**Data Mining Lab**

**Winter Semester 2019/2020**

**Topic: #12 Cost-sensitive boosting**

**Handed on out:** 17.11.2019

**Submission date:** 28.02.2020

**Mentor:** Eirini Ntoutsi

**Team:**

**(1)** Arab Oghli, Omar 10006018

**(2)** Kanafani, Hussain

**(3)** Mughrabi, Muaid

**Introduction:**

\* What is the problem you are trying to solve?

Classification of data with imbalanced class distributions using different algorithms.

\* Why it is important?

The problem is critical in many domains, such as medical diagnosis, network insturion, fraud detection, and ensuring nonbiased personal screening.

\* Why existing approaches cannot solve it and a new one is needed?

Accuracy is less meaningful measure in the context of the class imbalance problem, thus rendering accuracy improvement as unimportant.

**Method:**

\* What is the method about?

The used method supplements the known AdaBoost algorithm by assigning a cost to each instance indicating its identification importance.

**Goal:**

Understanding the basic method (related work), replicating the methods suggested by the paper to analize and understand their impact, and trying to extend the provided solution.

**Tasks:**

**T1. Get familiar with the proposed method (all three variations of Adaboost plus the original Adaboost)**

**T2. Get familiar with the related work in the field. You should be able to describe the current state of the art. How do you assess the contribution of the new methods?**

AdaBoost, AdaCost.

**T2. How the proposed method improves the state of the art? Provide 3 positive and 3 negative aspects of the method**

Positive:

1. Pays more attention to the minority class.
2. Data independent. Does not create excess data that needs verification (such as SMOTE).
3. toDo

Negative:

1. Can create a Tradeoff between the accuracy and regarding the minority class.
2. Initial weights can be regarded as parameter that needs fine tuning.
3. Not suitable for real-valued classification. As the number of classes and their initial weights should be known In advance.

**T3. Implementation of the method. If available, please ensure that you understand how it works, you can validate its correctness and that is in accordance to the paper (if not, understand where/why).**

**T4. Experimental evaluation**

**T4.1. Evaluation aspects: what do you want to evaluate, why and what sort of experiments do you need?**

F-mesuare, gmean. Accuracy can not be used as a measure due to data imbalance. Trying different imbalance ratios with different cost setups. Choosing different cost setups is planned to be done as suggested in the paper.

**T4.2. Justified selection of datasets. Please include some not included in the original paper**

As proposed in the paper. The selected dataset should contain some sort of imbalance between its labels. Medical diagnosis datasets are therefore good candidates. Additionally, we will use a bank marketing dataset that contains more instances than the proposed datasets.

Another dataset is used also as an example to multi-class problem.

**T4.3. Justified selection of competitors and baselines.**

Due to time constraints, only simple to implement competitors are in the point of interest. AdaCost and AdaBoost are therefore the most appropriate candidates as comparison baseline. The work can also be extended to consider other competitors as CSB2.

AdaBoost is the original boosting algorithm and should be therefore considered -possibly with accuracy/ regarding minority class tradeoff-, while AdaCost provides an older state-of-the-art baseline.

**T4.4. Justified selection of evaluation measures.**

As mentioned earlier accuracy can not be used as an evaluation measure. The following list describes the used evaluation measures, each with its own learning objective:

G-mean: to balance identify ability between two classes.

F-measure: achieve high recognition success of the small class.

**T4.5. Sensitivity of the method to parameters**

Imbalance ratio, cost setup are the main parameters to be considered.

**T4.6. Except for the quantitative evaluation, can you also show some qualitative results? What sort of results? What do we want to see?**

Accuracy tradeoff. If the data shows such behavior, it can be plotted.

**T5. Extensions: Suggest 3 possible directions for extension/improvement. Can you pursue such an extension in the context of the project? Yes/No and why. If yes, add another task below and break it down into subtasks if possible.**

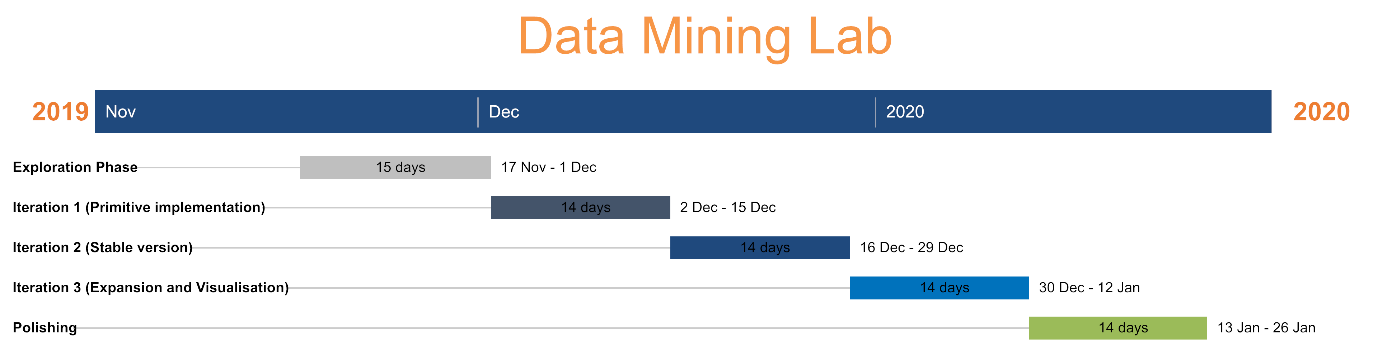
1. ML Pipeline: increases reusability and flexibility of the project.
2. Using Different weak learners. Decision Tree, linear classifier, etc..
3. toDo

**T6. What do you think of the problem, method, solution, experimental findings?**

**T7. Presentation**

**T8. Poster**

**Timetable:** //Please add a Gantt chart with the tasks, deadlines and persons involved.

****

**Literature: //what other resources except for the original paper you have used? Please cite properly and explain how they are relevant for your project.**

**Communication protocol:**

**How/when do we synchronize on progress/ problems etc? How do we ensure continuation in the reporting progress (what should be reported from meeting to meeting, we cannot start every time from scratch).**

Gitlab provides a project management tool. Issues and milestones are used to plan the project in both short and long term. Team members meet at least 1x week to discuss the workflow of the project -scrum like meetings-.

**How do you internally organize the work?**

First come first served principle, regarding the team members skills and wishes.