Mondrian Confidence Machine

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, Alex Gammerman

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project

Working Paper #4

March 28, 2003

Project web site:
http://vovk.net/kp

Abstract

Mondrian Confidence Machine (MCM) is an on-line prediction algorithm that, given a split of all examples into a finite number of types k and for each type a significance level δ_k, outputs as its prediction the set of labels deemed possible at the level δ_k. MCM includes as special cases Transductive Confidence Machine (TCM) and Inductive Confidence Machine (ICM) and is designed to take care of such issues as different risks of false positive and false negative predictions, conditional inference, and a slow teacher. In this paper we generalize known results about TCM and ICM showing that each MCM is type-wise well-calibrated, in the sense that predictions at significance levels δ_k will be wrong with relative frequency at most δ_k for each type k in the long run. Our experimental results show advantages of MCM over the previously known algorithms.

Contents

1 Introduction

We are interested in the problem of "predicting with confidence": instead of just point predictions output by the majority of machine-learning algorithms, we would to have some indication of how likely different labels are. In this paper (as in [4] but unlike [5]) we formalize this problem as that of computing "predictive regions" (sets of labels). In the simplest case, we are given a *significance level* $\delta > 0$ (the probability of error we are prepared to tolerate) and the goal is to compute predictive regions, ideally consisting of just one label, containing the true label with probability $1-\delta$. It was found recently [4] that Transductive Confidence Machine (first introduced in [5]) when applied in the on-line fashion always achieves this goal provided the examples are independent and identically distributed (i.i.d.). This is illustrated in §5.2.

In this paper we consider a slightly different problem. All possible examples are split into several types k (e.g., different types can correspond to different labels, or kinds of objects, or just be determined by the ordinal number of the example). A prediction algorithm takes as input a set of significance levels δ_k, one for each type k, and for each new object outputs as its prediction a predictive region. There are two natural desiderata for such algorithms:

- they should be *type-wise well-calibrated*, in the sense that in the long run the predictions for examples of type k are wrong with relative frequency (at most) δ_k;
- if the first desideratum is satisfied, they should *perform* well, in the sense that the number of *uncertain* (containing more than one label) predictions should be as small as possible or, if the number of uncertain predictions cannot be further improved, the number of empty predictions is as large as possible.

This paper constructs what we call a "Mondrian Confidence Machine" (MCM), and shows (in §3), without using any assumptions beyond i.i.d. (the standard assumption saying that the examples are generated independently from the same distribution), that it is well-calibrated in a strong non-asymptotic sense: the conditional probability of error given that the current type is k and given all the preceding types and errors is always δ_k. (This is the type-wise version of a result proven in [4].) In this paper we do not

deal formally with the second desideratum, but we will see in §5 that MCM produces reasonable results on benchmark data sets.

MCM is a generalization of TCM and ICM; it solves the following three practical problems.

- The problem of "asymmetric classification" (§4.3). MCM allows different significance levels to be specified for each possible classification of an object. This might be useful in, e.g., distinguishing useful messages and spam in the problem of e-mail filtering: classifying a useful message as spam is a more serious error than vice versa.

- The problem of conditional inference (raised by Cox [1] and treated in §§4.3–4.4). Even in the situation where there is only one significance level, we would often like our predictions to be well-calibrated within each class and not just globally. Alternatively, we might want our predictions to be well-calibrated within the set of examples with a particular attribute.

- The problem of a slow teacher, who discloses the true label not immediately but only after some delay (dealt with in §4.5 and, at a much deeper level, in [3]).

In §5 we demonstrate that the second problem can indeed be real with experiments on real-life data sets, the USPS handwritten digits data set and the thyroid medical data set.

2 Type-wise region predictors

Our basic protocol is as follows. Nature outputs pairs

$$(x_1, y_1), (x_2, y_2), \dots \tag{1}$$

called *examples*. Each example (x_i, y_i) consists of an *object* x_i and its *label* (or *class*, or *classification*) y_i; e.g., the objects can be hand-written digits and y_i their classifications (numbers from 0 to 9). The objects are elements of a measurable space $\mathbf{X}$ called the *object space* and the labels are elements of a measurable space $\mathbf{Y}$ called the *label space*. We will use the notation $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ for the *example space*; therefore, the infinite data sequence (1) will be an element of the measurable space $\mathbf{Z}^\infty$. We assume that the data

sequence (1) is output according to P^∞ for some probability distribution P in $\mathbf{Z}$, but no further assumptions will be made.

We are given a division of the Cartesian product $\{1,2,\dots\}\times\mathbf{Z}$ into *types*: a function

$$\kappa : \{1,2,\dots\}\times\mathbf{Z}\to K$$

maps each pair (n,z) (z is an example and n will be, in our applications, the ordinal number of this example in the data sequence $z_1,z_2,\dots$) to its type; K is the finite set of possible types. It is required that the elements $\kappa^{-1}(k)$ of each type k form a rectangle $A\times B$, for some $A\subseteq\{1,2,\dots\}$ and $B\subseteq\mathbf{Z}$. The function κ will be fixed in most of this paper; it will be called the *taxonomy*.

We are interested in algorithms for predicting, at every trial n, the label y_n given the object x_n and all the previous examples, from (x_1,y_1) to (x_{n-1},y_{n-1}). Since we are interested in prediction with confidence, our algorithms are given an extra input $\{\delta_k\in(0,1) : k\in K\}$, a *significance level* δ_k for each class k; the complementary value $1-\delta_k$ is called the *confidence level*. Formally, we define a *type-wise region predictor* to be a function

$$\Gamma : \mathbf{Z}^*\times\mathbf{X}\times(0,1)^K\to 2^{\mathbf{Y}} \tag{2}$$

($(0,1)^K$ is the set of all functions of the type $K\to(0,1)$ and $2^{\mathbf{Y}}$ is the set of all subsets of $\mathbf{Y}$; the argument $\delta : K\to(0,1)$ will be written as subindex) which, for every significance levels $\delta_1\geq\delta_2$ (such inequalities are always understood component-wise), every positive integer n, and every *incomplete data sequence*

$$x_1,y_1,\dots,x_{n-1},y_{n-1},x_n \tag{3}$$

(we often ignore unnecessary parentheses, such as those around (x_i,y_i)) satisfies

$$\Gamma_{\delta_1}(x_1,y_1,\dots,x_{n-1},y_{n-1},x_n)\subseteq\Gamma_{\delta_2}(x_1,y_1,\dots,x_{n-1},y_{n-1},x_n). \tag{4}$$

Intuitively, given the incomplete data sequence (3) and a significance level δ, the region predictor Γ predicts that

$$y_n\in\Gamma_\delta(x_1,y_1,\dots,x_{n-1},y_{n-1},x_n),$$

and the smaller δ the more emphatic the prediction; condition (4) is a natural requirement of consistency.

For any infinite data sequence

$$\omega = (x_1, y_1, x_2, y_2, \dots), \tag{5}$$

significance level $\delta : K \to (0,1)$ and positive integer n, we define the number of errors that Γ makes on examples of type k at the significance level δ on the sequence ω during the first n trials to be

$$\begin{aligned}\operatorname{Err}_n^k(\Gamma_\delta, \omega) := \#\big\{i = 1, \dots, n : \\ \kappa(z_i) = k \;\&\; y_i \notin \Gamma_\delta\left(x_1, y_1, \dots, x_{i-1}, y_{i-1}, x_i\right)\big\},\end{aligned}$$

where $\#B$ (or $|B|$) stands for the size of the set B. Similarly, the number of uncertain predictions is given by

$$\begin{aligned}\operatorname{Unc}_n^k(\Gamma_\delta, \omega) := \#\big\{i = 1, \dots, n : \\ \kappa(z_i) = k \;\&\; \left|\Gamma_\delta\left(x_1, y_1, \dots, x_{i-1}, y_{i-1}, x_i\right)\right| > 1\big\},\end{aligned}$$

and the number of empty predictions is given by

$$\begin{aligned}\operatorname{Emp}_n^k(\Gamma_\delta, \omega) := \#\big\{i = 1, \dots, n : \\ \kappa(z_i) = k \;\&\; \Gamma_\delta\left(x_1, y_1, \dots, x_{i-1}, y_{i-1}, x_i\right) = \emptyset\big\}.\end{aligned}$$

The number of times Nature output an example of type k (i.e., the number of times $\kappa(i, z_i) = k$) before and including trial n is

$$\operatorname{Num}_n^k := \#\big\{i = 1, \dots, n : \kappa(z_i) = k\big\}.$$

Sometimes we will also need the individual prediction results

$$\operatorname{err}_n^k(\Gamma_\delta, \omega) := \begin{cases} 1 & \text{if } y_n \notin \Gamma_\delta \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\Gamma_\delta := \Gamma_\delta\left(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n\right).$$

In this paper we usually consider *randomized region predictors*, which depend, additionally, on a random number $\tau_n \in [0,1]$.

3 Mondrian Confidence Machine

Mondrian Confidence Machine (MCM) is a way of transition from what we call a "Mondrian strangeness measure" to a region predictor. A family of measurable functions $\{A_n : n \in \mathbb{N}\}$, where $\mathbb{N}$ is the set of all positive integers, $A_n : \mathbf{Z}^n \to \mathbb{R}^n$ for all $n \in \mathbb{N}$, and $\mathbb{R}$ is the set of all real numbers (equipped with the Borel σ-algebra), is called a *Mondrian strangeness measure* if, for any $n \in \mathbb{N}$, any $(z_1, \dots, z_n) \in \mathbf{Z}^n$, any permutation π of $\{1, \dots, n\}$ that does not change the types of the examples in $(z_1, \dots, z_n)$ (in the sense that $\kappa(z_{\pi(i)}) = \kappa(z_i)$ for $i = 1, \dots, n$), and any $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$,

$$(\alpha_1, \dots, \alpha_n) = A_n(z_1, \dots, z_n) \implies (\alpha_{\pi(1)}, \dots, \alpha_{\pi(n)}) = A_n(z_{\pi(1)}, \dots, z_{\pi(n)}). \quad (6)$$

In other words,

$$A_n : (z_1, \dots, z_n) \mapsto (\alpha_1, \dots, \alpha_n) \quad (7)$$

is called a Mondrian strangeness measure if every α_i is determined by the example z_i, the sequence $\kappa(z_1), \dots, \kappa(z_n)$ and, for each type $k \in K$, the multiset $\{\!|z_i : \kappa(z_i) = k|\!\}$,

The *MCM* associated with the Mondrian strangeness measure A_n is the following randomized region predictor:

$$\Gamma_\delta\left(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n, \tau_n\right) \quad (8)$$

is defined to be the set of all labels $y \in \mathbf{Y}$ such that

$$\frac{\#\{j : \alpha_{i_j} > \alpha_n\} + \tau_n \#\{j : \alpha_{i_j} = \alpha_n\}}{m} > \delta, \quad (9)$$

where j ranges over $1, \dots, m$ and $i_1, \dots, i_m$ is the set

$$\{i = 1, \dots, n : \kappa(i, z_i) = \kappa(n, z_n)\}$$

sorted in the ascending order (in particular, $i_m = n$) and

$$(\alpha_1, \dots, \alpha_n) := A_n((x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y)). \quad (10)$$

In general, an *MCM* is the MCM associated with some Mondrian strangeness measure.

Theorem 1 *For any MCM Γ, any significance level $\delta : K \to (0,1)$, and any probability distribution P in $\mathbf{Z}$, the image of $(P \times U)^\infty$ under the mapping*

$$(\omega \in \mathbf{Z}^\infty, \tau \in [0,1]^\infty) \mapsto (k_n, e_n)_{n=1}^\infty \\ := (\kappa(n, z_n), \mathrm{err}_n(\Gamma_\delta, \omega, \tau))_{n=1}^\infty$$

is a probability distribution satisfying the following properties: k_n are independent random elements taking values in K; the conditional probability that $e_n = 1$ given $k_1, e_1, \ldots, k_{n-1}, e_{n-1}, k_n$ is always δ_{k_n}.

Theorem 1 also implies

Corollary 1 *Every MCM Γ is* precisely type-wise well-calibrated *in the sense that, for each class $k \in K$,*

$$\mathrm{Num}_\infty^k = \infty \Longrightarrow \lim_{n\to\infty} \frac{\mathrm{Err}_n^k(\Gamma_{1-\delta}, \omega, \tau)}{\mathrm{Num}_n^k} = \delta_k \tag{11}$$

for $(P \times U)^\infty$-almost all $\omega \in \mathbf{Z}^\infty$ and $\tau \in [0,1]^\infty$.

Corollary 1 uses the natural notation

$$\mathrm{Num}_\infty^k := \lim_{n\to\infty} \mathrm{Num}_n^k .$$

The adverb "precisely" in its statement indicates that (11) is an equality; we will say that a region predictor is *type-wise well-calibrated* if it satisfies (11) with "=" replaced by "$\le$" and "lim" replaced by "lim sup".

4 Special cases

The only important aspect of a taxonomy κ is the equivalence relation it induces ((n', z') and (n'', z'') are κ-equivalent if $\kappa(n', z') = \kappa(n'', z'')$) and not the chosen labels $\kappa(n, z)$ for the equivalence classes. There are many different ways to split the rectangle $\mathbb{N} \times \mathbf{Z}$ into rectangles (as evidenced by Piet Mondrian's numerous compositions; see Figure 1).

In this section we will consider several simple special cases.

First we introduce a natural partial order on taxonomies. We say that a taxonomy κ_1 is *finer* than another taxonomy κ_2 if, for all pairs (n', z') and (n'', z''),

$$\kappa_1(n', z') = \kappa_1(n'', z'') \Longrightarrow \kappa_2(n', z') = \kappa_2(n'', z'').$$

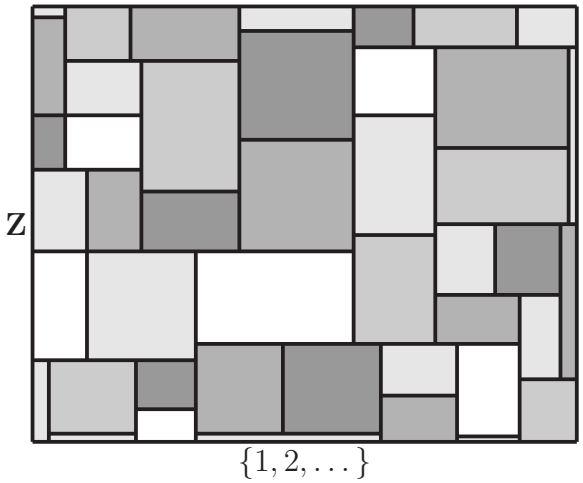

Figure 1: A random taxonomy (cf. Composition with Color Planes and Gray Lines 1 by Piet Mondrian, 1918)

We will say that κ_1 and κ_2 are *equivalent* if each of them is finer than the other; we will sometimes identify equivalent taxonomies. TCM corresponds to the crudest (i.e., constant, see Figure 2) taxonomy. Since we are only interested in taxonomies with a finite number of types, we have the following proposition.

Proposition 1 *Let a taxonomy κ_1 be finer than a taxonomy κ_2. If a region predictor is (precisely) well-calibrated w.r. to κ_1, it is (precisely) well-calibrated w.r. to κ_2.*

A taxonomy effectively partitions the example space $\mathbf{Z}$ into rectangular groups. By considering different partitions we can construct specialized versions of the MCM to cope with different drawbacks of previously known algorithms. Essentially the only difference between the versions of the MCM is the method of calculating the p-values. For example, calculating the TCM's p-values in our on-line protocol we compare the strangeness measure of a new example against the strangeness measures of all examples observed up to that point. In contrast, finer models of MCM compare the strangeness measure of a new example with the previously observed examples of the same type.

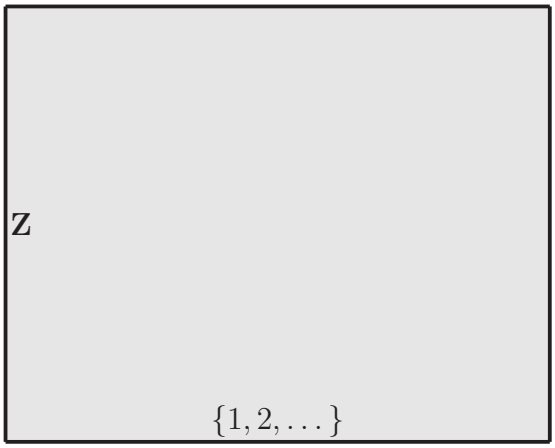

Figure 2: TCM

4.1 Transductive Confidence Machine

TCM is an MCM corresponding to the least fine taxonomy shown in Figure 2. It is shown in [4] that TCM is well-calibrated; this is a special case of Theorem 1.

4.2 Inductive Confidence Machine

ICM is also a special case of MCM; for a full description, see [4]. The corresponding taxonomy is shown in Figure 3. Again the result of [4] that ICM is well-calibrated is a special case of Theorem 1.

4.3 Class-conditional inference and asymmetric classification

In this case the types are determined by labels; we would like to have guarantees about error frequency for each individual label, and perhaps also have different significance levels for different labels. The corresponding taxonomy is shown in Figure 4, where it is assumed that $\mathbf{Y} = \{y^{(1)}, \dots, y^{(L)}\}$. In principle, each class k can have its own significance level δ_k (*asymmetric classification*), but we are also interested in the case where all δ_k coincide (*class-conditional inference*).

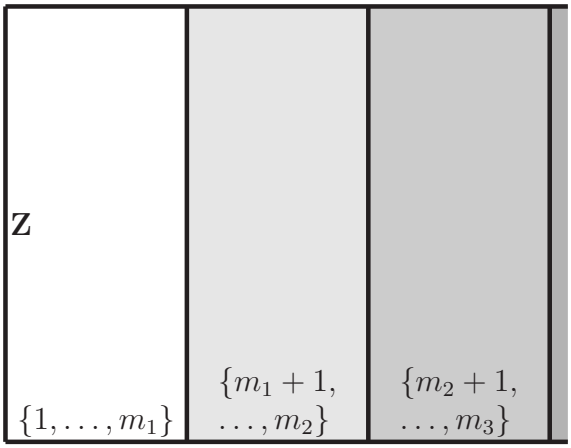

Figure 3: ICM

4.4 Attribute-conditional inference

The notion of MCM allows the user to respect the conditionality principle ([1], [2], §2.3) if he chooses to do so. Consider the following standard example (slightly modified) due to Cox [1]. Suppose we have two instruments for measuring an unknown bit; at each trial one instrument is used once, and the instrument to use is chosen at random (tossing a fair coin). Instrument 1 is more accurate, with the probability of mistake equal to 1%, whereas the probability of mistake for instrument 2 is 5%. Formally, each object is a pair $x = (i, b)$, where $i \in \{1, 2\}$ is the instrument used and $b \in \{0, 1\}$ is the result of the measurement; the label $y \in \{0, 1\}$ is the true bit.

It is clear that, asymptotically, at level 99.5% the optimal TCM will predict objects $(1, \ldots)$ with certainty and will not predict objects $(2, \ldots)$ at all (in the sense that its predictions will be the set $\{0, 1\}$ of all labels). Therefore, it will be well-calibrated. At level 97% the optimal TCM will asymptotically predict all objects with certainty; so it will also be well-calibrated.

In both cases conditional validity is problematic (as argued by Cox); it does not prevent, however, the predictions from being well-calibrated. But the situation becomes even worse if we want to have two different significance levels for objects $(1, \ldots)$ and $(2, \ldots)$: if we take 0.5% for $(1, \ldots)$ and 3% for $(2, \ldots)$, type-wise well-calibratedness is lost.

The taxonomy for Cox's example is shown in Figure 5, where "Instrument 1" stands for the set of examples $((1, \ldots), \ldots)$ and "Instrument 2"

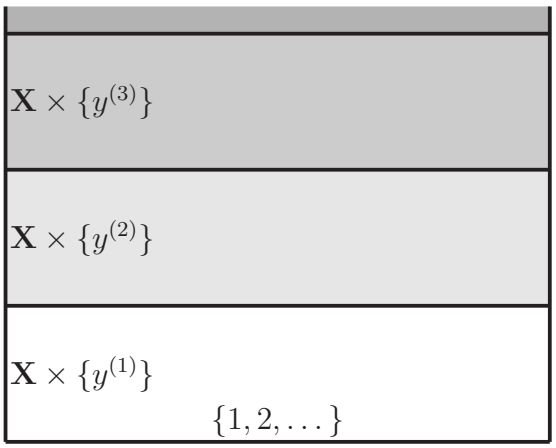

Figure 4: Asymmetric inference

stands for the set of examples $((2, \dots), \dots)$.

4.5 Slow teacher

The usual scenario of on-line learning is unrealistic in that in practice we can expect a delay between giving a prediction and finding out the true label. (And, indeed, if there is no delay in obtaining the true label, there is no need in the prediction.) In this subsection we will show that, if the delay is constant, a simple MCM will be well-calibrated and asymptotically optimal in this situation of a"slow teacher".

Let D be the delay. Define $\kappa(n, z) := n \mod D$ (this is illustrated in Figure 6 for $D = 3$). Theorem 1 implies that this MCM is well-calibrated. For more detail and a more efficient algorithm, see [3].

5 Experimental results

As we mentioned in §1, MCM makes it possible to deal with the challenges of asymmetric classification, conditional inference, and a slow teacher. It is obvious that asymmetric classification and a slow teacher are important issues not addressed by the approaches described in literature. In this section we will describe experimental results demonstrating conditional properties of TCM (and, even more so, ICM) that in many applications may be interpreted

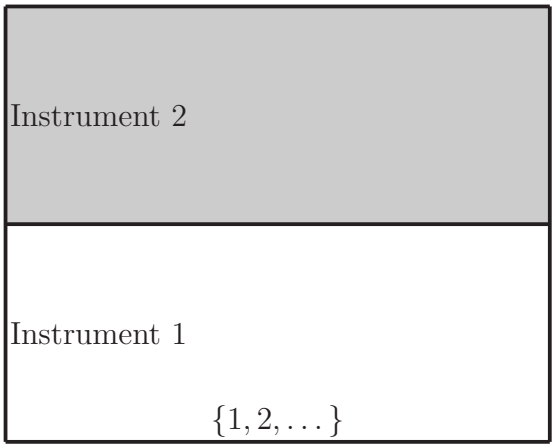

Figure 5: Cox's example

as failures. As the strangeness measure we will always use the 1-Nearest Neighbor measure

$$\alpha_i := \frac{\min_{j \neq i: y_j = y_i} d(x_i, x_j)}{\min_{j \neq i: y_j \neq y_i} d(x_i, x_j)},$$

where d is the Euclidean distance (i.e., an object is considered strange if it is in the middle of objects labeled in a different way and is far from objects labeled in the same way); the objects will be vectors in a Euclidean space.

5.1 Data sets

We chose two high-dimensional real-life data sets from very different problem domains, image recognition and medical diagnostics. In our experiments we combined the training and test sets, and randomly permuted the examples, to make sure the i.i.d. (or at least exchangeability) assumption holds. We use the performance measure mentioned earlier, counting the number of erroneous, uncertain and empty predictions. In Figures 7–13 we will plot Err_n, $n\delta$, Unc_n and Emp_n against n for different data sets and confidence levels $1 - \delta$.

The first data set is the well-known USPS data set of handwritten digits, consisting of scanned 16×16 grayscale images of handwritten numerals 0–9 taken from postcodes gathered by the US Postal Service. There are 256 continuous attributes corresponding to the intensity of each pixel in the image,

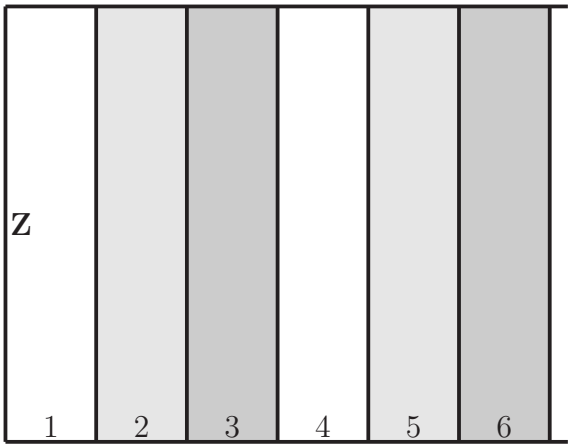

Figure 6: Slow teacher

and ten possible classification labels. With the training and test sets combined there are 9298 examples in total. The data set is reasonably balanced in the proportion of examples for each class.

The other data set is the thyroid disease records, supplied by the Garavan Institute and J. Ross Quinlan. The problem is to determine whether a patient referred to the clinic is hypothyroid. We used the "ann-thyroid" data downloaded from the UCI web site

```
ftp://ftp.ics.uci.edu/pub/
machine-learning-databases/
thyroid-disease/
```

Each record has 21 attributes in total (15 Boolean and 6 continuous) corresponding to various symptoms and measurements taken from each patient. The data set contains 7200 examples in total and is highly unbalanced in its representation of the 3 possible classes corresponding to medical diagnoses (2.30% of examples are in the class "primary hyperthyroid", 5.11% of examples in the class "compensated hyperthyroid", and 92.59% examples in the class "normal").

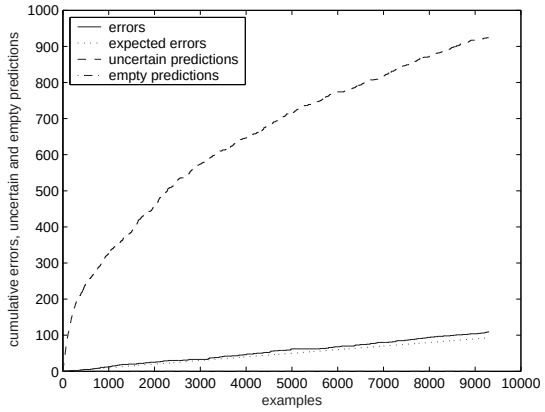

Figure 7: The performance of TCM on the USPS data set at the 99% confidence level.

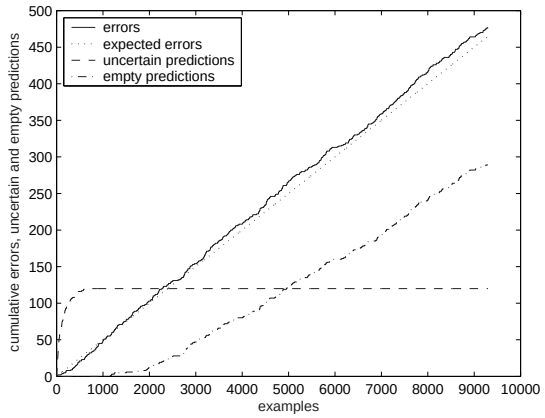

Figure 8: The performance of TCM on the USPS data set at the 95% confidence level.

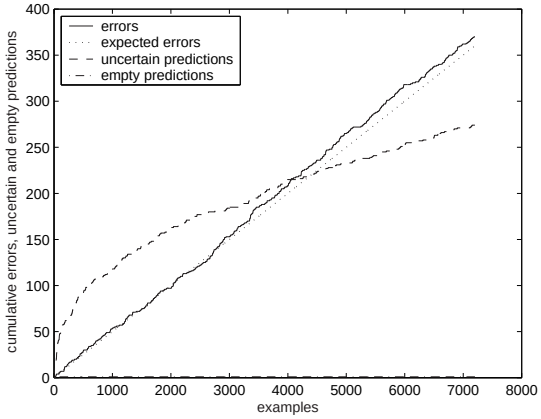

Figure 9: The performance of TCM on the thyroid data set at the 95% confidence level.

5.2 Transductive Confidence Machine

We will be interested in how the MCM's performance compares with the TCM's performance (since ICM is just a computationally efficient modification of TCM, and its performance is usually not as good). Figures 7 and 8 show that TCM is well-calibrated on the USPS data set; since the USPS data set is so clean, we can see that when the confidence level becomes too easy (95%) the algorithm starts producing empty predictions (always leading to an error and serving as a warning that the current example is difficult). Figure 9 shows that the unconditional TCM is well-calibrated on the thyroid data set.

5.3 Class-conditional performance

Figures 10–13 show results of experimenting with the USPS and thyroid data sets with all confidences levels set to 95%. Figure 10 shows that the TCM is not well-calibrated at the 95% confidence level on the USPS "5" digits, giving 11.73% of errors. (Since we randomly permute the data and our algorithms are randomized, all such figures are subject to moderate statistical fluctuations: typical standard deviation is 1%. All reported results are for programs using MATLAB's generator of random numbers with seed 0.) In contrast the class conditional MCM gives 5.31% of errors (Figure 11). Fig-

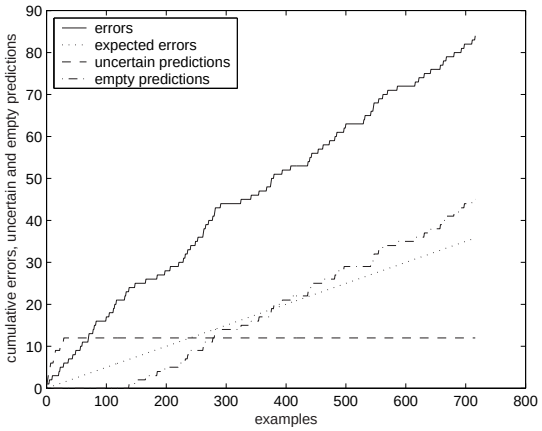

Figure 10: The performance of TCM on the USPS data sets for the "5" digit images at the 95% confidence level.

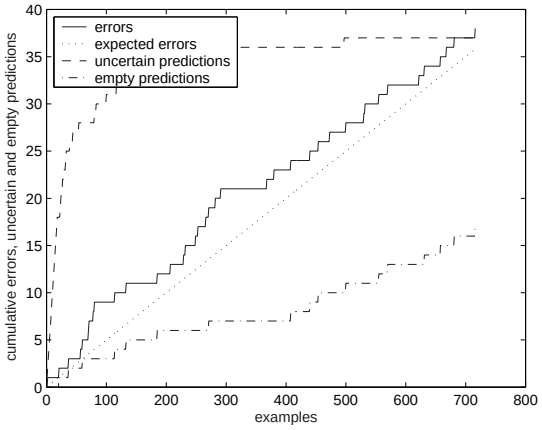

Figure 11: The performance of the class-conditional MCM on the USPS data sets for the "5" digit images at the 95% confidence level.

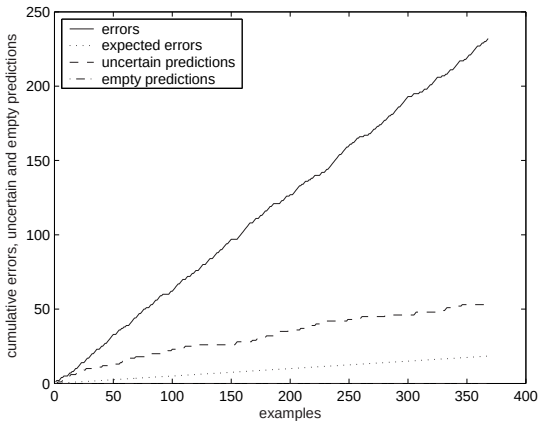

Figure 12: The performance of TCM on the thyroid data set's class "compensated hyperthyroid".

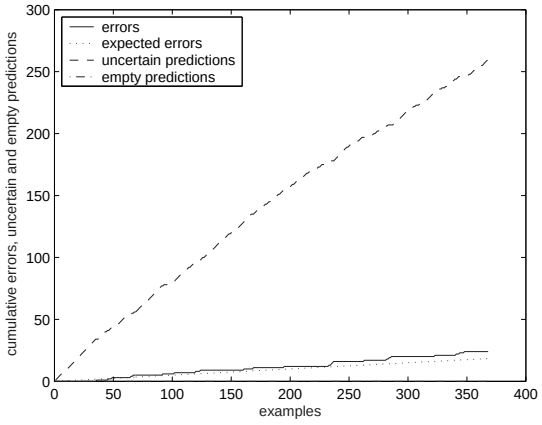

Figure 13: The performance of class-conditional MCM on the thyroid data set's class "compensated hyperthyroid".

Table 1: Percentage of errors at the 95% confidence level and the corresponding p-values (upper if the percentage of errors is above 5% and lower if it is below 5%) in experiments on the USPS data set.

class	size	errors	errors (%)	p-value
0	1553	13	0.84	3.35×10^{-20}
1	1269	12	0.95	1.02×10^{-15}
2	929	52	5.60	0.22
3	824	69	8.37	2.87×10^{-5}
4	852	90	10.56	4.29×10^{-11}
5	716	84	11.73	8.68×10^{-13}
6	834	23	2.76	9.24×10^{-4}
7	792	36	4.55	0.31
8	708	67	9.46	6.80×10^{-7}
9	821	31	3.78	0.06

Table 2: Percentage of errors at the 95% confidence level and the corresponding p-values (upper if the percentage of errors is above 5% and lower if it is below 5%) in experiments on the thyroid data set. The classes are: "primary hyperthyroid" (0), "compensated hyperthyroid" (1), "normal" (2).

class	size	errors	errors (%)	p-value
0	166	40	24.10	7.87×10^{-17}
1	368	232	63.04	1.14×10^{-201}
2	6666	98	1.47	6.10×10^{-54}

ures 10 and 11 show that this correction in the errors results in an increased frequency of uncertain predictions; there is also a decrease in the number of empty predictions.

It is natural to expect certain digits to be more easily confused with other digits, and others easier to discriminate. For example with TCM, "0" digits give on average a lower level of observed errors (0.84%) as compared with the expected 5% (and 4.83% as given by MCM).

The results for the thyroid data set are even more extreme. In Tables 1 and 2 we give the p-values computed under the null hypothesis that the probability of error is 5%. (By "upper p-value" we mean the probability that the percentage of errors equals or exceeds the observed one, and similarly for "lower p-values".) Some of the p-values are extremely small; the evidence against the null hypothesis is statistically highly significant.

5.4 Attribute-conditional performance

In our experiments we have not seen as gross failures in the TCM's attribute-conditional performance as those described in the previous subsection. It is natural to test the attribute-conditional performance of TCM on the thyroid data set since the USPS data set does not have any natural attributes to conditions on: all attributes in it are of the same nature (the brightness level of a pixel) and continuous. There are several natural attributes to condition on in the thyroid data set, such as sex, the presence of tumor, etc., but TCM's conditional performance was reasonable (the greatest anomaly was for conditioning on the "tumor" attribute, where TCM's error rate was around 7%).

6 Conclusion

Our results have demonstrated another flexible dimension to the confidence machine framework. This could be especially useful for medical applications allowing the supervisor of the learning task to break prospective patients into groupings of her choosing and to specify individual significance levels for each.

Acknowledgments

This work was partially supported by EPSRC (grant GR/R46670/01), BBSRC (grant 111/BIO14428), and EU (grant IST-1999-10226).

References

[1] David R. Cox. Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29:357–372, 1958.

[2] David R. Cox and David V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.

[3] Daniil Ryabko, Vladimir Vovk, and Alex Gammerman. Online region prediction with real teachers, On-line Compression Modelling project, `http://vovk.net/kp`, Working Paper #5, March 2003.

[4] Vladimir Vovk. On-line Confidence Machines are well-calibrated, On-line Compression Modelling project, `http://vovk.net/kp`, Working Paper #1. In *Proceedings of the Forty Third Annual Symposium on Foundations of Computer Science*, pages 187–196. IEEE Computer Society, 2002.

[5] Vladimir Vovk, Alex Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453, San Francisco, CA, 1999. Morgan Kaufmann.

A Appendix: Proof sketch of Theorem 1

In the statement of Theorem 1 we did not describe explicitly the probability distribution generating the classes $k_n = \kappa(n, z_n)$; it is clear that, for each $k \in K$, $k_n = k$ with probability

$$P\{z \in \mathbf{Z} : \kappa(n, z) = k\}, \tag{12}$$

and in conjunction with the statement of Theorem 1 this completes the description of the probability distribution generating the sequence $k_1, e_1, k_2, e_2, \ldots$. Let Q be the probability distribution in $K \times \{0, 1\}$ according to which a pair $(k, e) \in K \times \{0, 1\}$ is asserted to be generated: first k

is generated with probability (12) and then e is set to 1 with probability δ_k (and to 0 with probability $1-\delta_k$). Our goal is to prove that (k_n, e_n), $n = 1, 2, \dots$, are generated according to Q^∞.

We only explain the basic idea of the proof. To show that $(k_1, e_1, \dots, k_N, e_N)$ is distributed as Q^N (it is easy to get rid of the assumption of a fixed horizon N), we use the standard idea of reversing the time. We can imagine that the sample $(z_1, \dots, z_N)$ is generated in two steps: first, the multiset $\{\!|z_1, \dots, z_N|\!\}$ is generated from some probability distribution (namely, the image of P under the mapping $(z_1, z_2, \dots) \mapsto \{\!|z_1, \dots, z_N|\!\}$), and then the actual sample $(z_1, \dots, z_N)$ is chosen randomly from the set of all orderings of the multiset $\{\!|z_1, \dots, z_N|\!\}$. Already the second step ensures that, conditionally on knowing $\{\!|z_1, \dots, z_N|\!\}$ and k_N (and, therefore, conditionally on knowing k_N alone), the bit e_N is distributed as $B_{\delta_{k_N}}$. Indeed, roughly speaking (i.e., ignoring ties and borderline effects), e_N will be 1 if α_N is among the $\mathrm{Num}_N^{k_N}\,\delta$ largest α_i of its type, and the probability of this is δ_{k_N} since all type-preserving permutations are equiprobable; when z_N is disclosed, the value e_N will be settled; conditionally on knowing $\{\!|z_1, \dots, z_N|\!\}$, z_N and k_{N-1} (and, therefore, knowing $\{\!|z_1, \dots, z_{N-1}|\!\}$ and k_{N-1}), e_{N-1} will be 1 with probability $\delta_{k_{N-1}}$, and so on.

The details are similar to those in [4].

On-line Compression Modelling Project
Working Papers

1. *On-line confidence machines are well-calibrated*, by Vladimir Vovk, April 2002.

2. *Asymptotic optimality of Transductive Confidence Machine*, by Vladimir Vovk, May 2002.

3. *Universal well-calibrated algorithm for on-line classification*, by Vladimir Vovk, November 2002.

4. *Mondrian Confidence Machine*, by Vladimir Vovk, David Lindsay, Ilia Nouretdinov and Alex Gammerman, March 2003.

5. *Testing exchangeability on-line*, by Vladimir Vovk, Ilia Nouretdinov and Alex Gammerman, February 2003.

6. *Criterion of calibration for Transductive Confidence Machine with limited feedback*, by Ilia Nouretdinov and Vladimir Vovk, April 2003.

7. *Online region prediction with real teachers*, by Daniil Ryabko, Vladimir Vovk and Alex Gammerman, March 2003.

8. *Well-calibrated predictions from on-line compression models*, by Vladimir Vovk, April 2003.