On Windowing as a subsampling method for Distributed Data Mining

David Martínez-Galicia

Director: Alejandro Guerra-Hernández
Co-directors: Nicandro Cruz-Ramírez, Xavier Limón

Universidad Veracruzana
Centro de investigación en Inteligencia Artificial
Sebastían Camacho No. 5
Xalapa, Veracruz, Mexico (91000)

Universidad Veracruzana

Introduction

- Data Mining (DM) consists of applying analysis algorithms that produce models to predict or describe the data [1].

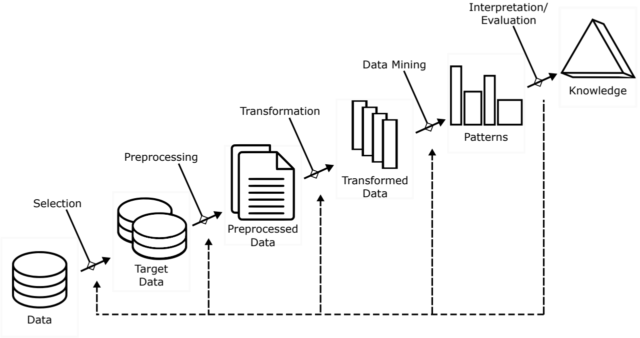


Figure: Knowledge Discovery on Databases (KDD) process.

Universidad Veracruzana

Introduction

- Distributed Data Mining (DDM) concerns the application of DM procedures trying to optimize the available resources in distributed environments [2].

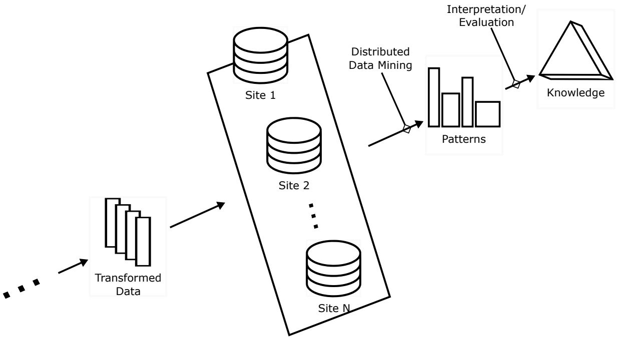


Figure: Distributed Data Mining (DDM).

Universidad Veracruzana

Scope

This work studies three points necessary to adopt Windowing as a subsampling technique in distributed environments:

1. Method generalization.
2. Sub-sampling characterization.
3. Model description.

Universidad Veracruzana

Windowing

- Technique proposed by John Quinlan that induces models from large datasets selecting a small sample from the training instances [3].

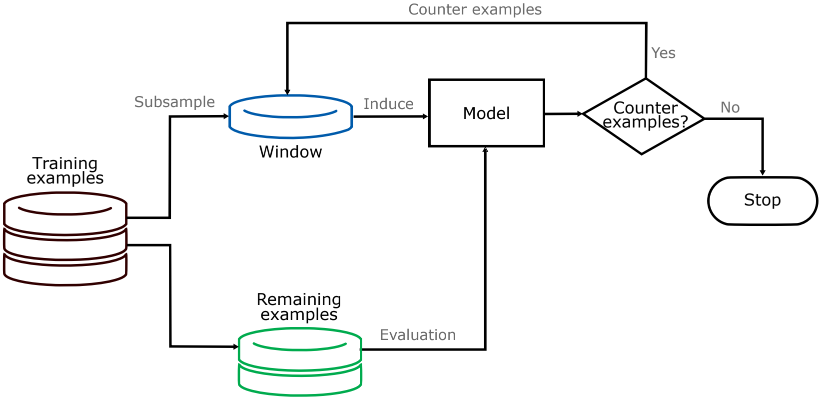
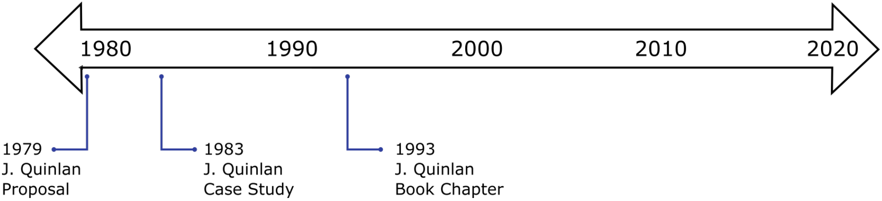


Figure: Windowing diagram.

Universidad Veracruzana

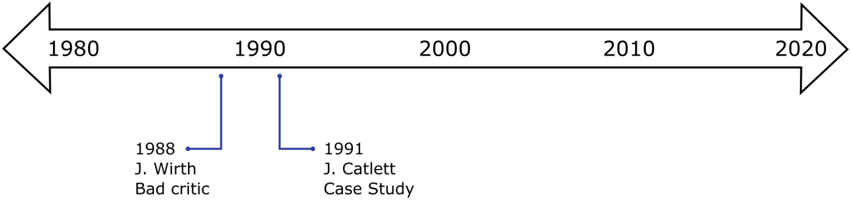
Related Work I



- J. Quinlan based his research in the hypothesis that it is possible to generate an accurate decision tree to explain a large dataset, even when a small part of the examples is selected for induction [3].

Universidad Veracruzana

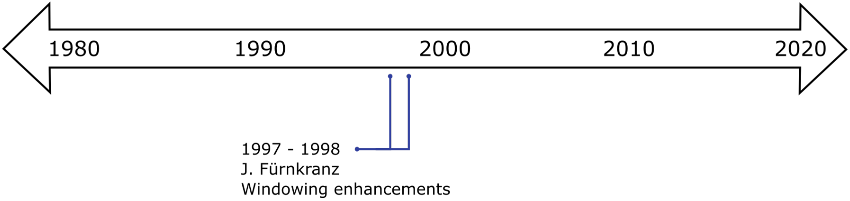
Related Work II



- J. Wirth and J. Catlett publish an early critic [4] about the costs of Windowing where they suggest avoiding its use in noisy domains because it considerably increases the CPU requirements.

Universidad Veracruzana

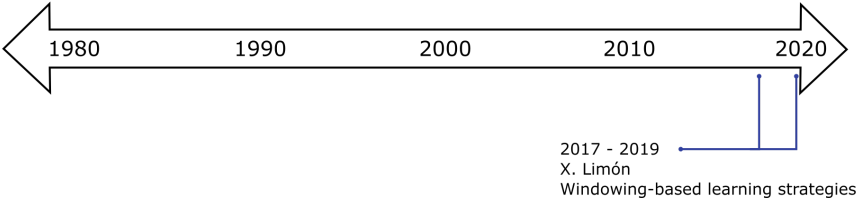
Related Work III



- J. Fürnkranz focused his research in new mechanisms to optimize the time convergence, the levels of accuracy and the performance in noisy domains [5].

Universidad Veracruzana

Related Work IV



- X. Limón *et al.* introduce a new framework for DDM, where they propose different Windows-based strategies that are capable to perform aggressive samplings [6].

Hypothesis

- Windowing exhibits consistent behavior through the use of different Machine Learning models in DDM scenarios, i.e., models with high levels of accuracy are induced from small samples.
- In these scenarios, it is possible to obtain gains in terms of performance, model complexity and data compression, against traditional sub-sampling methods.

Universidad Veracruzana

Objectives

General objective: Studying the behavior of Windowing through the use of different Machine Learning models.

Specific objectives:

1. Measuring the correlation between the model accuracy and the percentage of instances.
2. Suggesting metrics that measure informational features to compare the samples and the induced models.
3. Comparing Windowing with other sub-sampling techniques to observe the advantages of its use.
4. Characterizing the operation of this technique on different types of datasets.
5. Providing a wide description about Windowing behavior and the best conditions to make use of it.

Universidad Veracruzana

Justification I

Johannes Fürnkranz [7] has argued that this method offers three advantages:

1. It copes well with memory limitations, reducing considerably the number of examples to induce a model of acceptable accuracy.
2. It offers an efficiency gain by reducing the time of convergence, especially when using a rule learning algorithm, as Foil.
3. It offers an accuracy gain, particularly in noiseless datasets, possibly because learning from a sample may result in a less over-fitting theory.

Universidad Veracruzana

Justification II

Articles related to JaCa-DDM [8, 6] have shown:

1. A strong correlation between the accuracy of the learned Decision Trees and the percentage of examples used to induce them.
2. The performed reductions are as big as the 90% of the available training examples.

Universidad Veracruzana

Contributions

1. The empirical evidence that the use of Windowing can be generalized to other Machine Learning algorithms.
2. A methodology that involves different Theory Information metrics to characterize the data transformation performed by a sampling.
3. The implementation of the proposed metrics available in a digital repository. [1]
4. Two papers as result of our participation in MICAI.
   - Windowing as a Sub-Sampling Method for Distributed Data Mining. Mathematical and Computational Applications, 25(3), 39. MDPI AG.
   - Towards Windowing as a Sub-Sampling Method for Distributed Data Mining. Research in Computing Science Journal. In press.

Universidad Veracruzana

[1] https://github.com/DMGalicia/Thesis-Windowing

Methodology

The methodological design of this work includes 3 experiments to study:

1. The Windowing generalization.
2. The sample characterization (comparison with traditional samplings).
3. The study of the evolution of the windows.

JaCa-DDM [2] is adopted to run the experiments.

Universidad Veracruzana

[2] https://github.com/xl666/jaca-ddm

Counter Strategy

JaCa-DDM defines a set of Windowing-based strategies using J48, the Weka implementation [9] of C4.5. Due to the great similarity with the Windowing's original formulation, the Counter strategy is selected.

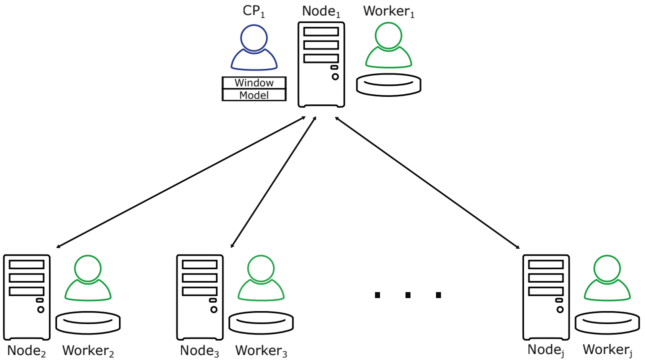


Figure: Counter strategy

Datasets

Experiments are tested on 15 datasets selected from the UCI [10] and MOA [11] repositories.

Dataset	#Instances	#Attributes	Attrib. Type	Missing Val.	#Classes
Adult	48842	15	Mixed	Yes	2
Australian	690	15	Mixed	No	2
Breast	683	10	Numeric	No	2
Diabetes	768	9	Mixed	No	2
Ecoli	336	8	Numeric	No	8
German	1000	21	Mixed	No	2
Hypothyroid	3772	30	Mixed	Yes	4
Kr-vs-kp	3196	37	Numeric	No	2
Letter	20000	17	Mixed	No	26
Mushroom	8124	23	Nominal	Yes	2
Poker-lsn	829201	11	Mixed	No	10
Segment	2310	20	Numeric	No	7
Sick	3772	30	Mixed	Yes	2
Splice	3190	61	Nominal	No	3
Waveform5000	5000	41	Numeric	No	3

Universidad Veracruzana

On Windowing generalization I

This experiment seeks to:

- Corroborate the correlation reported in literature.
- Provide evidence about the generalization of Windowing.
- Characterize the sampling with informational properties.

Decision trees (j48) and other 4 Weka models are induced by running a 10-fold stratified cross-validation on each dataset.

Universidad Veracruzana

On Windowing generalization II

Weka algorithms:

- Naive Bayes: A probabilistic classifier based on Bayes' theorem [12].
- jRip: An inductive rule learner based on RIPPER [13].
- Multilayer-Perceptron: A perceptron trained by backpropagation [14].
- SMO: An implementation for training a support vector classifier [15].

In order to measure the performance of models, their accuracy is defined as the percentage of correctly classified instances:

$$\frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Universidad Veracruzana

On Windowing generalization III

- Kullback-Leibler divergence (D_{KL}) [16] is defined as:

$$D_{KL}(P_{DS} \parallel P_{Window}) = \sum_{c \in Class} P_{DS}(c) \; log_2 \left(\frac{P_{DS}(c)}{P_{Window}(c)} \right) \tag{2}$$

- Sim_1 [17] is a similarity measure between datasets defined as:

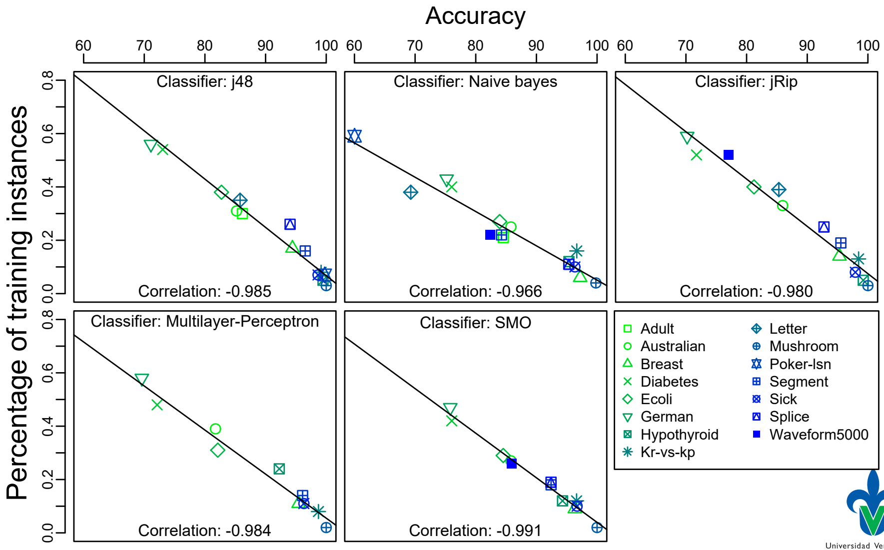
$$sim_1(Window, DS) = \frac{|Item(Window) \cap Item(DS)|}{|Item(Window) \cup Item(DS)|} \tag{3}$$

- *Red* [18] measures redundancy in a dataset in terms of conditional population entropy (CPE):

$$Red = 1 - \frac{CPE}{\sum_{a \in Attrs} log_2 \; |dom(a)|} \tag{4}$$

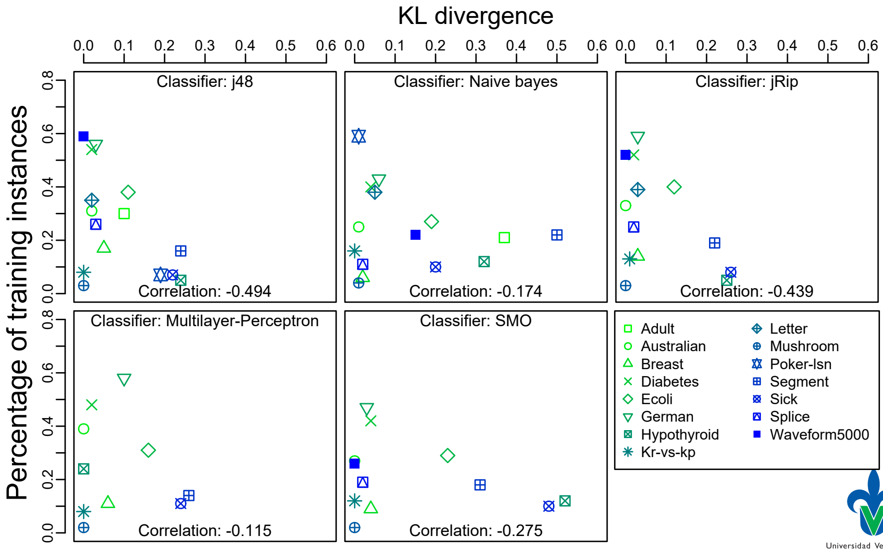
Universidad Veracruzana

Results: Generalization I

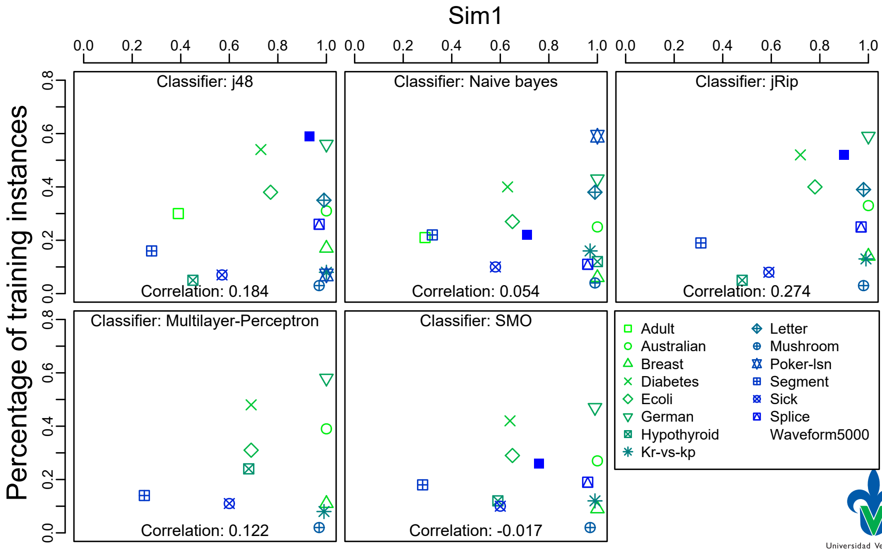


Universidad Veracruzana

Results: Generalization II

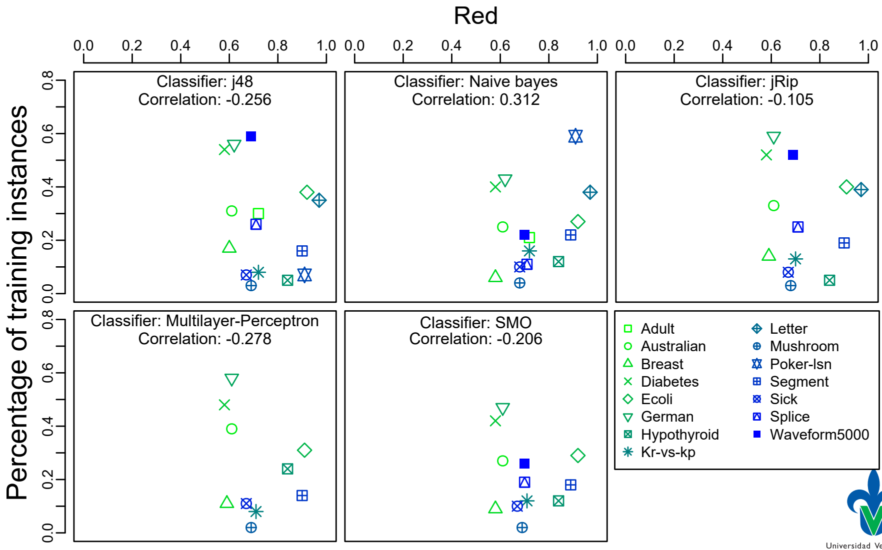


Results: Generalization III



Universidad Veracruzana

Results: Generalization IV



Universidad Veracruzana

Comparing Windowing with subsampling techniques I

This experiment seeks to:

- Obtain a deeper understanding of the informational properties of the computed models, as well as those of the samples.
- Compare Windowing with traditional sampling techniques.

For this, decision trees (j48) are adopted as classifiers.

Comparing Windowing with subsampling techniques II

- The Area Under the ROC Curve (AUC) defined as the probability of a random instance to be correctly classified:

$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{5}$$

- The Minimum Description Length (MDL) defined the sum of the length of the model $L(H)$, and the length of the data when encoded using the theory as a predictor for the data $L(D|H)$ [19]:

$$MDL = L(H) + L(D|H) \tag{6}$$

Universidad Veracruzana

Comparing Windowing with subsampling techniques III

The metrics are used to compare the window and the model computed by Windowing, against those obtained as follows:

- Without sampling, using all the available data to induce the model.
- By Random sampling, using samples of the size of the windows.
- By Stratified random sampling, using samples of the size of the windows.
- By Balanced random sampling, using samples of the size of the windows.

10 repetitions of 10-fold stratified cross-validation are run on each dataset.

Universidad Veracruzana

Statistical Test

- The comparison of A algorithms on D datasets is realized following the method proposed by Demšar[20].
- It is based on the use of the Friedman[21, 22] test with a corresponding post-hoc test (Nemenyi test).
- The null-hypothesis states that if the performance of the algorithms is similar, their ranks should be equal.

$$R_a = \frac{1}{D} \sum_{d \in D} R_a^d \tag{7}$$

Universidad Veracruzana

Statistics

Friedman	Iman and Davenport
$\chi_F^2 = \frac{12D}{A(A+1)}\left[\sum_a R_a^2 - \frac{A(A+1)^2}{4}\right]$	$F_f = \frac{(D-1)\times\chi_F^2}{D\times(A-1)-\chi_F^2}$
Distributed according to χ_F^2 with $A-1$ degrees of freedom.	Distributed according to the F-distribution with $A-1$ and $(A-1)(D-1)$ degrees of freedom.

- If the null hypothesis of similar performances is rejected, then the Nemenyi post-hoc test is realized for pairwise comparisons.

Universidad Veracruzana

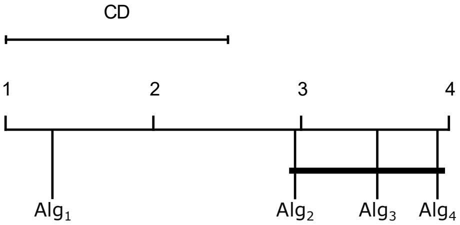
Post-hoc Test

- The performance of two classifiers is significantly different if their corresponding average ranks differ by at least the critical difference:

$$CD = q_\alpha \sqrt{\frac{A(A+1)}{6D}} \tag{8}$$

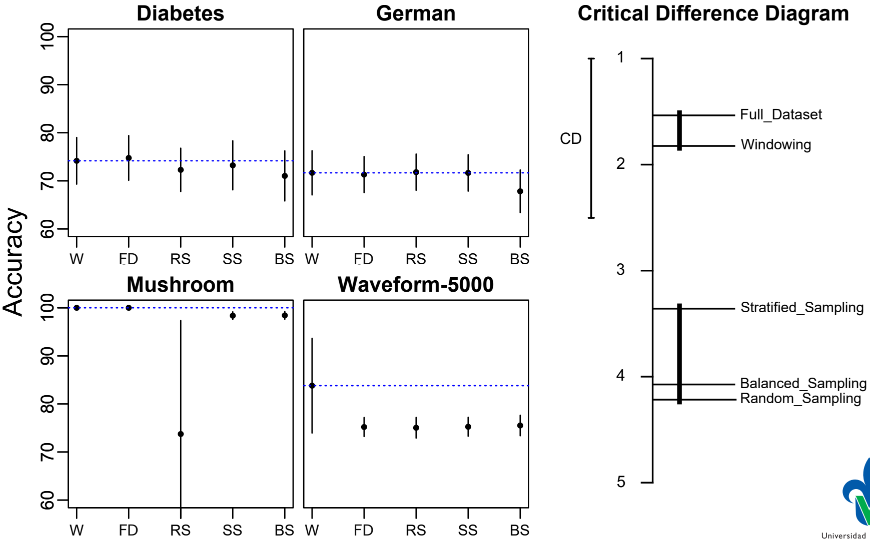
Critical values q_α are based on the studentized range divided by $\sqrt{2}$.

- Results can be visually represented with a Critical Difference diagram.

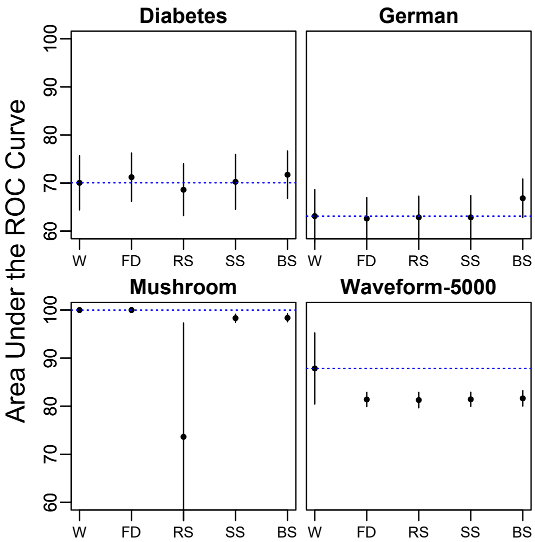


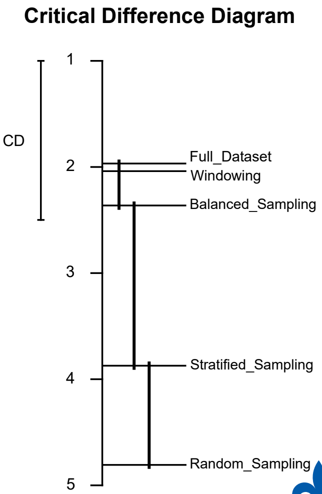
Universidad Veracruzana

Results: Accuracy

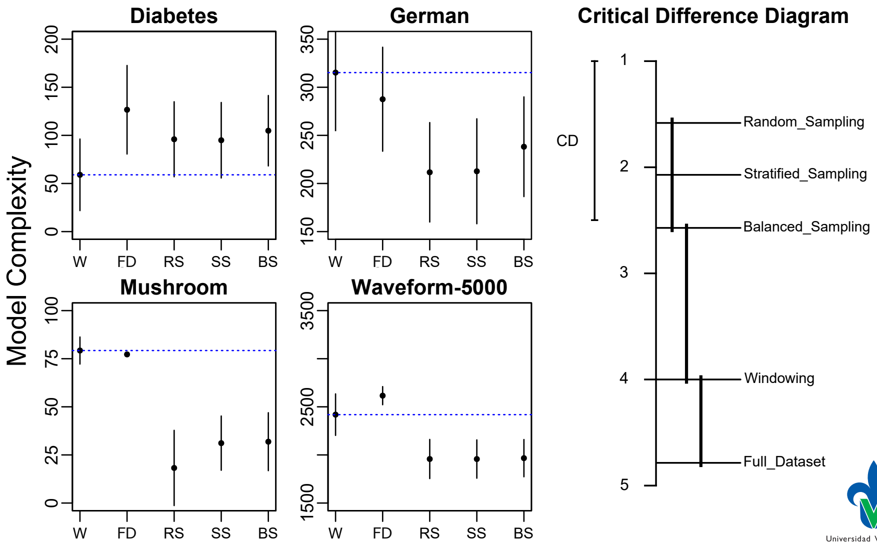


Results: Area Under the ROC curve



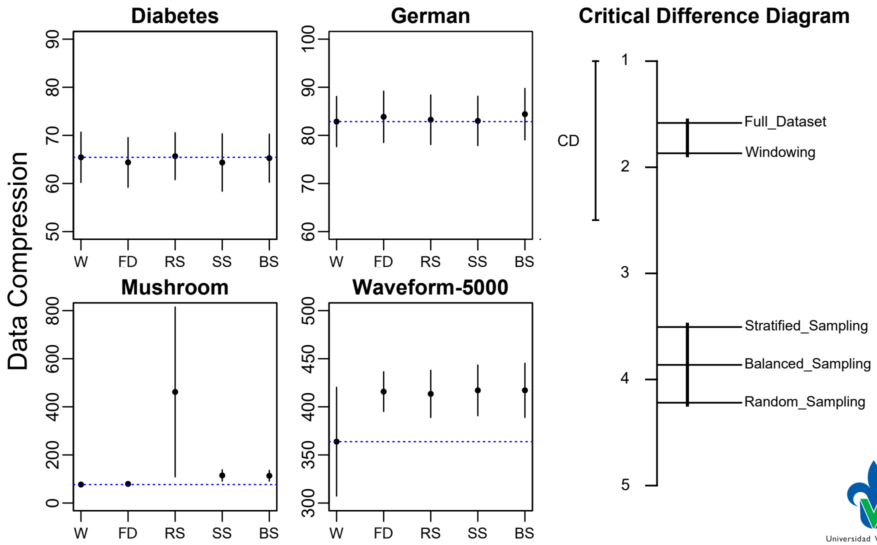


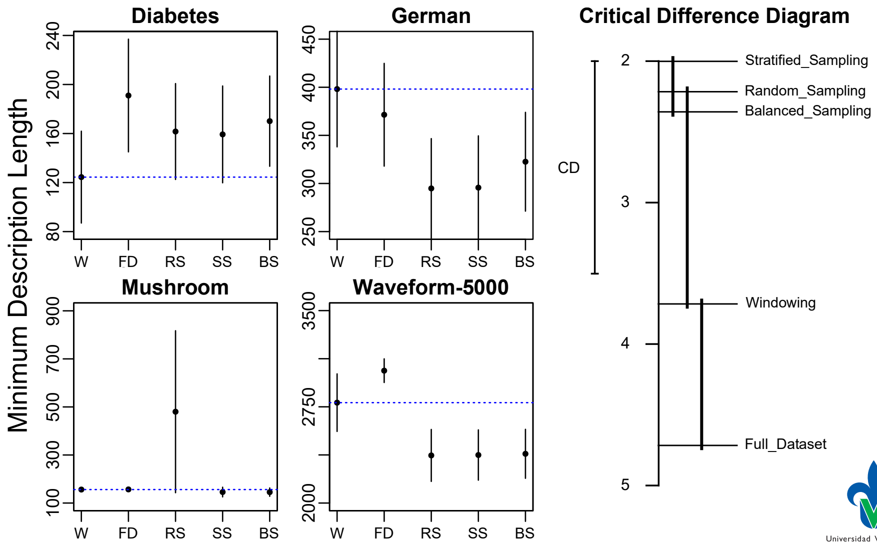
Results: Model complexity



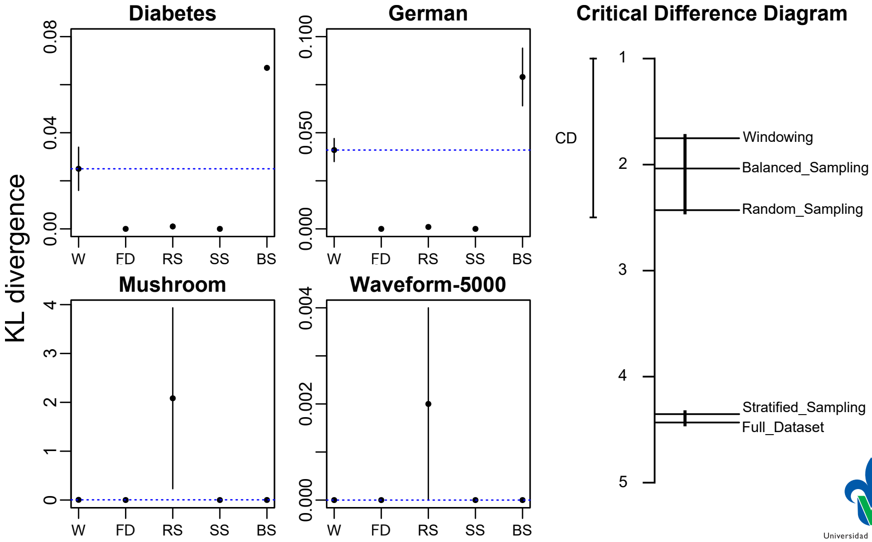
Universidad Veracruzana

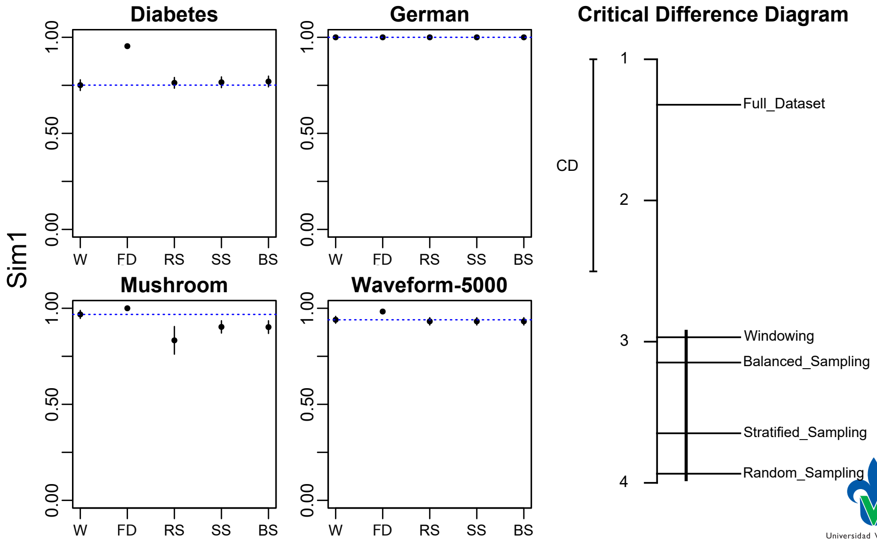
Results: Data compression




Minimum Description Length
Diabetes
German
Mushroom
Waveform-5000
W FD RS SS BS
Critical Difference Diagram
CD
Stratified_Sampling
Random_Sampling
Balanced_Sampling
Windowing
Full_Dataset

Results: KL Divergence




Sim1
Diabetes
German
Mushroom
Waveform-5000
1.00
0.50
0.00
W
FD
RS
SS
BS
Critical Difference Diagram
CD
1
2
3
4
Full_Dataset
Windowing
Balanced_Sampling
Stratified_Sampling
Random_Sampling

Window evolution over time

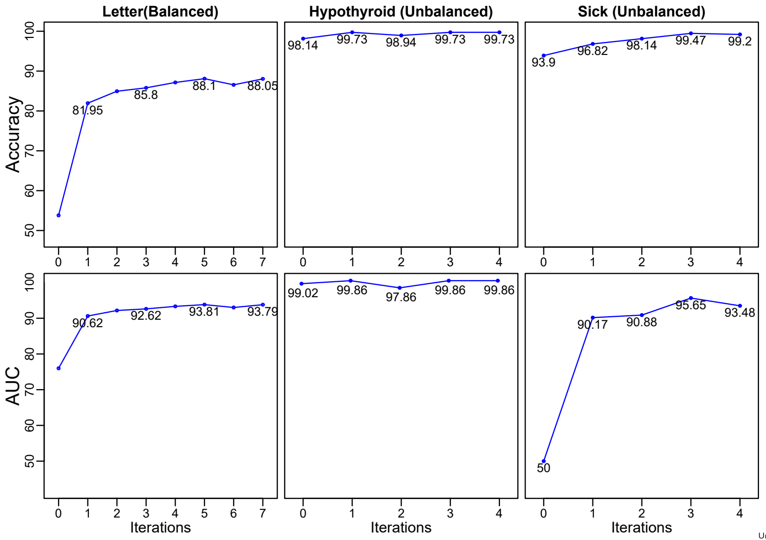
This experiment aims to yield a full description about the evolution of the windows and their effects on the model.

For this:

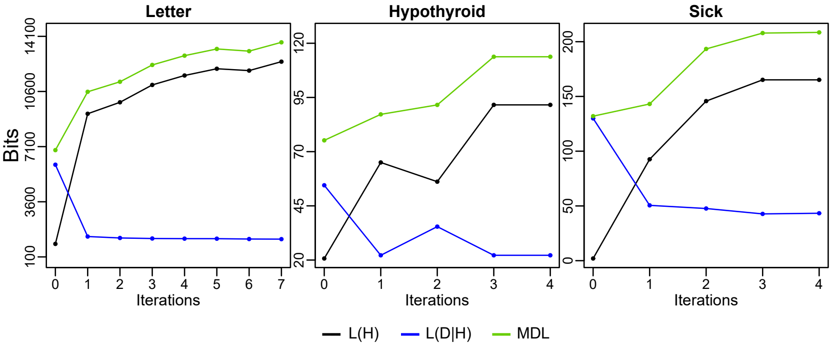
- Counter was modify in order to save the window evolution.
- A 10-fold stratified cross-validation is run by every dataset.
- Metrics in experiments A and B were calculated every iteration.
- Decision trees (j48) are adopted as classifiers.

Universidad Veracruzana

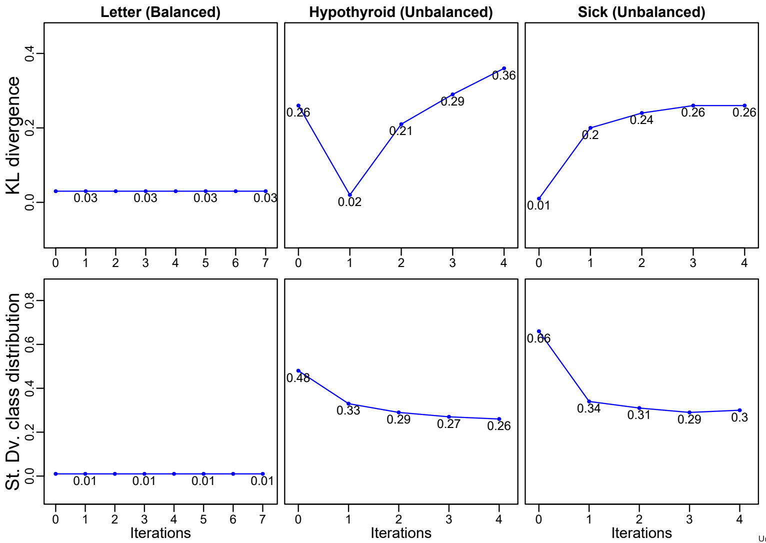
Results: Evolution of performance



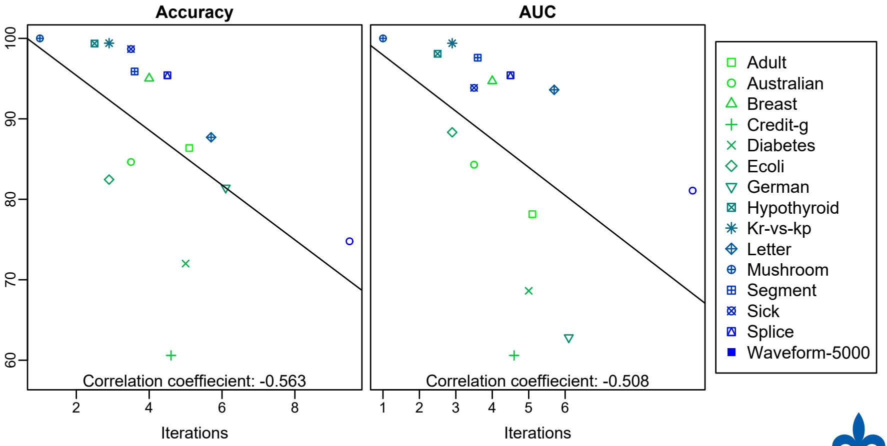
Results: Evolution of MDL



Results: Evolution of the class distribution



Results: Iterations vs. Accuracy



Universidad Veracruzana

Conclusions

Counter, the Windowing-based learning strategy, not only supplies a natural workflow for distributed scenarios, but it also offers some benefits:

- **A homogeneous behavior beyond decision trees.** It allows the induction of accurate models while performing an aggressive sampling.
- **The determination of an appropriate sample size.** This problem is often tackled most of the time by trial and error.
- **Decision trees with better data compression.** Models tend to be larger but more accurate than traditional samplings.
- **Samples with more balanced class distributions.** This behavior is restricted by the number of instances and their relevance.

Universidad Veracruzana

Future Work

This work suggests future lines of research on Windowing, including:

1. Optimizing the search model process.
2. Adopting metrics for detecting relevant data.
   - PhD proposal: Detection of noisy, redundant, and relevant data to improve the Windowing performance.
   - Maillo *et al.* [23] review multiple metrics to describe redundancy, complexity, and density of a problem and also propose two data big metrics.
3. Dealing with datasets of higher number of dimensions.

Universidad Veracruzana

References I

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, p. 27–34, Nov. 1996.

L. Zeng, L. Li, L. Duan, K. Lu, Z. Shi, M. Wang, W. Wu, and P. Luo, "Distributed data mining: A survey," *Information Technology and Management*, vol. 13, 12 2012.

J. R. Quinlan, "Induction over large data bases," Tech. Rep. STAN-CS-79-739, Computer Science Department, School of Humanities and Sciences, Stanford University, Stanford, CA, USA, 5 1979.

J. Wirth and J. Catlett, "Experiments on the costs and benefits of windowing in ID3," in *Machine Learning, Proceedings of the Fifth International Conference on Machine Learning, Ann Arbor, Michigan, USA, June 12-14, 1988* (J. E. Laird, ed.), pp. 87–99, Morgan Kaufmann, 1988.

Universidad Veracruzana

References II

J. Fürnkranz, "More efficient windowing," in *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference, AAAI 97, IAAI 97, July 27-31, 1997, Providence, Rhode Island, USA* (B. Kuipers and B. L. Webber, eds.), pp. 509–514, AAAI Press / The MIT Press, 1997.

X. Limón, A. Guerra-Hernández, N. Cruz-Ramírez, and F. Grimaldo, "Modeling and implementing distributed data mining strategies in JaCa-DDM," *Knowledge and Information Systems*, vol. 60, no. 1, pp. 99–143, 2019.

J. Fürnkranz, "Integrative windowing," *Journal of Artificial Intelligence Research*, vol. 8, pp. 129–164, 1998.

X. Limón, A. Guerra-Hernández, N. Cruz-Ramírez, H. G. Acosta-Mesa, and F. Grimaldo, "A windowing strategy for distributed data mining optimized through GPUs," *Pattern Recognition Letters*, vol. 93, pp. 23–30, 7 2017.

I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Toools and Techniques*.

Burlington, MA., USA: Morgan Kaufmann Publishers, 2011.

Universidad Veracruzana

References III

D. Dua and C. Graff, "UCI machine learning repository," 2017.

A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "Moa: Massive online analysis," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1601–1604, 2010.

G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*, (San Mateo), pp. 338–345, Morgan Kaufmann, 1995.

W. W. Cohen, "Fast effective rule induction," in *Twelfth International Conference on Machine Learning*, pp. 115–123, Morgan Kaufmann, 1995.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation*, p. 318–362.

Cambridge, MA, USA: MIT Press, 1986.

J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning* (B. Schoelkopf, C. Burges, and A. Smola, eds.), MIT Press, 1998.

Universidad Veracruzana

References IV

S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

S. Zhang, C. Zhang, and X. Wu, *Knowledge Discovery in Multiple Databases*. Advanced Information and Knowledge Processing, London, UK: Springer-Verlag London, Limited, 2004.

M. Møller, "Supervised learning on large redundant training sets," *International Journal of Neural Systems*, vol. 4, no. 1, pp. 15–25, 1993.

J. Rissanen, "Stochastic complexity and modeling," *The Annals of Statistics*, vol. 14, pp. 1080–1100, 1986.

J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, p. 1–30, 2006.

M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

Universidad Veracruzana

References V

M. Friedman, "A comparison of alternative tests of significance for the problem of *m* rankings," *Ann. Math. Statist.*, vol. 11, pp. 86–92, 03 1940.

J. Maillo, I. Triguero, and F. Herrera, "Redundancy and complexity metrics for big data classification: Towards smart data," *IEEE Access*, vol. 8, pp. 87918–87928, 2020.

Universidad Veracruzana

Conditional Population Entropy

$$CPE = -\sum_{i=1}^{n_c} p(c_i) \sum_{a=1}^{n_a} \sum_{v=1}^{n_{v_a}} p(x_{a,v}|c_i) \cdot log_2 p(x_{a,v}|c_i)$$

Where:

- n_c is the number of classes, n_a is the number of attributes.
- n_{v_a} is the number of values for the attribute a.
- c_i stands for the $i - th$ class.
- $x_{a,v}$ represents the $v - th$ value of attribute a.

Universidad Veracruzana

Counter configuration

Parameter	Value
Maximum number of rounds	10 - 15
Initial percentage for the window	0.20
Validation percentage for the test	0.25
Change step of accuracy every round	0.35

Universidad Veracruzana

Auto-adjust stop procedure

The *changeStep* parameter defines a threshold. If the accuracy of the current model compared with the accuracy of the previous model surpasses this parameter, then other round is computed, otherwise, the process stops.

Universidad Veracruzana