Проект Data Mining Framework

# Въведение

Основната задача в проекта е да се направи система за експериментиране и тестване на data mining алгоритми. За целта трябва да се решат следните по-малки задачи:

* модул за прочитане на данни от различни формати:
  + аrff и name/data формат ( в този формат са повечето от данните в <http://archive.ics.uci.edu/ml/>)
  + За в бъдеще да се направи и да чете от база данни
* модул, който прави стратифицирана крос валидация на данните, Този модул има 2 функции:
  + Функция getTrainingSet - като вход ще влизат:
    - На колко групи да е крос валидацията
    - прочетените данни
    - Random seed
    - номер на исканото обучаващо множество  
      **Изхода ще е обучаващо множество**
  + Функция getTestingSet - като вход ще влизат:
    - На колко групи да е крос валидацията
    - прочетените данни
    - Random seed
    - номер на исканото тестващо множество  
      **Изхода ще е тестващо множество**
* модул, който използвайки горните модули тества алгоритъм за класификация:

нека всеки алгоритъм за класификация имплементира интерфейса Classifier.

public interface IClassifier {

public void buildClassifier(ArrayList<Instance> instances);

public int classifyInstance(Instance instance);

}

Модула използва модула за крос валидация и обучава и тества алгоритъма 10 пъти. Събира резултатите и ги предоставя на модула за извеждане на резултати

* модул за извеждане на резултати – трябва да предлага няколко възможности:
  + извеждане на статистически резултати в конзола или файл
  + извеждане на графични резултати – по всеки два атрибута как се разпределят примерите като класове – прилага се различно отбелязване за правилните и грешните примери.
* Разработка на алгоритмите за машинно самообучение:
  + Nearest Neighbour
  + kNN
  + kNN варианти
  + Decision Tree
  + Neural Network(Backpropagation) - \* ако остане време
* Модул за тестване за статистически по добър алгоритъм над определени бази данни.
  + Използва се модула за зареждане на данни, крос валидация и за тестване на алгоритъм и се изпълнява T-pair test
  + Преди разработката на този модул ще се наложи research относно алгоритъма!
  + Вход:
    - Списък от бази данни и списък от алгоритми
  + Изход
    - таблица „Алгоритъм-База Данни”, в която са отбелязани резултатите от тестовете за всички алгоритми и данни и са отбелязани статистически по добрите резултати.
* Модул за тестване на алгоритми за Машинно самобучение без учител
  + Разработване на алгоритъма kMeans
  + Разработка на алгоритъма Expectation Maximization
  + Разработка на Online kMeans
* Разработка на графичен интерфейс
* Разработка на модул за динамично добавяне на нови алгоритми за класификация и/или модификация на старите.