GENERAL CONFERENCE
FAIR Future Artificial Intelligence Research
23|24 SET 2024
NAPOLI

SPOKE 6 / WP 6

DISTRIBUTED HETEROGENEOUS TRANSFER LEARNING

Paolo Mignone, Gianvito Pio, Michelangelo Ceci

INTRODUCTION

In many machine learning applications, it is **difficult or expensive to obtain training data** described through the <u>same feature space</u> and following the <u>same data distribution</u> of the examples where the predictive model will be applied.

Transfer Learning

Learn a predictive function for a **target domain** by exploiting also data from a separate, but related domain, called **source domain.**

The adoption of **transfer learning** techniques also increases the **sustainability of the training process**, since:

- may **reduce the human resources** required to gather labeled data
- may **reduce the computational resources**, by reusing models already trained in other contexts

BACKGROUND

A ***domain*** is defined as $D = \{F, P(X)\}$, where F is a feature space, X is a set of observations, $P(X)$ is the marginal probability distribution over X.

A ***task*** is defined as $T = \{Y, f\}$, where Y is the output space of the prediction task; f is a predictive function learned from a set of training examples in the form $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in Y$.

$D_s = \{F_s, P(X_s)\} \rightarrow$ **source domain** $\quad T_s = \{Y_s, f_s\} \rightarrow$ **source task**

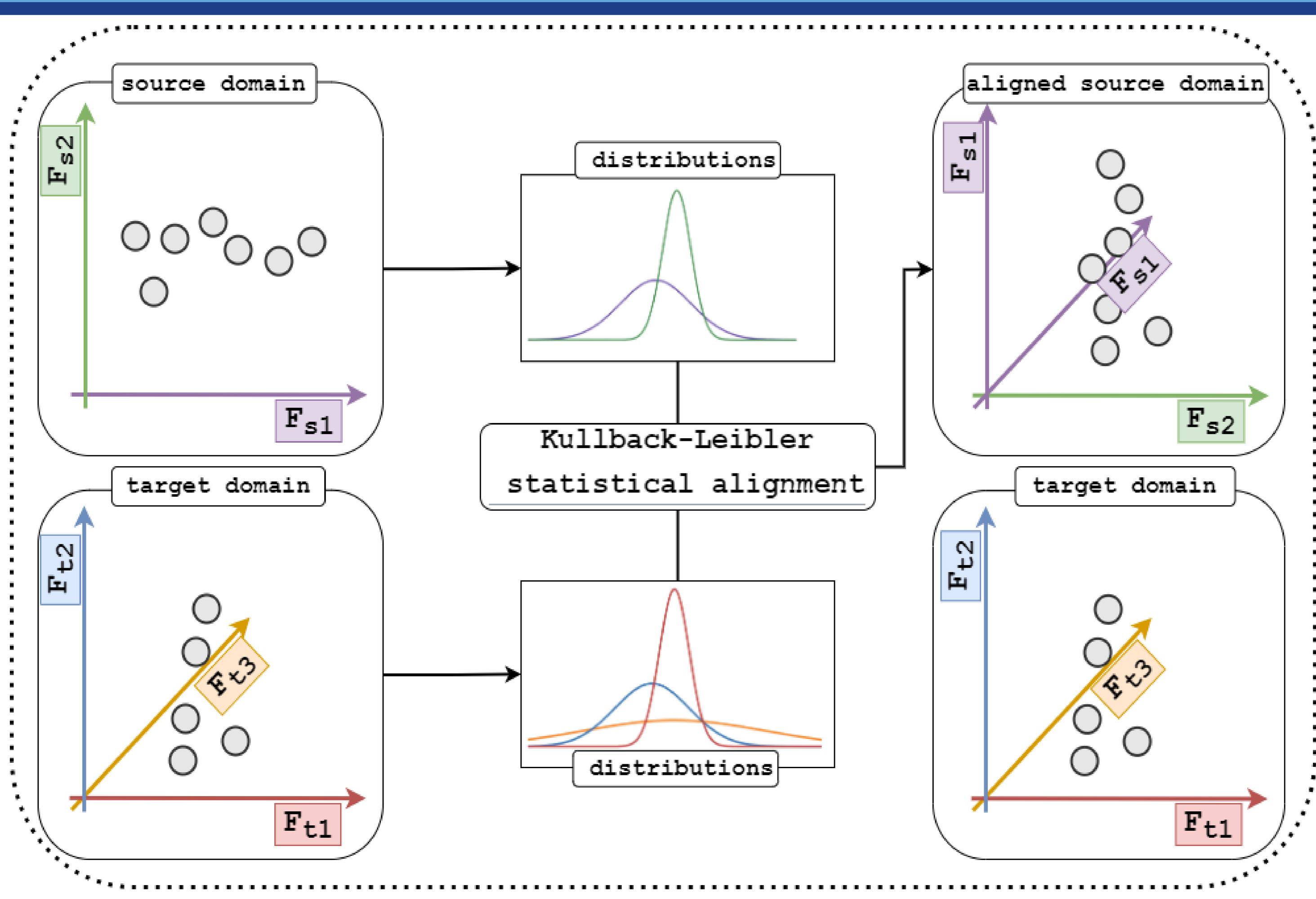
$D_t = \{F_t, P(X_t)\} \rightarrow$ **target domain** $\quad T_t = \{Y_t, f_t\} \rightarrow$ **target task**

Considered setting: Heterogeneous Transfer Learning

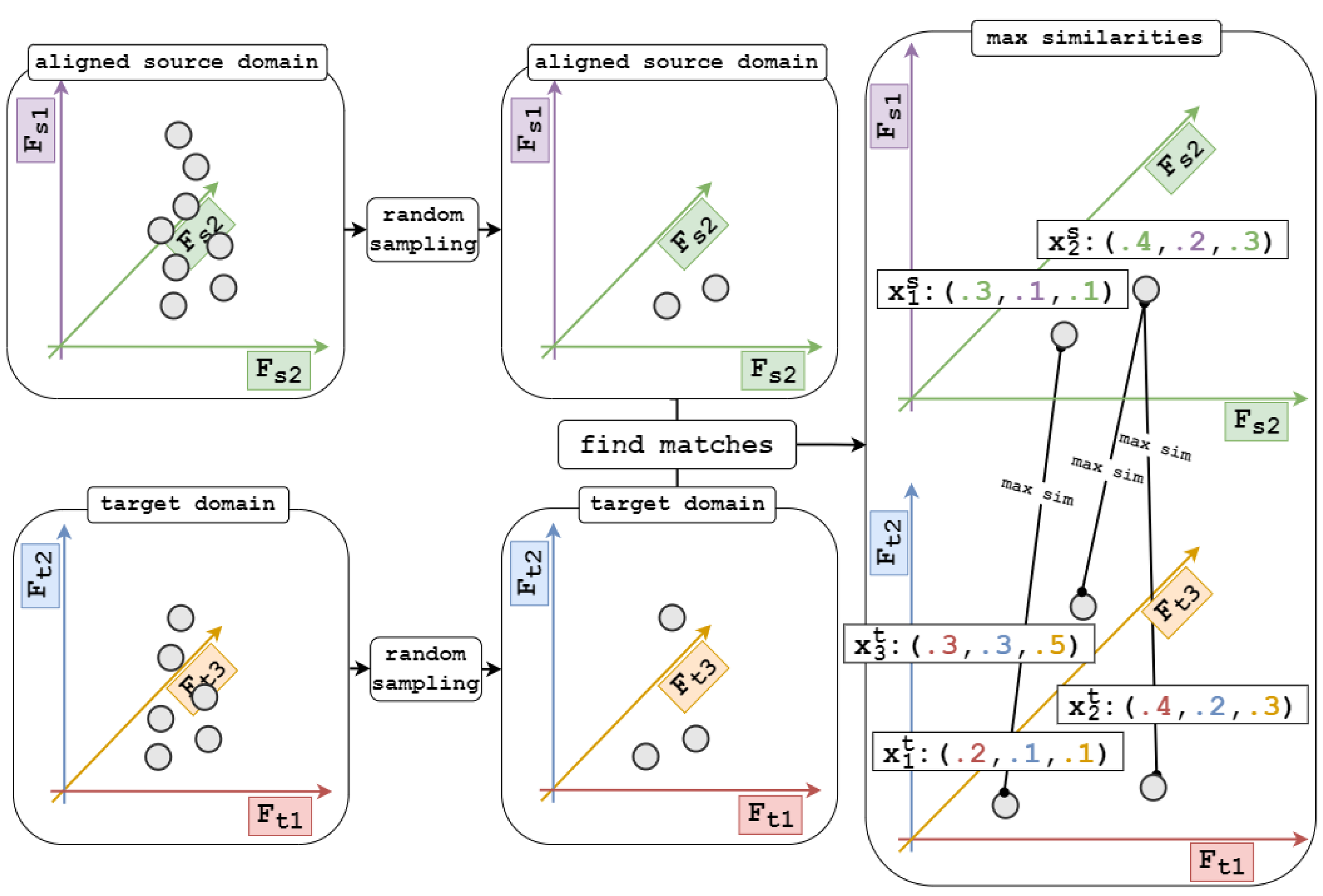
$F_s \neq F_t$ and $P(X_s) \neq P(X_t)$

THE PROPOSED METHOD STEAL

STAGE 1 – FEATURE ALIGNMENT



STAGE 2 – SOURCE-TARGET INSTANCE MATCHING

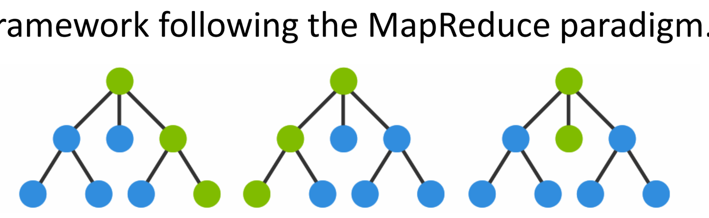


hybrid instances

x^t_1: (.2,.1,.1)	x^s_1: (.3,.1,.1)
x^t_2: (.4,.2,.3)	x^s_2: (.4,.2,.3)
x^t_3: (.3,.3,.5)	x^s_2: (.4,.2,.3)

STAGE 3 – DISTRIBUTED MODEL TRAINING

All the stages are implemented using the Apache Spark framework following the MapReduce paradigm.



The final predictive model is learned from the obtained hybrid dataset using a distributed version of Random Forests available in Apache Spark.

DATASETS

GENE REGULATORY NETWORK RECONSTRUCTION (BIOINFORMATICS)

TASK: LINK PREDICTION
SETTING: POSITIVE-UNLABELED LEARNING
TARGET DOMAIN: HOMO SAPIENS ORGANISM
SOURCE DOMAIN: MOUSE ORGANISM

	HET		HOM		HOM-RED	
	Human	Mouse	Human	Mouse	Human	Mouse
Positive interactions	235,706	14,613	235,706	14,613	4,714	4,714
Unlabeled interactions	235,706	235,706	235,706	235,706	4,714	4,714
Gene features	174	161	6	6	6	6
Gene-pair features	348	322	12	12	12	12

Available Dataset: https://data.d4science.net/xQ7P

CEREBRAL STROKE DETECTION (MEDICAL)

TASK: BINARY CLASSIFICATION
SETTING: SUPERVISED LEARNING
TARGET DOMAIN: CEREBRAL STROKE IN HOSPITAL PATIENTS
SOURCE DOMAIN: SEPSIS IN HOSPITAL PATIENTS

	FULL		REDUCED	
# instances	Stroke	Sepsis	Stroke	Sepsis
relapsed_stroke/died	643	11,735	643	1,173
single_stroke/survived	41,288	117,657	643	1,173
total	41,931	129,392	1,286	2,346

Available Dataset: https://data.d4science.net/eEn3

ENERGY CONSUMPTION FORECASTING

TASK: REGRESSION
SETTING: SUPERVISED LEARNING
TARGET DOMAIN:
ENERGY CONSUMPTION OF A SET OF CLIENTS
SOURCE DOMAIN:
ENERGY CONSUMPTION OF ANOTHER SET OF CLIENTS

Fold	Training period	Testing period	Source Training instances	Target Training instances	Testing instances
1	2010-2011	2012-2019	924	912	14,688
2	2010-2012	2013-2019	1,848	1824	12,852
3	2010-2013	2014-2019	2,772	2,736	11,016
4	2010-2014	2015-2019	3,696	3,648	9,180
5	2010-2015	2016-2019	4,620	4,560	7,344
6	2010-2016	2017-2019	5,544	5,472	5,508
7	2010-2017	2018-2019	6,468	6,384	3,672
8	2010-2018	2019	7,392	7,296	1,836

RESULTS

GENE REGULATORY RECONSTRUCTION – PU LEARNING

HETEROGENEOUS

Method	AUR@K	Impr. over T
T (no transfer)	0.610	-
STEAL	0.679	11.3%

HOMOGENEOUS

Method	AUR@K	Impr. over T
T (no transfer)	0.533	-
S (optimal feature alignment)	0.544	2.1%
T+S (optimal feature alignment)	0.551	3.4%
STEAL	0.680	27.6%

REDUCED

Method	HOM-RED
JGSA	0.500
TJM	0.554
BDA	0.558
JDOT	0.540
STEAL	0.589

CEREBRAL STROKE DETECTION - CLASSIFICATION

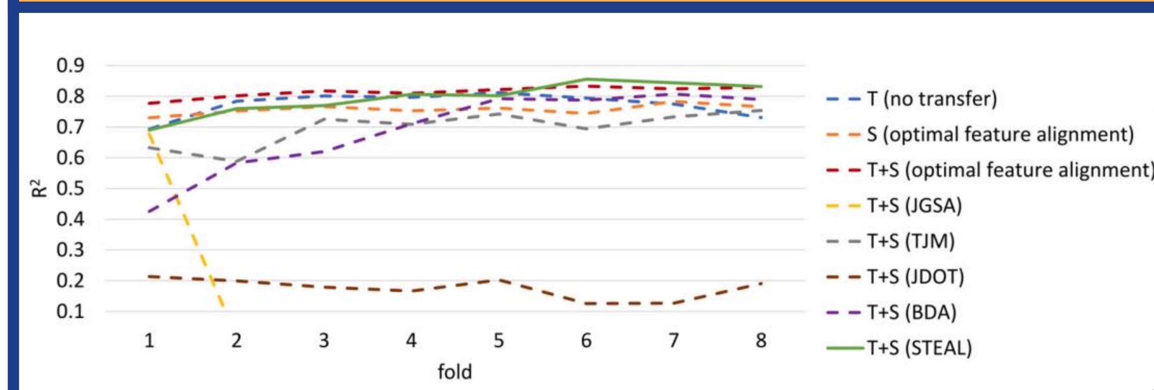
HETEROGENEOUS

Method	Acc	Macro Prec	Macro Rec	Macro F1-score	Weighted Prec	Weighted Rec	Weighted F1-score
T (no transfer)	0.902	0.526	0.647	0.528	0.975	0.902	0.935
STEAL	0.942	0.530	0.597	0.540	0.974	0.942	0.957

REDUCED

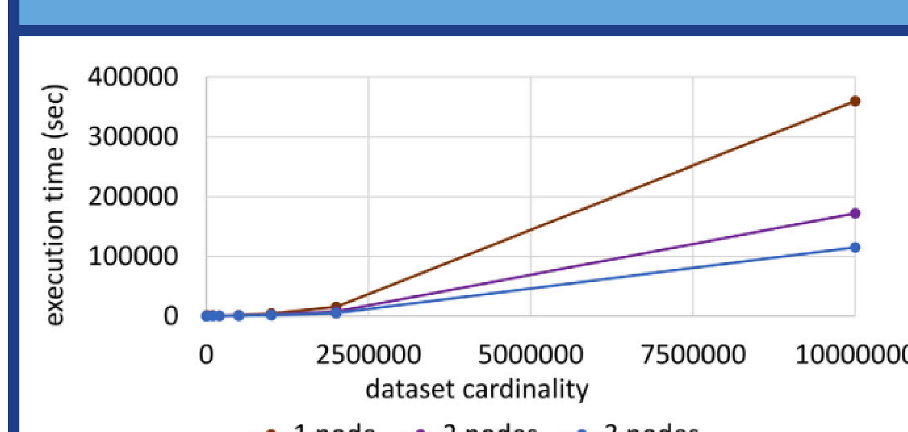
Method	Acc	Macro Prec	Macro Rec	Macro F1-score	Weighted Prec	Weighted Rec	Weighted F1-score
JGSA	0.575	0.589	0.575	0.557	0.589	0.575	0.557
TJM	0.655	0.656	0.655	0.654	0.656	0.655	0.654
BDA	0.674	0.675	0.674	0.673	0.675	0.674	0.674
JDOT	0.551	0.551	0.551	0.550	0.551	0.551	0.550
STEAL	0.768	0.773	0.768	0.767	0.773	0.768	0.767

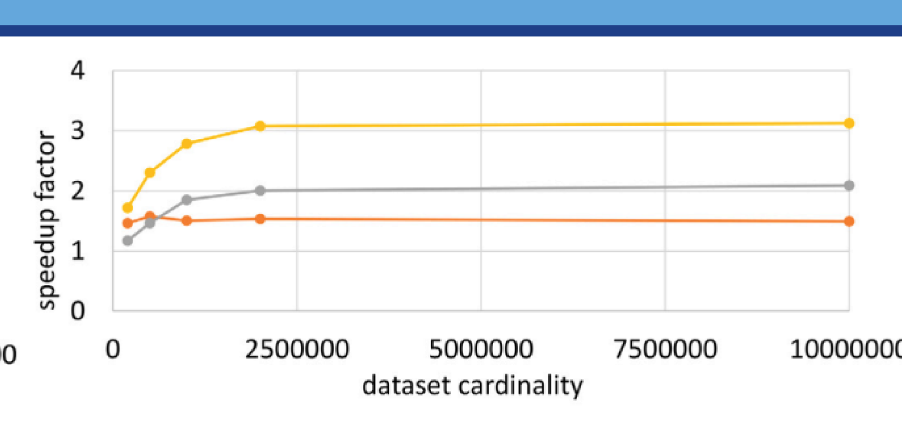
ENERGY CONSUMPTION FORECASTING - REGRESSION



Method	RMSE	R^2
T (no transfer)	154.622	0.723
S (optimal feature alignment)	159.418	0.707
T+S (optimal feature alignment)	142.928	0.764
JGSA	447.778	-1.809
TJM	174.461	0.647
BDA	174.213	0.639
JDOT	275.516	0.125
STEAL	148.008	0.745

SCALABILITY ANALYSIS





REFERENCE

Paolo Mignone, Gianvito Pio, Michelangelo Ceci, *Distributed Heterogeneous Transfer Learning*, Big Data Research, Volume 37, 2024, 100456, ISSN 2214-5796, https://doi.org/10.1016/j.bdr.2024.100456
Available Software: https://figshare.com/articles/software/STEAL/19482533

Funded by the European Union NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research