Clus Active-Learning – User Guide – version 1.0

Felipe Kenji Nakano (felipekenji.nakano@kuleuven.com)

https://itec.kuleuven-kulak.be/?page_id=4701
June 30, 2020

Contents

1 Introduction

Clus is a machine learning framework which employs predictive clustering trees to address supervised and semi-supervised tasks [3]. We have extended the original framework to include an Active Learning module. The main results can be found at our paper: " Active Learning for Hierarchical Multi-Label Classification (to appear in Data Mining and Knowledge Extraction)"

For a more detailed description of how Active Learning works and predictive clustering trees, we address the reader to the following references:

- Settles, B. (2009). Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences.
- Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. Machine learning, 73(2), 185.

The remainder of this manual is described as following: Section 2 presents instruction on how to get started with Clus-AL. Next, Section 3 brings an example on how to run Clus-AL.

2 Getting Started

The newest version of Clus-AL is available at:
`https://itec.kuleuven-kulak.be/?page_id=4701`

The framework is developed using the programming language Java, and that is the only requirement to run Clus-AL.

Further, we introduce the settings file used as input.

2.1 Settings File

As its input, Clus-AL requires a separate settings file describing each parameter. The settings file is divided by sections, such as Data, General and Hierarchical. For the Active Learning module, we have created an Active section. Please find below the description of each Active parameter.

> - ActiveDataset = File in the .arff format representing the unlabelled data. Mind that, experiments are simulated, thus this file should contain the labels.
> - BudgetPerIteration = Numerical value representing the amount of extra budget added per iteration. Leftovers are carried over.
> - ActiveLearningAlgorithm = String value representing the Active Learning algorithm. The full list is available at Section CITAR.
> - BatchSize = Numerical value representing the number of instance-label pairs given to the Oracle every iteration.

> - LabelCost = Array of numerical values representing the cost associated to querying a instance-label pair in such level, e.g. [1,2,3,4] for a 4-level hierarchy.
> - WriteActiveQueriedInstances = Boolean value which determines whether the queried instances will be outputted.
> - WriteActiveTrainPredictions = Boolean value which determines whether the prediction probabilities of the train dataset will be outputted.
> - WriteActiveTestPredictions = Boolean value which determines whether the prediction probabilities of the test dataset will be outputted.
> - WriteActiveTrainError = Boolean value which determines whether the error of the train dataset will be outputted.
> - WriteActiveTestError = Boolean value which determines whether the error of the test dataset will be outputted.
> - Iteration = Numerical value which determines the number of iterations. If not specified, Clus-AL will run until there are no instance-label pairs available anymore.

2.2 Active Learning Algorithms

- Random = Select instance-label pairs without any criteria;
- UncertaintySampling = Select the instance-label pairs with minimum value using Equation 1, proposed by [1];
- HCAL = Algorithm proposed by [5]. This algorithm uses two parameters: MaxIterations and PopulationSize. They are set in the Active section;
- QueryByCommittee = Select the instance-label pairs with maximum values using Equation 2, proposed by [2];
- SSMAL = Algorithm proposed by [4]. This algorithm uses an alpha parameters determined by setting Alpha in the Active section;
- QueryByCommitteeHierarchy = Select the instance-label pairs with maximum values using Equation 3.

$$Unc(\mathbf{x}_i, y_j) = |P(y_j|\mathbf{x}_i) - 0.5| \tag{1}$$

$$Var(\mathbf{x}_i, y_j) = \frac{\sum_f (P_f(y_j|\mathbf{x}_i) - \overline{P(y_j|\mathbf{x}_i)})^2}{F - 1} \tag{2}$$

$$H\text{-}QBC(\mathbf{x_i}, y_j) = Var(\mathbf{x_i}, y_j) + \frac{\sum_{y_a}^{Anc(y_j)} Var(\mathbf{x_i}, y_a) + \sum_{y_d}^{Desc(y_j)} Var(\mathbf{x_i}, y_d)}{|Anc(y_j)| + |Desc(y_j)|} \tag{3}$$

3 Example

To perform a single run, download Clus-AL and unzip its contents. Next, run the following command:

```
java -jar Clus.jar -active -hsc -forest settings.s
```

All three parameters “-active", “-hsc" and “-forest" are necessary to run Clus-AL.

- “-active" tells Clus to run the Active Learning framework;
- “-hsc" tells Clus to run Clus-HSC version. Recall that this is necessary since Label-Based methods require an underlying local classifier;
- “-forest" tells Clus to use random forests as base classifiers;

Considering the settings file described in Table 1, Clus-AL will run for 3 iterations, selecting 10 instance-label pairs per iteration using the H-QBC algorithm, spending a maximal of 10 Budget and the error of the test datasets will be reported for each iteration. A straightforwardly usable version of such settings file is available at the website.

References

[1] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*, pages 148–156, 1994.

[2] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[3] Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. Decision trees for hierarchical multi-label classification. *Machine learning*, 73(2):185, 2008.

[4] J. Wu, C. Ye, V. S. Sheng, J. Zhang, P. Zhao, and Z. Cui. Active learning with label correlation exploration for multi-label image classification. *IET Computer Vision*, 11(7):577–584, 2017.

[5] Yifan Yan and Sheng-Jun Huang. Cost-effective active learning for hierarchical multi-label classification. In *IJCAI*, pages 2962–2968, 2018.

Table 1: Example of settings file

		#Mandatory/Default	#Values
[General]			
Verbose	=	-1	int: {-1} Set this value to -1 to avoid unnecessar
[Data]			
File	=	labeled.arff	String: path to file to already labeled dataset
TestSet	=	test.arff	String: path to test dataset
[Active]			
ActiveDataset	=	unlabelled.arff	String: String: path to the unlabeled dataset (pr
BudgetPerIteration	=	10	int: higher than 0
ActiveLearningAlgorithm	=	QueryByCommitteeHierarchy	One of the algorithms described in Section 2.2
BatchSize	=	10	int higher than 0
LabelCost	=	[1,1,1,1,1,1]	One value per level of the hierarchy
WriteActiveTestError	=	True	Output the test error
[Hierarchy]			
Type	=	Tree	Only trees are available in the moment
HSeparator	=	/	Character used to split hierarchical labels
ClassificationThreshold	=	[0,2,4,6,8...100]	values from 0 to 100 with steps of 2
[Hierarchy]			
MinimalWeight	=	5	