Decision trees for stream data mining – new results

Leszek Rutkowski – leszek.rutkowski@iisi.pcz.pl
Lena Pietruczuk – lena.pietruczuk@iisi.pcz.pl
Maciej Jaworski – maciej.jaworski@iisi.pcz.pl
Piotr Duda – piotr.duda@iisi.pcz.pl

Czestochowa University of Technology
Institute of Computational Intelligence

Table of contents

Decision trees for stream data mining – new results

[1]

- Leszek Rutkowski, Lena Pietruczuk, Piotr Duda, Maciej Jaworski, *Decision trees for mining data streams based on the McDiarmid's bound*, IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6, pp. 1272–1279, 2013.

[2]

- Leszek Rutkowski, Maciej Jaworski, Lena Pietruczuk, Piotr Duda, *Decision trees for mining data streams based on the Gaussian approximation*, IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 108-119, 2014.

[3]

- Leszek Rutkowski, Maciej Jaworski, Lena Pietruczuk, Piotr Duda, *The CART decision tree for mining data streams*, Information Sciences, vol. 266, pp. 1 – 15, 2014.

The Hoeffding's inequality– incorrectly used for stream data mining (1)

The commonly known algorithm called 'Hoeffding's Tree' was introduced by P. Domingos and G. Hulten in [4]. The main mathematical tool used in this algorithm was the Hoeffding's inequality [5]

Theorem: *If* $X_1, X_2, \dots, X_n$ *are independent random variables and* $a_i \leq X_i \leq b_i$ $(i = 1, 2, \dots, n)$, *then for* $\epsilon > 0$

$$P\{\bar{X} - E[\bar{X}] \geq \epsilon\} \leq e^{-2n^2\epsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2}$$

where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } E[\bar{X}] \text{ is expected value of } \bar{X}.$$

[4] P. Domingos and G. Hulten, "Mining high-speed data streams", Proc. 6th ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining, pp. 71-80, 2000.
[5] W. Hoeffding, "Probability inequalities for sums of bounded random variables", Journal of the American Statistical Association, vol. 58, issue 301, pp. 13-30, March 1963.

The Hoeffding's inequality– incorrectly used for stream data mining (2)

If $X_i, i = 1, \ldots, n,$ are random variables of range R then the Hoeffding's bound takes the form

$$P\left\{\bar{X} - E[\bar{X}] \leq \sqrt{\frac{R^2 \ln 1/\delta}{2n}}\right\} \geq 1 - \delta$$

The Hoeffding's inequality– incorrectly used for stream data mining (3)

The Hoeffding's inequality is wrong tool to solve the problem of choosing the best attribute to make a split in the node using Gini index (e.g. the CART algorithm) or information entropy (e.g. the ID3 algorithm). This observation follows from the fact that:

- The most known split measures based on information entropy or Gini index, can not be presented as a sum of elements.
- They are using only frequency of elements.
- The Hoeffding's inequality is applicable only for numerical data.

The Hoeffding's inequality– incorrectly used for stream data mining (4)

Therefore the idea presented in [4] violates the assumptions of the Hoeffding's theorem (see [5]) and the concept of Hoeffding Trees has no theoretical justification.

[4] P. Domingos and G. Hulten, "Mining high-speed data streams", Proc. 6th ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining, pp. 71-80, 2000.
[5] W. Hoeffding, "Probability inequalities for sums of bounded random variables", Journal of the American Statistical Association, vol. 58, issue 301, pp. 13-30, March 1963.

The Hoeffding's inequality– incorrectly used for stream data mining (5)

In our papers [1], [2] and [3] we challenge all the results presented in the world literature (2000-2014) based on the Hoeffding's inequality. In particular, we challenge the following result:

Attribute a_{MAX1} is better to make a split than attribute a_{MAX2} with probability $1 - \delta$ if

$$G(a_{MAX1}) - G(a_{MAX2}) > \epsilon_H$$

where

$$\epsilon_H = \sqrt{\frac{R^2 \ln 1/\delta}{2n}}$$

and $G(.)$ is a split measure.

New techniques to derive split measures

In our papers [1], [2], [3]:

a) we study the following decision trees and split measures:

- Information gain (ID3)
- Gini gain (CART)

b) we propose two different techniques to derive split measures:

- The McDiarmid's inequality
- The Gaussian approximation

c) we replace incorrectly obtained bound ϵ_H (see the previous slide) by new bounds shown on the next slide

d) we determine the number of data elements sufficient enough to make a split

Comparison of our results with the Hoeffding's bound

	Information gain	Gini index gain
Hoeffding's bound	$\epsilon_H = \sqrt{\frac{R^2 \ln 1/\delta}{2N}}$ **Incorrectly obtained**	$\epsilon_H = \sqrt{\frac{R^2 \ln 1/\delta}{2N}}$ **Incorrectly obtained**
McDiarmid's bound	$\epsilon_{M_1} = C_{Gain}(K,N)\sqrt{\frac{\ln 1/\delta}{2N}}$ $C_{Gain}(K,N) = 6(K\log_2 eN + \log_2 2N) + 2\log_2 K$ **see [1]**	$\epsilon_{M_2} = 8\sqrt{\frac{\ln 1/\delta}{2N}}$ **see [1]**
Gaussian approximation	$\epsilon_{G_1} = z_{(1-\delta)}\sqrt{\frac{2Q(C)}{N}}$ $Q(C) = \frac{\log^2_2 e^2}{Ce^2} + \frac{\log^2_2 e^2}{e^2} + \frac{\log^2_2 e}{e} + \frac{\log^2_2 (2C)}{4}$ **see [2]**	$\epsilon_{G_2} = z_{(1-\delta)}\sqrt{\frac{2Q(K)}{N}}$ $Q(K) = 5K^2 - 8K + 4$ **see [3]**

Where:

- N denotes the number of data elements
- K denotes the number of classes
- $z_{(1-\delta)}$ denotes the $(1-\delta)$-th quantile of the standard normal distribution $N(0;1)$
- $C = \frac{1-th}{th}$ and th is a threshold value

Remarks

I. For information gain the value of ϵ_{G_1} is much better than the value of ϵ_{M_1}, however it can be applied only to a two class problems. For details see [2].

II. For Gini index the value of ϵ_{G_2} gives better results than ϵ_{M_2} for $K = 2$. For $K > 2$ the value of ϵ_{M_2} is more efficient. For details see [3].

How much data is enough to make a split?

	Information gain	Gini index gain
Hoeffding's bound	$N > \frac{\mathbf{R^2 \ln 1/\delta}}{2(\epsilon_H)^2}$ **Incorrectly obtained**	$N > \frac{\mathbf{R^2 \ln 1/\delta}}{2(\epsilon_H)^2}$ **Incorrectly obtained**
McDiarmid's bound	Can be determined numerically **see [1]**	$N > \frac{\mathbf{32 \ln 1/\delta}}{(\epsilon_{M_2})^2}$ **see [1]**
Gaussian approximation	$N > (z_{1-\delta})^2 \frac{\mathbf{2Q(C)}}{(\epsilon_{G_1})^2}$ $Q(C) = \frac{\log^2_2 e^2}{Ce^2} + \frac{\log^2_2 e^2}{e^2} + \frac{\log^2_2 e}{e} + \frac{\log^2_2 (2C)}{4}$ **see [2]**	$N > \frac{\mathbf{2Q(K)}(z_{1-\delta})^2}{(\epsilon_{G_2})^2}$ $Q(K) = 5K^2 - 8K + 4$ **see [3]**