Opportunities for parallelization

Some material based on the Mining Massive Datasets book

- reading data
- writing data
- statistical operations
- data transformation
- data aggregation
- feature construction
- SQL-like operations (it depends!)

Opportunities for parallelization

```
## K-Means Clustering

1. Choose the number of clusters(K) and obtain the data points
2. Place the centroids c_1, c_2, ..... c_k randomly
3. Repeat steps 4 and 5 until convergence or until the end of a fixed
number of iterations
4. for each data point x_i:
       - find the nearest centroid(c_1, c_2 .. c_k)
       - assign the point to that cluster
5. for each cluster j = 1..k
       - new centroid = mean of all points assigned to that cluster
6. End
```

Opportunities for parallelization

Algorithm 3: Apriori algorithm

```
F_1 = {frequent items of size 1};
for (k = 1; F_k != φ; k++) do begin
    C_{k+1} = apriori-gen(F_k); // New candidates generated from F_k
    for all transactions t in database do begin
        C'_t = subset(C_{k+1}, t); // Candidates contained in t
        for all candidate c ∈ C'_t do
            c.count++; // Increment the count of all candidates
             in C_{k+1} that are contained in t
        end
        F_{k+1} = {C ∈ C_{k+1} | c.count ≥ minimum suport}
        //Candidates in C_{k+1} with minimum support
    end
end
```

Opportunities for parallelization

Algorithm 2: C4.5 Algorithm

1. Check for **base cases**.
2. For each attribute ***a***

   find the **normalized information gain** from splitting on ***a***.
3. Let ***a_best*** be the attribute with the

   highest normalized information gain.
4. Create a **decision node** that splits on ***a_best***.
5. Recur on the sublists obtained by splitting on ***a_best***, and add

   those nodes as children of **node**.

Opportunities for parallelization

- cross-validation?
- hyper parameter tuning
- batch training

A little more about Interpretability and Explainability

Map-Optimize-Learn: Predicting Cardiac Pathology in Children and Teenagers with a Deep Learning Based Tabular Learning Method

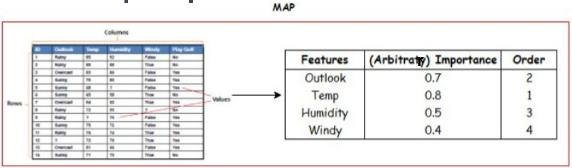
Mário T. R. Serra Neto, **Inês Dutra,** Marco Antonio F. Molinetti
Departamento de Ciência de Computadores
Faculdade de Ciências, Universidade do Porto
CINTESIS@RISE and CRACS INESC TEC

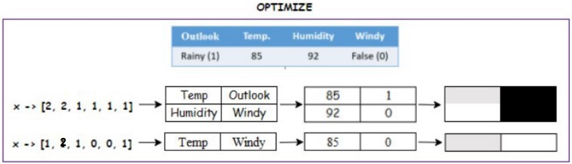
MOL: Map-Optimize-Learn

- Main goal: take advantage of the power of CNNs
  - But using tabular data

A little more about Interpretability and Explainability

MOL: Map-Optimize-Learn





M. T. R. Serra Neto, M. A. F. Mollinetti, I. Dutra, (2021). **Data Domain Change and Feature Selection to Predict Cardiac Pathology with a 2D Clinical Dataset and Convolutional Neural Networks (**Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(18), 15883-15884. https://doi.org/10.1609/aaai.v35i18.17938

M. T. R. Serra Neto, I. Dutra and M. A. F. Mollinetti, "**Map-Optimize-Learn: Predicting Cardiac Pathology in Children and Teenagers with a Deep Learning Based Tabular Learning Method,**" *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 2022, pp. 1-8, doi: 10.1109/IJCNN55064.2022.9889788.

A little more about Interpretability and Explainability

Map

- Feature selection
  - Filter
    - Use statistical metrics or information gain
  - Embedded
    - Pre-compute feature rank/importance/relevance
  - Wrapper
    - Feature selection wrapped in the ML algorithm

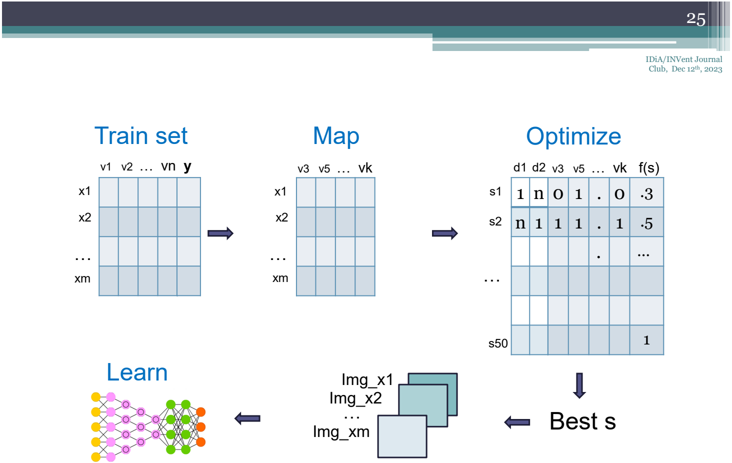
Optimize

- Swarm intelligence
  - Particle swarm optimization (PSO) – global w
  - Evolutionary PSO (EPSO) –w per particle
  - Ant-Bee Colony optimization (ABC)

  - Initialization of each vector in the population: (i is a particle, j a variable of particle i, lj and uj, bounds, Ø - random)

$$x_{ij}^{t=0} = l_j + \phi * (u_j - l_j)$$

A little more about Interpretability and Explainability



Experiments

- SI algorithms:
  - population size: 50
  - parameters:
    - PSO: w = 0.8; c1 = 1.8; c2 = 1.8
    - EPSO τ = 0.8
    - communication probability = 0.9
    - ABC: the maximum limit value is the average value between number of features and number of solutions.
- Validation: stratified 10-fold cross-validation
- Evaluation: balanced accuracy
- Comparison: Wilcoxon signed-rank test

A little more about Interpretability and Explainability

Other models

Algorithm	Parameters	Values
ANN	Training algorithm	[LBFGS, ADAM]
	Learning rate	[from 10^{-1} to 10^{-10}]
	Hidden layer neurons	[from 5 to 50 with a step of 5]
KNN	Number of neighbors	[from 2 to 10]
	Distance metric	[Euclidean, Manhattan]
	Weight metric	[Uniform, Distance]
LR	Penalty function	[L1, L2]
	Gamma	[Log space of 20 values from -4 to 4]
RF	Max. features in the best split	[1, 3, 10]
	Min. number of splits	[2, 3, 10]
	Min. samples to be in a leaf	[1, 3, 10]
	Number of estimators	[100, 300, 500]
XGB	Gamma	[0.5 to 3.0 with a 0.5 step]
	Sub samples	[0.6, 0.8, 1.0]
	Samples by tree	[0.6, 0.8, 1.0]
	Maximum depth	[2 to 5]
LGBM	Maximum number of leaves	[31, 127]
	Min. data in a leaf	[30, 50, 100, 300, 400]
	L1 and L2 regularization	[0.1, 1, 1.5]
CNN	Learning rate	[from 10^{-1} to 10^{-4}]

A little more about Interpretability and Explainability

Results

Algorithm	FS Str	Map Str	Opt Str	Test	# Feat
CNN	MOL	ANOVA	ABC	0.6442	12
CNN	**MOL**	**Distance**	**ABC**	**0.6442**	**12**
CNN	MOL	Fisher	ABC	0.6442	12
CNN	MOL	Gain Ratio	ABC	0.6442	12
CNN	MOL	Mutual Info.	ABC	0.6442	12
CNN	MOL	ANOVA	PSO	0.6442	12
CNN	**MOL**	**Distance**	**PSO**	**0.9225**	**12**
CNN	MOL	Fisher	PSO	0.6442	12
CNN	MOL	Gain Ratio	PSO	0.6442	4
CNN	MOL	Mutual Info.	PSO	0.9256	12
CNN	MOL	ANOVA	EPSO	0.9232	4
CNN	**MOL**	**Distance**	**EPSO**	**0.9250**	**4**
CNN	MOL	Fisher	EPSO	0.6442	4
CNN	MOL	Gain Ratio	EPSO	0.6442	4
CNN	MOL	Mutual Info.	EPSO	0.9209	4
ANN	**E**	-	-	**0.9038**	**6**
KNN	E	-	-	0.7310	6
LR	E	-	-	0.9030	9
RF	E	-	-	0.5178	1
XGB	E	-	-	0.8888	10
LGBM	E	-	-	0.9000	11
ANN	W	-	-	0.9038	5
KNN	W	-	-	0.9019	5
LR	**W**	-	-	**0.9058**	**5**
RF	W	-	-	0.8711	5
XGB	W	-	-	0.9042	5
LGBM	W	-	-	0.9012	12
ANN	S	-	-	**0.9050**	13
KNN	S	-	-	0.7457	13
LR	S	-	-	0.9030	13
RF	S	-	-	0.8990	13
XGB	S	-	-	0.8888	13
LGBM	**S**	-	-	**0.9059**	**13**
TabNet	-	-	-	0.9147	-

A little more about Interpretability and Explainability

Results

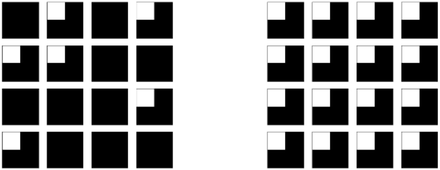
Algorithm	FS Str	Map Str	Opt Str	Test	# Feat
CNN	MOL	ANOVA	ABC	0.6442	12
CNN	**MOL**	**Distance**	**ABC**	**0.6442**	**12**
CNN	MOL	Fisher	ABC	0.6442	12
CNN	MOL	Gain Ratio	ABC	0.6442	12
CNN	MOL	Mutual Info.	ABC	0.6442	12
CNN	MOL	ANOVA	PSO	0.6442	12
CNN	**MOL**	**Distance**	**PSO**	**0.9225**	**12**
CNN	MOL	Fisher	PSO	0.6442	12
CNN	MOL	Gain Ratio	PSO	0.6442	4
CNN	MOL	Mutual Info.	PSO	0.9256	12
CNN	MOL	ANOVA	EPSO	0.9232	4
CNN	**MOL**	**Distance**	**EPSO**	**0.9250**	**4**
CNN	MOL	Fisher	EPSO	0.6442	4
CNN	MOL	Gain Ratio	EPSO	0.6442	4
CNN	MOL	Mutual Info.	EPSO	0.9209	4
ANN	E	-	-	**0.9038**	**6**
KNN	E	-	-	0.7310	6
LR	E	-	-	0.9030	9
RF	E	-	-	0.5178	1
XGB	E	-	-	0.8888	10
LGBM	E	-	-	0.9000	11
ANN	W	-	-	0.9038	5
KNN	W	-	-	0.9019	5
LR	**W**	-	-	**0.9058**	**5**
RF	W	-	-	0.8711	5
XGB	W	-	-	0.9042	5
LGBM	W	-	-	0.9012	12
ANN	S	-	-	**0.9050**	13
KNN	S	-	-	0.7457	13
LR	S	-	-	0.9030	13
RF	S	-	-	0.8990	13
XGB	S	-	-	0.8888	13
LGBM	**S**	-	-	**0.9059**	**13**
TabNet	-	-	-	0.9147	-

A little more about Interpretability and Explainability

Results

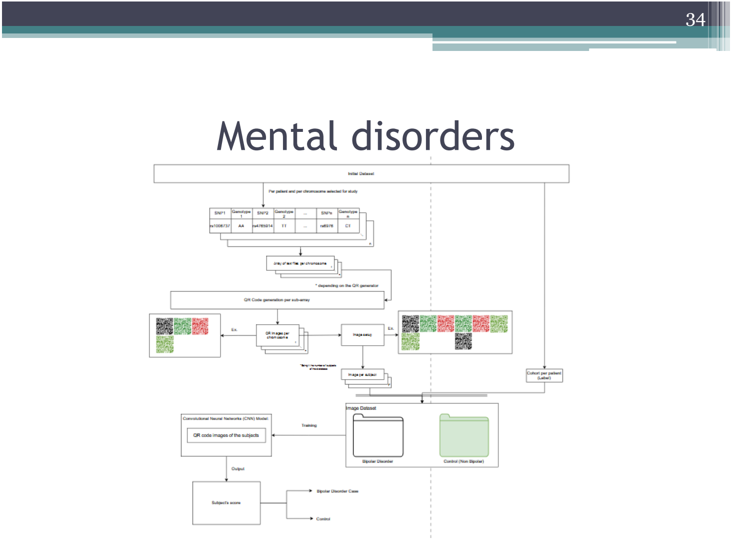
- Best combination MOL+Distance+EPSO
  - Balanced accuracy: 0.925

Examples of controls | Examples of cases



- Best model in the literature naïve Bayes: 0.93 (acc not balanced)
- Better than tabnet: 0.91 (balanced accuracy)

A little more about Interpretability and Explainability



A little more about Interpretability and Explainability

Collaboration

Data

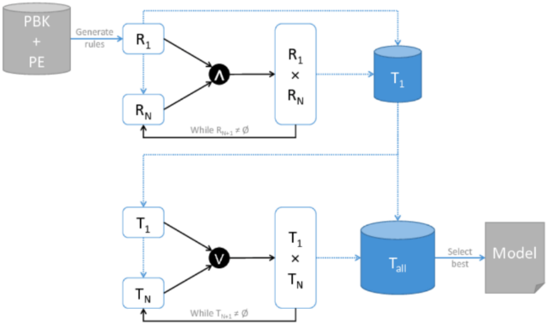
- WTCCC

Transforming patient genetic sequences to QR codes

Alberto Pinheira, Manuel Casal-Guisande, Alberto Comesaña-Campos, Inês Dutra, Camila Nascimento and Jorge Cerqueiro-Pequeño, Proposal and Definition of a Novel Intelligent System for the Diagnosis of Bipolar Disorder based on the use of Quick Response codes containing Single Nucleotide Polymorphism data. Technological Ecosystems for Enhancing Multiculturality (TEEM 2023).

A little more about Interpretability and Explainability

Probabilistic ILP naive search



Guided search: SkILL
Stochastic Inductive Logic Learner

- Fitness pruning
- Estimation pruning
- Prediction pruning

- J. Côrte-Real, I. Dutra, R. Rocha. Pruning strategies for the efficient traversal of the search space in PILP environments. Knowledge and Information Systems, 2021, 63(12):3183-3215.
- J. Côrte-Real, A. Dries, I. Dutra and R. Rocha. Improving Candidate Quality of Probabilistic Logic Models 34th International Conference on Logic Programming (ICLP 2018) - Technical Communications. Oxford, UK, July 2018
- J. Côrte-Real, I. Dutra, and R. Rocha. Estimation-based search space traversal in PILP environments. In J. Cussens and A. Russo, editors, 26th International Conference on Inductive Logic Programming (ILP 2016), volume 10326 of LNAI
- Joana Côrte-Real and Theofrastos Mantadelis and Inês Dutra and Ricardo Rocha and Elizabeth Burnside "SkILL - a Stochastic Inductive Logic Learner", in "14th IEEE International Conference on Machine Learning and Applications (ICMLA 2015)", IEEE, December 2015

A little more about Interpretability and Explainability