Testing Exchangeability On-Line

Vladimir Vovk, Illia Nouretdinov and Alex Gammerman
Computer Learning Research Center, Royal Holloway, University of London
ICML 2003

Presented by
Shen-Shyang Ho

Content:

1. Problem Setting
2. What is Exchangeability?
3. A Brief History of Martingales.
4. Informal definitions (Properties) of Martingales
5. Testing exchangeability :
   (a) Definitions
   (b) Theorem
   (c) Steps
   (d) Constructing Power Martingale
   (e) Test
6. Experimental Result

Problem Setting

1. Given a data stream $z_1, z_2, \cdots$, we are interested in testing the hypothesis of randomness/exchangeability **on-line**.
2. After observing each new example z_n Learner is required to output a number M_n reflecting the strength of evidence found against the hypothesis.
3. What if the data is **high dimensional**? No research on this!!
4. Usual statistical techniques do not work!!

What is Exchangeability? - (1)

1. When successive examples are chosen randomly from a distribution Q in a fixed space of possible examples Z, the examples are **independent**(in the sense of probability theory) and **all have the distribution** Q. They are **independent and identically distributed** (i.i.d and also the randomness assumption).
2. If the distribution of the sequence of examples $z_1, \cdots, z_n$ under Q is **invariant under any permutation of the indices**, the distribution is **exchangeable**.
3. The probability of the sequences does not change under reordering, i.e. all sequences with the identical occurrence of examples have same probability (i.e. $\frac{1}{n!}$).

What is Exchangeability? - (2)

1. Exchangeability is not equivalent to i.i.d.
2. However, if a sequence is i.i.d, it is exchangeable.
3. The condition of exchangeability is stronger than the assumption of identical distribution of the individual random variables in the sequence.
4. Exchangeability is weaker than the assumption that they are i.i.d .

A Brief History of Martingales - (1)

1. Martingale: A class of betting strategies popular in 18th century.
2. Assume a game in which the gambler wins his stake if a coin comes up heads and loses it if the coin comes up tails.
3. The strategy: the gambler double his bet after every loss, so that the first win would recover all previous losses plus win a profit equal to the original stake (say 1).
4. Since a gambler with **infinite wealth** is guaranteed to eventually flip heads, the martingale betting strategy was seen as a sure thing by those who practiced it.
5. Unfortunately, none of these practitioners in fact possessed infinite wealth, and the exponential growth of the bets would **quickly bankrupt those foolish enough to use the martingale after even a moderately long run of bad luck**.

A Brief History of Martingales - (2)

1. The concept of martingale in probability theory was introduced by Paul Pierre Levy (1935),
2. Much of the original development of the theory was done by Doob (1949, Application of the theory of martingales; 1953, Stochastic Process).
3. Part of the motivation for works on Martingale was to show the **impossibility of successful betting strategies**.

Informal Definition (Properties) of Martingales

A super-martingale which is a sequence of random variables $M_0, M_1, \cdots$ such that for all $n = 0, 1, \cdots, M_n$ is a measurable function of $z_1, z_2, \cdots, z_n$ (with the base case, M_0 a constant) and

$$M_n \geq E(M_{n+1} | M_1, \cdots, M_n)$$

For a martingale

$$M_n = E(M_{n+1} | M_1, \cdots, M_n)$$

In case you are confused later:

- Randomized transducer $\Rightarrow$ Martingale
- Deterministic transducer $\Rightarrow$ Super-martingale

Testing Exchangeability: Definitions

1. Nearest Neighbor (NN) individual strangeness measure:

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

2. SVM individual strangeness measure:
   (a) Lagrange multipliers
   (b) Distance from the hyperplane.

Testing Exchangeability: Definitions

1. A **randomized transducer** is a function f of the type $(Z \times [0,1])^* \to [0,1]$. It is called "transducer" as it can be regard as mapping each input sequence $(z_1, \theta_1, z_2, \theta_2, \cdots)$ in $(Z \times [0,1])^\infty$ into the output sequence $(p_1, p_2, \cdots)$ of "p-values" defined by $p_n = f(z_1, \theta_1, \cdots, z_n, \theta_n)$, $n = 1, 2, \cdots$.

2. Given an individual strangeness measure, for each sequence $(z_1, \theta_1, \cdots, z_n, \theta_n) \in (Z \times [0,1])^*$ define

$$f(z_1, \theta_1, \cdots, z_n, \theta_n) = \frac{\#\{i : \alpha_i > \alpha_n\} + \theta_n \#\{i : \alpha_i = \alpha_n\}}{n}$$

where α_i, $i = 1, 2, \cdots$, are computed from z_i using the individual strangeness measure. Each randomized transducer f that can be obtained in this way is called a **randomized confidence transducer**.

Testing Exchangeability: Definitions and Theorem

A function f is a **randomized E/U-transducer** if the output p-values $p_1, p_2, \cdots$ are always distributed according to the uniform distribution in $[0,1]^\infty$, provided the input examples examples $z_1, z_2, \cdots$ are generated by an exchangeable probability distribution in Z^∞.

Theorem: Each randomized confidence transducer is a randomized E/U transducer.

Testing Exchangeability: Steps

Testing exchangeability consists of two main steps:

1. Extract randomized (deterministic) E/U transducers from individual strangeness measure.
2. Construct the randomized (exchangeability) power (super-)martingale from the p-values (using Section 9, Vovk, 1993).

A deterministic transducer has similar definitions and results.

Testing Exchangeability: Constructing power martingale

Define a family of martingales, indexed by $\epsilon \in [0, 1]$, called the **randomized power martingale** by:

$$M_n^{(\epsilon)} = \prod_{i=1}^{n} \left(\epsilon p_i^{\epsilon-1}\right)$$

where p_n are the p-values output by a randomized confidence transducer, will be a non-negative **randomized exchangeability martingale** with initial value 1.

To eliminate the dependency on ϵ, use

$$M_n = \int_0^1 M_n^{(\epsilon)} d\epsilon$$

called the **simple mixture** of $M_n^{(\epsilon)}$.

Testing Exchangeability: Constructing power martingale

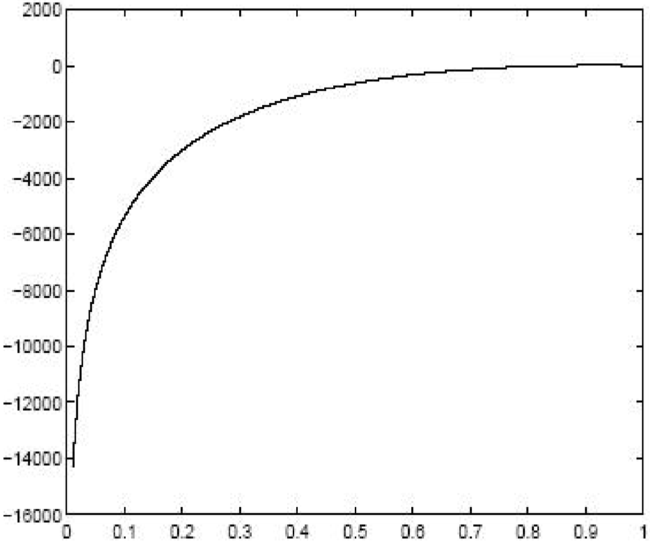

Figure 1. The final values, on the logarithmic (base 10) scale, attained by the randomised NN power martingales $M_n^{(\epsilon)}$ on the (full) USPS data set.

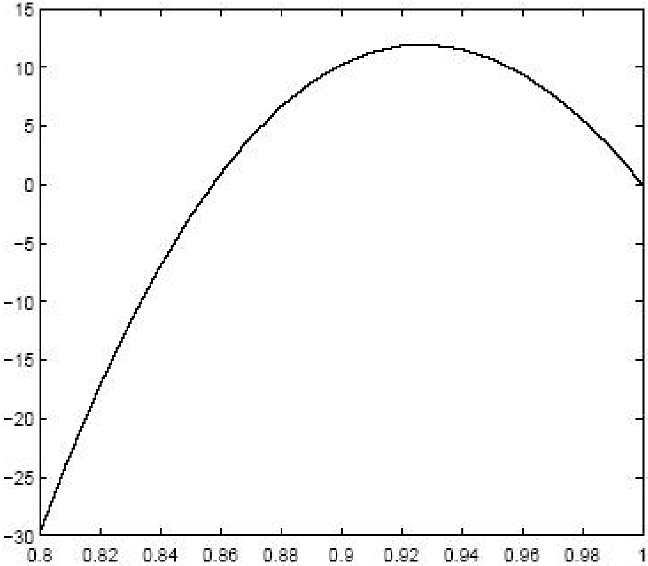

Figure 2. The final values for a narrower range of the parameter ϵ.

Testing Exchangeability: The Test

When randomized transducer is used, a martingale is constructed; when a deterministic transducer is used, a super-martingale is constructed. According to Doob's inequality,

$$\mathcal{Q}\{\exists n : M_n \geq C\} \leq \frac{1}{C}$$

for both the martingale and super-martingale (both are denoted M_n here) and C is an arbitrary positive constant, if M_n ever takes a large value, our belief in $\mathcal{Q}$ is undermined.

Experimental Results: USPS dataset

1. 7291 training examples and 2007 test examples.
2. merge the two sets in the order of training examples then test examples.
3. each example is an image of 16×16 matrix of pixels
4. labels: 0 to 9.

Experimental Results:

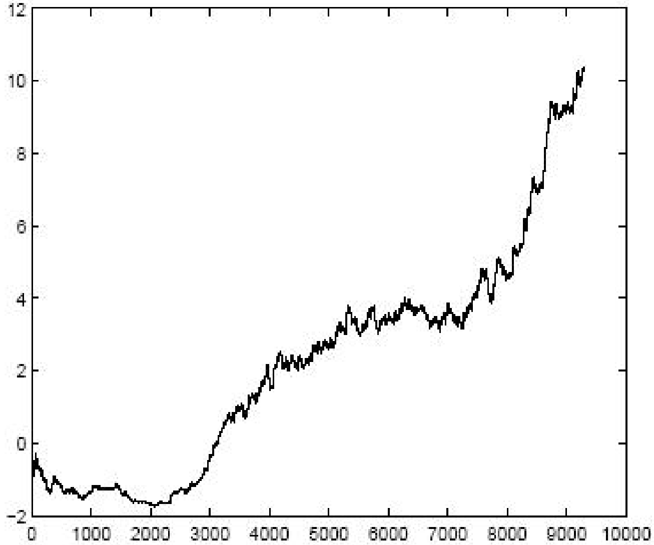

Figure 3. On-line performance of the randomised NN SM on the USPS data set. The growth is shown on the logarithmic (base 10) scale: $\log M_n$ is plotted against n. The final value attained is 2.18×10^{10}.

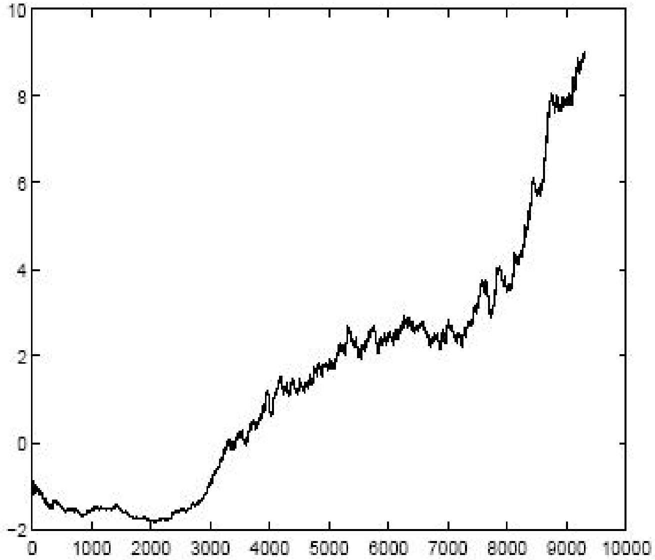

Figure 4. On-line performance of the deterministic NN SM on the USPS data set. The final value is 9.13×10^8.

Experimental Results:

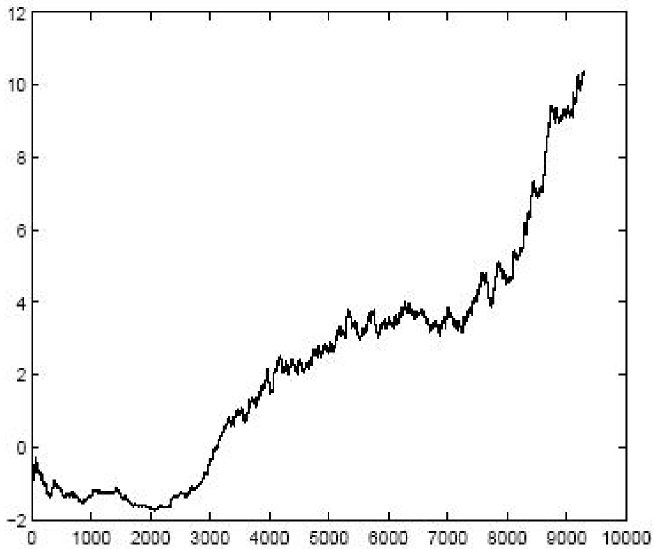

Figure 3. On-line performance of the randomised NN SM on the USPS data set. The growth is shown on the logarithmic (base 10) scale: $\log M_n$ is plotted against n. The final value attained is 2.18×10^{10}.

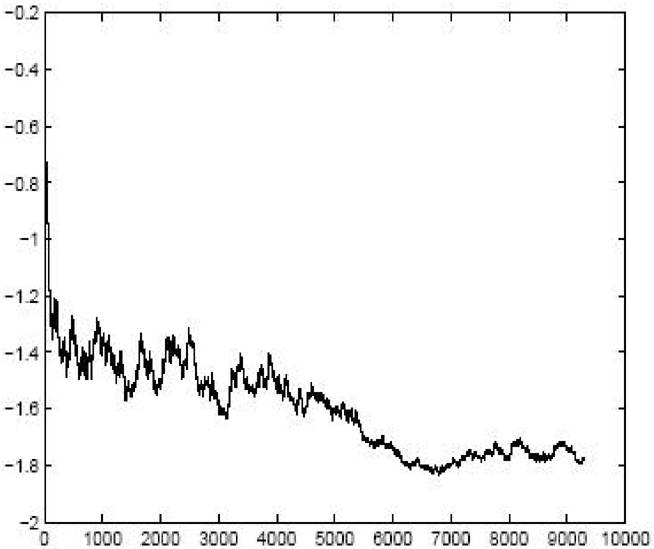

Figure 5. On-line performance of the randomised NN SM on a randomly permuted USPS data set. The final value is 0.0117.

Conclusion/Advantage/Problem:

Advantage:

1. One-pass algorithm
2. Only one sliding window needed.
3. Usable on high dimensional data.

Problem:

1. How large (compared to the initial martingale) should the martingale be before we say that the distribution changes?
2. How long (the rate of increase) before we say that there is a distribution change?
3. A confidence measure for the change?
4. An error bound for the change detection?
5. What if we have unlabeled data streams?

Reference

1. V.G. Vovk, A Logic of Probability, with Application to the Foundations of Statistics, Journal of the Royal Statistical Society, Series B (Methodological), Vol. 55, 2, 317-351, 1993.
2. Glenn Shafer and Vladimir Vovk, Probability and Finance: It's only a Game!, 2001
3. Vladimir Vovk, Alexander Gammerman and Glenn Shafer, Algorithmic learning in a random world (Manuscript), 2004.