

# Towards heterogeneity - homogeneity index of a test with respect to a training set (Extended Abstract)

Michał Grabowski  
michal.grabowski@twp.olsztyn.pl  
College of Economics and Computer Science TWP  
Olsztyn

Karol Wierzchołowski  
karol@wierzcholowski.net  
University of Warmia and Mazury  
Department of Mathematics and Computer Science  
Olsztyn

**Abstract** *We propose the index supporting the following heuristics: a test should divide a training set to non-similar (scattered) sets and the elements within any set in the partition should be similar (concentrated). The index is an entropy-driven mixture of local information given by the training set and of global information given by the whole space of all possible examples (like metric or similarity function on sets of examples). The index was tested by the decision tree induction method. The results of the introductory experiments show that the idea is worth of further study.*

## 1. Introduction

In the decision tree induction the following information is given at each level of recursion:

- current training set  $T$ ,
- current set  $S$  of accessible tests,
- optional category  $c$ .

At each stage of decision tree construction a test should be chosen from  $S$ , which is to be placed into the node under construction. The test-choosing criterion is the essence of decision tree construction. Any test divides the current training set into sets, let us say,  $T_1, \dots, T_n$ . We inherit from  $T$  a partition of every set  $T_i$  into categories. A classical approach chooses a test that minimizes the weighted sum of entropies of partitions of  $T_1, \dots, T_n$ . We add to several used criteria a new one based on the following heuristics:

- the sets  $T_1, \dots, T_n$  should be as non-similar as possible,
- the elements within any set  $T_i$  should be as similar as possible.

We can recalculate that, eventually, Lopes de Mantras[4] gives the criterion that takes the test minimizing the following ratio:

$$\frac{\text{weighted sum of entropies of partitions of } T_i \text{ into categories}}{\text{entropy of the partition of } T \text{ into } T_1, \dots, T_n}$$

Therefore the sets  $T_i$  in their external relation are taken into account and they should be more or less equal with respect to cardinality while each set  $T_i$  should possess a prevailing category. This criterion has an advantage in some situations, nevertheless it makes use of local information given by the training set  $T$  only.

Another criterion measures statistical independence between category distribution in  $T$  and partition of  $T$  into subsets  $T_1, \dots, T_n$ . More precisely, the two random variables:

$$\begin{aligned} g(x) &= \text{category of } x \text{ in } T \\ h(x) &= i, \text{ where } x \in T_i \end{aligned}$$

should be as statistically independent as possible. Again, this criterion makes use of local information given by training set  $T$  only, i.e. the appropriate probabilities are calculated as relevant frequencies in  $T$ .

We shall define the new criterion that makes use of global information hidden in the whole space of all possible examples: a metric on this space and induced by this metric similarity function on sets of examples. We use tolerance function on sets given by Doherty, Łukaszewicz, Skowron, Szalas[5].

Usually, the starting set of accessible tests is of cardinality at least as big as the cardinality of the set of attributes, therefore quite big. Decision tree induction algorithm can not use all accessible tests, since the resulting tree would be unacceptably big (even exponential with respect to the number of tests). Thus, at every stage of decision tree construction, algorithm can choose from a very big set of possible tests but the number of tests finally used in the resulting tree is relatively small. That is why the test-choosing criterion is important. Better algorithms choose tests which are more essential for classification and it results in better classification error.

In section 2 we introduce definitions of similarity on sets induced by environment structure, entropy of a test and finally, hh-index of a test.

In section 3 we show how to employ our approach to “alive” records, i.e. we define a metric on records and show that the metric can induce needed environments structure on the set of all records. We define an algorithm of decision tree construction using hh-index. The results of introductory experiments are given here.

In section 4 some final remarks are given.

## 2. Technicalities

Let  $X$  be a space of examples,  $C$  be a set of categories of a notion,  $T$  be a training set of labeled with categories in  $C$  examples. A test is a function  $t: X \rightarrow \{1, 2, \dots, n\}$ .

We assume that a structure of environments of examples is given: for every  $x \in X$  we are given a set  $n(x)$  of examples in  $X$  similar to the element  $x$ . Usually, the environments  $n(x)$  are induced by similarity functions on  $X$  or by metrics on  $X$ . In both cases some threshold is involved.

Now we define similarity function on sets of examples following Doherty, Łukaszewicz, Szałas, Skowron[5].

For any subsets  $U, V$  of the example space  $X$  we define the degree of inclusion of  $U$  in  $V$ :

$$v(U, V) = \frac{|\{x \in U \mid \exists y \in V y \in n(x)\}|}{|U|}$$

Thus  $v(U, V)$  is the proportion of elements in  $U$ , for which there are similar elements in  $V$ . Now we are in a position to define similarity of sets  $U, V$ :

$$\tau(U, V) = \min(v(U, V), v(V, U))$$

Notice that  $\tau(U, V) \in [0, 1]$ .

For any test  $t: X \rightarrow \{1, 2, \dots, n\}$  we define the following.

For  $i = 1, 2, \dots, n$   $T_i = \{x \in T \mid t(x) = i\}$ . We assume that  $n \geq 2$ .

For  $1 \leq i < j \leq n$ :

$$p_{ij} = \tau(T_i, T_j)$$

$$\mathbf{p}_{ij} = \frac{p_{ij}}{\sum_{1 \leq i < j \leq n} p_{ij}}, \text{ - normalized similarity } p_{ij} \text{ of } T_i, T_j.$$

Let  $m = |\{(i, j) \mid 1 \leq i < j \leq n\}| = n \cdot \frac{(n-1)}{2}$ .

Average similarity between sets  $T_1, T_2, \dots, T_n$ :

$$avs = \frac{\sum_{1 \leq i < j \leq n} p_{ij}}{m}$$

$$MaxE = -\log\left(\frac{1}{m}\right) \text{ - maximal entropy for } m \text{ probabilities.}$$

Now we give a few proposals of scattering index of a test with respect to a training set. Unfortunately, it seems that there is no one canonical index. That is why we have the word ‘‘Towards’’ in the title. Canonical situations are very rare in data exploration. The first proposal is the following.

$$\mathfrak{R}(T, t) = \begin{cases} f(avs) \cdot \left( MaxE + \sum_{1 \leq i < j \leq n} \mathbf{p}_{ij} \cdot \log(\mathbf{p}_{ij}) \right) & \text{if } n > 2 \\ \tau(T_i, T_j) & \text{if } n = 2 \end{cases}$$

The reasoning behind the above formula is as follows. The sum  $\sum \mathbf{p}_{ij} \cdot \log(\mathbf{p}_{ij})$  resembles (formally it is the same) entropy and it measures how much the normalized similarities  $\mathbf{p}_{ij}$  are different. Greater value of this sum indicates that the similarities  $\mathbf{p}_{ij}$  are more or less the same (i.e. the similarities between sets  $T_1, T_2, \dots, T_n$  are more or less the same). Thus ‘‘good’’ similarities between sets  $T_1, T_2, \dots, T_n$  maximize the sum. In order to have minimizing criterion, we take the formula  $MaxE - (-\sum \mathbf{p}_{ij} \cdot \log(\mathbf{p}_{ij})) = MaxE + \sum \mathbf{p}_{ij} \cdot \log(\mathbf{p}_{ij})$ . Finally, we want to have an impact of average similarity between sets  $T_1, T_2, \dots, T_n$  on heterogeneity index  $\mathfrak{R}(T, t)$  and we come to the proposed formula. The case  $n=2$  is treated separately. Function  $f$  is to represent an influence of average similarity

between sets  $T_1, \dots, T_n$  on heterogeneity index . We are of the opinion that function  $f$  should decrease index in greater degree for small  $avs$  and in smaller degree for greater  $avs$ . For the moment we propose two functions  $f$ :

$$f(avs) = avs \cdot avs$$

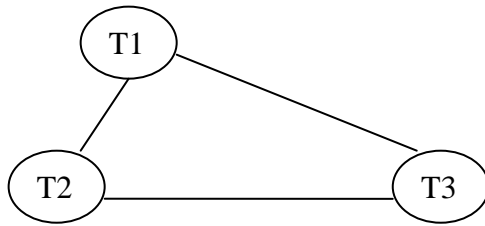
$$f(avs) = \frac{1}{(1 - \log(avs))}$$

Another proposals:

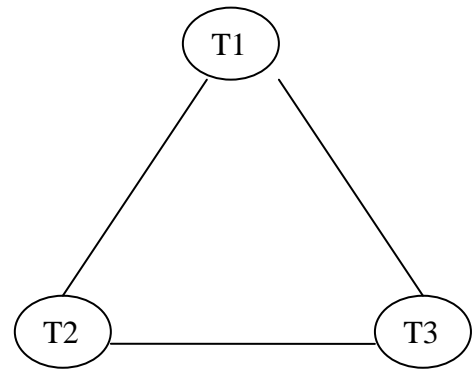
$$\mathfrak{R}(T, t) = \frac{avs}{\left(-\sum_{1 \leq i < j \leq n} p_{ij} \cdot \log(p_{ij})\right)}$$

$$\mathfrak{R}(T, t) = \frac{avs}{\left(-\sum_{1 \leq i < j \leq n} \mathbf{p}_{ij} \cdot \log(\mathbf{p}_{ij})\right)}$$

All indexes support the following intuition:



worse scattering, greater index,



better scattering, smaller index,

where distances in the figure between  $T_i, T_j$  represent similarities between  $T_i, T_j$ .

Now we define the entropy of a test  $t$  with respect to a training set  $T$ .

Let for  $c \in C$ :  $T_{ic} = \{x \in T_i \mid x \text{ in } T \text{ has category } c \}$ .

$$E(T, t) = \sum_{1 \leq i \leq n} \left( \frac{|T_i|}{|T|} \right) \cdot \left( - \sum_{c \in C} \left( \frac{|T_{ic}|}{|T_i|} \right) \cdot \log \left( \frac{|T_{ic}|}{|T_i|} \right) \right)$$

as it is defined in Cichosz[1]. We accept  $E(T, t)$  as homogeneity index since if every set  $T_i$  has a prevailing category than it can be considered homogeneous. We are aware that this is an unsatisfactory and temporary proposal and that homogeneity issue has to be analyzed further (final remark 3).

Now we combine scattering index and entropy of test into **hh-index** of test with respect to training set T. Let a weight  $0 \leq \alpha \leq 1$  of test's entropy is given.

$$hh(T, t) = \alpha \cdot E(T, t) + (1 - \alpha) \cdot \mathfrak{R}(T, t)$$

Test-choosing criterion:

choose test  $t \in S$  such that  $hh(T, t) \rightarrow \min$ .

### 3. Metric on records and results of small experiments

In order to apply our hh-index we have to show how to generate structure of environments of similar records. We shall define a metric on records at first.

Assume that we are given records

$$r_1 = (a_{11}, a_{12}, \dots, a_{1n})$$

$$r_2 = (a_{21}, a_{22}, \dots, a_{2n}).$$

We assume, that every attribute is either numerical or nominal.

#### Distance between values of numerical attributes

Let  $\delta$  be the standard deviation calculated from appropriate column of data set T and let a, b be some attribute values.

$$d(a, b) = \frac{|a - b|}{\delta}$$

In this way we make distance  $d(a, b)$  independent of measurement units of data (Koronacki, Mielniczuk[3]). Thus we count distance in units being  $\delta$ .

#### Distance between values of nominal attributes

For this moment we do not have a really good idea how to measure distance between values of nominal attributes. We give a bit ad hoc proposal.

We assume that there are finitely many values  $w_1, w_2, \dots, w_k$  in the domain of the attribute.

- (a) If the domain of the attribute is linearly ordered then distance is measured as distance with respect to this order.
- (b) In opposite case let  $p_1, p_2, \dots, p_k$  be frequencies of occurrences of values  $w_1, w_2, \dots, w_k$  in the training set. We can interpret  $p_1, p_2, \dots, p_k$  as probabilities of appearing of values  $w_1, w_2, \dots$  in data set. We define the following distance:

$$d(w_i, w_j) = \frac{1}{k} + |p_i - p_j| \text{ if } i \neq j$$

$$d(w_i, w_i) = 0$$

Distance between records is defined in ordinary Euclidian way:

$$d(r_1, r_2) = \sqrt{\sum_{1 \leq i \leq n} d(a_{1i}, a_{2i})^2}$$

If we are given a threshold  $q$  then we define environments of similar records:

$$n(r) = \{r' \mid d(r, r') \leq q\}$$

Therefore, the similarity function between sets of records is ready to be used, since its definition depends on similarity environments only.

### hh-algorithm

input data:

- training set T,
- threshold on the number of used tests along a path: p
- threshold on the radius of similarity environments: q
- weight of test's entropy:  $\alpha$

actions:

1. Define distance function d between records
2. Define similarity function  $\tau(U,V)$  between sets of records (the threshold q is involved here)
3. Compute the decision tree D by slightly modified standard schema:
  - halt, when the number of tests used exceeds the given threshold p,
  - at each stage of construction choose test minimizing hh-index with respect to the current training set

output:

probability of wrong classification by the computed tree D

We have performed some introductory experiments.

Table 1. Heart Disease

NT	Random	Entropy	HH – algorithm									
			57,5	107,1	156,6	206,2	255,7	305,2	354,8	404,3	453,9	503,4
2	0,6963	0,5741	0,7815	0,7556	0,7037	0,7444	0,7444	0,7444	0,7111	0,7111	0,6815	0,663
3	0,6593	0,5704	0,7741	0,7963	0,7852	0,8	0,7963	0,7667	0,7778	0,7889	0,7185	0,7185
4	0,6741	0,5444	0,7481	0,7889	0,8185	0,8222	0,7741	0,7593	0,7593	0,7667	0,7185	0,7185
5	0,6481	0,5593	0,7444	0,7556	0,7852	0,7667	0,7704	0,7556	0,7444	0,7593	0,7333	0,7333
6	0,7222	0,5815	0,7741	0,7778	0,8037	0,7889	0,7963	0,7741	0,7593	0,7667	0,7185	0,7185
7	0,6333	0,5704	0,7963	0,7926	0,8037	0,7852	0,7815	0,7741	0,7704	0,7815	0,7407	0,7407
8	0,6852	0,5593	0,7444	0,8	0,8037	0,8111	0,8148	0,8037	0,763	0,7741	0,7593	0,7593
9	0,6852	0,5741	0,7519	0,8148	0,8222	0,7963	0,7889	0,7778	0,7889	0,7963	0,7889	0,7889
10	0,6815	0,5815	0,7481	0,8037	0,8148	0,8185	0,8333	0,8111	0,7667	0,7704	0,7704	0,7704
11	0,6407	0,5481	0,7778	0,7889	0,8037	0,7741	0,7815	0,7667	0,7519	0,7778	0,737	0,737
12	0,6481	0,5889	0,7593	0,8185	0,7963	0,8	0,8	0,7926	0,7593	0,7778	0,7593	0,7185
13	0,6667	0,5778	0,7519	0,8148	0,8074	0,7926	0,8	0,7815	0,7444	0,7593	0,6926	0,6926

The first column contains the limits on number of tests used along a path in the constructed tree.

The second column contains the total accuracies (1 – classification error) of classification of the algorithm with random choice of a test at each stage of construction.

The third column contains the total accuracies of classification of standard algorithm by entropy.

The first row of the HH-algorithm part of the table represents the consecutive radiuses (with respect to our metric on records)  $q$  of similarity environments.

The consecutive rows contain total accuracies of classification given by hh-algorithm with similarity function on sets of records induced by radiuses  $q$  of similarity environments from the first row. Zero is taken in hh-index as the weight  $\alpha$  of test's entropy, therefore the heterogeneity index is involved in classification, only.

We have taken three small data sets: *Heart-diseases* from RSES packet, *Wine* from UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/>), *Diabetes* from RSES packet.

In every case, we divide into 10 parts the sorted sequence of all possible distances between records of input data set. The maximal number of each part is taken as the consecutive radius  $q$  of similarity environments. In all these three cases the qualitative information is almost the same (observations 1. and 2.).

1. In each case the algorithm with random choice of test is better than standard algorithm by entropy. If there is no methodological mistake in experiment, it may indicate that entropy criterion gives over-fitting. Maybe, we have used cross-validation in improper way. We shall analyze this signal in more details in the near future.
2. In average, hh-algorithm gives the best accuracy near the mean radius and is better than the remaining two algorithms.

We end this section with basic information on *Heart-diseases* data set.

Number of examples: 270.

Number of attributes: 13.

Minimal distance between records: 8,00.

Maximal distance between records: 503,40.

Tests: "less or equal" with respect to values from the training data set.

#### 4. Final remarks

1. Our proposal is not canonical in at least three directions.
  - (a) We have no unique heterogeneity index.
  - (b) We have no unique homogeneity index.
  - (c) There are different possibilities of defining similarity function on sets of examples. We do not have to follow tolerance function on sets of Doherty, Łukaszewicz, Skowron, Szałas[5].
2. The big data sets with many attributes must be tested with several proposals of similarity functions on sets, several heterogeneity, homogeneity indexes. Algorithms must be compared with respect to accuracy and to the error variance instead of accuracy only.
3. The similarity environments should be implanted into homogeneity index since otherwise we lose information hidden in similarity environments structure. For example, we could take the minimal number of environments covering the set  $T_i$  as homogeneity

index of  $T_i$ . But we are faced then with the problem of relevant rescaling of that number in order to join it with our heterogeneity index.

4. The several entropy-driven indexes measuring the degree of separation between clusters are defined on the base of probabilities of appearing of the  $i$ -th record in  $s$ -th cluster, see Fonseca, Cardoso[2]. We use entropy in different context. The similarities between sets are used instead of probabilities. We hope that our index can be used also as an index measuring the quality of clustering, when a structure of similarity environments is given on the space of examples.

## **Bibliography**

1. Cichosz P., Machine Learning (in polish) ,ISBN 83-204-2544-1, WNT, 2000
2. Fonseca J. R. S., Cardoso M. G. M. S., Mixture-model cluster analysis using information-theoretical criteria, Intelligent Data Analysis, vol. 11, No 2, 2007
3. Koronacki J., Mielniczuk J., Statistics (in polish) , ISBN 83-204-3242-1, 2006
4. Lopez de Mantaras R., A distance-based attribute selection measure for decision tree induction. Machine Learning, 1991, 6, s. 81-92
5. Doherty P., Łukaszewicz W., Skowron A., Szalas A., Knowledge representation techniques, ISBN 3-540-33518-8, Springer, 2006