Tjala algorithm

The word Tjala is the name of honey ants in the Pitjantjatjara language, one of the Australian Indigenous languages. Some ants in their colonies store a sweet nutrient in their abdomens, becoming the size of grapes and feeding other ants. They are a traditional delicacy of Australian First Nations people. They stand out from other ants and are actively used by humans.

Definitions:

Term – part of the signature, beginning with the word IF. Term consists of attribute, operator, and value.

Attribute – a variable that characterizes the sample. An attribute in one rule is used no more than once to avoid the following situation: IF ENS0001 = 5 AND ENS0001 = 15…

Operator – sign(s) of relations. Used operators: >,<.

Value – the number to which the attribute in the sample is equal.

Node (aka vertex) – a separate term in the signature

Edge – imaginary connection between successive nodes

Antecedent – part of the signature before THEN

Consequent – part of the signature after THEN

Pruning – Removing terms from a signature to improve the quality of the signature

The vector lm shows the relation of the given sample to each cluster. It is a list of all clusters in the hierarchy (excluding the root cluster) with values of 1 or 0 for the given sample: 1 if the sample is in the cluster and 0 if not.

The vector li is the proportion of samples in a given cluster that are covered by a particular signature

*v ̅* – the mean value of the vectors in the given cluster for the set of samples covered by the given term

varmax – the sum of the largest and lowest variance values among all terms

w(l) – the weight of the cluster. This corresponds to the position of the cluster in the hierarchy. A higher weight indicates a higher position. The weight values are in the range 0 < w(l) < 1. For the highest and lowest clusters, the weights are set to 0.99 and 0.01, respectively. The weights for all other clusters are evenly distributed within this range.

The signatures are created until the iteration limit is exhausted or the signatures stop improving (convergence is reached; the best signature no longer changes). It is important to understand whether the signature is included in the list of signatures (decision list) because of the limit of iterations or if it is best possible. Therefore, this information is displayed in the output.

The algorithm

0. Request the user to input the maximum number of terms in signature (Max number of terms), the maximum number of uncovered samples (Max uncovered samples), the maximum number of iterations (Max number of iterations), the maximum number of iterations during which a signature does not change to stop iterations and recognize the signature as the best (Rule convergence), the minimum number of samples covered by a signature (Min samples per rule), and the number of ants (Ant colony size).

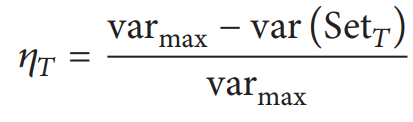
N.B. The uploaded version has predefined settings.

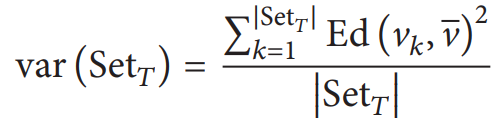
Iteration

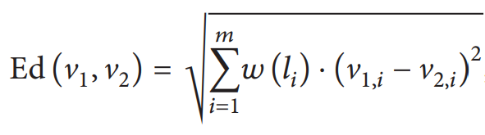
1. Generate all possible terms for the samples

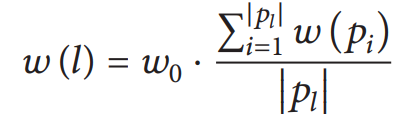
2. Calculate the quality of each term

The quality of the term is calculated using the following formula:

, where varmax is the sum of the best and worst variance values of all terms and

, where

, where 𝑤(𝑙𝑖) is the weight of the 𝑖th cluster label and V1,𝑖 and V2,𝑖 represent the 𝑖th cluster value of two samples, respectively, and

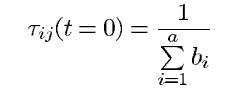
, where 𝑤0 is set to 0.75, 𝑝𝑙 is the parent cluster label set of the cluster label 𝑙, and 𝑤(𝑝𝑖) is the weight associated with the 𝑖th parent cluster label of the cluster label 𝑙.

All formulas are adopted from [1].

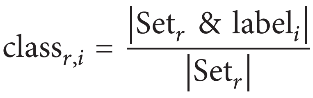
3. An ant creates a signature

Terms are added to the signature one by one until the limit of the number of terms in the signature is reached. The probability of choosing a term for a signature is determined by the following formula: the product of the pheromone and the heuristic of the node, divided by the sum of the same products of all terms not used in this signature.

The initial amount of pheromones is equal in all nodes, inversely proportional to the number of values of all attributes, and is determined by the following formula:

 (adopted from [2]),

where a is the total number of attributes, and bi is the number of possible values of the attribute Ai.

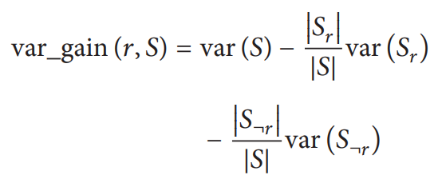
The consequent of a signature is determined by the following formula:  (adopted from [1]),

where Set𝑟 represents the set of samples covered by signature 𝑟, which generates a vector of length 𝑚 (𝑚 is the number of cluster labels) as a result of that signature; label𝑖 is the 𝑖th component of the cluster vector; |Set𝑟 & label𝑖| is the number of samples belonging to the 𝑖th cluster of the cluster hierarchy covered by signature 𝑟. The cluster 𝑖, a vector of length 𝑖, represents the proportion of samples that are covered by signature 𝑟 in a particular 𝑖th cluster.

4. Signature pruning

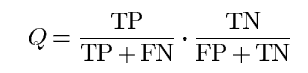
Pruning proceeds as follows:

The quality is calculated for the whole signature and for each variant with one deleted term. In the latter case, the consequent of the signature is also recalculated. If deleting a term does not reduce the quality of the signature, this term is to be deleted. The operation is iterated until no terms can be deleted without reducing the quality or only one term remains in the signature.

The quality of the signature is equal to the variance gain and determined by the following formula:  (adopted from [1]),

where var\_gain (*r,S*) is the variance gain of signature 𝑟 relative to 𝑆; 𝑆 is the training set; (𝑆𝑟) is the set of samples covered by the signature *r*, and (𝑆¬𝑟) is the set of samples not covered by signature *r*.

5. Update Pheromones

The quality of the signature for the purposes of updating pheromones is determined by the formula:  (adopted from [2]), where

TP – true positives, the number of samples covered by the signature that have the cluster predicted by the signature.

FP – false positives, the number of samples covered by the signature that have a cluster different from the cluster predicted by the signature.

FN – false negatives, the number of samples not covered by the signature but that have the cluster predicted by the signature.

TN – true negatives, the number of samples not covered by the signature and that do not have the cluster predicted by the signature.

Q is between 0 and 1.

The addition of pheromones into the nodes occurs according to the following formula:  (adopted from [2]), where R is a set of terms present in the signature.

The amount of pheromone in each node that is not used in the signature, as a result of evaporation, is reduced to the following value:

(1 – ∑pheromones in signature nodes)/number of unused nodes

6. Compare the generated signature with the best signature of the iteration (ruleib). If the generated signature is of superior quality, it becomes the best signature for the iteration.

7. Iterate until the limit of iterations (= number of ants) or convergence is reached.

8. Add the best signature of the iteration to the list of signatures.

9. Remove samples covered by the signatures from further analysis.

10. Iterate from step 1 until the maximum number of uncovered samples is reached or the best new signature after the end of the iteration does not cover the minimum number of samples.

11. Display the list of signatures in the following form:

Rule #1:

(max iterations)

IF CHRM1 > 265.0 THEN:

Can clusters:

Cluster 1: 84.4% covered

...............

Cannot clusters:

Cluster 0: 15.5% covered

----------

Scale = 10, CHRM1 < 26.5: 74.3% covered in cannot clusters

Scale = 20, CHRM1 < 13.25: 61.800000000000004% covered in cannot clusters

Scale = 30, CHRM1 < 8.833333333333334: 52.7% covered in cannot clusters

Scale = 40, CHRM1 < 6.625: 47.7% covered in cannot clusters

Scale = 50, CHRM1 < 5.3: 44.900000000000006% covered in cannot clusters

Scale = 60, CHRM1 < 4.416666666666667: 42.5% covered in cannot clusters

Scale = 70, CHRM1 < 3.7857142857142856: 38.7% covered in cannot clusters

Scale = 80, CHRM1 < 3.3125: 38.7% covered in cannot clusters

Scale = 90, CHRM1 < 2.9444444444444446: 33.800000000000004% covered in cannot clusters

Scale = 100, CHRM1 < 2.65: 33.800000000000004% covered in cannot clusters

----------

Where:

"Rule #1" is the number of the signature in the order of generation.

"IF CHRM1 > 265.0 THEN" is a signature.

"(Max iterations/convergence)" indicates whether convergence or the iteration limit was reached.

"Cannot clusters" are clusters that were defined by the user to not be covered by signatures with a user-defined inclusion threshold.

"84.4% coverage" is the proportion of samples within the cluster that satisfy this signature.

Clusters in the consequent of the signature are displayed in descending order of the proportion of samples covered by the signature in each cluster.

"Scale = 10, CHRM1 < 26.5: 74.3% covered in cannot clusters" indicates the proportion of samples in cannot clusters that would be covered by the term if the value of the term were 10 times smaller. It shows the differentiating ability of the given signature.