femto-st SCIENCES & TECHNOLOGIES

New convex-based metamorphic relations and large-scale ML model evaluation

The 37th International Conference on Testing Software and Systems

17 September 2025

Jessy COLONVAL[1] and Fabrice BOUQUET[1]

[1]Université Marie et Louis Pasteur, CNRS, insitut FEMTO-ST (UMR 6174), F-25000 Besançon, France

CNRS UNIVERSITÉ MARIE & LOUIS PASTEUR SUP MICRO TECH ENSMM utbm université de technologie Belfort-Montbéliard

Table of contents

Context

Oracle problem

Oracle hypothesis: it's possible to determine if an output is correct

➡ FALSE

Context

Oracle problem

Oracle hypothesis: it's possible to determine if an output is correct

➡ FALSE

In reality:

- no oracle;
- too difficult too implement.

Context

Oracle problem

Oracle hypothesis: it's possible to determine if an output is correct
➡ FALSE

In reality:

- no oracle;
- too difficult too implement.

Machine learning models have an oracle problem!

Context

Oracle problem

Metamorphic relations:

- No verification of each input / output;
- Verification of properties between input / output.

➡ **Ligthen the problem of the oracle!**

Table of contents

State-of-the-art metamorphic relations

Identity relation — MR 1

When two models have the same:

- training algorithm;
- meta-parameters;
- data sets.

State-of-the-art metamorphic relations

Identity relation — MR 1

When two models have the same:

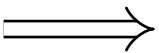
- training algorithm;
- meta-parameters;
- data sets.

➡ **Outputs must be identical!**

State-of-the-art metamorphic relations

Points shuffle relation — MR 2

attr0	attr1	class
2.91	0.59	0
2.14	−0.47	0
⋮	⋮	⋮
0.19	0.29	1
1.12	0.95	2
⋮	⋮	⋮
0.01	−1.42	3
3.67	2.13	3

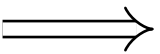


attr0	attr1	class
2.91	0.59	0
1.12	0.95	2
⋮	⋮	⋮
0.19	0.29	1
2.14	−0.47	0
⋮	⋮	⋮
0.01	−1.42	3
3.67	2.13	3

State-of-the-art metamorphic relations

Points shuffle relation — MR 2

attr0	attr1	class
2.91	0.59	0
2.14	−0.47	0
⋮	⋮	⋮
0.19	0.29	1
1.12	0.95	2
⋮	⋮	⋮
0.01	−1.42	3
3.67	2.13	3



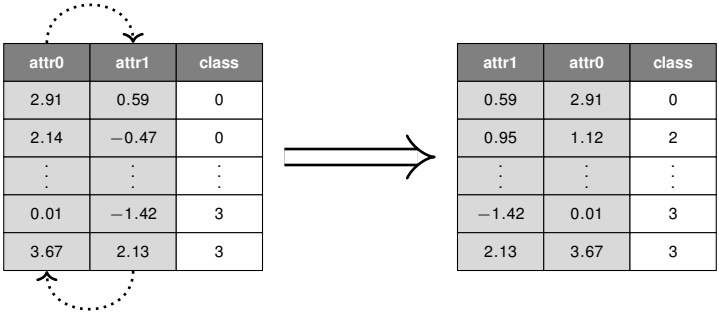
attr0	attr1	class
2.91	0.59	0
1.12	0.95	2
⋮	⋮	⋮
0.19	0.29	1
2.14	−0.47	0
⋮	⋮	⋮
0.01	−1.42	3
3.67	2.13	3

Outputs must be identical!

Except when part of points are used.

State-of-the-art metamorphic relations

Attributes shuffle — MR 3

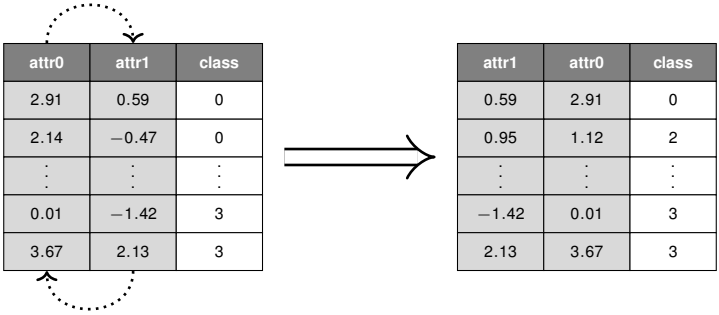


attr0	attr1	class
2.91	0.59	0
2.14	−0.47	0
⋮	⋮	⋮
0.01	−1.42	3
3.67	2.13	3

attr1	attr0	class
0.59	2.91	0
0.95	1.12	2
⋮	⋮	⋮
−1.42	0.01	3
2.13	3.67	3

State-of-the-art metamorphic relations

Attributes shuffle — MR 3



attr0	attr1	class
2.91	0.59	0
2.14	−0.47	0
⋮	⋮	⋮
0.01	−1.42	3
3.67	2.13	3

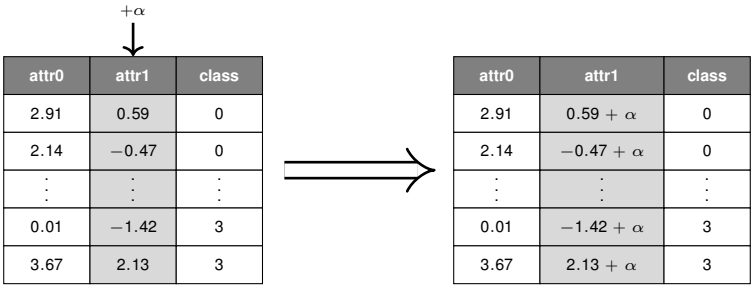
attr1	attr0	class
0.59	2.91	0
0.95	1.12	2
⋮	⋮	⋮
−1.42	0.01	3
2.13	3.67	3

Outputs must be identical!

Except when part of attributes are used.

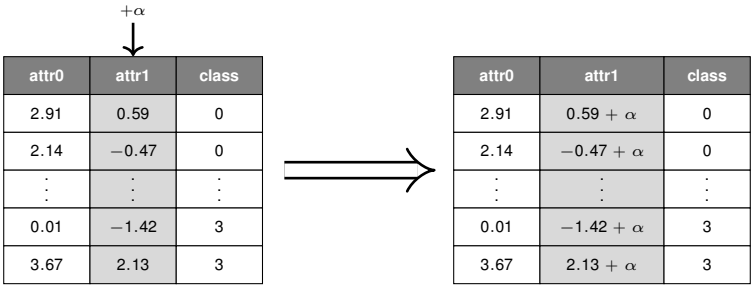
State-of-the-art metamorphic relations

Transformation relation — MR 4



State-of-the-art metamorphic relations

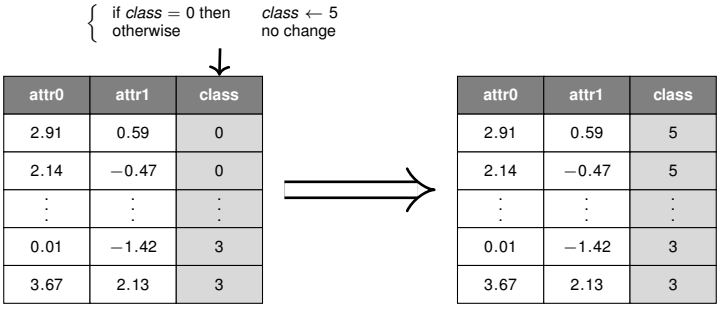
Transformation relation — MR 4



Outputs must be identical!

State-of-the-art metamorphic relations

Class permutation relation — MR 5



State-of-the-art metamorphic relations

Class permutation relation — MR 5

attr0	attr1	class
2.91	0.59	0
2.14	−0.47	0
⋮	⋮	⋮
0.01	−1.42	3
3.67	2.13	3

attr0	attr1	class
2.91	0.59	5
2.14	−0.47	5
⋮	⋮	⋮
0.01	−1.42	3
3.67	2.13	3

Outputs must be identical!

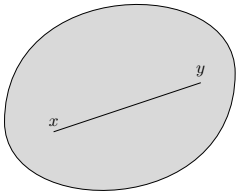
Table of contents

Convex-based metamorphic relations

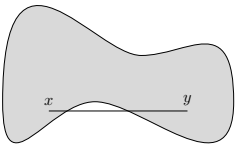
Convex definition

Definition

A part H of $\mathbb{R}^n$ said to be convex if, for all pairs (x, y) of elements of H, the segment $[x, y]$ is entirely contained within H. In other words, H is convex when $\forall x, y \in H$ and $\forall \lambda \in [0; 1], \lambda x + (1 - \lambda)y \in H$ [10].



(a) A convex

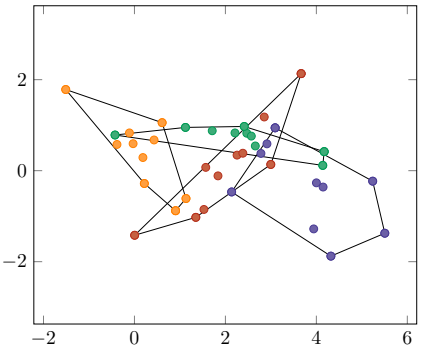


(b) A non-convex

Convex-based metamorphic relations

How they are created?

- Using the algorithm **QuickHull** *n*D (*QHull*[1]).
- Approximation of influence areas of classes.



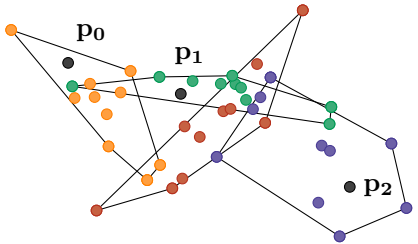
[1] http://www.qhull.org/

Convex-based metamorphic relations

Membership relation — MR 6

Hypothesis:

- Convex = one area of class.
- Point inside ➡ similar characteristics.
  ➡ identical class.

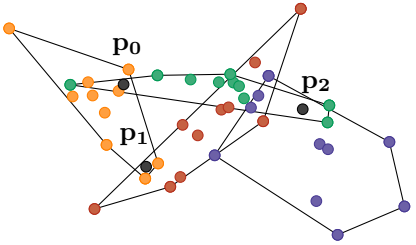


Convex-based metamorphic relations

Superposition relation — MR 7

Observations:

- Shared areas.
- Which choice will the models make? ➡ We don't choose!

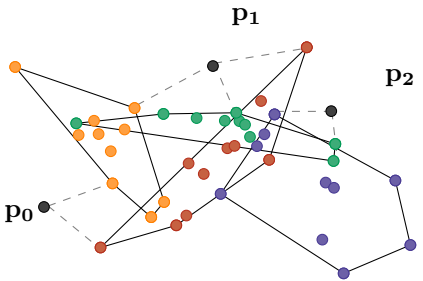


Convex-based metamorphic relations

Attachment relation — MR 8

Hypothesis:

- Model extrapolation.
- Nearest convex(es) ⇒ similar characteristics.
  ⇒ identical class.



Convex-based metamorphic relations

Limits & Robustness

Tests evaluate values:

- within the limits;
- at the limits;
- off limits.

Convex-based metamorphic relations

Limits & Robustness

Tests evaluate values:

- within the limits; ➡ in the domain;
- at the limits; ➡ borders of the domain;
- off limits. ➡ outside the domain.

Convex-based metamorphic relations

Limits & Robustness

Tests evaluate values:

- within the limits; ➡ in the domain;
- at the limits; ➡ borders of the domain;
- off limits. ➡ outside the domain.

Models must:

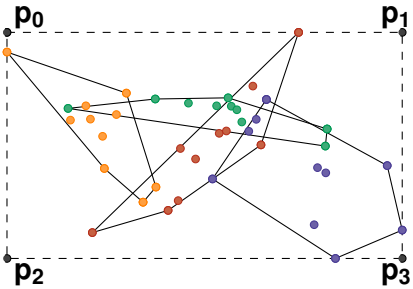
➡ Do not produce an error: **Robustness — MR 9**

➡ Do not produce an error when the value is outside the domain: **Boundary Robustness — MR 10**

➡ Respect the relation **Boundary Attachment — MR 11**

Convex-based metamorphic relations

Illustration of the use of limits

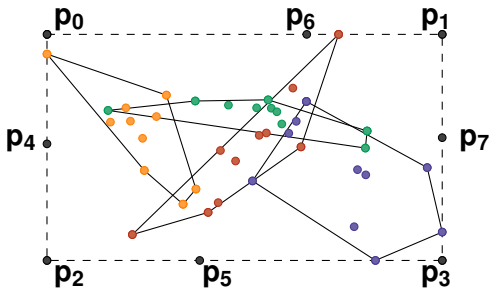


Values:

- all at the limits;

Convex-based metamorphic relations

Illustration of the use of limits

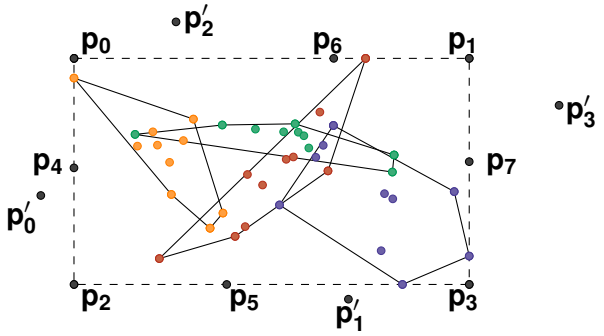


Values:

- all at the limits;
- only one at the limits;

Convex-based metamorphic relations

Illustration of the use of limits



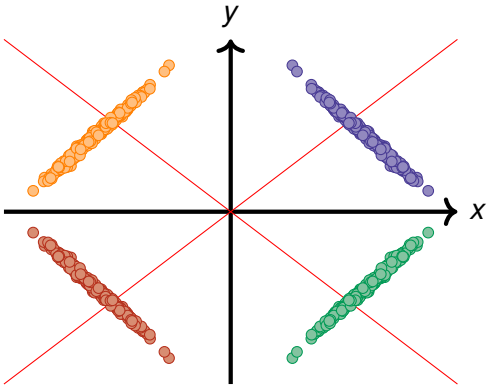
Values:

- all at the limits;
- only one at the limits;
- one out of limits.

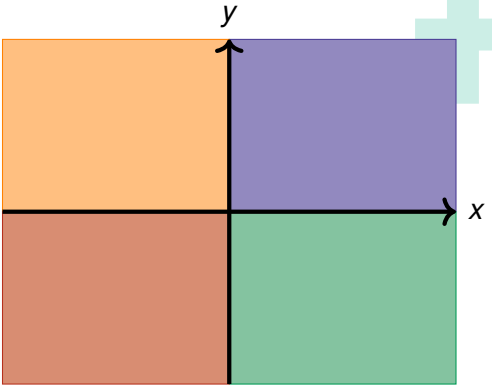
Table of contents

Non-convex metamorphic relations

Precision relation — MR 12



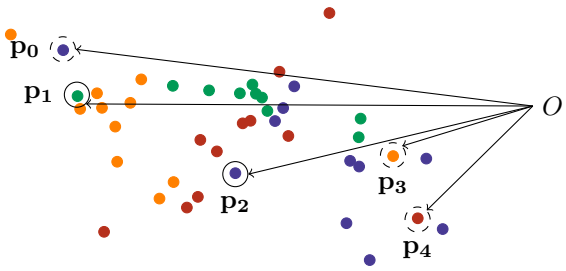
(a) Data set



(b) Expected predictions

Non-convex metamorphic relations

Outliers relation — MR 13

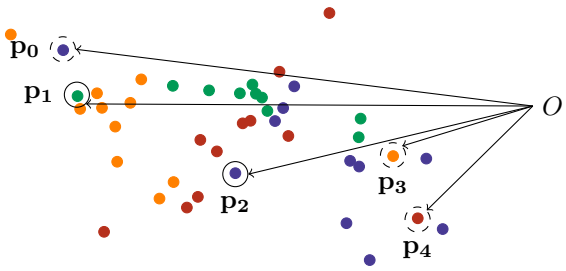


Origin:

- added to the dataset;
- already existing.

Non-convex metamorphic relations

Outliers relation — MR 13



Origin:

- added to the dataset;
- already existing.

Models shoudn't predict them correctly!

Table of contents

Experimentations

Synthetic data sets

Points	Attributes	Classes	Noises	Distribution	Ratio
s – 250	**s** – 10	**b**	**n**	**ho**	**y**
s – 250	**m** – 100	**p** – 5	**y** - 60%	**he**	**n**
s – 250	**l** – 1 000	**b**	**y** – 10%	**ho**	**n**
m – 25 000	**s** – 10	**b**	**y** – 60%	**he**	**y**
m – 25 000	**m** – 100	**p** – 5	**y** – 10%	**ho**	**y**
m – 25 000	**l** – 1 000	**b**	**n**	**he**	**y**
l – 125 000	**s** – 10	**p** – 5	**y** – 10%	**he**	**y**
l – 125 000	**m** – 100	**b**	**y** – 60%	**ho**	**y**
l – 125 000	**m** – 100	**b**	**n**	**ho**	**y**

Experimentations

Algorithms

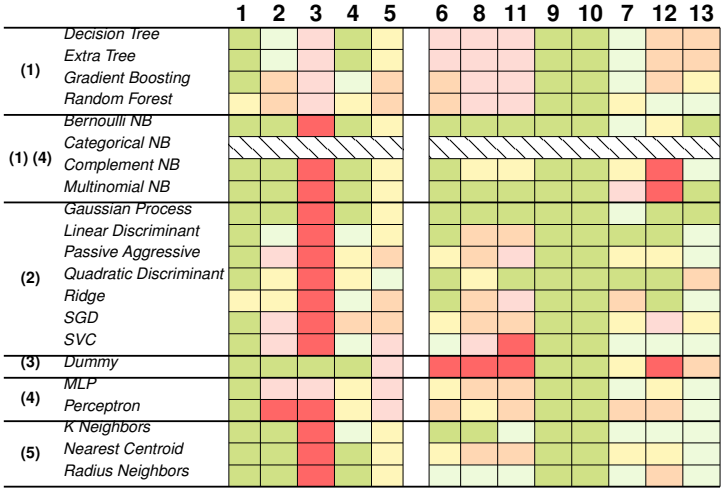
	Paradigm (1)	(2)	(3)	(4)	(5)
Bernoulli NB	■			■	
Categorical NB	■			■	
Complement NB	■			■	
Decision Tree Classifier	■				
Extra Tree Classifier	■				
Dummy Classifier			■		
Gaussian Process Classifier		■			
Gradient Boosting Classifier	■				
K Neighbors Classifier					■
Linear Discriminant Analysis		■			
MLP Classifier				■	
Multinomial NB	■			■	
Nearest Centroid					■
Passive Aggressive Classifier		■			
Perceptron				■	
Quadratic Distriminant Analysis		■			
Radius Neighbors Classifier					■
Random Forest Classifier	■				
Ridge Classifier		■			
SGD Classifier		■			
SVC		■			

5 **paradigms** :

- (1) decision trees;
- (2) support vector machines;
- (3) overall distribution of classes;
- (4) neural networks;
- (5) neighborhoods.

Experimentations

Results



70 276 **models**

Analysis :

- 2 failures - **MR 1** ;
- robustness ✓ ;
- too many failures - **MR 5**.

Table of contents

Conclusion

- 5 state-of-the-art relations;
- 8 new metamorphic relations;
  - 6 based on convexes;
  - 2 not.
- evaluation of 21 algorithms with 70 276 models;
- revealed the probable existence of bugs.

Thanks for your attention!