CS699 SC1 – Summer 2020
Project Assignment

There are three options. You need to choose one of these options.

The project must be performed by a team of two students. Every team must present their project.

Option 1

The goal of option 1 is to give students an opportunity to perform a **classification** data mining task.

You choose a real world dataset, define your own data mining goal (a **classification**), and perform necessary data mining tasks to achieve the goal. It is strongly suggested that you choose a data mining goal that has a potential for practical use. It is also strongly suggested that you find a "fresh" dataset, which, to the best of your knowledge, was rarely used by other people You should not select a synthetically generated dataset and you should not use a dataset from UCI Machine Learning Repository. You must also avoid using a dataset on the Kaggle website that has been used by many people. You may want to check government (federal, state, or municipal) websites.

Once you build data mining models, you must evaluate the data mining result using appropriate performance measures.

The following specifies minimum requirements. You can choose a larger dataset and you can perform additional tasks not mentioned in the requirements if you want.

- The project must be "**classification**." If you are interested in other types of data mining, indicate it in your proposal. I will review it and may approve it.

- Dataset minimum requirements
  - At least 20 attributes
  - At least 300 tuples

  If you are interested in a certain dataset but it does not meet the above requirements, then indicate that in your proposal. I will review it and may approve it.

- Data mining minimum requirements
  - You need to consider at least four attribute selection methods (that are implemented on Weka) plus a set of attributes chosen by yourself.
  - You need to build classifier models using at least five different classifier algorithms for each chosen set of attributes. So, you need to build and test total at least 25 classifier models.

- You may try any data preprocessing/preparation/transformation to increase the performance of your classifier models.

- Model testing
  - Once you complete data preprocessing, you must **split your dataset into a training dataset and a test dataset**. You must make sure that the class distribution is preserved in both datasets.
  - You build your models from the training dataset and you **test your models on the test dataset**.

- Performance comparison
  - Compare performance of all 25 classifier models you built using the following performance measures: accuracy, TP rates, FP rates, ROC curve (or area under curve), and other measures if you want.
  - Choose one model that you think is the best for your data mining goal. You need to justify why you chose that model.
  - Since you need to use at least five attribute selection algorithms and at least five classifier algorithms, you need to compare at least twenty five classifier models.

Schedule and Deliverables

(Only one member of each team needs to submit proposal and final project report.)

1. Proposal
   a. Due: 6/2
   b. Include the names of your team.
   c. Dataset description: You must include the source of your dataset and detailed description of it. Your description must include the names and meanings of all attributes as well as the number of tuples and the number of attributes.
   d. Clearly state your data mining goal (e.g., I want to predict whether a new customer will buy a computer or not).
   e. **Clearly indicate which attribute is the class attribute**.
   f. You also need to submit your dataset.
2. Final project report due: 7/21
   You must submit all project documentation as described below. This is a hard deadline and there will be a 10% late penalty per day after the deadline.
3. Project report
   a. A project report should include:
      (1) Cover page
      (2) Statement of your data mining goal
      (3) Detailed description of the dataset
      (4) Detailed description of data mining tool(s) or algorithm(s) you used
      (5) Detailed description of data mining procedure (the procedure you actually followed) including all data preprocessing you performed. You must show the attributes selected by each attribute selection method and the attributes you chose.

(6) Data mining result and evaluation:
    a. You must include all performance measures, including confusion matrices, from Weka's output window for all 25 models.
    b. Justification for your selection of the best model
(7) Discussion and conclusion, including what you learned from this project.

b. In your report, you must clearly state what each team member did for this project.
c. Your report must be at least 10 pages long (with 12pt font and single spaced).

4. When you submit your project report, you also need to submit **all datasets**, including
    a. Initial dataset
    b. Initial training and test dataset
    c. The training and test datasets that were used for your best model
    d. Other intermediate dataset(s) if needed
5. Other deliverables may be required based on the nature of your individual project, which will be determined after I have more information about your project.
6. Presentation:
    a. Each team will have 15 – 20 minutes for presentation.
    b. All students must be present in the class during the presentations. If you do not attend a presentation (when other teams are presenting), 5 points will be deducted for each missed presentation.

<u>Grading</u>

- Project overall and project report: 70 points
- Presentation: 20 points
- Participation: 10 points

<u>Project overall and report (70)</u>

- Project report is due 7/21. There is no grace period and there will be a late penalty of 10 points per day if you submit late.
- Whether the data mining result is practically usable. If your dataset was not used by other people (to the best of your and my knowledge), your project has potential for some practical use, and the performance of your model is reasonably good, then you may get an extra credit up to 10 points.
- Technical soundness of your approach. Otherwise, up to 10 points will be deducted.
- The performance of your best classification model. Note that there is no performance threshold which is used to grade your project. This is because different datasets and different data mining goals can result in different performance. I will use my own judgement considering your dataset and your data mining goal. If the performance of your models is very low (e.g., 60% or lower accuracy), then you must try to increase the performance and/or try to explain why it is so low. If you do not address such a low performance in one way or another, up to 10 points will be deducted.

- Whether all necessary components are included in the documentation. Otherwise, up to 15 points will be deducted.
- Organization of your documentation. If your documentation is poorly organized, up to 10 points will be deducted.
- Whether your discussion and conclusion is substantive and technically and logically sound. Otherwise, up to 10 points will be deducted.

Presentation (20)

- Presentations will be done on 7/28 and 8/4.
- The order of presentation will be determined alphabetically by the last name of your team members.
- Presentation slides are due as follows:
    - Teams presenting on 7/28: 7/24
    - Teams presenting on 8/4: 7/31
    - If you submit late, there will be 1point late penalty per day.

Your presentation will be graded based on the following criteria.

- Whether the presentation accurately represents what you did. Otherwise, up to 2 points will be deducted.
- Whether presentation material is well organized in describing what you did. Otherwise, up to 2 points will be deducted.
- Whether graphs and/or tables were well utilized to present the result. Otherwise, up to 2 points will be deducted.
- Whether questions are properly answered. Otherwise, up to 2 points will be deducted.

Participation (10)

- If a student misses a presentation, 5 points will be deducted for each missed presentation.

Important

It is very important that I should be able to reproduce your data mining model and data mining result based on your documentation. So, the description of your data mining procedure, including all preprocessing you performed, must be detailed and accurate. If I cannot reproduce your model and result, you will lose up to **40 points**.

Option 2

Option 2 is an experiment to determine whether a bagging method and a boosting method increase the performance of classifier models. Follow the instruction given below.

- Select 20 datasets for classification.

- Select 5 classification algorithms.
- For each dataset *D* and each classifier algorithm *A*, perform the following:
  - Run *A* on *D*, with 10-fold cross-validation chosen as the test method, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC*.
  - Run *Bagging* with *A* on *D*, with 10-fold cross-validation chosen as the test method, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC* for each class.
  - Run *AdaBoostM1* with *A* on *D*, with 10-fold cross-validation chosen as the test method, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC* for each class.
- You must repeat the above 100 times (20 datasets x 5 classifier algorithms).
- Then, organize your result, present your result (as a table, graph, or any other format), and draw your conclusion. Try to be creative when you present your result so that your result may be effectively conveyed to readers of your report. Remember that your goal is to determine whether those ensemble methods increase classifier performance.

Schedule and Deliverables

(Only one member of each team needs to submit deliverables)

1. Proposal
   - Due: 6/2
   - Submit all datasets you chose.
   - Description of all datasets:
     For each dataset, you must include:
     - The name of the dataset
     - The number of tuples and the number of attributes
     - Names and meanings of all attributes
     - Name of the class attribute and class distribution
     - Source of the dataset
   - Names of the classification algorithms you chose
2. Project report
   - Due: 7/21
   - Your project must include:
     - Cover page
     - If you performed any preprocessing on any dataset, you need to describe in detail the preprocessing you performed and you also need to submit the final dataset that was created after the preprocessing.
     - Result of the experiment: You need to present your result using tables, graphs, charts, or in other visual format so that readers of your report can easily and effectively understand your result.
     - Discussion and conclusion

Grading

- Project overall and project report: 70 points
- Presentation: 20 points
- Participation: 10 points

Project overall and report (70)

- Project report is due 7/21. There is no grace period and there will be a late penalty of 10 points per day if you submit late.
- If the whole or part of the experiment is not technically sound/correct, up to 20 points will be deducted.
- Whether all necessary components are included in the documentation. Otherwise, up to 15 points will be deducted.
- Organization of your documentation. If your documentation is poorly organized, up to 10 points will be deducted.
- Whether your discussion and conclusion is substantive and technically and logically sound. Otherwise, up to 10 points will be deducted.
- If the presentation of your result is considered "excellent" you will get extra 10 points.

Presentation (20)

- Presentations will be done on 7/28 and 8/4.
- The order of presentation will be determined alphabetically by the last name of your team members.
- Presentation slides are due as follows:
    - Teams presenting on 7/28: 7/24
    - Teams presenting on 8/4: 7/31
    - If you submit late, there will be 1point late penalty per day.

Your presentation will be graded based on the following criteria.

- Whether the presentation accurately represents what you did. Otherwise, up to 2 points will be deducted.
- Whether presentation material is well organized in describing what you did. Otherwise, up to 2 points will be deducted.
- Whether graphs and/or tables were well utilized to present the result. Otherwise, up to 2 points will be deducted.
- Whether questions are properly answered. Otherwise, up to 2 points will be deducted.

Participation (10)

- If a student misses a presentation, 5 points will be deducted for each missed presentation.

Option 3

This option is an experiment to compare an undersampling method and an oversampling method to handle unbalanced datasets. Follow the instruction given below.

- Select at least 10 unbalanced datasets for classification. Make sure that the class attribute is a binary attribute and the fraction of the minority class is no more than 20%.
- Select 5 classification algorithms.
- For each dataset *D* and each classifier algorithm *A*, perform the following:
  - Split D into a training dataset D_{tr} and a test dataset D_{ts}. Use about 2/3 as the training dataset and 1/3 as the test dataset. Make sure that the class distribution is preserved.
  - Build a classifier model using the algorithm *A* from the training dataset D_{tr}, and test the model on the test dataset D_{ts}, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC* for each class.
  - From the training dataset D_{tr}, create an undersampled dataset $D_{tr\text{-}us}$.
  - Build a classifier model using the algorithm *A* from the undersampled training dataset $D_{tr\text{-}us}$, and test the model on the test dataset D_{ts}, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC* for each class.
  - From the training dataset D_{tr}, create an oversampled dataset $D_{tr\text{-}os}$.
  - Build a classifier model using the algorithm *A* from the overrsampled training dataset $D_{tr\text{-}os}$, and test the model on the test dataset D_{ts}, and collect the following performance measures: *TPR*, *FPR*, *F-measure*, and *AUC* for each class.
- You must repeat the above at least 50 times (at least 10 datasets x 5 classifier algorithms).
- Then, organize your result, present your result (as a table, graph, or any other format), and draw your conclusion. Try to be creative when you present your result so that your result may be effectively conveyed to readers of your report. Remember that your goal is to determine whether undersampling is better or oversampling is better for unbalanced dataset.

Schedule and Deliverables

(Only one member of each team needs to submit deliverables)

1. Proposal
   - Due: 6/2
   - Submit all datasets you chose.
   - Description of all datasets:
     For each dataset, you must include:
     - The name of the dataset
     - The number of tuples and the number of attributes
     - Names and meanings of all attributes
     - Name of the class attribute

    - Show which class is the minority class and which class is the majority class, and also show the ratio of the two.
    - Source of the dataset
  - Names of the classification algorithms you chose
2. Project report
  - Due: 7/21
  - Your project must include:
    - Cover page
    - If you performed any preprocessing on any dataset, you need to describe in detail the preprocessing you performed and you also need to submit the final dataset that was created after the preprocessing.
    - Result of the experiment: You need to present your result using tables, graphs, charts, or in other visual format so that readers of your report can easily and effectively understand your result.

    Discussion and conclusion

<u>Grading</u>

- Project overall and project report: 70 points
- Presentation: 20 points
- Participation: 10 points

<u>Project overall and report (70)</u>

- Project report is due 7/21. There is no grace period and there will be a late penalty of 10 points per day if you submit late.
- If the whole or part of the experiment is not technically sound/correct, up to 20 points will be deducted.
- Whether all necessary components are included in the project report. Otherwise, up to 15 points will be deducted.
- Organization of your documentation. If your documentation is poorly organized, up to 10 points will be deducted.
- Whether your discussion and conclusion is substantive and technically and logically sound. Otherwise, up to 10 points will be deducted.
- If the presentation of your result is considered “excellent” you will get extra 10 points.

<u>Presentation (20)</u>

- Presentations will be done on 7/28 and 8/4.
- The order of presentation will be determined alphabetically by the last name of your team members.
- Presentation slides are due as follows:
    - Teams presenting on 7/28: 7/24
    - Teams presenting on 8/4: 7/31

- If you submit late, there will be 1point late penalty per day.

Your presentation will be graded based on the following criteria.

- Whether the presentation accurately represents what you did. Otherwise, up to 2 points will be deducted.
- Whether presentation material is well organized in describing what you did. Otherwise, up to 2 points will be deducted.
- Whether graphs and/or tables were well utilized to present the result. Otherwise, up to 2 points will be deducted.
- Whether questions are properly answered. Otherwise, up to 2 points will be deducted.

Participation (10)

- If a student misses a presentation, 5 points will be deducted for each missed presentation.