Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut für Verteilte Systeme
Fachgebiet Wissensbasierte Systeme (KBS)

Lecturer: Prof. Dr. Eirini Ntoutsi

Teaching assistant: Vasileios Iosifidis 1.2.2018

Data Mining II

Winter semester 17/18

Project 3: Semi-Supervised Learning

In this project we, deal with classification under limited labels. We will explore two different methods of semi-supervised learning: Self-Learning and Co-Training.

The main idea in Self-Learning is to use the labeled set L to build a classifier, then iteratively apply the model to the unlabeled corpus U and in each iteration, expand the training set with instances from the unlabeled corpus which were predicted with high confidence, according to a threshold δ, by the classifier.

Co-Training assumes that the feature space, lets denote it by X, can be split into two parts, X = (X^1, X^2), called “views". Co-Training algorithm trains two classifiers $Φ^1$, $Φ^2$ each working exclusively on one view, X^1, X^2 , respectively. Initially, both classifiers are trained over the labeled set L, but each on its own view, X^i, i ∈ {1, 2}. The unlabeled data are utilized as follows: At each iteration, $Φ^1$ classifies a few unlabeled instances for which he is more confident about and appends them to the training set. Similarly for $Φ^2$. Co-Training then updates both classifiers with this additional “pseudo-labeled" data. More information of the algorithm can be found here[1] (Table 1).

<u>Dataset Census Income:</u>

Census dataset[2] contains demographic information on people in the US. There are 8 categorical attributes (such as education, occupation, marital status, etc.) and 6 numerical attributes (such as age, capital gain, etc.). Each person is classified according to whether they receive an annual salary of more than $50K or less (class attribute).

<u>Dataset Sem Eval 2017:</u>

SemEval2017[3] is a twitter dataset annotated by humans experts for the task of sentiment classification. It contains 3 classes (positive/negative/neutral) but in this project we will focus on binary classification so we have deleted neutral class. File is comma ‘,’ separated: first column contains class label and second column contains tweet’s text.

- **Task 3.1: Self-Learning Classification for SemEval2017 dataset**

  In this task, you have to apply Self-Learning for text classification. You will use SemEval2017 for this task. You must use a probability classifier. For SemEval2017 report:

  - number of total and distinct words.
  - number of tweets per class.
  - plot #tweets per #words (x-axis number of words, y-axis number of tweets).

  This dataset contains class labels for all instances. Use Stratified Sampling to obtain 10% instances of each class (consider this subset as your **initial training set**) while the remaining 90% will be your unlabeled set (do not discard the class label).

  Using Self Learning try to annotate the unlabeled dataset by accepting only 2.000 in each iteration (must select the best candidates) using as confidence threshold δ = 0.65%

  In each iteration report accuracy and f1 scores and plot them in the end (in order to do that use the class labels which you ignored in the unlabeled dataset).

  Using the same training (sub) set, repeat the exercise and plot the results (accuracy/f1 scores) for δ = 0.80% and δ = 0.95% .

  Report the number of instances per class for each δ (plot them in each iteration for each δ). What do you observe while δ increases ? Explain the differences.

- **Task 3.2: Co-Training Classification for Census Income dataset**

  In this task you will use Census income dataset. You have to find 2 different and independent views of the features space (try to separate them equally, 7 attributes each, use Feature Selection methods[4]). Explain how you chose these 2 feature spaces.

  You have, again, to separate the dataset to 10% training set and 90% unlabeled data (using stratified sampling). Using Co-Training algorithm[1] (in Table 1 you can see the pseudo code of the algorithm, set p = 500, n = 200 and u=4*p + 4*n, consider as positive class '<=50' and negative '>50') try to annotate the unlabeled dataset and report in each iteration accuracy and f1 scores for each classifier.

- **Task 3.3: Self-Learning Classification for Census Income Dataset**

  Apply Self-Learning in Census Income dataset but now use Self Learning instead of Co-Training. In each iteration accept the best 500 positive and 200 negative instances instead of setting a threshold δ.

  Plot the results in each iteration.

  Compare results of task 3.2 with Self-Learning results.

Notes:

- You are free to choose any tool/language you prefer, but you must provide the code of the different steps.
- You can form groups of 1-4 people (max. 4 !!!). All group members should be listed in the report. All group members receive the same grade (for the given project).
- Deadline: 10/03/2018 (23:59 Berlin Time). Any submissions afterwards will not be evaluated.

Deliverables:

- A report on the project.
- Source code and experimental results (as e.g., a zip).
- Email the aforementioned to **iosifidis@L3S.de** with subject "Data Mining 2 - Project 3".

References

1. https://www.cs.cmu.edu/~avrim/Papers/cotrain.pdf
2. Census dataset: http://archive.ics.uci.edu/ml/datasets/Census+Income
3. Available through StudIp (SemEval2017.csv.zip)
4. http://scikit-learn.org/stable/modules/feature_selection.html