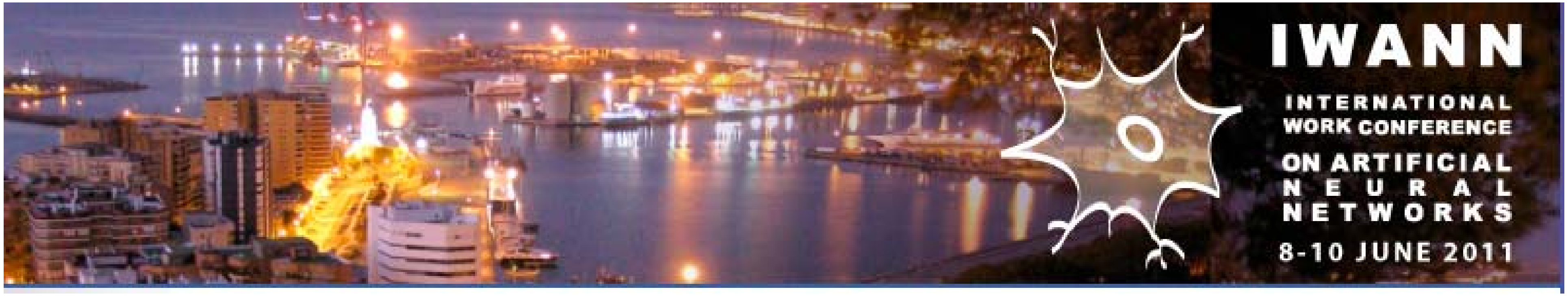


Dataset Shift in Classification: Approaches and Problems

Francisco Herrera

Research Group on Soft Computing and Information Intelligent Systems (SCI^{2}S)
http://sci2s.ugr.es
Dept. of Computer Science and A.I.
University of Granada, Spain

Email: herrera@decsai.ugr.es
http://decsai.ugr.es/~herrera

Why is difficult to learn from DATA?

Intrinsic data characteristics

Imbalanced data sets

Overlapping

Small disjuncts/rare data sets

Density: Lack of data

Noise and Bordelne data

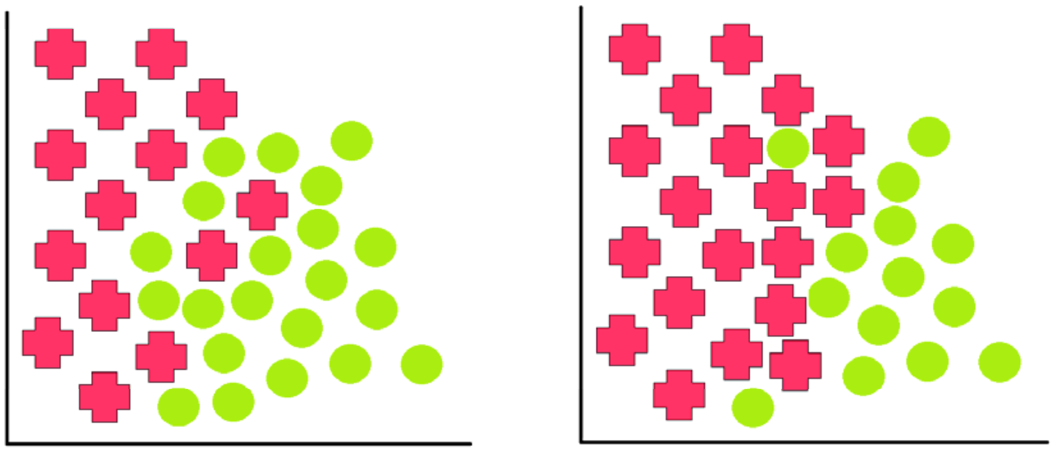
…

Dataset shift

Dataset Shift

when training and test

distributions are different



Contents

I. The Problem of Dataset Shift

II. Dataset Shift: A Terminology Review

III. Characterizing the change. Types of Dataset Shift

IV. Causes of Dataset Shift

V. Dataset Shift: Some approaches

VI. A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

VII. Dataset Shift: Final comments

Contents

I. The Problem of Dataset Shift

II. Dataset Shift: A Terminology Review

III. Characterizing the change. Types of Dataset Shift

IV. Causes of Dataset Shift

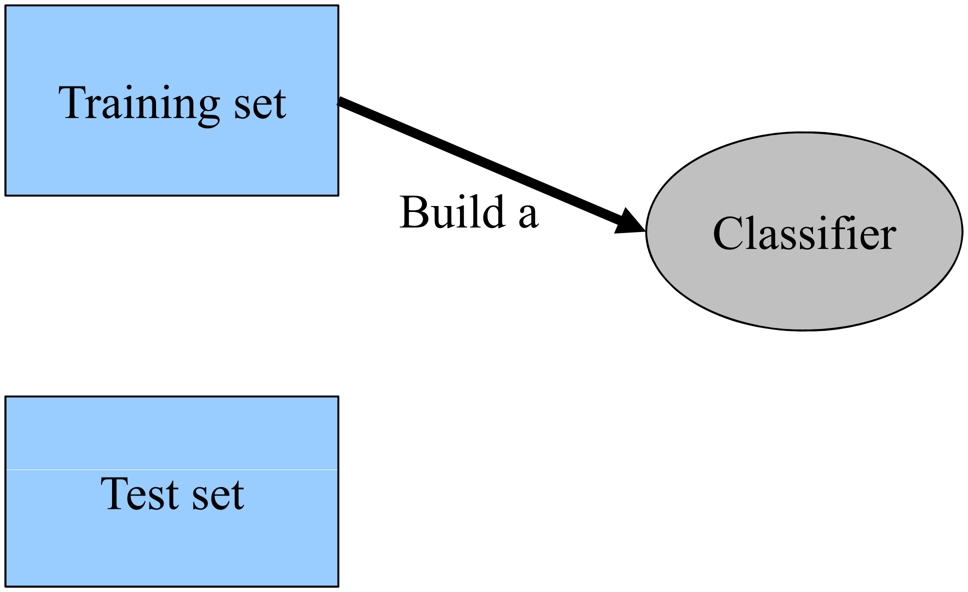
V. Dataset Shift: Some approaches

VI. A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

VII. Dataset Shift: Final comments

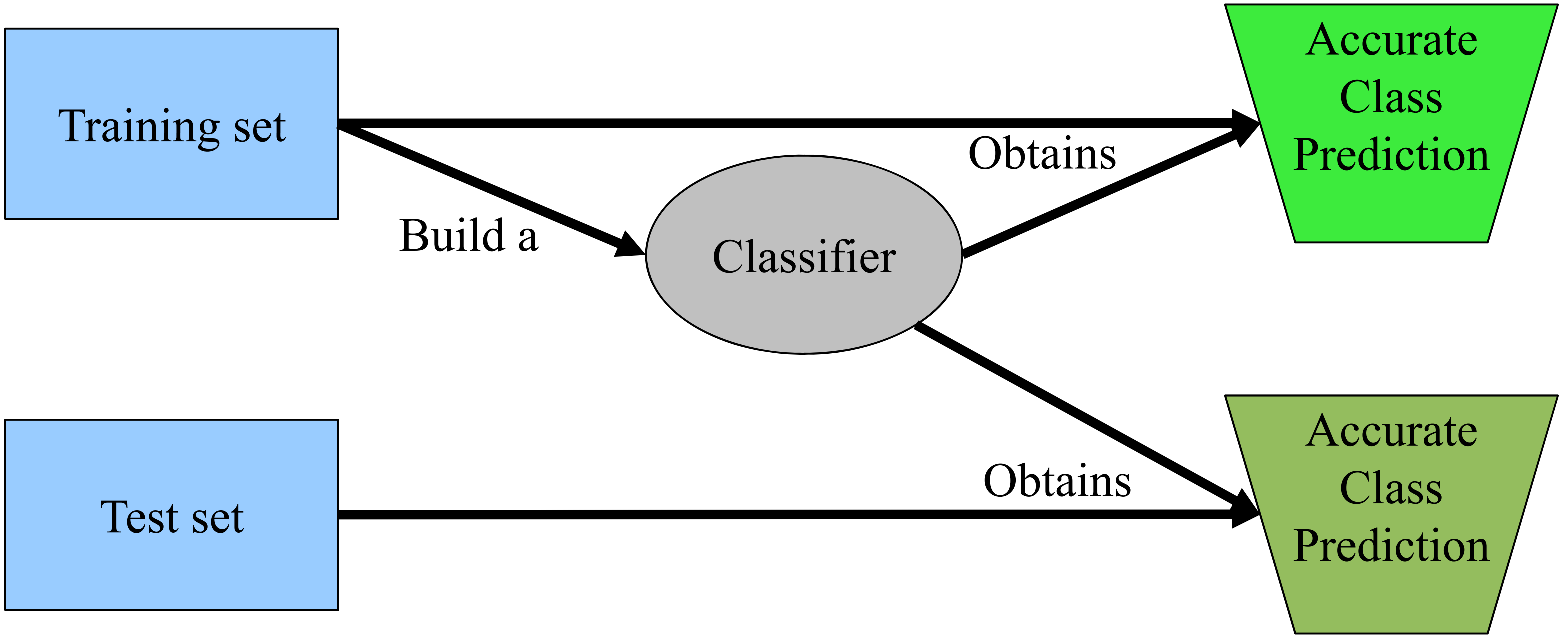
The Problem of Dataset Shift

- **Basic assumption in classification:**



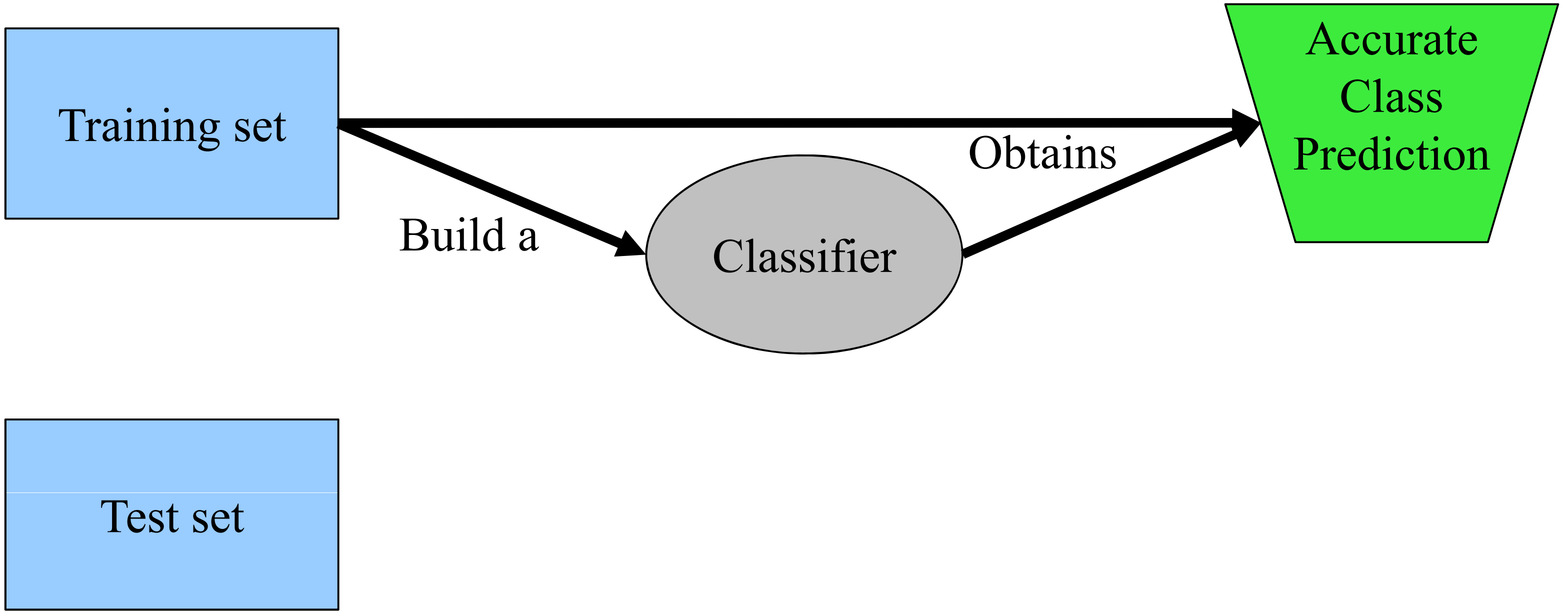
The Problem of Dataset Shift

- **Basic assumption in classification:**



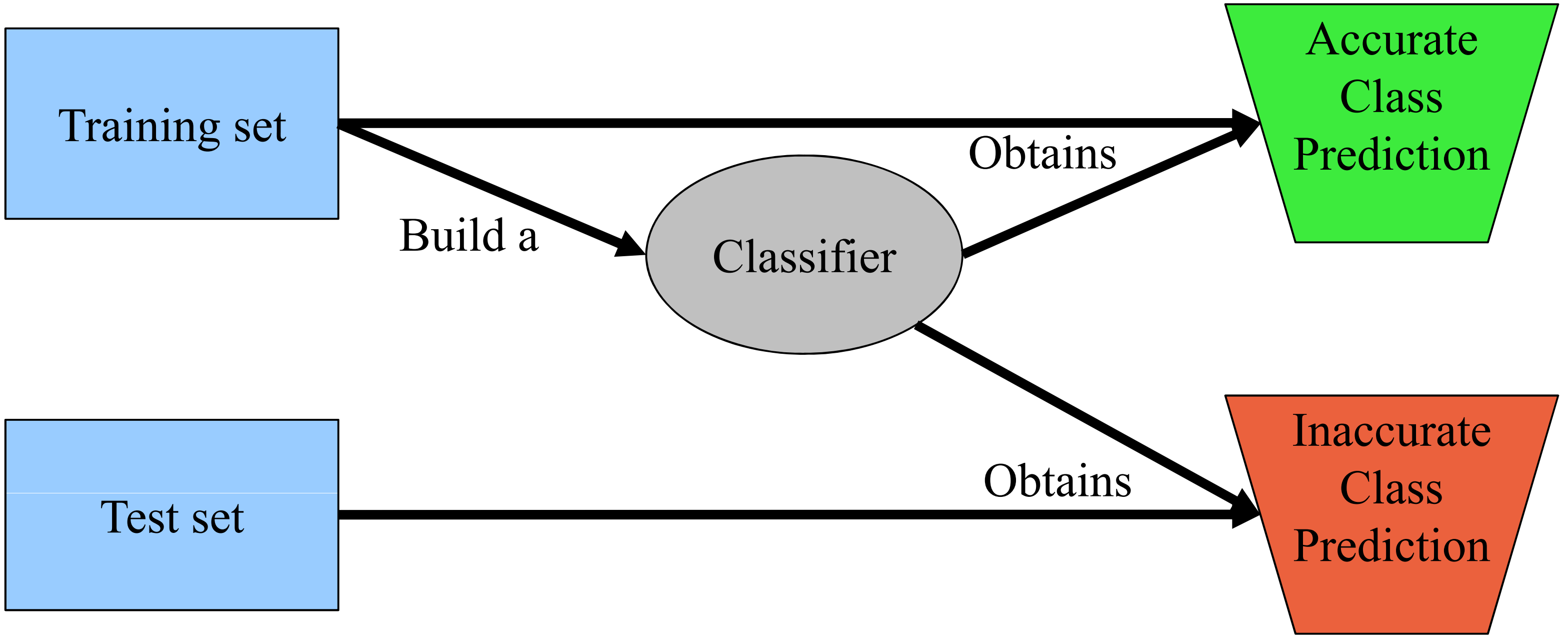
The Problem of Dataset Shift

- **But sometimes....**



The Problem of Dataset Shift

- **But sometimes....**



- **The classifier has an overfitting problem.**
- **Is there a change in data distribution between training and test sets (Data fracture)?.**

The Problem of Dataset Shift

- **The classifier has an overfitting problem.**
  - **Change the parameters of the algorithm.**
  - **Use a more general learning method.**

- **There is a change in data distribution between training and test sets (Dataset shift).**
  - **Train a new classifier for the test set.**
  - **Adapt the classifier.**
  - **Modify the data in the test set …**

Contents

I. The Problem of Dataset Shift

II. Dataset Shift: A Literature Review

III. Characterizing the change. Types of Dataset Shift

IV. Causes of Dataset Shift

V. Dataset Shift: Some approaches

VI. A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

VII. Dataset Shift: Final comments

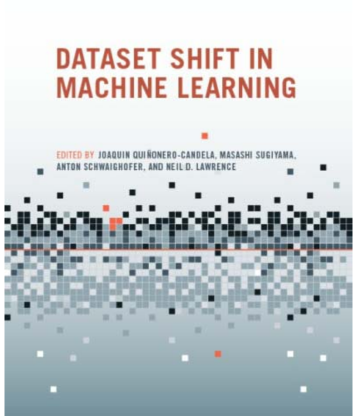
Dataset Shift: A literature review

Lack of a standard term

- Data fracture (Cieslak & Chawla)
- Dataset shift (Quiñonero et al.)
- Changing environments (Alaiz-Rodriguez & Japkowitz)

The term “Dataset Shift” was first used in the book

J. Quiñonero Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. **Shift in Machine Learning.** The MIT Press, 2009.


DATASET SHIFT IN MACHINE LEARNING
EDITED BY JOAQUIN QUIÑONERO-CANDELA, MASASHI SUGIYAMA, ANTON SCHWAIGHOFER, AND NEIL D. LAWRENCE

the first compilation on the field, where it was defined as “cases where the joint distribution of inputs and outputs differs between training and test stage.”

Dataset Shift: A literature review

As an example, the following terms have been used in the literature to refer to Dataset Shift:

“Concept shift” or “concept drift” where the idea of different data distributions is associated with changes in the class definitions (i.e. the “concept” to be learned) (Widmer, Jubat, 1996)

“Changes of classification”, where it is defined as “In the change mining problem, we have an old classifier, representing some previous knowledge about classification, and a new data set that has a changed class distribution.” (Wang, Zhow, Fu, Yu, Jeffrey, Yu, 2003).

Dataset Shift: A literature review

As an example, the following terms have been used in the literature to refer to Dataset Shift:

“Changing environments”, defined as “The fundamental assumption of supervised learning is that the joint probability distribution p(x/d) will remain unchanged between training and testing. There are, however, some mismatches that are likely to appear in practice.” (Alaiz-Rodríguez, Japkowicz, 2008)

“Contrast mining in classification learning”, a slightly different take on the issue: “Given two groups of interest, a user often needs to know the following. Do they represent different concepts? To what degree do they differ? What is the discrepancy and where does it originate from?” (Yang, Wu, Zhu, 2008)

Dataset Shift: A literature review

As an example, the following terms have been used in the literature to refer to Dataset Shift:

“Fracture points”, defined as “fracture points in predictive distributions and alteration to the feature space, where a fracture is considered as the points of failure in classifiers’ predictions - deviations from the expected or the norm.” (Cieslak, Chawla, 2009)

“Fractures between data”, defined as the case where “we have data from one laboratory (dataset A), and derive a classifier from it that can predict its category accurately. We are then presented with data from a second laboratory (dataset B). This second dataset is not accurately predicted by the classifier we had previously built due to a fracture between the data of both laboratories.” (Moreno-Torres, Llorà, Goldberg, Bhargava, 2011).

Dataset Shift: A literature review

As an example, the following terms have been used in the literature to refer to Dataset Shift: "Concept shift" or "concept drift", "Changes of classification", "Changing environments", "Contrast mining in classification learning", "Fracture points" and "Fractures between data".

Such inconsistent terminology is a disservice to the field as it makes literature searches difficult and confounds the discussion of this important problem. We recommend the term ***Dataset Shift*** for any situation in which training and test data follow distributions that are in some way different. Formally, we define it as

Definition 1. ***Dataset shift appears when training and test joint distributions are different.*** That is, when $P_{tra}(y, x) \neq P_{tst}(y, x)$

J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N.V. Chawla, F. Herrera A Unifying view of Dataset Shift in Classification. Pattern Recognition, 2011, In press.

Contents

I. The Problem of Dataset Shift

II. Dataset Shift: A Literature Review

III. Characterizing the change. Types of Dataset Shift

IV. Causes of Dataset Shift

V. Dataset Shift: Some approaches

VI. A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

VII. Dataset Shift: Final comments

Characterizing the change. Types of Dataset Shift

Characterizing the change:

Covariate shift

Prior probability shift

Concept shift

Characterizing the change. Types of Dataset Shift

Notation:

- A set of features or covariates x
- A target variable y (the class variable)
- A joint distribution P(x,y)

Focus on the problems:

1. Prediction problem: Given X predict Y: X → Y
2. Class label causally determines de values of the covariates Y → X.

The joint distribution P(y,x) can be written as:

P(y/x) P(x) in X → Y problems.

P(x/y) P(y) in Y → X problems.

Characterizing the change. Types of Dataset Shift

Covariate shift

The term covariate shift was first defined ten years ago by (Shimodaira, 2000), where it refers to changes in the distribution of the input variables x.

Covariate shift is probably the most studied type of shift, but there appears to be some confusion in the literature about the exact definition of the term. There are also some equivalent names, such as “population drift”, “Change in the data distributions”, “Differing training and test distributions”, ...

Characterizing the change. Types of Dataset Shift

Covariate shift

(Storkey, 2009) defines covariate shift as something that occurs "when the data is generated according to a model P(y|x)P(x) and where the distribution P(x) changes between training and test scenarios." This seems to capture the essence of the term as it is most commonly used.

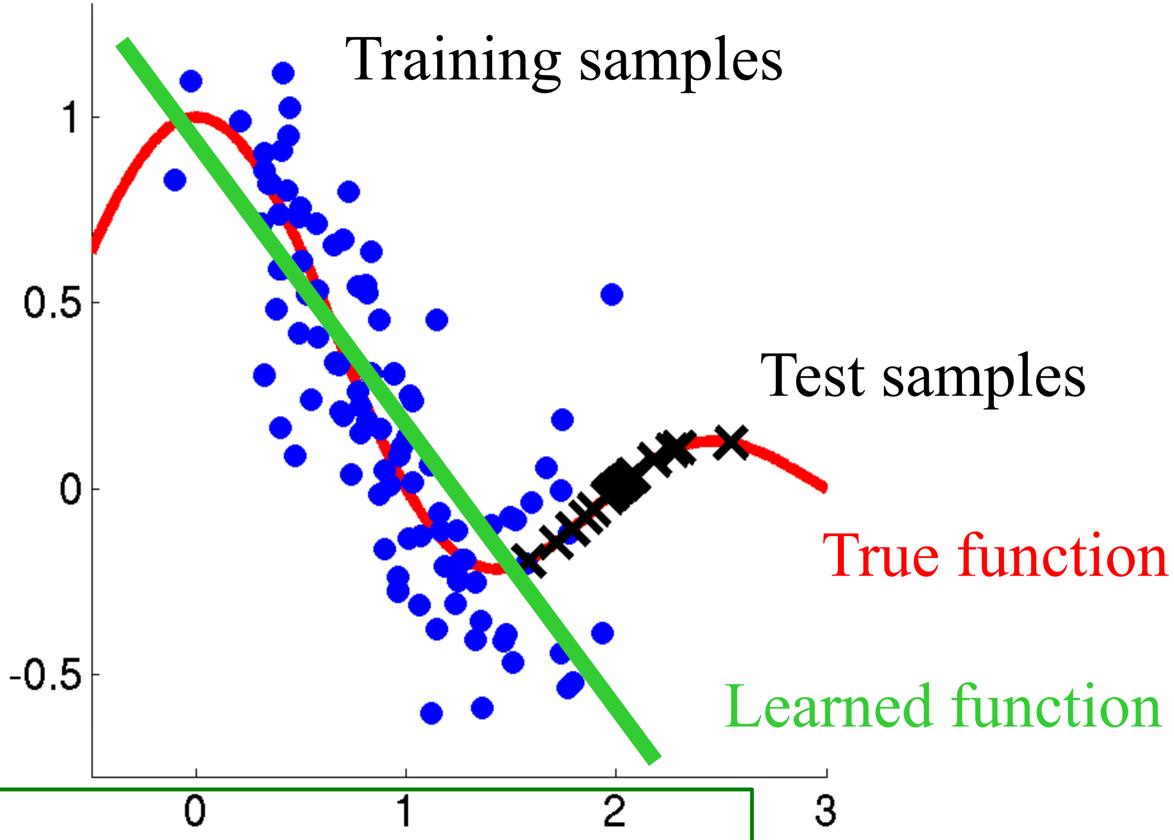
We propose the following as a consistent formal definition.

Definition 2. Covariate shift appears only in X→Y problems, and is defined as the case where $P_{tra}(y|x) = P_{tst}(y|x)$ and $P_{tra}(x) \neq P_{tst}(x)$.

Characterizing the change. Types of Dataset Shift

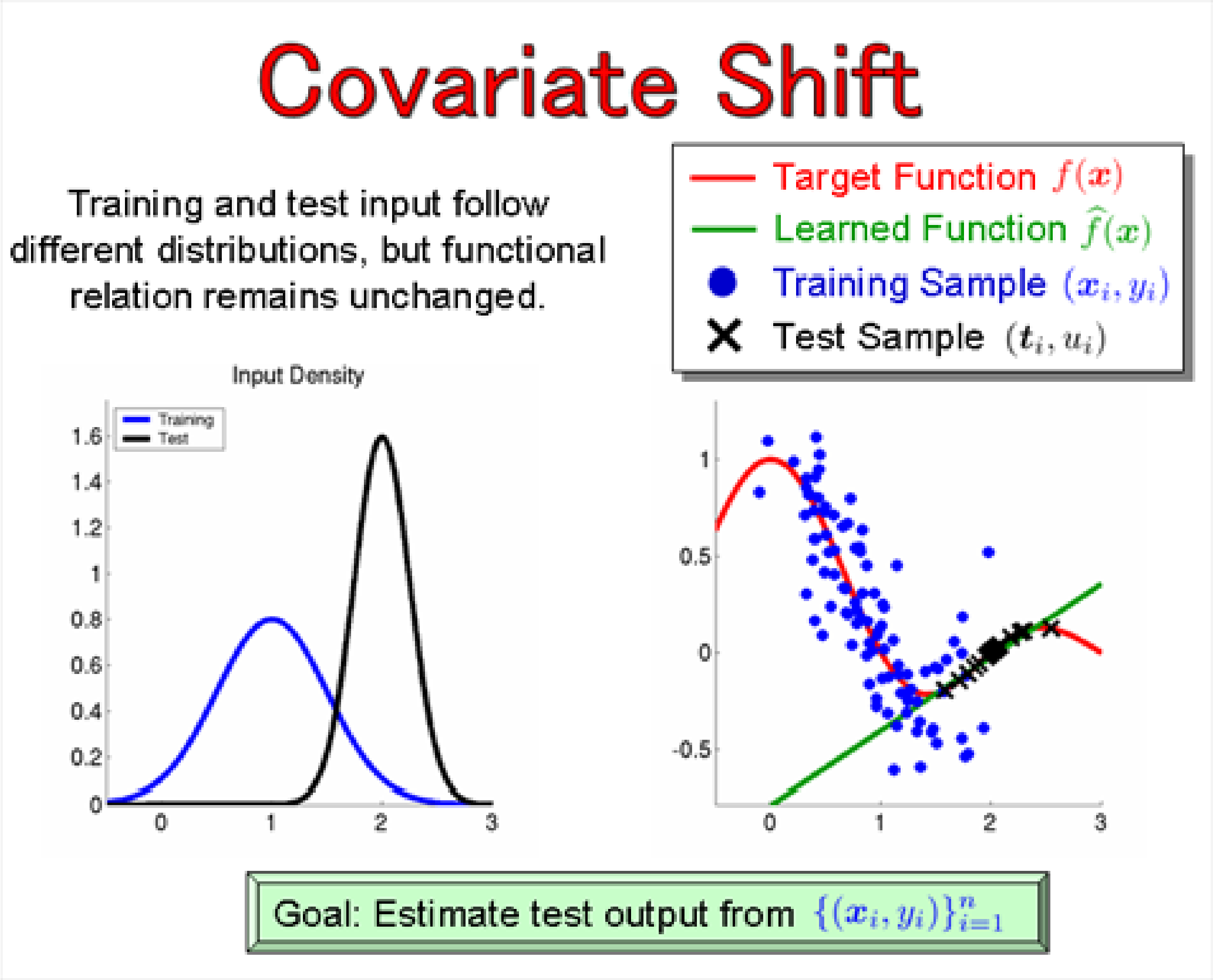
Covariate shift

To illustrate the effect of covariate shift, let’s focus on linear extrapolation



© Masashi Sugiyama,
http://sugiyama-www.cs.titech.ac.jp/~sugi/figs/covariate-shift.png

Characterizing the change. Types of Dataset Shift



Covariate Shift

Training and test input follow different distributions, but functional relation remains unchanged.

Goal: Estimate test output from $\{(x_i, y_i)\}_{i=1}^{n}$

© Masashi Sugiyama,
http://sugiyama-www.cs.titech.ac.jp/~sugi/figs/covariate-shift.png

Characterizing the change. Types of Dataset Shift

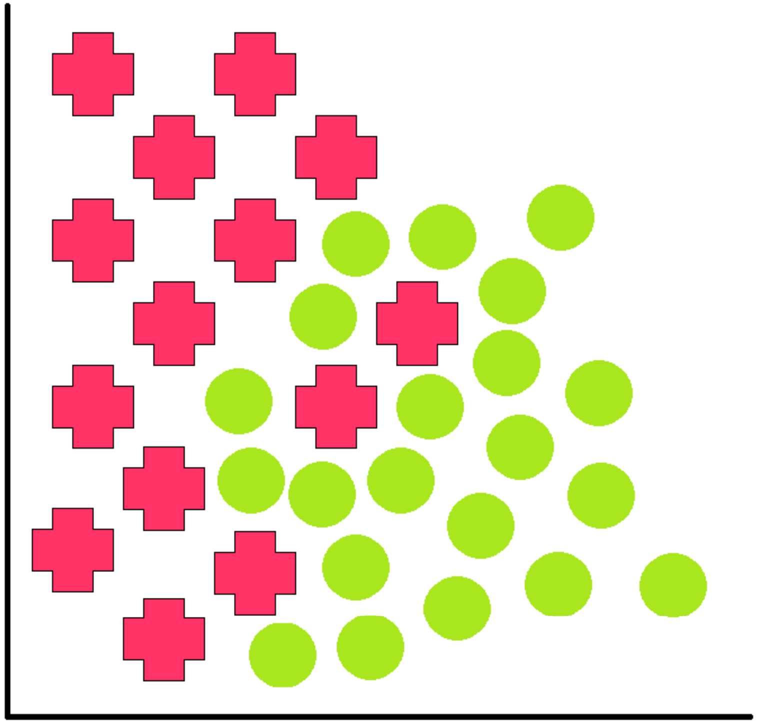
Prior probability shift

Prior probability shift refers to changes in the distribution of the class variable y. It also appears with different names in the literature and the definitions have slight differences between them.

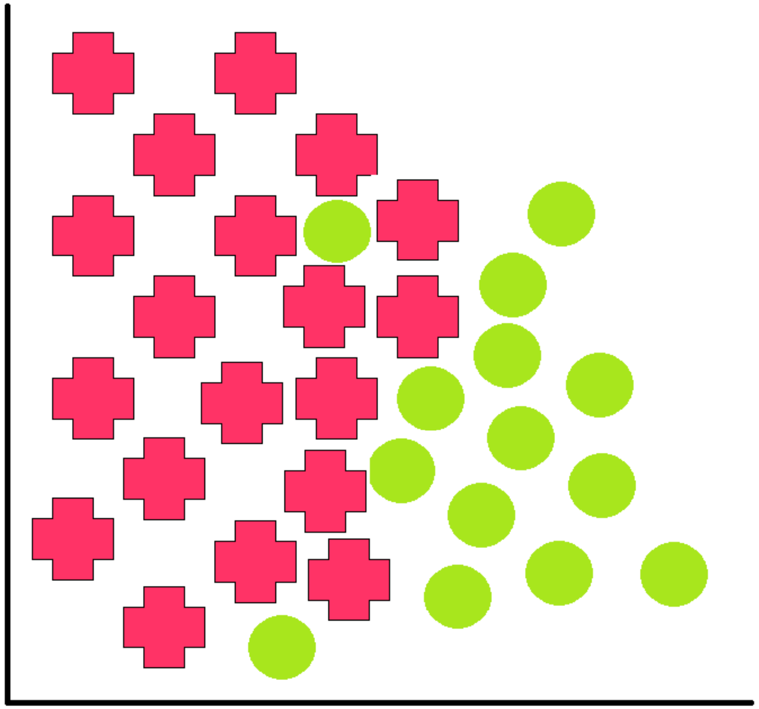
Definition 3. Prior probability shift appears only in Y→X problems, and is defined as the case where $P_{tra}(x|y) = P_{tst}(x|y)$ and $P_{tra}(y) \neq P_{tst}(y)$.

Characterizing the change. Types of Dataset Shift

Prior probability shift



Training



Test

Characterizing the change. Types of Dataset Shift

Concept shift

Concept shift happens when the relationship between the input and class variables change, which presents the hardest challenge among the different types of Dataset Shift that have been tackled so far (referred to as ‘concept drift’ in the literature) . Formally, we define it as:

Definition 4. Concept shift is defined as

- $P_{tra}(y|x) \neq P_{tst}(y|x)$ and $P_{tra}(x) = P_{tst}(x)$ in X→Y problems.
- $P_{tra}(x|y) \neq P_{tst}(x|y)$ and $P_{tra}(y) = P_{tst}(y)$ in Y→X problems.

Contents

I. The Problem of Dataset Shift

II. Dataset Shift: A Literature Review

III. Characterizing the change. Types of Dataset Shift

IV. Causes of Dataset Shift

V. Dataset Shift: Some approaches

VI. A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

VII. Dataset Shift: Final comments

Causes of Dataset Shift

We comment on some of the most common causes of Dataset Shift:

Sample selection bias and non-stationary environments.

These concepts have created confusion at times, so it is important to remark **that these terms are factors that can lead to the appearance of some of the shifts explained, but they do not constitute Dataset Shift themselves.**

Causes of Dataset Shift

Sample selection bias the discrepancy in distribution is due to the fact that the training examples have been obtained through a biased method, and thus do not represent reliably the operating environment where the classifier is to be deployed (which, in machine learning terms, would constitute the test set).

Non-stationary environments. It appears when the training environment is different from the test one, whether it is due to a temporal or a spatial change.

Causes of Dataset Shift

Sample selection bias: The term Sample selection bias refers to a systematic flaw in the process of data collection or labeling which causes training examples to be selected non-uniformly from the population to be modeled.

The term has been used as a synonym of covariate shift (which is not correct), but also on its own as a related problem to Dataset Shift.

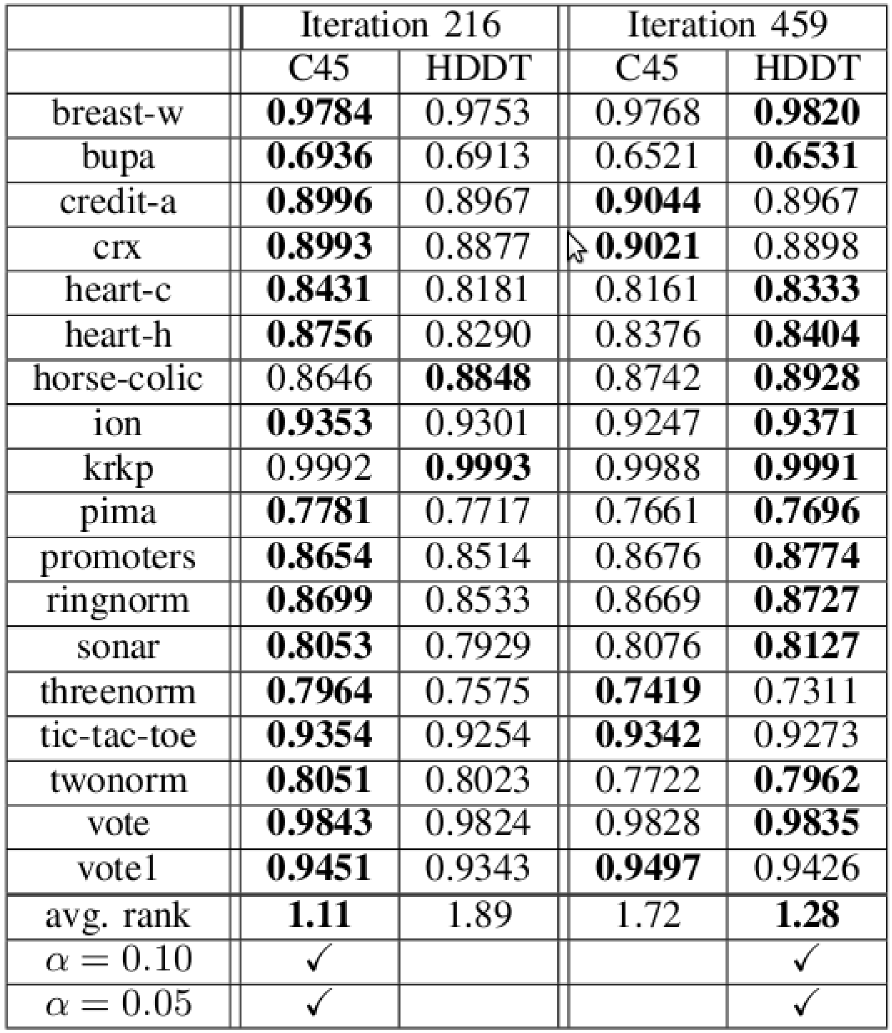
Definition 5. **Sample selection bias, in general, causes the data in the training set to follow** $P_{tra} = P(s = 1|x, y)$, while the data in the test set follows $P_{tst} = P(y, x)$. Depending on the type of problem, we have:

$P_{tra} = P(s = 1|y, x)P(y|x)P(x)$ and $P_{tst} = P(y|x)P(x)$ in X→Y problems.
$P_{tra} = P(s = 1|y, x)P(x|y)P(y)$ and $P_{tst} = P(x|y)P(y)$ in Y→X problems.

where s is a binary selection variable that decides whether a datum is included in the training sample process ($s = 1$) or rejected from it ($s = 0$).

Causes of Dataset Shift

Sample bias selection: Influence of partitioning on classifiers' performance



	Iteration 216		Iteration 459	
	C45	HDDT	C45	HDDT
breast-w	**0.9784**	0.9753	0.9768	**0.9820**
bupa	**0.6936**	0.6913	0.6521	**0.6531**
credit-a	**0.8996**	0.8967	**0.9044**	0.8967
crx	**0.8993**	0.8877	**0.9021**	0.8898
heart-c	**0.8431**	0.8181	0.8161	**0.8333**
heart-h	**0.8756**	0.8290	0.8376	**0.8404**
horse-colic	0.8646	**0.8848**	0.8742	**0.8928**
ion	**0.9353**	0.9301	0.9247	**0.9371**
krkp	0.9992	**0.9993**	0.9988	**0.9991**
pima	**0.7781**	0.7717	0.7661	**0.7696**
promoters	**0.8654**	0.8514	0.8676	**0.8774**
ringnorm	**0.8699**	0.8533	0.8669	**0.8727**
sonar	**0.8053**	0.7929	0.8076	**0.8127**
threenorm	**0.7964**	0.7575	**0.7419**	0.7311
tic-tac-toe	**0.9354**	0.9254	**0.9342**	0.9273
twonorm	**0.8051**	0.8023	0.7722	**0.7962**
vote	**0.9843**	0.9824	0.9828	**0.9835**
vote1	**0.9451**	0.9343	**0.9497**	0.9426
avg. rank	**1.11**	1.89	1.72	**1.28**
$\alpha = 0.10$	✓			✓
$\alpha = 0.05$	✓			✓

- **Classifier performance results over two separate iterations of random 10-fold cross-validation.**
- **A consistent random number seed was used across al datasets within an iteration.**

Raeder, Hoens & Chawla

Consequences of Variability in Classifier Performance Estimates., **ICDM '10 Proceedings of the 2010 IEEE International Conference on Data Mining**

Wilcoxon test: Clear differences for both algorithms

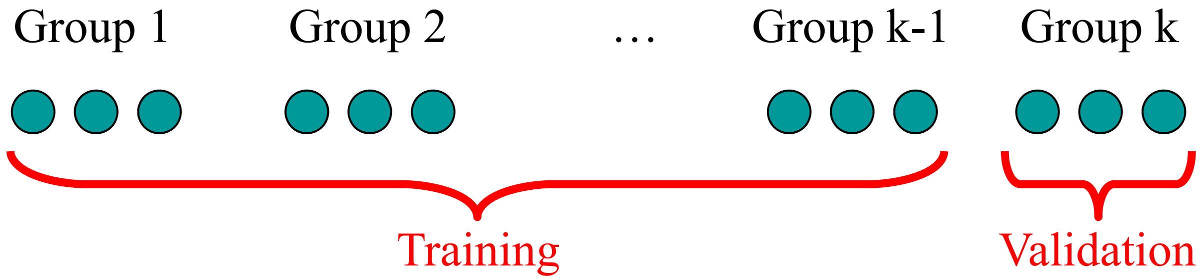
Causes of Dataset Shift

Challenges in correcting the dataset shift generated by the sample selection bias

Sample selection bias.

Cross-Validation

- Divide training samples into K groups.
- Train a learning machine with k-1 groups.
- Validate the trained machine using the rest.
- Repeat this for all combinations and output the mean validation error.



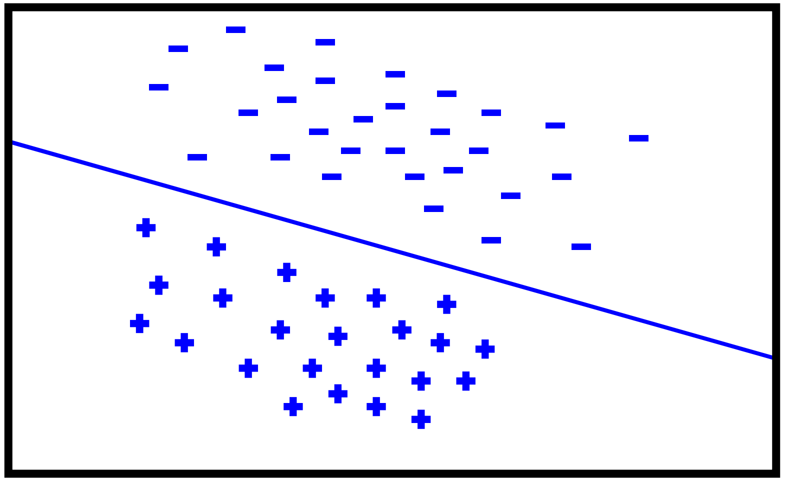
$$\widehat{f}(\boldsymbol{x}) \qquad \left(\widehat{f}(\boldsymbol{x}_t) - y_t\right)^2$$

- **CV is almost unbiased without covariate shift.**
- **But, it is heavily biased under covariate shift!**

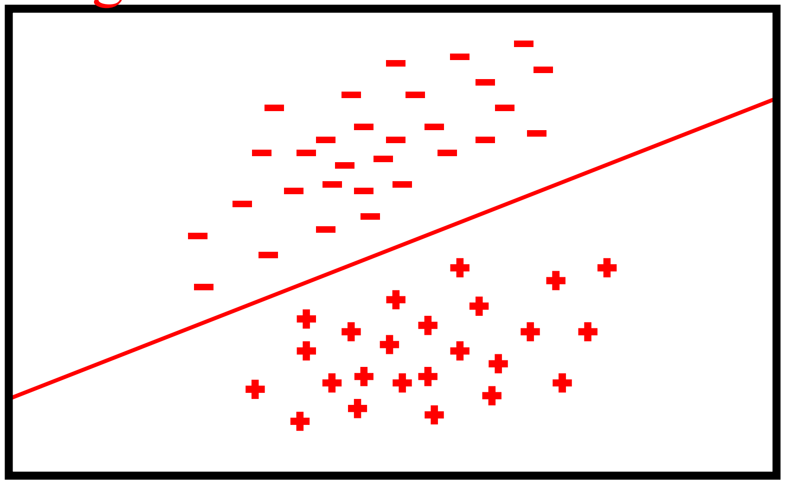
Causes of Dataset Shift

Challenges in correcting the dataset shift generated by the sample selection bias

source domain



target domain

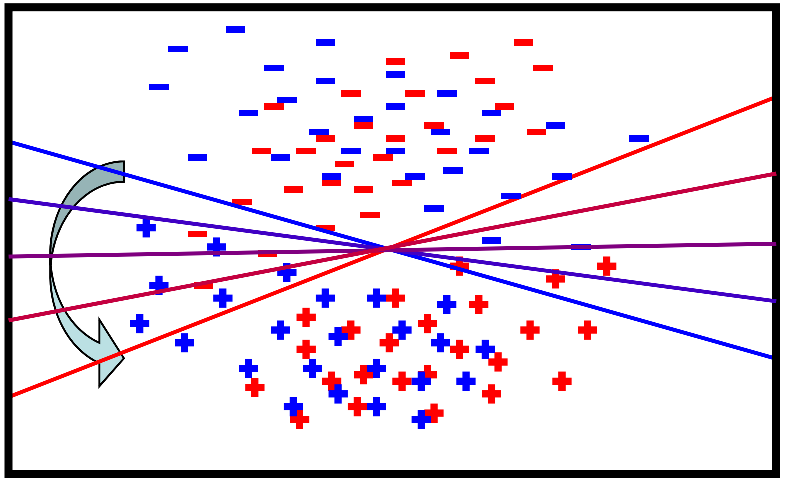


Causes of Dataset Shift

Challenges in correcting the dataset shift generated by the sample selection bias

source domain

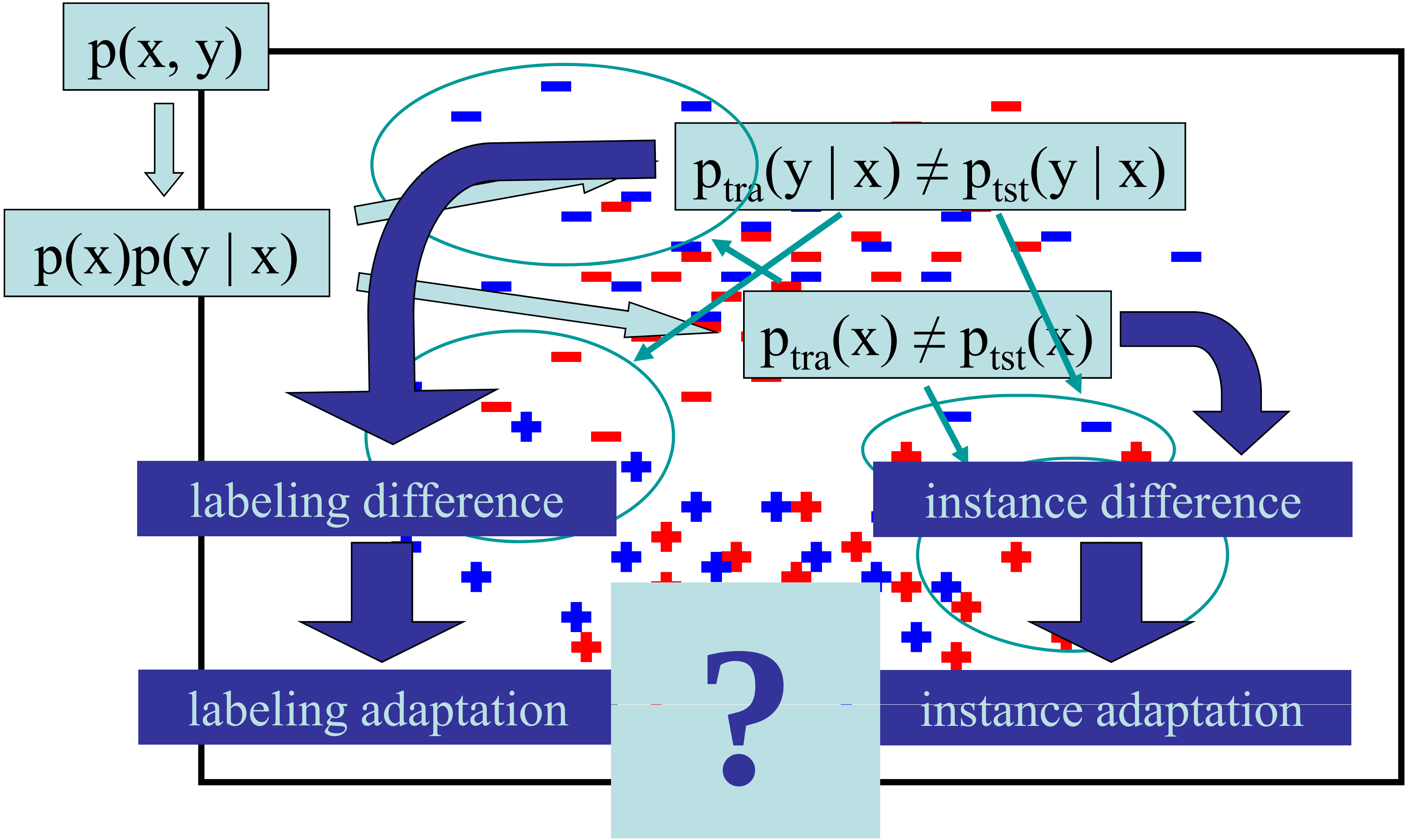
target domain



Causes of Dataset Shift

Challenges in correcting the dataset shift generated by the sample selection bias

Where Does the Difference Come from?



$p(x, y)$

$p(x)p(y \mid x)$

$p_{tra}(y \mid x) \neq p_{tst}(y \mid x)$

$p_{tra}(x) \neq p_{tst}(x)$

labeling difference

instance difference

labeling adaptation

instance adaptation

?

Causes of Dataset Shift

Challenges in correcting the dataset shift generated by the sample selection bias

Sample Selection Bias/Covariance Shift

Main Idea: Re-weighting (important sampling) the source domain data.

To correct sample selection bias:

$$\theta^* = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n_S} \frac{\mathcal{P}(x_{T_i})}{\mathcal{P}(x_{S_i})} l(x_{S_i}, y_{S_i}, \theta)$$

weights for source domain data

How to estimate $\frac{\mathcal{P}(x_{T_i})}{\mathcal{P}(x_{S_i})}$?

One straightforward solution is to estimate $P(X_S)$ and $P(X_T)$, respectively. However, estimating density function is a hard problem.

Causes of Dataset Shift

Challenges in correcting the dataset shift generated by the sample selection bias

Sample Selection Bias/Covariance Shift

Kernel Mean Match (KMM) [Huang et al. NIPS 2006]

Main Idea: KMM tries to estimate $\beta_i = \frac{\mathcal{P}(x_{S_i})}{\mathcal{P}(x_{T_i})}$ directly instead of estimating density function.

It can be proved that β_i can be estimated by solving the following quadratic programming (QP) optimization problem.

$$\min_{\beta} \quad \frac{1}{2}\beta^T K \beta - \kappa^T \beta$$

$$s.t. \quad \beta_i \in [0, B] \;\; and \;\; \left|\sum_{i=1}^{n_S} \beta_i - n_S\right| \leq n_S \epsilon$$

To match means between training and test data in a RKHS

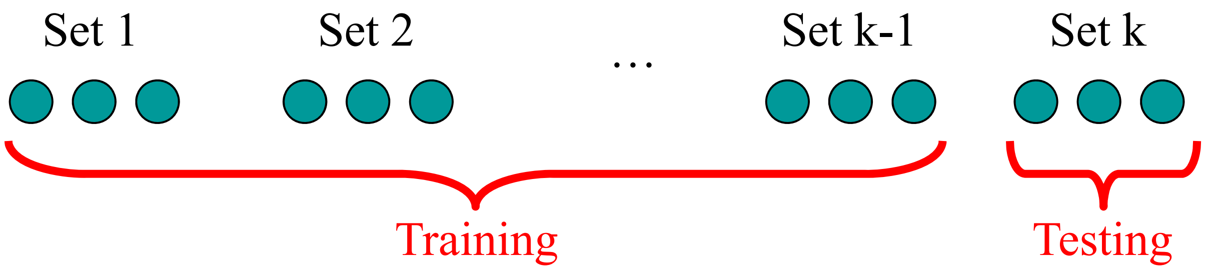
Theoretical Support: Maximum Mean Discrepancy (MMD). The distance of distributions can be measured by Euclid distance of their mean vectors in a RKHS.

Causes of Dataset Shift

Challenges in correcting the dataset shift generated by the sample selection bias

Sample selection bias. Importance-Weighted CV (IWCV)

- **When testing the classifier in CV process, we also importance-weight the test error.**



$$\widehat{f}(\boldsymbol{x})$$

$$\frac{p_{test}(\boldsymbol{x}_t)}{p_{train}(\boldsymbol{x}_t)}\left(\widehat{f}(\boldsymbol{x}_t) - y_t\right)^2$$

- **The use of IVLS mitigates the problem of inconsistency**
- **IWCV gives almost unbiased estimates of generalization error even under covariate shift**

M. Sugiyama, M. Krauledat, and K. Müller. Covariate shift adaptation by importance weighted cross validation. The Journal of Machine Learning Research, 8:985–1005, 2007.

Causes of Dataset Shift

More on the dataset shift generated by the sample selection bias

Partitioning and Dataset shift

To introduce minimal shift:

1) Pick a random example.
2) Assign it to the current fold.
3) Find the closest unused example of the same class. Move the current fold to the next one in order.
4) Repeat steps 2-3 until you are done with all examples of the chosen class, then repeat for the other classes.

Xinchuan Zeng & Tony Martinez. *Distribution-balanced stratified cross-validation for accuracy estimation.* Journal of Experimental & Theoretical Artificial Intelligence, 12:1, 2000, 1 – 12.

To introduce maximal shift:

1) Pick a random example.
2) Assign it to the current fold.
3) Find the closest unused example of the same class.
4) Repeat steps 2-3 until you have included n/K examples in the fold, then move the current fold to the next one in order.

- K is the number of folds, while n is the number of examples of the current class in the dataset.

Causes of Dataset Shift

More on the dataset shift generated by the sample selection bias

Partitioning and Dataset shift

Results: 150 partitions (50 random, 50 minimun and maximun dataset shift)

	C45	LDA	PDFC	QDA	RIPPER	SVM
1NN	39 / 109 / 2	100 / 48 / 2	0 / 7 / 143	150 / 0 / 0	0 / 83 / 67	0 / 15 / 135
C45		96 / 54 / 0	0 / 11 / 139	150 / 0 / 0	0 / 38 / 112	0 / 37 / 113
LDA			0 / 2 / 148	150 / 0 / 0	0 / 28 / 122	0 / 12 / 138
PDFC				150 / 0 / 0	38 / 112 / 0	14 / 136 / 0
QDA					0 / 0 / 150	0 / 0 / 150
RIPPER						1 / 114 / 35

Legend: Row wins / Tie / Column wins

Causes of Dataset Shift

More on the dataset shift generated by the sample selection bias

Partitioning and Dataset shift

Results: 50 partitions minimun shift

	C45	LDA	PDFC	QDA	RIPPER	SVM
1NN	25 / 25 / 0	50 / 0 / 0	0 / 5 / 45	50 / 0 / 0	0 / 30 / 20	0 / 0 / 50
C45		49 / 1 / 0	0 / 0 / 50	50 / 0 / 0	0 / 1 / 49	0 / 0 / 50
LDA			0 / 0 / 50	50 / 0 / 0	0 / 0 / 50	0 / 0 / 50
PDFC				50 / 0 / 0	11 / 39 / 0	0 / 50 / 0
QDA					0 / 0 / 50	0 / 0 / 50
RIPPER						0 / 33 / 17

Legend: Row wins / Tie / Column wins

Causes of Dataset Shift

More on the dataset shift generated by the sample selection bias

Partitioning and Dataset shift

Results: 50 partitions random

	C45	LDA	PDFC	QDA	RIPPER	SVM
1NN	13 / 37 / 0	50 / 0 / 0	0 / 1 / 49	50 / 0 / 0	0 / 22 / 28	0 / 0 / 50
C45		45 / 5 / 0	0 / 0 / 50	50 / 0 / 0	0 / 1 / 49	0 / 0 / 50
LDA			0 / 0 / 50	50 / 0 / 0	0 / 0 / 50	0 / 0 / 50
PDFC				50 / 0 / 0	15 / 35 / 0	0 / 50 / 0
QDA					0 / 0 / 50	0 / 0 / 50
RIPPER						0 / 32 / 18

Legend: Row wins / Tie / Column wins

Causes of Dataset Shift

More on the dataset shift generated by the sample selection bias

Partitioning and Dataset shift

Results: 50 partitions maximun shift

	C45	LDA	PDFC	QDA	RIPPER	SVM
1NN	1 / 47 / 2	0 / 48 / 2	0 / 1 / 49	50 / 0 / 0	0 / 31 / 19	0 / 15 / 35
C45		2 / 48 / 0	0 / 11 / 39	50 / 0 / 0	0 / 37 / 13	0 / 37 / 13
LDA			0 / 2 / 48	50 / 0 / 0	0 / 28 / 22	0 / 12 / 38
PDFC				50 / 0 / 0	12 / 38 / 0	14 / 36 / 0
QDA					0 / 0 / 50	0 / 0 / 50
RIPPER						1 / 49 / 0

Legend: Row wins / Tie / Column wins

Causes of Dataset Shift

Non-stationary environments. In real-world applications, it is often the case that the data is not (time- or space-) stationary.

One of the most relevant non-stationary scenarios involves adversarial classification problems, such as spam filtering and network intrusion detection.

This type of problem is receiving an increasing amount of attention in the machine learning field; and usually copes with non-stationary environments due to the existence of an adversary that tries to work around the existing classifier's learned concepts. In terms of the machine learning task, this adversary warps the test set so that it becomes different from the training set, thus introducing any possible kind of Dataset Shift.

Contents

I. The Problem of Dataset Shift

II. Dataset Shift: A Literature Review

III. Characterizing the change. Types of Dataset Shift

IV. Causes of Dataset Shift

V. Dataset Shift: Some approaches

VI. A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

VII. Dataset Shift: Final comments

Dataset Shift: Some approaches

Methods for determining the existence Dataset Shift between two datasets

Dataset shift solvers

Prior probability shift analysis

Dataset Shift: Some approaches

Methods for determining the existence Dataset Shift between two datasets

Significant proposals in the literature have focused on determining the existence and/or shape of Dataset Shift between two datasets:

- Correspondence tracing (Wang et al, 2003)
- Conceptual equivalence (Yang et al. 2008) (Software available)
- A statistical framework to analyze the changes in data distributions (Cieslak and Chawla, 2009) (Software available)

The software is under preparation.

Dataset Shift: Some approaches

Methods for determining the existence Dataset Shift between two datasets

Correspondence tracing

- Requires a rules-based classifier, where each sample is classified by exactly one rule.
- Build a new rules-based classifier from the test dataset.
- For each example in dataset B, identify the old and new classifying rules.
- For each old rule, identify the corresponding new rules.

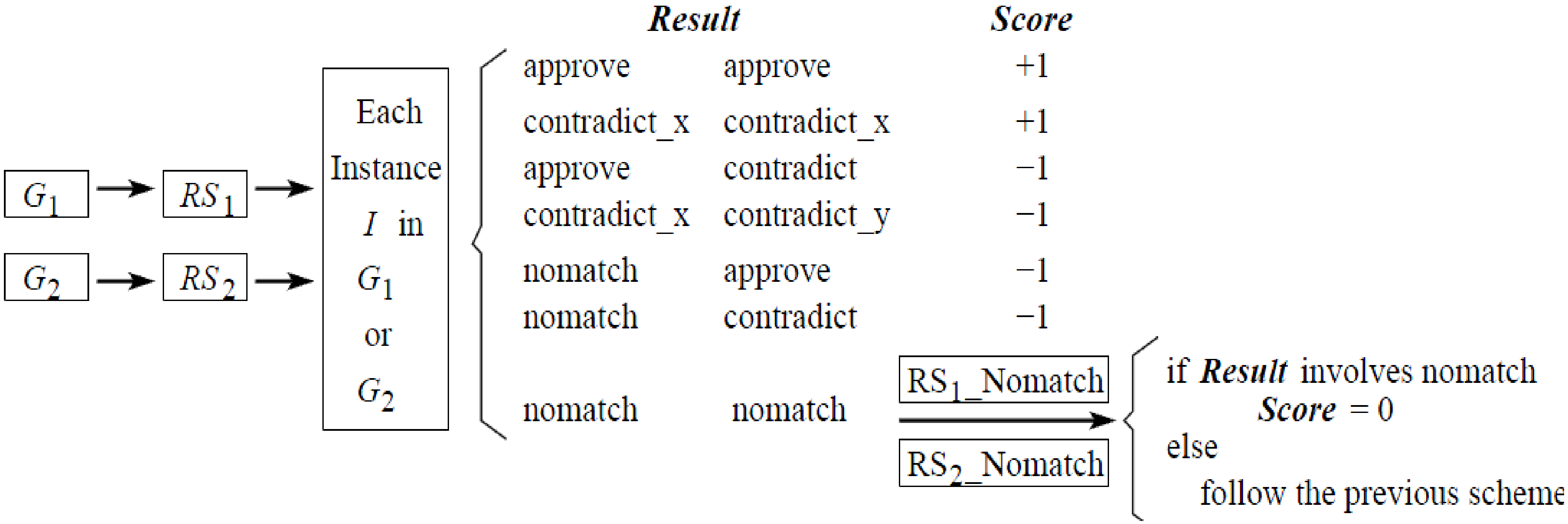
K. Wang, S. Zhou, C. A. Fu, J. X. Yu, F. Jerey, and X. Yu. Mining changes of classification by correspondence tracing. *In Proceedings of the 2003 SIAM International Conference on Data Mining (SDM 2003*), 2003.

Dataset Shift: Some approaches

Methods for determining the existence Dataset Shift between two datasets

Conceptual equivalence

- First work to step away from rule comparison, analyzing the data directly (classifier seen as a black box).
- Learn a representation of each group's concept, then ...



Result		***Score***
approve	approve	+1
contradict_x	contradict_x	+1
approve	contradict	−1
contradict_x	contradict_y	−1
nomatch	approve	−1
nomatch	contradict	−1
nomatch	nomatch	

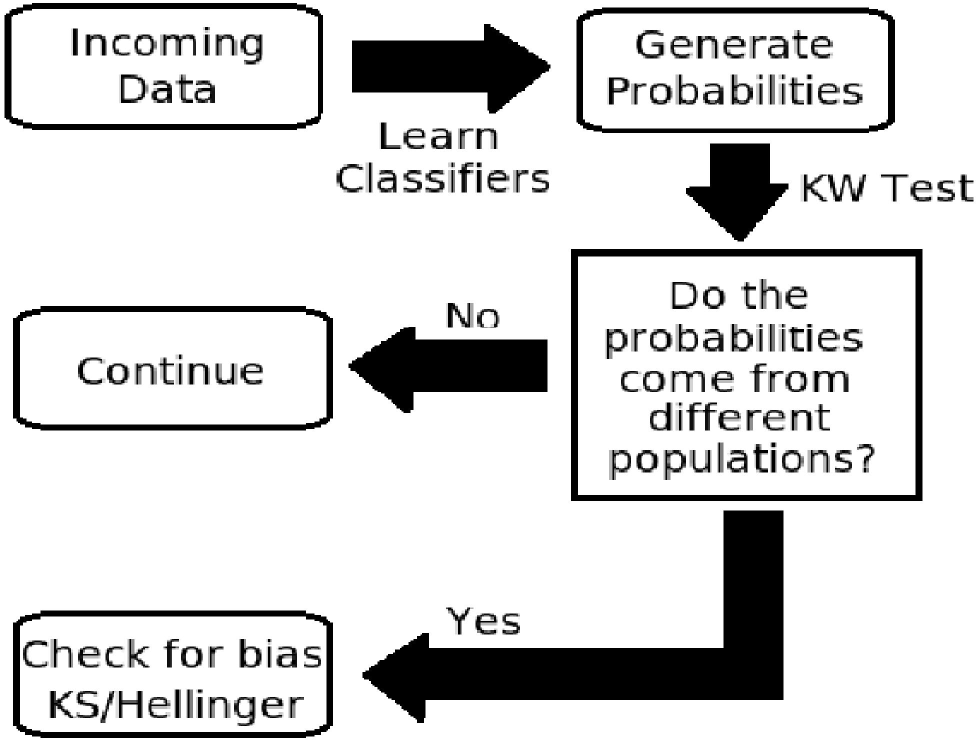
if ***Result*** involves nomatch
    Score = 0
else
    follow the previous scheme

Y. Yang, X. Wu, and X. Zhu. Conceptual equivalence for contrast mining in classification learning. *Data & Knowledge Engineering*, 67(3):413-429, 2008.

Dataset Shift: Some approaches

Methods for determining the existence Dataset Shift between two datasets

A statistical framework



Incoming Data
Learn Classifiers
Generate Probabilities
KW Test
Do the probabilities come from different populations?
No
Continue
Yes
Check for bias KS/Hellinger

D. A. Cieslak and N. V. Chawla. A framework for monitoring classifiers' performance: when and why failure occurs? *Knowledge and Information Systems*, 18(1):83-108, 2009.

T. Raeder, N. V. Chawla (July 2009). Model Monitor (M^2): Evaluating, Comparing, and Monitoring Models. *Journal of Machine Learning Research (JMLR), 10:1387--1390, 2009*

Dataset Shift: Some approaches

Dataset Shift solvers

Covariate shift has been extensively studied in the literature, and a number of proposals to work under it have been published. Some of the most important ones include:

Covariate shift solvers:

- Weighting the log-likelihood function (Shimodaira, 2000)
- Importance weighted cross validation (Sugiyama et al, 2007 JMLR) (Software available)
- Integrated optimization problem. Discriminative learning. (Bickel et al, 2009 JMRL) (Software available)
- Kernel mean matching (Gretton et al., 2009, Book:Dataset Shift…) (Software available)
- Adversarial search (Globerson et al, 2009, Book:Dataset Shift …) et al, 2011) (Software available)

The software is under preparation.

Dataset Shift: Some approaches

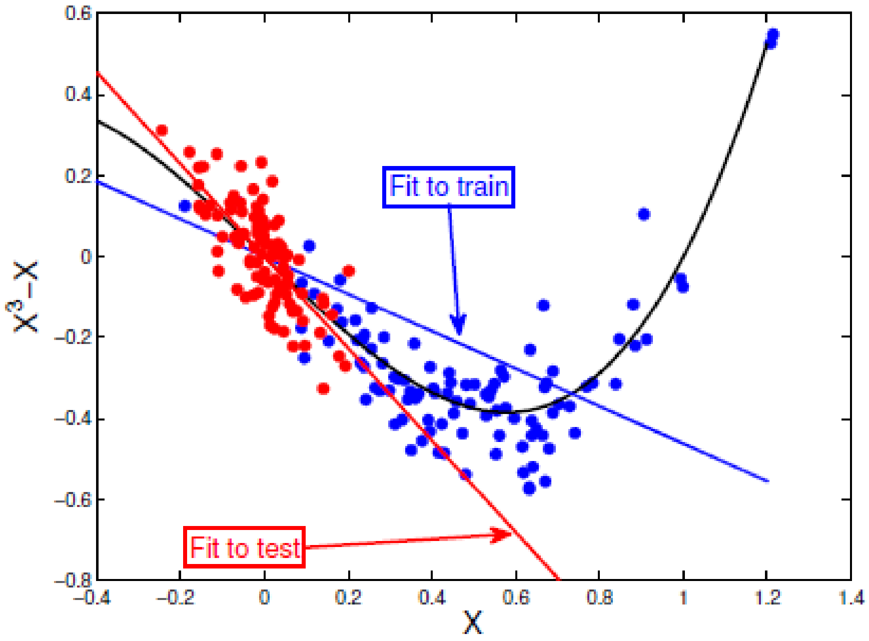
Importance weights

Covariate shift

A toy example

- Toy data [Shimodaira, 2000]
  - $\mathrm{P}_{\mathrm{tr}}(x) \sim \mathcal{N}(0.5, 0.5^2)$,
  - $\mathrm{P}_{\mathrm{te}}(x) \sim \mathcal{N}(0, 0.3^2)$

- $y = -x + x^3 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.3^2)$
- Linear regression



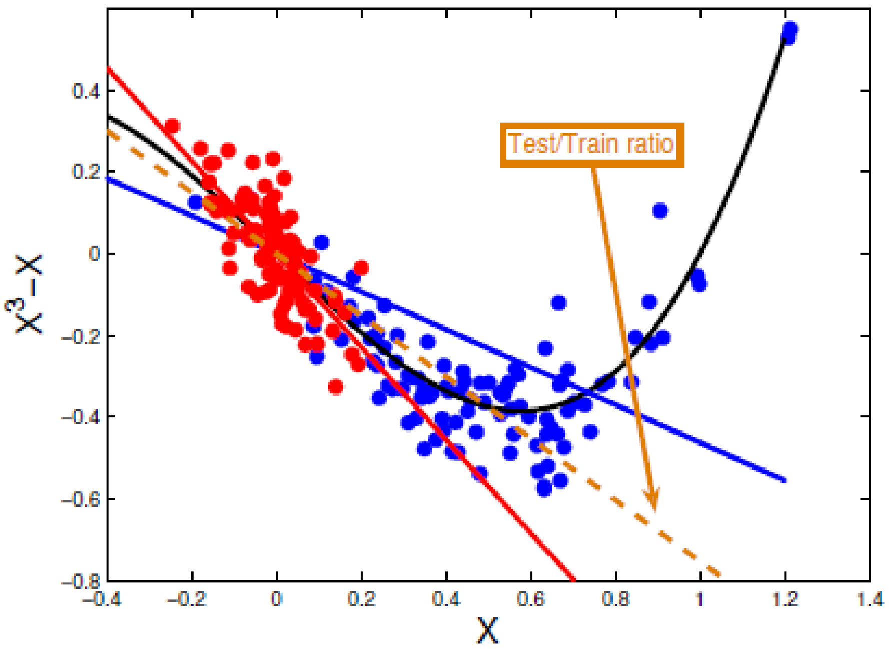
Dataset Shift: Some approaches

Importance weights

Covariate shift

Importance weighting

- Ridge regression, linear kernel
- Importance weighting improves performance



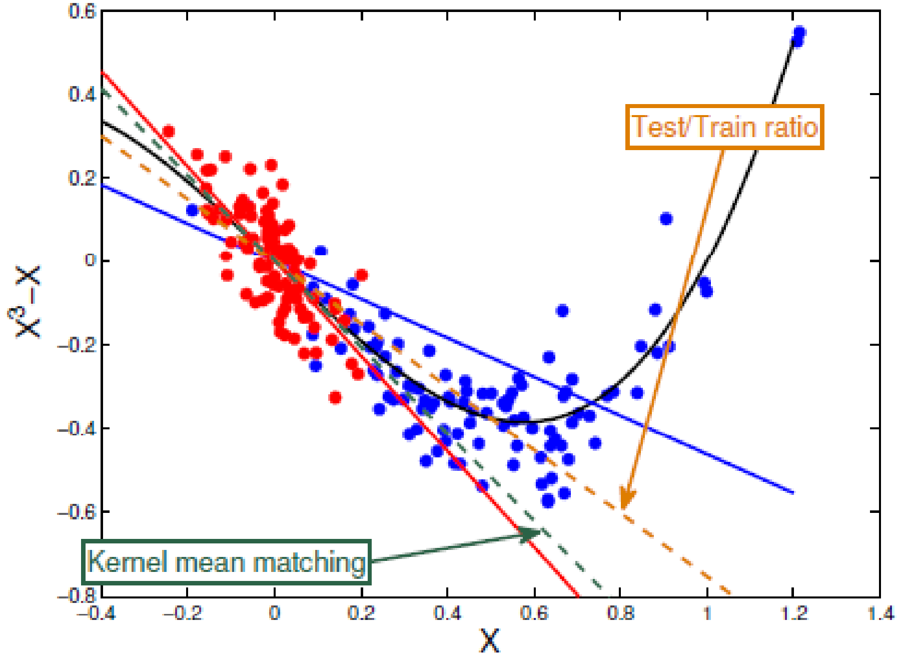
Dataset Shift: Some approaches

Importance weights

Covariate shift

KMM

- Compare KMM and importance sampling



Dataset Shift: Some approaches

Dataset Shift solvers

Other solvers:

- Subclass reestimation (Alaiz-Rdguez et al. 2009) (Software available)
- Repairing dataset shift approaches: GP-RFD: GP-based feature extraction technique to repair fractures between data originated in different biological laboratories (Moreno-Torres et al, 2011) (Software available)

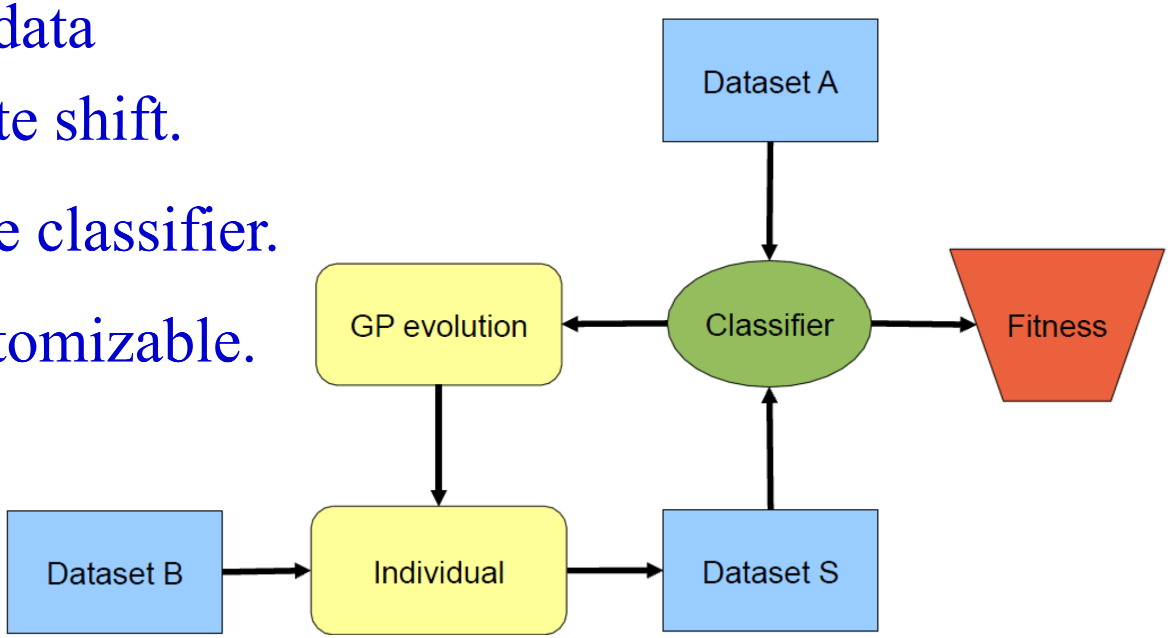
The software is under preparation.

Dataset Shift: Some approaches

Dataset Shift solvers

GP-RFD: Genetic Programming Repairing Fractures between Data

- ✓ Treats the classifier as a black box.
- ✓ Can mine any kind of data fracture, not just covariate shift.
- ✓ Adapts the data, not the classifier.
- ✓ Robust and highly customizable.



J.G. Moreno-Torres, X. Llorà, D.E. Goldberg and R. Bhargava. Repairing Fractures between Data using Genetic Programming-based Feature Extraction:
A Case Study in Cancer Diagnosis. *Information Sciences*, In Press, 2011.

Dataset Shift: Some approaches

Prior Probability Shift

Prior probability shift has also been studied deeply, with a multitude of proposals appearing in the literature. There are two main strategies when designing classifiers for expected prior probability shift conditions:

Adaptive approaches: These proposals train a classifier over the available data and then the adapt some of its parameters according to the (usually unlabeled) test data.

Robust approaches: Base the choice of classifier on some measure that is ideally transparent to changes in class distribution. The best known example would be ROC curve analysis.

Dataset Shift: Some approaches

Prior Probability Shift

Prior probability shift analysis

- ROC curve analysis works for problems of type b (Y → X), since there is no “concept drift” even if the class distribution change.
- Concept drift: Change in the class-conditional distributions = change in the model's true and false positive rates.
- In type a (X → Y) problems, concept drift exists and we need more complex models.

T. Flawcett and P. Flach. A response to Webb and Ting's “On the application of roc analysis to predict classification performance under varying class distributions”. *Machine Learning*, 58:33-38, 2005.

Contents

I. The Problem of Dataset Shift

II. Dataset Shift: A Literature Review

III. Characterizing the change. Types of Dataset Shift

IV. Causes of Dataset Shift

V. Dataset Shift: Some approaches

VI. A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

VII. Dataset Shift: Final comments

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

Dataset Shift: This is a common problem that can affect all kind of classification problems, and it often appears due to sample selection bias issues.

However, **the data-set shift issue is specially relevant when dealing with imbalanced classification**, because in highly imbalanced domains, the minority class is particularly sensitive to singular classification errors, due to the typically low number of examples it presents.

In the most extreme cases, a single misclassified example of the minority class can create a significant drop in performance.

Moreno-Torres, J. G., & Herrera, F. (2010). A preliminary study on overlapping and data fracture in imbalanced domains by means of genetic programming-based feature extraction. In *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA 2010) (pp. 501–506).*

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

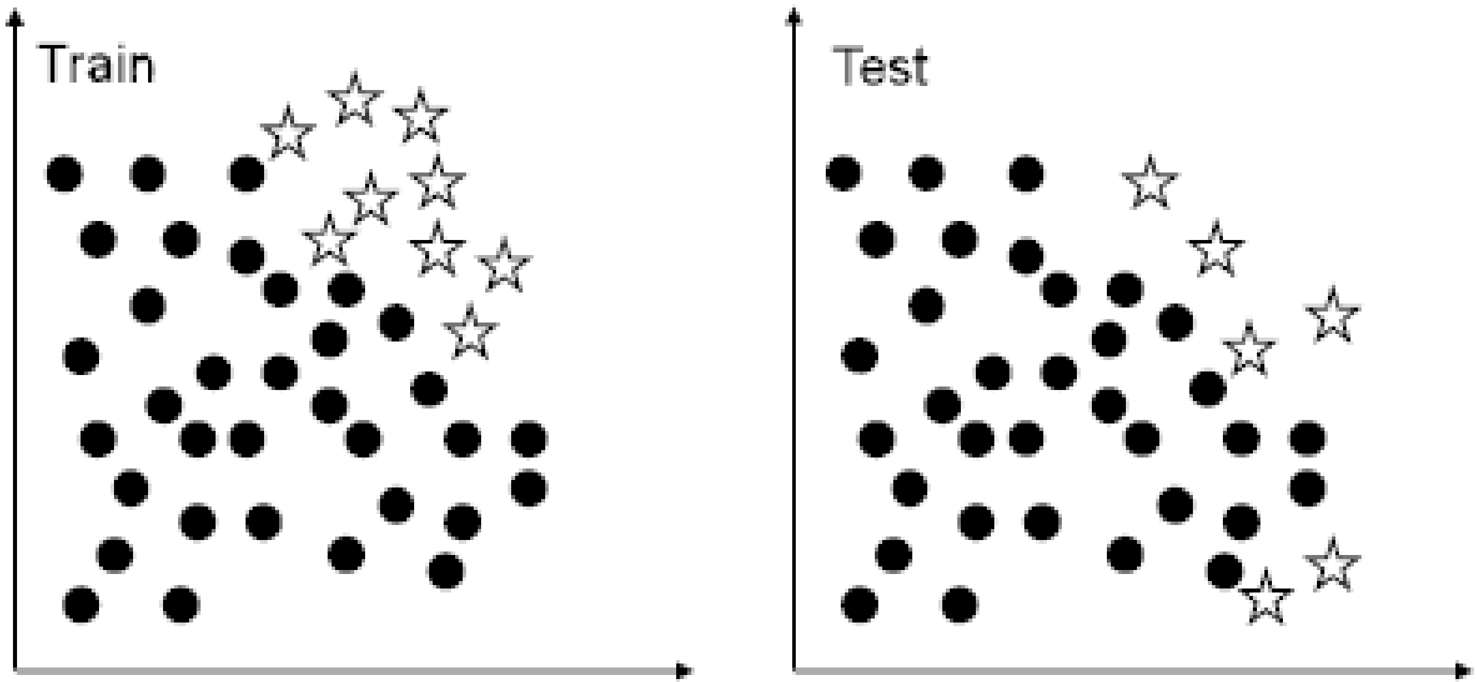


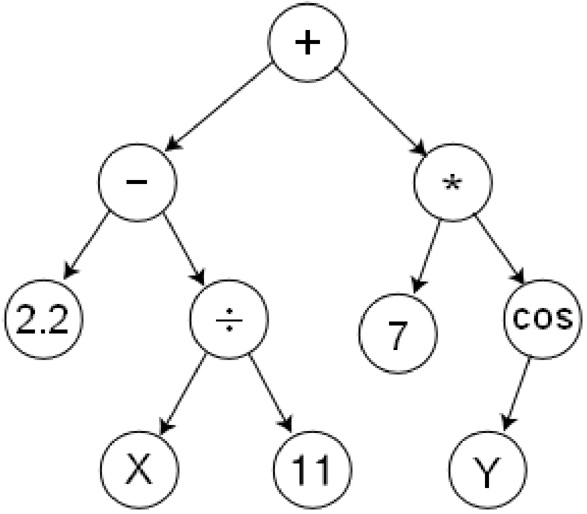
Figure 18: Example of the impact of data-set shift in imbalanced domains.

Moreno-Torres, J. G., & Herrera, F. (2010). A preliminary study on overlapping and data fracture in imbalanced domains by means of genetic programming-based feature extraction. In *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA 2010) (pp. 501–506).*

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

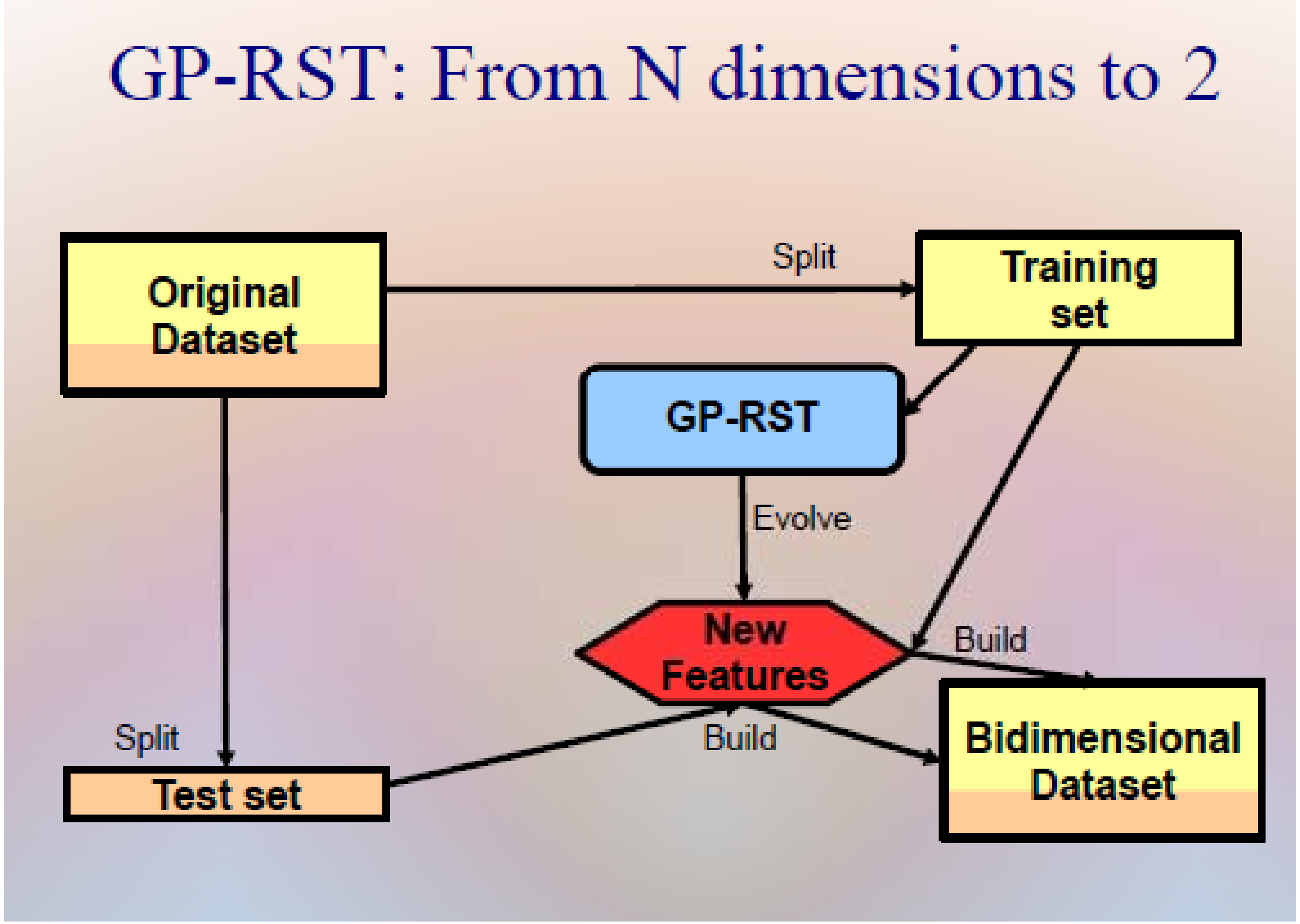
GP-RST: From N dimensions to 2

- Goal: obtain a 2-dimensional representation of a given dataset that is as separable as possible.
- Genetic Programming based: evolves 2 trees simultaneously as arithmetic functions of the previous N-dimensions.
- Evaluation of an individual dependant on Rough Set Theory measures.



$$\left(2.2-\left(\frac{X}{11}\right)\right)+\left(7*\cos(Y)\right)$$

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data



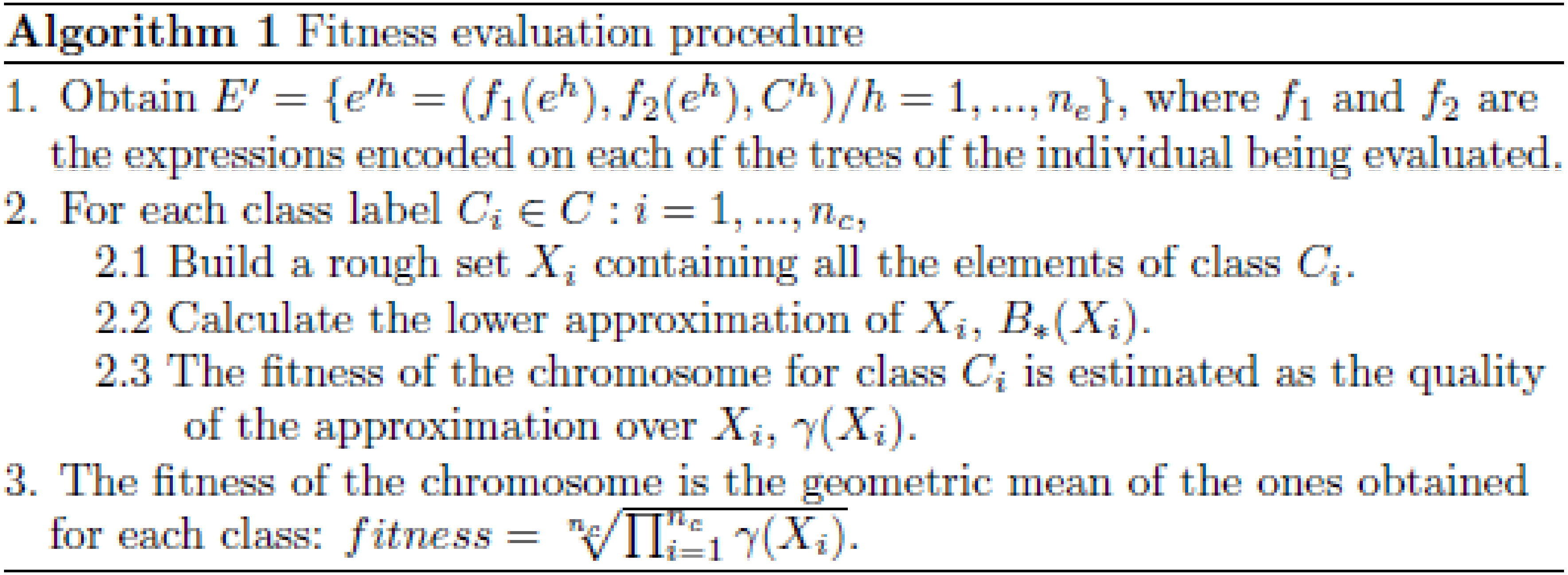
GP-RST: From N dimensions to 2

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

The quality of approximation $\gamma(x)$ is the proportion of the elements of a rough set that belong to its lower approximation.

$$B_*(X) = \{x \in X : R^{'}(x) \subseteq X\}$$

$$\gamma(x) = \frac{|B_*(X)|}{|X|}$$

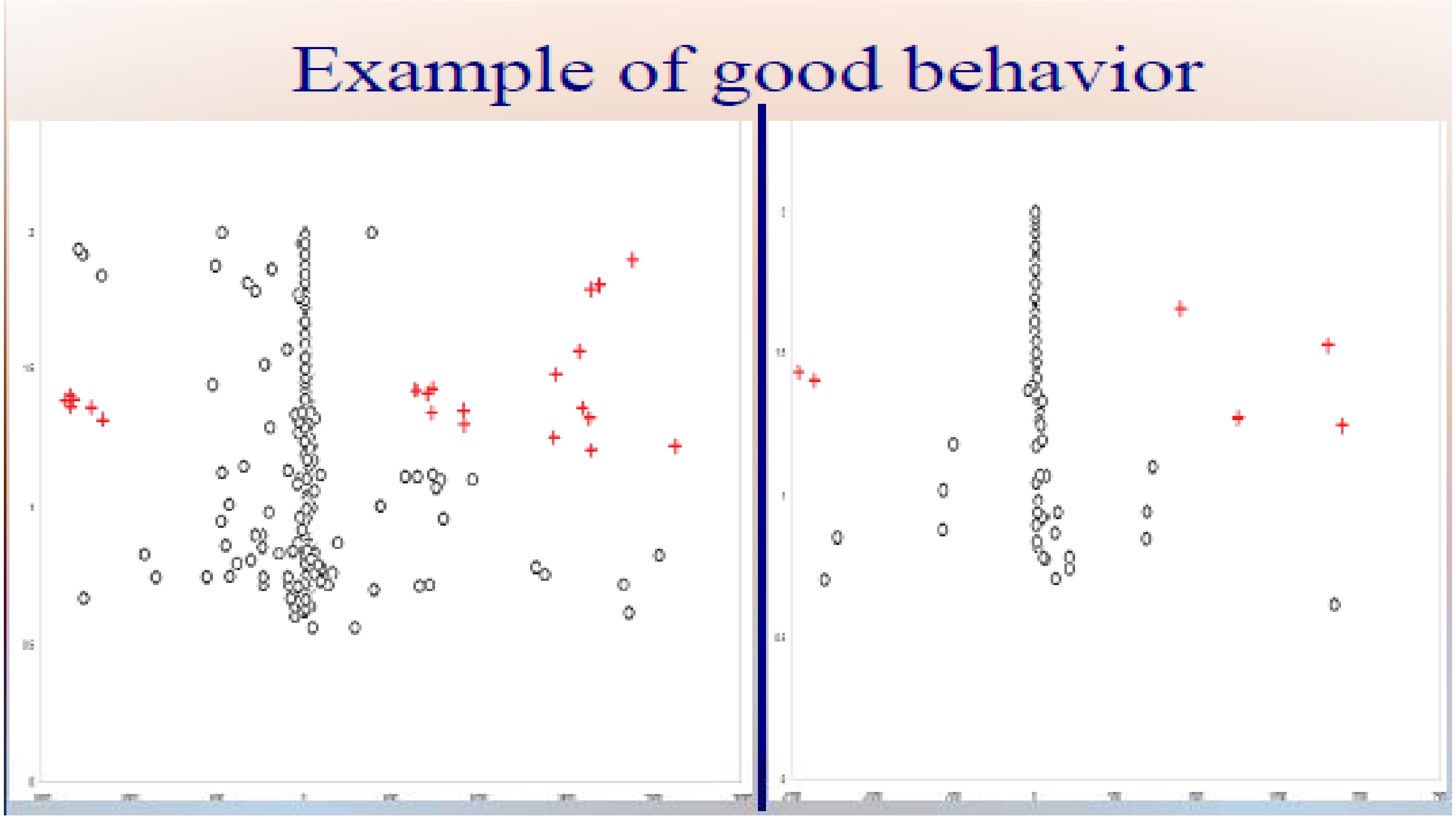


Algorithm 1 Fitness evaluation procedure

```
1. Obtain $E' = \{e'^h = (f_1(e^h), f_2(e^h), C^h)/h = 1, ..., n_e\}$, where $f_1$ and $f_2$ are
   the expressions encoded on each of the trees of the individual being evaluated.
2. For each class label $C_i \in C : i = 1, ..., n_c$,
     2.1 Build a rough set $X_i$ containing all the elements of class $C_i$.
     2.2 Calculate the lower approximation of $X_i$, $B_*(X_i)$.
     2.3 The fitness of the chromosome for class $C_i$ is estimated as the quality
           of the approximation over $X_i$, $\gamma(X_i)$.
3. The fitness of the chromosome is the geometric mean of the ones obtained
   for each class: $fitness = \sqrt[n_c]{\prod_{i=1}^{n_c} \gamma(X_i)}$.
```

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

Good behaviour. pageblocks 13v4, 1st partition.



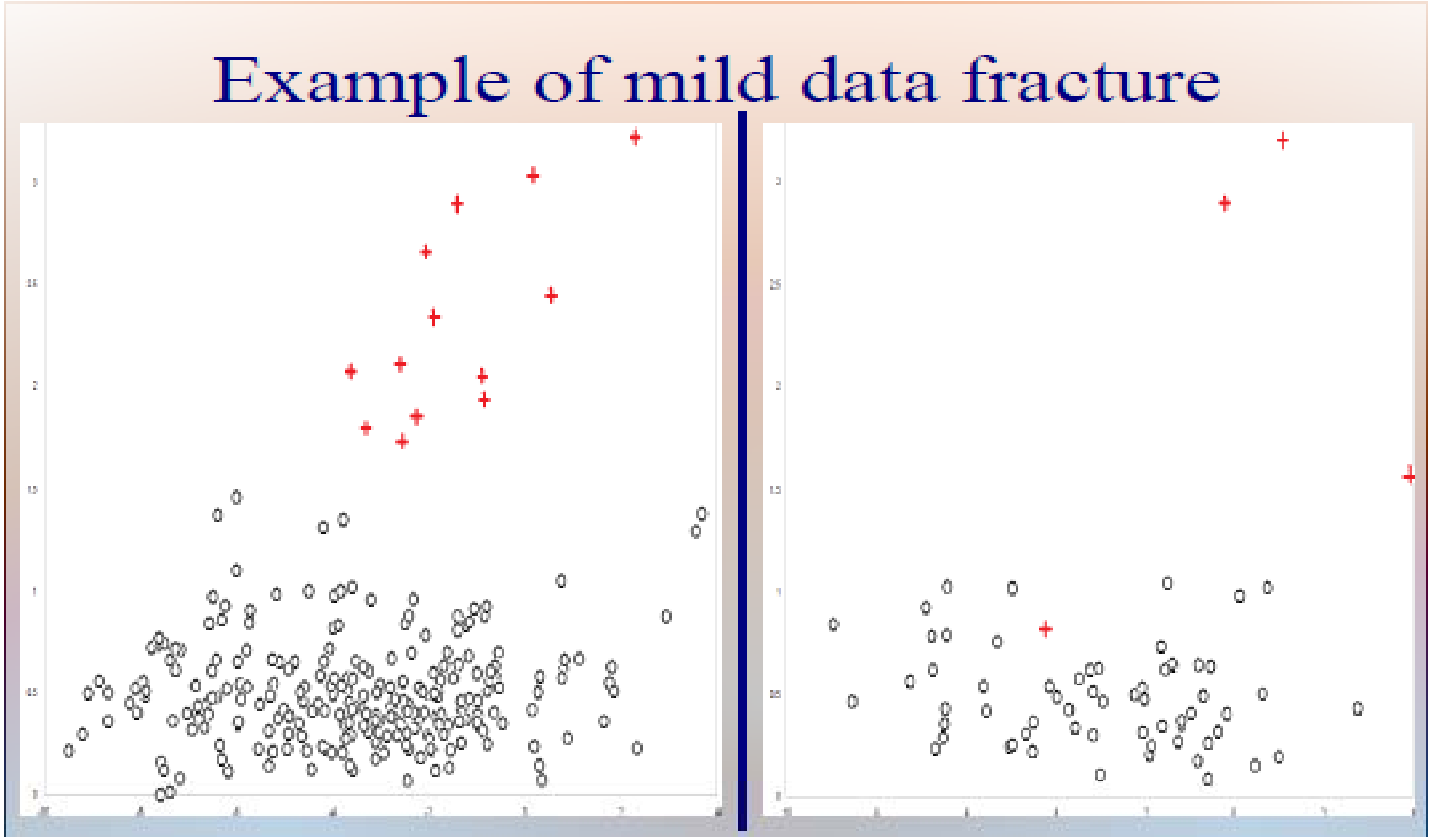
Example of good behavior

(a) Training set (1.0000)

(b) Test set (1.0000)

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

Dataset shift. ecoli 4, 1st partition.

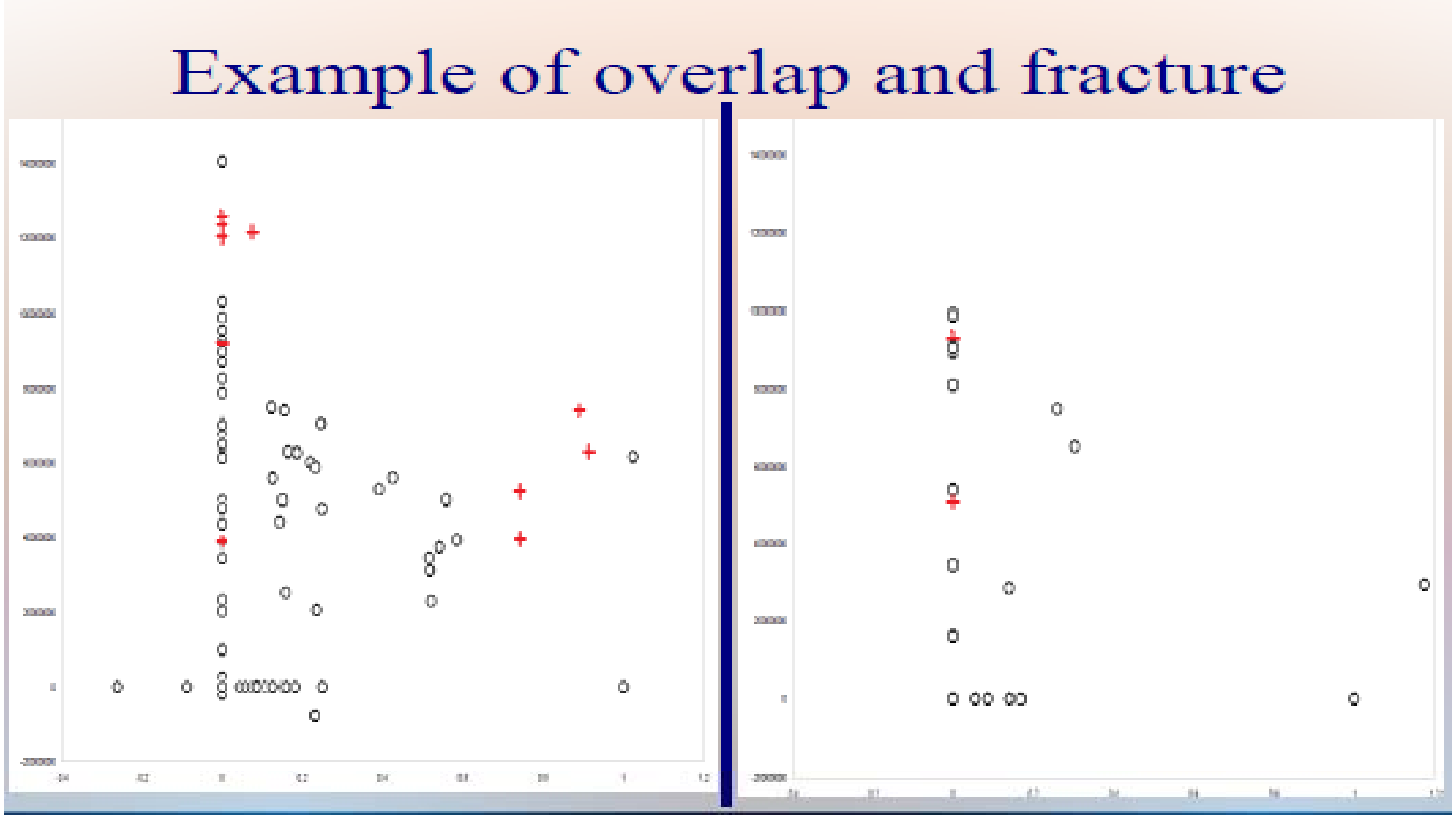


(a) Training set (0.9663)

(b) Test set (0.8660)

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

Overlap and dataset shift. glass 016v2, 4th partition.



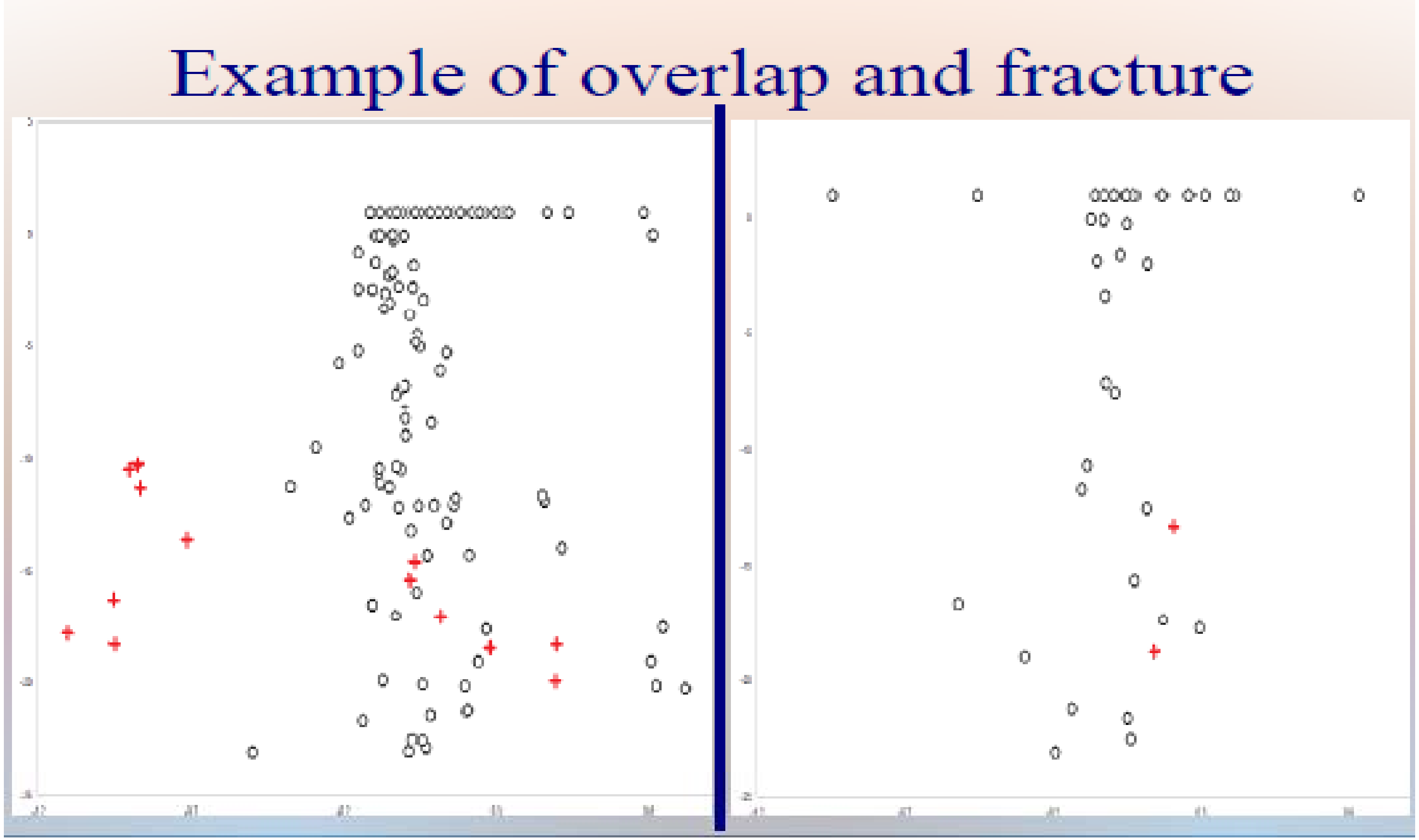
Example of overlap and fracture

(a) Training set (0.3779)

(b) Test set (0.0000)

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

Overlap and dataset shift. glass 2, 2nd partition



Example of overlap and fracture

(a) Training set (0.6794)

(b) Test set (0.0000)

A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

- Imbalanced classification problems are difficult when overlap and/or data fracture are present.
- Single outliers can have a great influence on classifier performance.
- This is a novel problem in imbalanced classification that need a lot of studies.

Contents

I. The Problem of Dataset Shift

II. Dataset Shift: A Literature Review

III. Characterizing the change. Types of Dataset Shift

IV. Causes of Dataset Shift

V. Dataset Shift: Some approaches

VI. A Genetic-Programming based Feature Selection and RST for Visualization of Fracture between Data

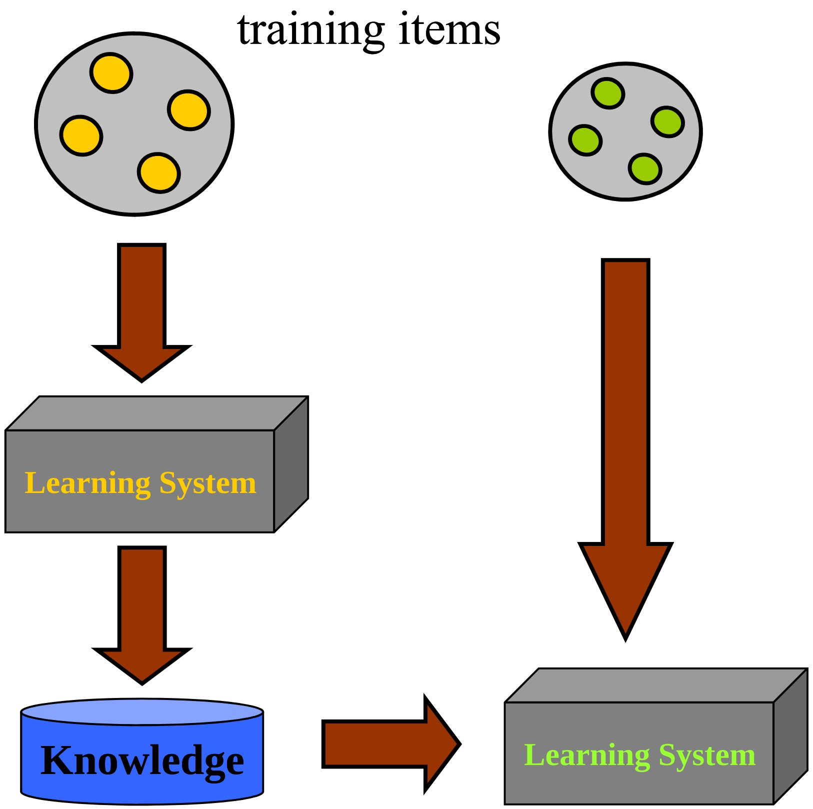
VII. Dataset Shift: Final comments

Dataset Shift: Final comments

- **Dataset shift is a very important and common problem in many real-world applications. Dataset shift happens all the time.**
- **There are some common generic causes**
- **Intertwined issue with other problems: imbalanced classification, transfer learning, …**

Transfer Learning (TL):
The ability of a system to recognize and apply knowledge and skills learned in previous tasks to novel tasks (in new domains)

Sinno Jialin Pan and Qiang Yang. **A Survey on Transfer Learning**
IEEE Transactions on Knowledge and Data Engineering, 22(10):1345-1359, Oct. 2010.



Dataset Shift: Final comments

- Dataset shift is a very important and common problem in many real-world applications. Dataset shift happens all the time.
- There are some common generic causes
- Intertwined issue with other problems: imbalanced classification, transfer learning, …
- There are a number of solutions in the literature, but they are mostly limited to work under specific scenarios.
- There need a lot of work for including in a common framework the proposals for determining the existence and/or shape of Dataset Shift between two datasets.
- It is also important to have a taxonomy for proposals to work under it have been published.

Acknowledgements

Joint work with José García Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla

J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N.V. Chawla, F. Herrera

A Unifiying view of Dataset Shift in Classification. Pattern Recognition, 2011, In press.

Dataset Shift in Classification

Thanks !!!