Multiprobabilistic Venn Predictors with Logistic Regression

Ilia Nouretdinov[1,*], Dmitry Devetyarov[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[2], Volodya Vovk[1], 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
[2] EGA Institute for Women's Health, University College London
[3] BioCentre and Department of Chemistry, University of Reading
[*] Address correspondence to I.Nouretdinov: ilia@cs.rhul.ac.uk.

Abstract. This paper describes the methodology of providing multi-probability predictions for proteomic mass spectrometry data. The methodology is based on a newly developed machine learning framework called Venn machines. They allow us to output a valid probability interval. We apply this methodology to mass spectrometry data sets in order to predict the diagnosis of heart disease and early diagnoses of ovarian cancer. The experiments show that probability intervals are valid and narrow. In addition, probability intervals were compared with the output of a corresponding probability predictor.

1 Introduction

Prediction of heart disease (HD) 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 accurate this prediction is. In this work we describe the methodology of proteomics mass spectrometry data analysis based on hedging predictions that is how strongly we believe in this prediction [3].

The framework of Venn machines was introduced in [5] 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 to each patient's diagnosis. Secondly, the only statistical assumption which is used is the exchangeability assumption which can be satisfied when the data sets are in random order.

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 hypothesis about the new label, we classify the new object into one of the categories, and then use frequencies of labels in the chosen category as predictable 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 [6] and SVM in [7]). 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.

The methodology is designed for the analysis of proteomic mass spectrometry data collected in the UKCTOCS trial (for more information see `www.ukctocs.org.uk`).

2 Venn Machines

Consider a training set consisting of object, x_i, label, y_i, pairs: $(x_1, y_1), \ldots, (x_n, y_n)$. To predict a label y_{n+1} for a new object $x_{n+1} = x_{\text{new}}$, we check different hypotheses $y_{n+1} = y$ each time including the pair $(x_{n+1}, y_{n+1}) = (x_{\text{new}}, y)$ into the set.

The idea of Venn machines is based on a *taxonomy function* $A_n, n \in \mathbb{N}$, which classifies the relation between an example and the set of the 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 to each example (x_i, y_i) its 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 conventional way of using Venn ideas was as follows. Categories 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'. Then empirical probabilities of an object within category τ to have a label y are found as

$$P_\tau(y') = N_\tau(y')/N_\tau. \tag{1}$$

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 τ^*. Then the empirical probabilities $P_{\tau^*}(y')$ are considered as probabilities of the object x_{n+1} to have a label y'. The idea of conformal predictors [5] allows us to construct several probability distributions (multi probability

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_{τ^*}, $N_{\tau^*}(y')$ and empirical probability distribution

$$P_{\tau^*(x_{n+1},y)}(y') = N_{\tau^*}(y')/N_{\tau^*}. \tag{2}$$

This distribution depends implicitly on the object x_{n+1} and its hypothetical label y. Trying all possible hypothesises of the label y_{n+1} being equal to y, we obtain a set of distributions $P_y(y') = P_{\tau^*(x_{n+1},y)}(y')$ 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_{τ^*} 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

$$[p^-_{\text{new}}, p^+_{\text{new}}] = [\min\{p_0(1), p_1(1)\}, \max\{p_0(1), p_1(1)\}] \ , \tag{3}$$

which is an estimation 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 percentage of cases between these examples should be between θ and 1 as well.

A Venn predictor is entirely defined by its taxonomy. In the next section we describe a taxonomy based on the logistic regression.

3 Logistic Regression

Logistic regression outputs 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 $P\{y_i = 1\} = 1/\left(1 + e^{-<x_i,b>}\right)$. The optimization goal for logistic regression is:

$$\sum_{i=1}^{n} \log\left(1 + e^{(-1)^{y_i}\langle x_i, b\rangle}\right) + a\,\langle b, b\rangle \to \min_b . \tag{4}$$

This formula is based on the maximum likelihood estimation for Logistic regression, with an added regularisation term $a\,\langle b, b\rangle$ 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 a 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:

$$p_{\text{new}} = \frac{1}{1 + e^{-\langle x_{\text{new}}, \hat{b} \rangle}} \ , \tag{5}$$

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 in Venn machine as an underlying algorithm. As earlier the aim is to predict labels y_i which are equal to 0 for the controls and 1 for the cases, by objects x_i — vectors of features, which are intensities of the most frequent peaks in the logarithm scale.

The probabilistic method of logistic regression allows us to create a new type of taxonomy. The *logistic taxonomy* $\tau_i, i = 1, \ldots, n+1$ 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. Each of them will contain approximately equal number of examples. We use $K = 5$ in this research (the dependence on the parameter K is discussed in Appendix). Let $L_0 = 0$ and $L_1, L_2, \ldots, L_K$ be the integers closest to $(n+1)/K, 2(n+1)/K, 3(n+1)/K, \ldots, n+1$. The category τ_i is then defined accordingly to the intervals formed by division points $p'_{L_0}, p'_{L_1}, \ldots, p'_{L_K}$ where value p_i falls: τ_i is the number $j \in \{1, \ldots, K\}$ such that $p'_{L_{j-1}} < p_i \leq p'_{L_j}$.

4 Mass Spectrometry Data

We would like to develop a methodology of providing multiprobability predictions for proteomic mass spectrometry data. Hence, we have to take into account the format of the data and peculiarities of the problem.

Mass spectra plots 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 moving window; baseline subtraction; normalization to make sure that the total

amounts of ions across different samples are the same. 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. Finally, the intensities of identified peaks are calculated. Detailed description of the preprocessing of mass spectrometry samples can be found in [2].

Thus, the data we apply in our methodology is represented as intensities of identified peaks. The peaks are usually sorted by their frequency: the more examples a peak is presented, the higher the rank of the 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 that 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.

The total number of samples is: 561 in HD data set (187 cases and 374 controls); 312 in OC data set (104 cases and 208 controls). 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 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.

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'. Each sample in OC data set is also assigned a level C of the biomarker CA125. We will not be trying to predict this time, but we need it to form cross-validation sets for the early diagnostics.

5 Results and Discussion

To demonstrate how the proposed methodology works in practice, we applied the designed algorithms to HD and OC proteomic data sets.

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.

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: of the most frequent peaks, CA125 value (for OC dataset), additional dimension constantly equal to 1.

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 x_i in the leave-one-out mode.

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	0.510
4	0	0.143–0.259	0.371
5	0	0.616–0.634	0.622

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

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 $[p^-_{\text{new}}, p^+_{\text{new}}]$. For example, Venn machines output prediction intervals [0.313, 0.321] and [0.616, 0.616] for probabilities that examples 1 and 2 are cases ($y = 1$). As prediction interval indicate, the correct labels for example 1 and 2 are 0 and 1, respectively. The table also includes predictions p_{new} 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.

First, we would like to demonstrate validity of Venn predictions: true probabilities of label distribution are covered or almost covered by the interval between lower and upper Venn prediction. Since we do not know true probabilities of label distribution, we compare empirical probabilities, that is, mean true labels with the mean 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 $p^-_{\text{new}}, p^+_{\text{new}}$ (two dot-dashed lines) and (3) cumulative direct predictions p_{new} (a dashed line). 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, Figure 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, direct predictions can be misleading (the cancer

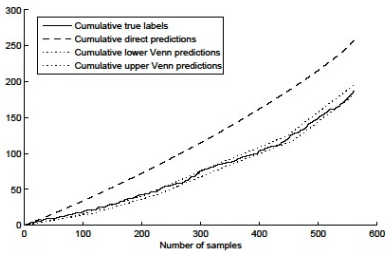

Fig. 1. Cumulative Venn and direct predictions for the HD data

probability is oversetimated, while Venn predictions cover true labels: an average true label lies between average low and upper bounds given by the Venn Machine.

It can be said that both algorithms relied on the assumption of the mechanism generating the data — the 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.

Even though the 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 bare in mind that when we force Venn machines to output a single prediction, they lose their theoretically proven property of validity. In this section we examine the accuracy of forced predictions when we are not able to use advantages of multiprobability predictions.

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 $p^{+}_{\text{new}} + p^{-}_{\text{new}} > 1$ for Venn prediction. This will also allow us to compare accuracy of Venn predictions with the direct predictions.

Given that the aim 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: the accuracy of the proposed methodology on samples in different time slots of the fixed interval (6 months).

Table 2 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)

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%

Table 2. Dynamics of Venn machine and logistic regression performance on the OC dataset

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.

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%.

Personal ID	Months in advance	Prediction interval
29	13	0.22–0.39
	10	0.59–0.71
	4	0.88–0.94
39	10	0.53–0.71
	4	0.44–0.94
	2	0.96–1.00
	1	0.97–1.00

Table 3. Dynamics of prediction intervals output by Venn machines for measurements taken from the same OC case

Table 3 shows the dynamics of prediction intervals output by Venn machines for samples 29 and 39. 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. This also means that the logistic regression as the underlying model becomes more adequate when the time to diagnostics is closer. Second,

the interval is moving towards 1. This implies that we have more trust in our prediction and the prediction is indeed correct.

6 Conclusion

This paper introduced the methodology of hedging predictions for proteomic mass spectrometry data. We applied the described methodology to the MALDI-TOF 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.

Acknowledgement

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.

References

1. Dawid, A.P., Probability Forecasting. Encyclopedia of Statistical Sciences. Wiley, New York. Vol. 7, 210–218 (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. Gammerman, A., Vovk, V. Hedging predictions in machine learning. Computer Journal. Vol. 50, Num. 2, 151–163 (2007).
4. Menon, U., Skates, S.J., Lewis, S. Rosenthal, A.N., Rufford, B., Sibley, K., Macdonald, N., Dawnay, A., Jeyarajah, A., Bast-Jr., R.C., Oram, D., Jacobs, I.J. Prospective study using the risk of ovarian cancer algorithm to screen for ovarian cancer. Journal of Clinical Oncology. Vol.23, 7919–7926 (2005).

5. Vovk, V., Gammerman, A., Shafer, G. Algorithmic learning in a random world. Springer, New York (2005).
6. Papadopoulos, H. Reliable Probabilistic Prediction for Medical Decision Support. In Proceedings of the 7th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI 2011), IFIP AICT 364, pp. 265–274. Springer (2011). doi: 10.1007/978-3-642-23960-1_32.
7. Zhou, C., Nouretdinov, I., Luo, Z., Adamskiy, D., Coldham, N., Gammerman, A. A Comparison of Venn Machine with Platt's Method in Probabilistic Outputs. 12th INNS EANN-SIG International Conference, EANN 2011 and 7th IFIP WG 12.5 International Conference, Artificial Intelligence Applications and Innovations, Corfu, Greece, September 15-18, 2011, Proceedings, Part II. IFIP AICT 364, p.483 (2011).
8. Tiss, A.; Timms, J.F.; Smith, C.; Devetyarov, D.; Gentry-Maharaj, A.; Camuzeaux, S.; Burford, B.; Nouretdinov, I.; Ford, J.; Luo, Z.; Jacobs, I.; Menon, U.; Gammerman, A.; Cramer, R. Highly accurate detection of ovarian cancer using CA125 but limited improvement with serum MALDI-TOF MS profiling. Internation Journal of Gynecol. Cancer, 2010, 20, pp.1518–1524.
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. The Computer Journal (2008).

Appendix. Dependence on the taxonomy parameter

In this work we used $K = 5$. On Figure 2 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.

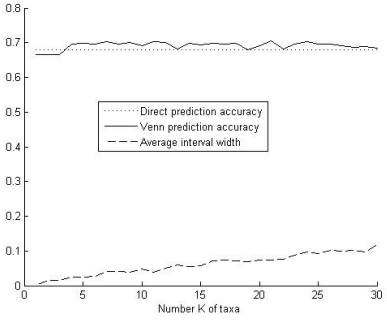

Fig. 2. Prediction accuracy and interval width for HD data for different K.