venn machines are given in Section 2. The main idea is as follows. We first divide examples into categories; the category assigned to an example may depend not only on the

example itself, but also on its relation to the rest of examples. For each possible new label, we classify the new object into one of the categories, and then use frequencies of labels in the chosen category as a distribution of the new object's label. Due to different hypotheses, the machine outputs several (two in the binary case) probability distributions (multiprobability distribution) for the new object rather than the single one.

Practically any known machine learning algorithm can be used as an underlying algorithm in this framework (such as Neural Network in [11] and SVM in [12, 13]). In this work we used logistic regression as an underlying algorithm. It is popular as a method that initially outputs probabilities and provides information about relative weight of features. We will compare Venn machines predictions with probability predictions output by logistic regression.

Venn machines are valid in the sense of agreeing with the observed frequencies (for details, see [3]). Among the first writers on frequentist probabilities we could name John Venn [4] and von Mises [5, 6]. The validity of Venn machines is based on special testing by supermartingales and is a generalization of the notion of valid probabilistic prediction.

The methodology is designed for the analysis of proteomic mass spectrometry data. The format of mass spectrometry data and how these data are obtained is described in Section 4. To demonstrate the methodology's usability, we applied it to MALDI-TOF data sets collected in the UKCTOCS trial.1 These proteomics data comprise three data sets: ovarian cancer (OC), breast cancer (BC), heart disease (HD). These are serum samples collected from patients diagnosed with the disease (we will call them cases) and healthy patients (they will be referred to as controls). The feature of OC data is that this data set represents serial samples collected over the period of 7 years. Thus, there are measurements taken from the same patients at different times. In addition, we would like to confirm peaks earlier identified as statistically significant for the considered diseases. In Section 5.4 we will demonstrate that different methods, in spite of a different nature of predictions, allow us to pinpoint the same informative mass spectrometry peaks.

2 Venn machines

First, we describe the general framework of Venn predictors.

Consider a training set consisting of pairs $(x1, y1) , \ldots , (xn, yn)$. To predict a label $yn+1$ for a new object $xn+1 = x\text{new}$, we check different hypotheses

$$y_{n+1} = y, \quad (1)$$

each time including the pair $(xn+1, yn+1) = (x\text{new}, y)$ into the training set.

The idea of Venn machines is based on a *taxonomy function An, n* ∈ N, which classifies the relationship between an example and the set of all other examples:

$$\tau_i = A_{n+1}((x_i, y_i), \{(x_1, y_1), \ldots ,$$

$$(x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots , (x_{n+1}, y_{n+1})\}).$$

Values τ_i are called categories and are taken from a finite set $T = \{\tau_1, \tau_2, \ldots , \tau_k\}$. Equivalently, a taxonomy function assigns each example (x_i, y_i) to a category τ_i, or in other words grouping all examples to a finite set of categories. This grouping should not depend on the order of examples within a sequence.

The most traditional Venn-type predictor is as follows. Categories (taxa) are formed using only the training set. For each non-empty category τ the following values are calculated: N_τ—the total number of examples from the training set assigned to category τ, and $N_\tau\,(y')$—the number of examples within category τ having label y'. Empirical probabilities of an object within category τ to have a label y are found as

$$P_\tau\,(y') = N_\tau\,(y')/N_\tau\,. \quad (2)$$

Now, given a new object x_{n+1} with unknown label y_{n+1}, one should assign it somehow to the most likely category of those already found using only the training set; let it be $\tau*$. Then the empirical probabilities $P_\tau*(y')$ are considered as probabilities of the object $xn+1$ to have a label y'. The idea of Venn's predictors allows us to construct several probability distributions (multiprobability distribution) of a label y' for a new object. First we consider a hypothesis that the label y_{n+1} of a new object x_{n+1} is equal to y, $(y_{n+1} = y)$. Then we add the pair (x_{n+1}, y) to the training set and apply to this extended sequence the taxonomy function A. This groups all the elements of the sequence to categories. Let $\tau*(x_{n+1}, y)$ be the category containing the pair (x_{n+1}, y). Now for this category we calculate, as previously, the values $N_\tau*$, $N_\tau*(y')$ and empirical probability distribution

$$P_{\tau*(xn+1,y)}\,(y') = N_{\tau*}\,(y')/N_{\tau*}\,. \quad (3)$$

This distribution depends implicitly on the object x_{n+1} and its hypothetical label y. Trying all possible labels $y_{n+1} = y \in Y$, we obtain a set of distributions $Py(y') = P\tau*(x_{n+1}, y)(y')$ for all possible labels y. These distributions in general will be different, as when changing the value of y we change (in general) grouping into categories, the category $\tau*(x_{n+1}, y)$, containing the pair (x_{n+1}, y), the numbers $N_{\tau*}$ and $N_{\tau*}(y')$. So we obtain, as the output of Venn predictors, as many distributions as the number of possible labels.

In the two-class problem ($Y = \{0, 1\}$), Venn predictors have two probability distributions, defined by $p_y(1) = P\{y_{n+1} = 1\}$. Thus, the output can be interpreted as the interval

$$[p^-_{\text{new}}, p^+_{\text{new}}] = [\min\{p_0(1), p_1(1)\}, \max\{p_0(1), p_1(1)\}] \quad (4)$$

which is an estimate of probability that $y_{n+1} = 1$. We will refer to p^-_{new} and p^+_{new} as *lower Venn prediction* and *upper Venn prediction*, respectively. They can be interpreted as lower and upper bounds for the probability. Thus if one sets a risk threshold Θ and takes all the predictions with lower Venn prediction not smaller than Θ then the expected proportion of cases between these examples should be between Θ and 1 as well.

A Venn predictor is entirely defined by its taxonomy. The taxonomy can be based on a certain prediction algorithm. For instance, the category of an example can be the label of its nearest neighbour

$$\tau_i = y_j \quad (5)$$

where

$$j = \arg\min_{j \neq i} | |x_i - x_j| | \,. \quad (6)$$

In the next section we use logistic regression as an underlying algorithm in the Venn machine framework.

3 Logistic regression

Logistic regression is one of the algorithms which output the probability distribution of a new label. It produces these distribution as follows.

Suppose each object out of the training set $x_1, \ldots, x_n$ is an m-dimensional vector, each with corresponding labels $y_1, \ldots, y_n \in Y = \{0, 1\}$.

The statistical model of logistic regression is based on the assumption that y_i is 1 with probability p_i and 0 with probability $1 - p_i$, where

$$p_i = 1/1 + e^{-(xi,b)} \quad (7)$$

and

$$1 - pi = 1/1 + e^{(xi,b)} . \quad (8)$$

The m-dimensional vector b is an unknown parameter of the model and can be interpreted as signed weights of different attributes. An additional value of 1 may be appended to each x_i to allow a free additive term to (x_i, b).

The optimization goal for logistic regression is:

$$\sum_{i=1}^{n} \log (1 + e^{(-1)yi\_xi,b)} + a(b, b) \to \min_{b} . \quad (9)$$

This formula is based on the maximum likelihood estimator for Logistic regression, with an added regularisation term $a(b, b)$ to ensure that a minimum always exists and avoid overfitting. In this work we always set $a = 0.1$. The above minimisation problem can be solved by the gradient descent method. Denote by $\hat{b}$ the solution of the optimisation problem above.

For a new object x_{new}, the probabilistic prediction based on logistic regression will be:

$$p_{\text{new}} = 1/1 + e^{-(x_{\text{new}}, \hat{b})} , \quad (10)$$

which estimates the maximum likelihood probability that $y_{\text{new}} = 1$ if the data are generated by a distribution from a logistic model. We will call p_{new} a *direct* prediction to distinguish it from multi-probabilistic predictions produced by a Venn machine.

3.1 Logistic regression as an underlying algorithm

Now we can describe how logistic regression can be plugged into Venn machine as an underlying algorithm. As earlier the aim is the prediction of labels y_i which are equal to 0 for controls and 1 for cases, given objects x_i—vectors of features, which are intensities of the most frequent peaks in the logarithmic scale. The probabilistic method of logistic regression allows us to create a new type of taxonomy. The *logistic taxonomy* τ_i, i = 1, . . . , n + 1 is defined as follows. The solution of the optimisation problem $\hat{b}$ is calculated for the whole set

$$(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y)$$

as a training set and is used to make direct predictions $p_1, \ldots, p_{n+1}$ on the same (training) examples. These predictions are not fair leave-one-out predictions, but it is correct to use them for taxonomy construction.

Let p'_i, $i = 1, \ldots, n + 1$, be direct predictions p_i sorted in the ascending order. Set a number of taxonomy categories K. We use $K = 5$ in this study (the dependence on the parameter K is discussed in Appendix C). Let $L_0 = 0$ and $L_1, L_2, \ldots, L_K$ be the integers closest to $(n + 1)/K$, $2(n + 1)/K$, $3(n + 1)/K$, . . .

, $n + 1$, respectively. The category τ_i is then defined as the only possible number $j \in \{1, \ldots, K\}$ such that $P'L_{j-1} < p_i \leq p'L_j$.

Thus, we divide the examples into several groups of approximately equal size being grouped by the similarity of their direct predictions. We construct categories of equal size, because the small size of categories will result in overfitting and will be punished by the large diameter of a probability intervals. On the other hand, large categories will result in underfitting and will be punished by producing predictions that are not close enough to 0 or 1. Note that the logistic assumption that there exists such b for which logistic regression model is exactly true is not necessary for the validity of Venn Machine.

4 Mass spectrometry data

We would like to develop a methodology which provides multiprobability predictions for proteomic mass spectrometry data. Hence, we have to take into account the format of the data and peculiarities of the problem. We used mass MALDI-TOF spectra data sets collected in the UKCTOCS trial for three problems: heart disease (HD), breast cancer (BC), ovarian cancer (OC). A detailed description of the pre-processing of mass spectrometry samples for OC can be found in [2]. For two others it is analogous. The difference of OC from others is that in addition to spectra, OC data examples contain serum cancer antigen 125 (CA125) known to be the marker. The diagnostic accuracy of CA125 is very high at the moment of diagnosis but for early diagnosis it needs improvement with other markers. Therefore mass spectrometry data are collected for OC as well. Samples were obtained from women at various timepoints prior to diagnosis. For example, in OC data set there are 179 samples (from 104 cases): 59 cases had one measurement, 26 had 2 measurements, 11 had 3 measurements, 5 had 4 measurements, and 3 had 5 measurements. Each measurement was supplied with two control examples. The controls are not serial: if a case has more than one measurement, the control examples are new for each of them.

For fair classification that is our aim, we never use more than one sample from the same patient, thus the maximal possible number of samples used together in our classification problems is:

– 312 for OC data set (104 cases and 208 controls)
– 561 for HD data set (187 cases and 374 controls)
– 162 for BC data set (54 cases and 108 controls)

For each measurement, a sample of polypeptides in serum (or plasma) is taken. They were processed with Matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) MS for high-throughput profiling of clinical samples. For a sample, the spectrometer produces two column ASCII files of m/z values (m/z means the mass divided by the charge number of ions, in which mass spectra are represented) and corresponding intensities for further processing and analysis. Mass spectra plots profiling can be noisy because of physical, electrical or chemical sources. Pre-processing is applied to mass spectra to get rid of these systematic artefacts. Auxiliary goals of pre-processing are to normalize the spectra from different samples and reduce the dimensionality of the data. Pre-processing can include the following steps: smoothing by averaging the intensities within a FIR filter (moving window); baseline subtraction; normalization in the form of division by the area under the curve. Data processing steps (data reduction, smoothing by averaging the intensities within a FIR filter (moving window), baseline subtraction, normalization, peak defining and peak alignment) were applied to the data, as previously described.

After the true signal is extracted from mass spectra, peaks are identified in each spectrum and then aligned, that is, peaks from different spectra get related to each other and are considered as one peak. Next, the intensities of identified peaks are calculated. The peaks appearing in >33% of all spectra (cases and controls) were used for subsequent data analysis. Finally, the intensities of these peaks were averaged across replicates for each sample.

Finally, the data to which we apply our methodology is represented as intensities of identified peaks. The peaks are sorted by their frequency: the higher the number of samples in which a peak is presented, the higher the rank of that peak. We usually consider a certain number of the most frequent peaks only. Thus, every object x_i is a vector of features, which are intensities of the most frequent peaks. Each sample of the OC data set is also assigned a level of OC biomarker CA125, which helps discriminate OC samples from healthy samples. For this reason, for OC we will make our diagnosis based not only on MALDI-TOF data, but also on CA125 levels. Originally, each case was accompanied by two controls matched on patient age, sample collection location and sample collection date/time, among other factors. For this reason, in each data set the number of controls is twice as greater than the number of cases.

Each case is assigned a non-negative value $T(\tau)$—time to diagnosis confirmed by histology/cytology for OC data and time to death for HD data. We will refer to this value as 'time to diagnosis/death'. We will not be trying to predict the time to diagnosis/ death, however, we need it to form cross-validation sets for the early diagnostics. Each sample in OC data set is also assigned a level C of the biomarker CA125.

4.1 Previous analysis of MS data

Each object, x_i, only comprises intensities of the most common peaks. It was shown in statistical analysis [8] that the information useful for discrimination between healthy and diseased samples is concentrated in peaks 2 and 3 in OC data. For this reason, each OC object comprises the five most common peaks. As for the HD data, we consider the peaks represented in at least 1/3 of samples (41 peaks) in this data set.

We will choose the time slots relying on the results of statistical analysis which was carried out on all data sets in the triplet setting [8]. In this analysis we considered time slots of fixed width (6, 9 or 12 months) and identified those where peak intensities (together with CA125 for OC data) provided statistically significant discrimination between healthy and diseased samples. For example, for HD these are 12-month windows starting with 0, . . . , 22months in advance of the date of death; for BC these are 12-month windows starting with 0, . . . , 8 months in advance of the diagnosis. For the purpose of this paper, we consider the union of these time slots, that is, all samples for HD data set and samples taken up to 20 months in advance of BC diagnosis. As it was mentioned earlier, each sample of the OC data set is assigned a level of OC biomarker CA125 and we are aiming to improve the predictive ability of this biomarker by complementing it with the information carried by peak intensities. In [8] it was demonstrated that CA125 provides statistically significant discrimination between OC and healthy samples for up to 9 months in advance of the day of diagnosis. For this reason, each sample in OC data set is also assigned a level C of the biomarker CA125, and the aim of the OC data analysis is to improve predictive ability of CA125 by adding information contained in mass spectrometry peaks.

Since in statistical analysis we considered 6-month time slots, and CA125 performed well for up to 9 months in advance of the day of diagnosis, in this paper we will carry out the analysis of OC samples

taken between 10 and 16 months in advance of the diagnosis—the first time slot when mass spectra showed statistically significant improvement in OC prediction.

5 Results and discussion

To demonstrate how the proposed methodology works in practice, we applied the designed categorized confidence machine to MALDI-TOF data: HD, BC and OC data sets observed above. In all experiments, we use leave-one-out mode: each example (x_i, y_i) is considered as if it were a new test example and all the remaining examples in the data are treated as the training set.

5.1 Multiprobability prediction and its advantages

We are applying Venn machines with the taxonomy based on logistic regression to MALDI-TOF datasets. Each object x_i is a vector comprising the following features: the most frequent peaks, value '1' for possible absolute term in logistic regression model and CA125 value for OC dataset. Since logistic regression also produces probability distributions, we can compare the results of the application of the Venn machine based on logistic regression and the probabilistic predictor of logistic regression itself. The experiments were applied to the same type of objects xi in the leave-one-out mode. Results of experiments for several controls and cases of HD data are shown in Table 1 for illustrative purposes. For each example, the table contains the true label y_{new}, and the probability interval [p^{-}_{new}, p^{+}_{new}]. For example, Venn machines output prediction interval [0.313, 0.321] for probability that example 1 is a case (y = 1). As prediction intervals indicate, the correct label is 0. The table also includes predictions p_{new} (Direct prediction) output by logistic regression for each example. Recall that we call these predictions direct predictions as opposed to Venn predictions output by Venn machines. The table demonstrates that both direct and Venn predictions can be correct or erroneous. The performance of such predictions will be analyzed in Section 5.3.

Table 1 Leave-one-out Venn predictions for HD data

No.	True label	Venn prediction	Direct prediction
1	0	0.313–0.321	0.508
2	1	0.616–0.616	0.689
3	0	0.321–0.330	0.51
4	0	0.143–0.259	0.371
5	0	0.616–0.634	0.622
6	0	0.321–0.321	0.484
7	1	0.313–0.321	0.516
8	0	0.616–0.634	0.558
9	0	0.143–0.161	0.333
10	1	0.616–0.625	0.703

First, we would like to demonstrate the validity of Venn predictions: true probabilities of the distribution of labels are covered or almost covered by the interval between lower Venn prediction and upper Venn prediction. Since we do not know the true probabilities of the distribution of labels, we compare empirical probabilities, that is: the mean of the true labels with the mean of the direct and Venn predictions. Figure 1 is a graphical representation of corresponding cumulative results. The horizontal axis shows the number of observed examples. The vertical axis shows the cumulative values of: (1) true labels y_{new} (a solid line); (2) lower and upper Venn predictions p^{-}_{new}, p^{+}_{new} (two dot-dashed

lines) and (3) cumulative direct predictions p_{new} (a dashed line). Cumulative value means the value summarized over the previous examples: cumulative true value 100 is the sum of true values of the examples 1, 2, . . . , 100. The examples are sorted according to direct predictions.

Firstly, the plot demonstrates validity of Venn machine outputs. Secondly, we can see that probability intervals output by Venn machines are narrow (0.025 on average for HD data); hence, they are almost as precise as single probabilities. Finally, Fig. 1 demonstrates that probability intervals can be more accurate than single probabilities produced by logistic regression. It can be seen from the Figure that the true labels are very different from the direct predictions but are only slightly above the upper Venn prediction up to approximately 210 examples and within the upper and lower Venn predictions for the remaining examples after this point. Thus, we have an example where direct predictions are misleading, while Venn predictions tend to cover true labels.

Similar plots for OC and BC data sets can be found in Appendix A. They also confirm the property of validity of Venn machines: the line corresponding to cumulative true labels is covered or almost covered by the space between lines for cumulative Venn predictions. For OC the area between cumulative Venn prediction lines is also narrow (with average probability interval width of 0.026) and almost covers the cumulative true label line, while the line representing cumulative direct predictions diverges from the true label line for up to 180 first examples. For BC, the average interval width is much larger (0.552), hence, probability intervals are not as precise and informative. However, the width will decrease to 0.142 if we take only 20 peaks, and to 0.03 if we use only peak 19 that is the most informative one (as we will see further).

It can be said that both algorithms relied on the assumption of the mechanism generating the data—logistic regression statistical model. However, probability predictions used this mechanism directly, and Venn machines deployed the mechanism when defining the taxonomy. As a result, since the statistical model does not hold true (the opposite can be guaranteed only for artificially generated data), probabilities output by logistic regression are different from empirical probabilities. In contrast, Venn machine's validity was not affected by the fact that the model is not correct. Hence, Venn machine predictions appeared to be more accurate than singleton probability predictions. Results of Venn Machines based on underlying methods other than logistic regression can be found in [15], pages 173–175.

5.2 Prediction dynamics in OC data: examples

The feature of the OC data set is that OC cases can have several measurements taken at different moments. For this reason, we can observe the change in probability intervals output by Venn machines for this data set. We will consider several OC cases that have measurements taken over a long period of time and will show how probability intervals output by Venn machines are changing when the patient is approaching the moment of diagnosis.

The experimental setting was the following: We selected patients with at least three measurements. For each measurement, we considered the earliest 6-month time slot containing the measurement (e.g., if a measurement was taken 6.5 months in advance, we consider the time slot from month 12 to month 6). We then used all measurements taken in this time slot as a training set. If an OC case has several measurements in this time slots, then we eliminate from the training set all samples except for the one closest to the moment of diagnosis. All other parameters (such as the number of peaks, W and V) were the same as in previous experiments.

Table 2 shows the dynamics of prediction intervals for the serial case 39 approaching the time of diagnosis. This case is selected as the only one having 4 serial measurements during the last year.

Recall that the controls were reselected for each measurement and they are not serial ones, therefore it is impossible to look at such dynamics for them. Each row corresponds to a single measurement. Column 2 demonstrates how early in advance this measurement was taken. These samples with multiple measurements illustrate two trends in probability interval change. First, the interval is getting narrower when the moment of diagnosis is approaching, which means that two probability distributions produced by Venn machines are getting closer to each other, and as a result, the overall prediction is getting more precise. Second, the interval is moving towards 1. This implies that we have more trust in our prediction and the prediction is indeed correct.

Table 2 Dynamics of prediction intervals output by Venn machines for measurements taken from the same OC case

Personal ID	Months in advance	Prediction interval
39	10	0.53–0.71
	4	0.44–0.94
	2	0.96–1.00
	1	0.97–1.00

5.3 Accuracy of Venn predictor and underlying algorithm

Even though Venn machines and logistic regression produce multiprobability and probability predictions, respectively, their outputs can be interpreted as usual bare predictions (forced predictions). However, we should bear in mind that when we force Venn machines to output a single prediction, we are not able to use advantages of multiprobability predictions, although for validity of them Venn Machine was originally designed in [3].

We can extract forced predictions out of Venn machines and logistic regression the most intuitive way: we classify a new sample as 1 (case) if and only if $p_{\text{new}} > 0.5$ for direct prediction or if $p^{+}_{\text{new}} + p^{-}_{\text{new}} > 1$ for Venn prediction. Accuracy means the percentage of predictions that are correct after this interpretation. This will also allow us to compare accuracy of Venn predictions with direct predictions.

The objective of the research for BC and OC data sets is to predict the disease as early as possible. For this reason we consider the dynamics of predictive ability of mass spectrometry peaks across the timeline: we consider accuracy of the proposed methodology on samples in different time slots of the fixed duration (6 months for OC and 12 months for BC).

Table 3 allows us to compare accuracy of the forced predictions by Venn machine and logistic regression for the OC data. The table demonstrates that Venn machines are comparable with logistic regression in terms of forced prediction accuracy: in time slots close to the moment of diagnosis Venn machine is slightly outperformed by logistic regression, then in months 5–7 they have equal accuracy, and in months 8–11 (time slots we are mostly interested in) Venn machine overperforms logistic regression. Venn machines produce predictions with accuracy higher than 73% up to 10 months in advance of the moment of diagnosis.

Table 3 Dynamics of Venn machine and logistic regression performance on the OC dataset

Time slot	Venn machine			Logistic regression		
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
0–6	90.2	95.6	87.5	93.6	85.3	97.8
1–7	88.1	91.1	86.6	92.9	83.9	97.3
2–8	76.6	59.6	85.1	87.9	78.7	92.6
3–9	83.3	58.3	95.8	83.3	69.4	90.3
4–10	75.3	59.3	83.3	82.7	66.7	90.7
5–11	79.7	52.2	93.5	79.7	56.5	91.3
6–12	81.7	55.0	95.0	81.7	55.0	95.0
7–13	70.6	35.3	88.2	70.6	35.3	88.2
8–14	82.4	52.9	97.1	78.4	47.1	94.1
9–15	75.0	45.0	90.0	71.7	35.0	90.0
10–16	73.8	67.9	76.8	67.9	25.0	89.3
11–17	66.7	50.0	75.0	64.3	17.9	87.5
12–18	59.5	32.1	73.2	61.9	7.1	89.3
13–19	66.7	33.3	83.3	65.6	13.3	91.7
14–20	64.0	24.0	84.0	65.3	12.0	92.0
15–21	65.0	40.0	77.5	71.7	20.0	97.5
16–22	33.3	0.0	50.0	63.3	10.0	90.0

For HD we consider the whole dataset rather than dynamics across the timeline, since it is sufficient to predict this disease at any moment to prevent the consequences. The accuracy of the application of Venn machines to the HD data is 69.9%. The accuracy is again comparable with the accuracy of underlying algorithms: 67.9%.

Accuracy on the BC data is better than the accuracy of the underlying algorithm (55.2 %) but still low (64.4 %). We can speculate that this fact can be explained the following way: we showed earlier [8] and will show in this paper that BC data set contains only one peak which carries statistically significant information (peak 19). When we consider more features, we include more peaks that carry more noise, which results in poor performance of Venn machines. The early diagnosis of BC also gets more complicated because of the following feature of the BC data set: all BC samples were taken at least three months in advance of the moment of diagnosis.

5.4 Identification of informative peaks

Our previous research of OC, BC and HD data [8], devoted to statistical analysis of the data, allowed us to determine statistically significant peaks which could be potential candidates for biomarkers. We identified certain mass spectrometry peaks that carry statistically significant information for the diagnosis of the diseases. Venn machines can also indirectly pinpoint informative peaks. Despite the different nature of these methods, Venn machines confirm mass spectrometry peaks that carry statistically significant information for discrimination between controls and cases.

Statistical analysis was carried out for the same data sets but in a different experimental setting: the data were normalised against such factors as age, sample collection time and location, storage and transportation conditions. All samples were grouped in triplets comprising one case and two controls matched by these factors. Thus, when making predictions, we had additional information about label distribution: we knew that exactly one sample is diseased in a triplet. Statistical analysis in triplet setting pinpointed mass spectrometry peaks that allowed to statistically significantly discriminate between healthy and diseased samples in different time slots.

We will consider the time slots when Venn machines produced high accuracy on the datasets. For OC we are especially interested in time slots starting from month 10, because this is the first time slot

when CA125 on its own does not provide statistically significant discrimination between cases and controls. We will again consider one time slot including all samples for HD.

Venn predictors produce an explicit ranking of peaks. It can be extracted from the coefficients of the optimal value for parameter *b* from (9) calculated in Venn predictions. The optimal parameter *b* is recalculated for each example and each of two hypotheses; to summarize the coefficients for all runs, we calculate the mean values of the coefficients. The most important features are those with the highest absolute value of their coefficients.

Table 4 shows the peak with the highest ranking produced in different time slots by Venn machines for OC and BC. The table demonstrates that peak 3 is the top OC peak in months 9–11, peak 19 is the highest ranked BC peak in months 0–11.

Table 4 Top peaks pinpointed by Venn machines in different time slots for OC and BC datasets

Month	Top peak number	
	OC	BC
0	4	19
1	2	19
2	2	19
3	2	19
4	2	19
5	1	19
6	2	19
7	2	19
8	2	19
9	3	19
10	3	19
11	3	19
12	5	2
13	5	1
14	2	15
15	2	15
16	3	15

Table 5 summarizes all peaks selected by different approaches: statistical analysis (known from [8]) and Venn machines. Those peaks are shown that were selected in time slots of high interest: slots starting with months 0–9 for BC, 10–11 for OC, whole dataset for HD. For HD, the three peaks with the highest ranking produced by Venn machines are given. The m/z-values of the peaks shown in the table are given in Table 6 in Appendix B. Table 5 demonstrates that Venn machines confirm the peaks identified as carrying statistically significant information in the triplet setting. For OC data in time slots starting with month 10 or 11 both methods select peak 3 (9297.8 Da). These are the time slots when CA125 on its own does not carry statistically significant information as shown in previous research [8]. OC peak 3 was also pinpointed in research on other datasets. It is similar to the UKOPS peak CTAPIII with m/z-value 9288 Da, which also has lower intensities for OC samples [14]. In addition, peak 3 coincides with peak 7 (m/z-value range 9294.7–9319.7 Da) in the pilot trial [9].

Table 5 Numbers of the most important peaks selected with different methods for heart disease, breast cancer and ovarian cancer datasets

Method	HD	BC	OC
Statistical analysis	7, 6, 4	19	2, 3
Venn machine	4, 6, 7	19	3

Corresponding m/z values are shown in Table 6.

Figure 2 shows the median dynamics of values logC (logarithmic intensity of CA125) versus logC − log I(3) (the same minus logarithmic intensity of the third by frequency peak) for case measurements. It allows to visualize and compare relative growth of CA125 on its own and in combination with peak 3. In OC data, cases can have several measurements taken at different moments. For each time moment, the latest available case measurement for each triplet group is taken into account. These measurements are averaged by median through all samples. The figure illustrates that the combination of CA125 with peak 3 starts to grow earlier than logC. However, the CA125 growth at the moments close to diagnosis is quicker due to the exponential growth of CA125.

For BC data we will observe the dynamics of peak 19 identified as a statistically significant peak, whose intensities are supposed to be lower for cases rather than for controls according to our research. In Fig. 3 a solid line represents the median dynamics of peak 19 for BC cases, a dashed line represents the peak 19 median calculated for all BC controls. The values in the figure were calculated for samples within 9-month window ending with the month shown on the horizontal axis.

One can see from Fig. 3 that peak 19 median intensity drops about 15 months in advance of the moment of the diagnosis, which confirms our hypothesis about predictive ability of peak 19 and explains the results we obtained using this peak when discriminating between BC cases and controls.

6 Conclusion

This paper introduced the methodology of hedging predictions for proteomic mass spectrometry data. We applied the described methodology to real-world MALDITOF data sets and demonstrated how it works. We empirically confirmed the validity of Venn machines and demonstrated that Venn machines can provide narrow probability intervals that are more accurate than the probabilities provided by its underlying algorithm. Even though Venn machines produce multiprobabilistic predictions, their output can be interpreted as predictions without hedging, similarly to the output of conventional machine learning methods. It was demonstrated that when forced to make single predictions, our methodology provides accuracy similar to the accuracy of the underlying algorithms. As a result, this methodology can provide high accuracy well in advance of the moment of the disease diagnosis. However, accuracy is not the main goal of Venn predictions. One may be interested in probabilistic estimates of risk instead of pure predictions. For this task, Venn Machine based on an underlying algorithm (logistic regression) requires weaker assumptions than logistic regression itself and its estimates are accurate in a wider class of situations. Finally, we confirmed mass spectrometry peaks identified as statistically significant in the analysis carried out on the same data sets.

References

1. Dawid, A.P.: Probability Forecasting. Encyclopedia of Statistical Sciences, vol. 7. pp. 210–218. Wiley, New York (1985)
2. Devetyarov, D., Nouretdinov, I., Burford, B., Luo, Z., Chervonenkis, A., Vovk, V., Waterfield, M., Tiss, A., Smith, C., Cramer, R., Gentry-Maharaj, A., Hallett, R., Camuzeaux, S., Ford, J., Timms, J., Menon, U., Jacobs, I., Gammerman, A.: Analysis of serial UKCTOCS-OC data: discriminating abilities of proteomics peaks. (Technical report). http://www.clrc.rhul.ac.uk/projects/proteomic3.htm (2008)
3. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)
4. Vovk, V., Shafer, G., Nouretdinov, I.: Self-Calibrating Probability Forecasting. (On-line compression modelling project. Working paper 9) http://vovk.net/cp/09.pdf (2003)
5. von Mises, R.: Grundlagen der wahrscheinlichkeitsrechnung. Math. Z. 5, 52–99 (1919)
6. von Mises, R.: Wahrscheinlichkeitsrechnung, Statistik und Wahrheit. Julius Springer, Wien (1928)
7. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
8. Timms, J.F., Menon, U., Devetyarov, D., Tiss, A., Camuzeaux, S., McCurry, K., Nouretdinov, I., Burford, B., Smith, C., Gentry-Maharaj, A., Hallett, R., Ford, J., Luo, Z., Vovk, V., Gammerman, A., Cramer,R; Jacobs, I.: Early detection of ovarian cancer in pre-diagnosis samples using CA125 and MALDI MS peaks. Cancer Genomics Proteomics 8(6), 289–305 (2011)
9. Gammerman, A., Vovk, V., Burford, B., Nouretdinov, I., Luo, Z., Chervonenkis, A.,Waterfield, M., Cramer, R., Tempst, P., Villanueva, J., Kabir, M., Camuzeaux, S., Timms, J., Menon, U., Jacobs, I.: Serum proteomic abnormality predating screen detection of ovarian cancer. Comput. J. 52(3), 326–333 (2009). On behalf of the British Computer Society
10. Gelman, A., Carlin, J.B., Stern, H.S. Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall/CRC (2003)
11. Papadopoulos, H.: Reliable probabilistic prediction for medical decision support. In: Artificial Intelligence Applications and Innovations IFIP Advances in Information and Communication Technology, vol. 364, pp. 265–274 (2011)
12. Zhou, C., Nouretdinov, I., Luo, Z., Adamskiy, D., Coldham, N., Gammerman, A.: A comparison of Venn machine with Platt's method in probabilistic outputs. In: 12th INNS EANN-SIG International Conference, EANN 2011 and 7th IFIP WG 12.5 International Conference, Artificial Intelligence Applications and Innovations. Corfu, Greece, 15–18 September 2011. Proceedings Part II. IFIP AICT, vol. 364, pp. 483–490 (2011)
13. Lambrou, A., Papadopoulos,H., Nouretdinov, I., and Gammerman,A.: Reliable probability estimates
based on support vector machines for large multiclass datasets. In: AIAI 2012 Workshops, IFIP AICT, vol. 382, pp. 182–191. Springer (2012). doi:10.1007/978-3-642-33412-2_19
14. Timms, J.F. , Cramer, R., Camuzeaux, S., Tiss, A., Smith, C., Burford, B., Nouretdinov, I., Devetyarov, D., Gentry-Maharaj, A., Ford, J., Luo, Z., Gammerman, A., Menon, U., Jacobs, I.: Peptides generated ex vivo from serum proteins by tumour-specific exopeptidases are not useful biomarkers in ovarian cancer. Clin. Chem. 56, 262–271 (2010)
15. Devetyarov, D. Confidence and Venn machines and their applications to proteomics. Doctoral thesis (2011). Available at http://digirep.rhul.ac.uk/file/4d74228e-3ca0-d6ca-469f-0ce0b22c122d/1/PhD_Thesis_Final_Dmitry_Devetyarov2011.pdf

Appendix A. Validity plots for Venn machines applied to BC and OC data

Figure 4 is the validity plot for OC data. Figures 5, 6, 7 are for BC data. They are done in the same format as the validity plot for HD data (Fig. 1 in the main text).

Appendix B. M/z-values of identified peaks

M/z values of identified peak can be found in Table 6.

Table 6 m/z-values of statistically significant peaks for heart disease, breast cancer and ovarian cancer datasets

Dataset	Peak number	m/z-value
HD	4	4211.1
	6	4055
	7	5338.3
BC	19	6637.8
OC	2	7772.1
	3	9297.8

Appendix C. Dependence on the taxonomy parameter

In this work we used K = 5. On Fig. 8 we show what happens if this parameter is changed. It can be seen that with the accuracy becomes satisfactory with at least 4–5 taxa. On the other hand when the number K of taxa increases the interval width becomes wider (less informative) without essential improvement of the accuracy. So the choice of K = 5 was reasonable enough.

Figures

Figure 1 Cumulative Venn and direct predictions for the HD data

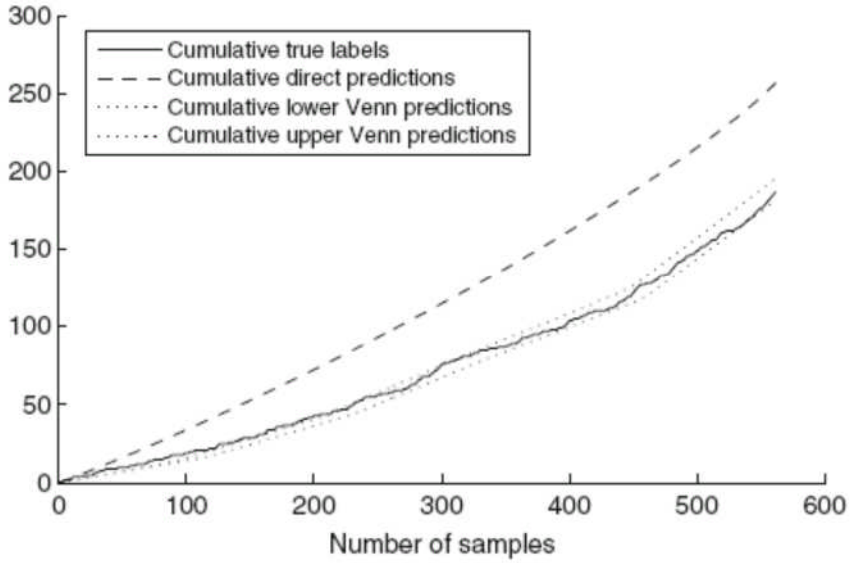


Figure 2 Median dynamics of rules log CA125 and log CA125 − log(Peak 3) (for OC cases only)

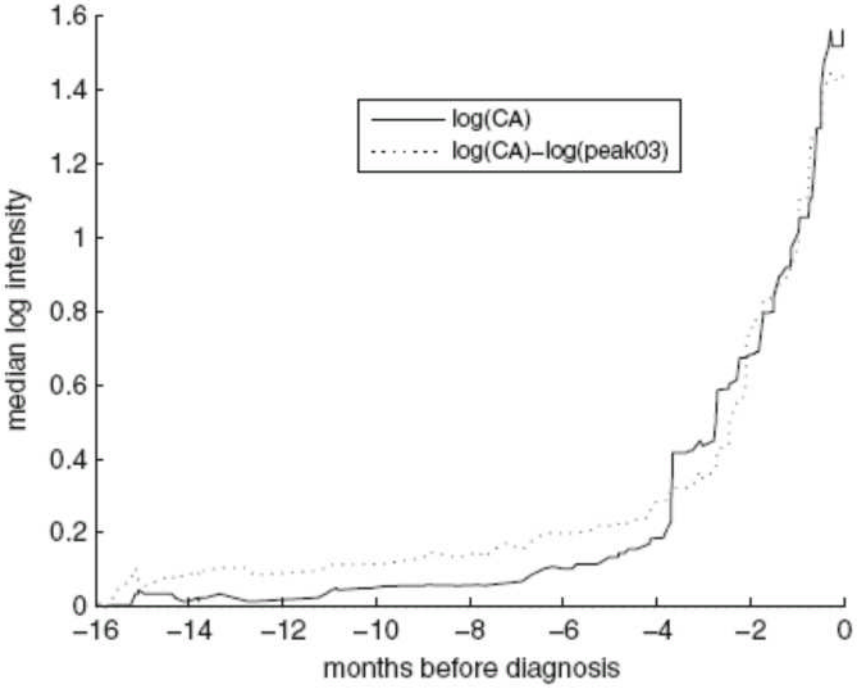


Figure 3 Median dynamics of peak 19 in BC data for cases and the median of peak 19 for controls

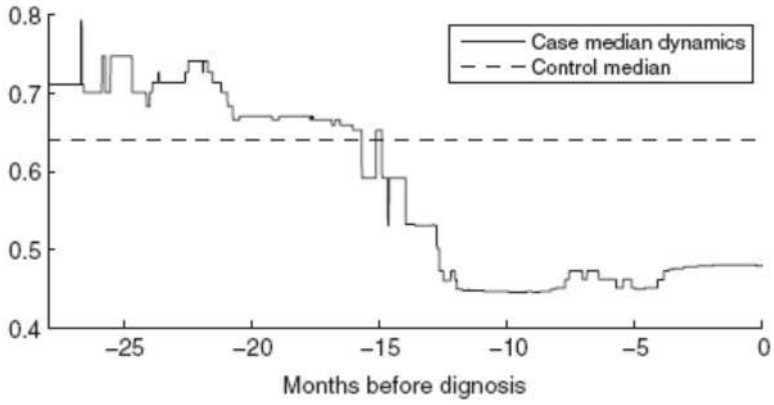


Figure 4 Cumulative Venn and direct predictions for the OC data (all samples)

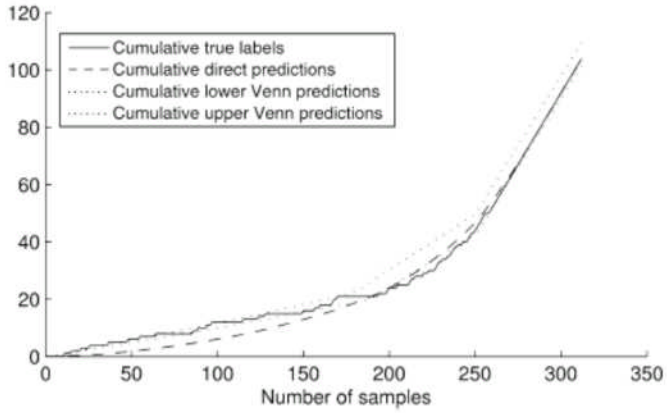


Figure 5 Cumulative Venn and direct predictions for the BC data (all samples, all 79 peaks)

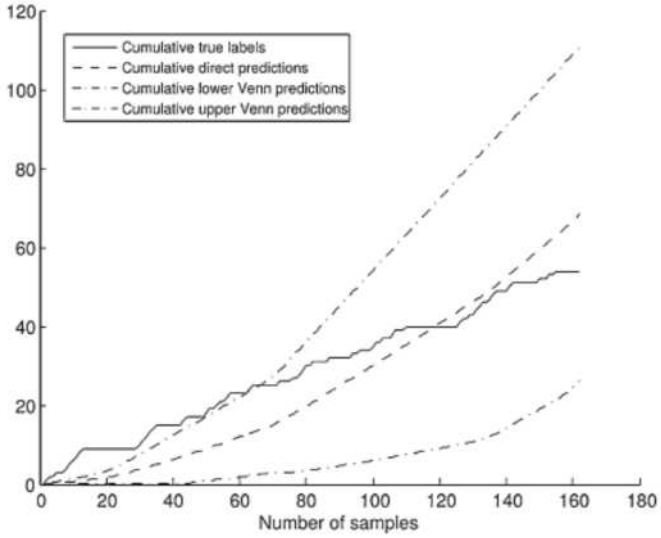


Figure 6 Cumulative Venn and direct predictions for the BC data (all samples, 20 peaks)

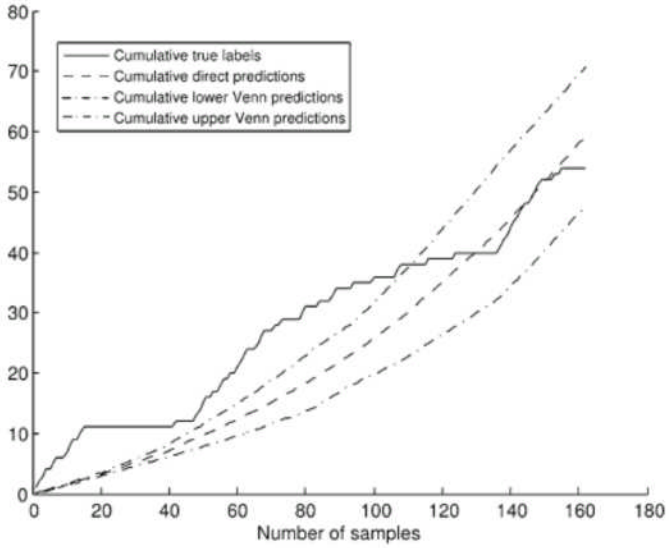


Figure 7 Cumulative Venn and direct predictions for the BC data (all samples, only peak 19)

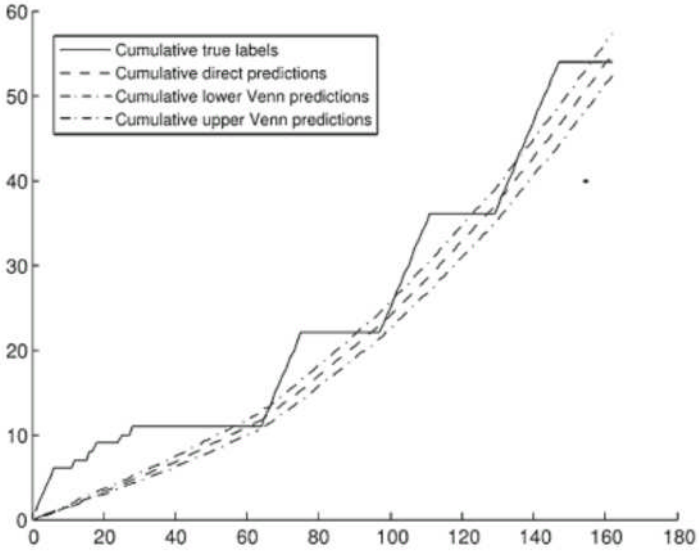
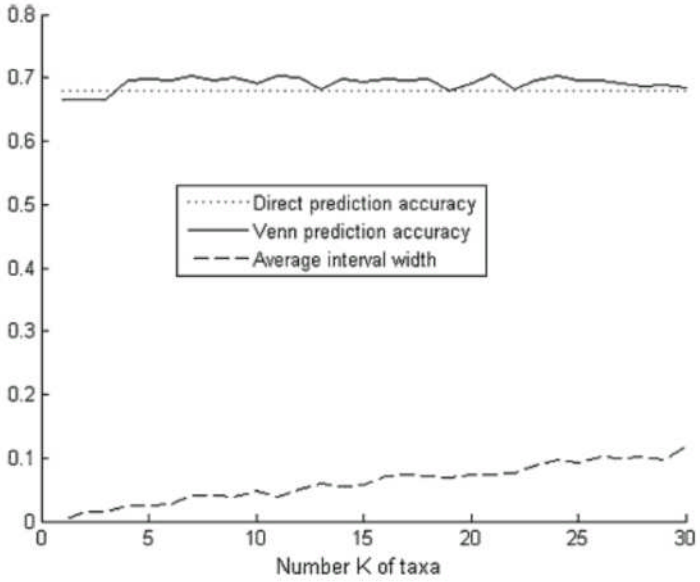


Figure 8 Prediction accuracy and interval width for HD data for different K


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
30
Number K of taxa
Direct prediction accuracy
Venn prediction accuracy
Average interval width