Chapter 1: Introduction

Introduction

Problem Statement

1. Human perspective bias on classifies an object
2. Time consuming to identify possible solution and decision to support business

Objectives

Develop a program that able to train a model and display with performance metrics

1. C, C++, C#, Java (or python discussed verbally in the class)
2. Include two modules, C4.5 algorithm and naïve Bayesian
3. Build classification model based on training set (defined in text file)
4. Evaluated by testing set (defined in text file)
5. Display classification result of each testing data
6. display metrics for evaluating performance of each trained classifier

Scope

User Scope

1. Allow user enter a file name for training set
2. Allow user to enter file name for testing set after training

System Scope

1. Module 1: C4.5 Classification Module
   1. Program will develop a tree model using C4.5
   2. Illustrate each stages of developing decision tree model
   3. Display measure for every iteration in building decision tree
   4. Measure for C4.5
      1. Info(D)
      2. InfoA(D)
      3. Gain(A)
      4. SplitInfoA(D)
      5. GainRatio(A)
   5. Illustrate resulting tree
2. Module 2: Naïve Bayesian
   1. Illustrate the computational process in classifying data tuples
      1. Ci = category, xi attribute value
      2. P(Ci), the prior probability of each of the categories
      3. P (xi | Ci), the posterior probability of each of the attribute values conditioned on each of the categories.
   2. Laplacian correction to avoid any computing probability values of zero
3. Module 3: Evaluation Module
   1. Accuracy
   2. Error Rate
   3. Sensitivity (Recall)
   4. Specificity
   5. Precision
   6. F and F1, Harmonic Mean of Precision and Recall

Chapter 2: Literature Review

Overview of Machine Learning concept and techniques

Literature on Decision Tree and its variants (eg. ID3, C4.5 and CART)

Literature on Bayes Theorem and Naïve Bayesian classification

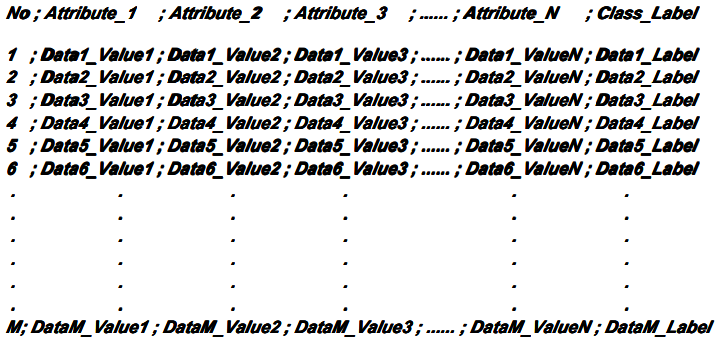
Chapter 3: Methodology

Detailed discussion on the selected methodology in relation to the design and development processes of the program

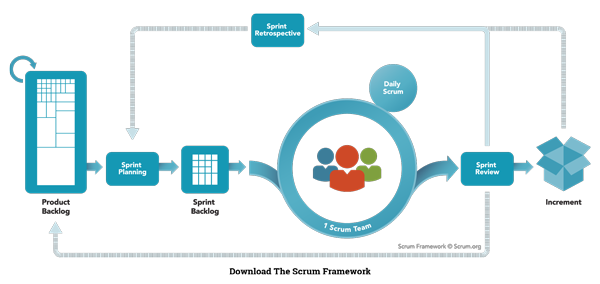
Dataset Design

The picture below showed the file data structure that use to store training set and testing set. The training set is used to create the classification model, and testing set is used to evaluate the performance of the trained model. The format of the file is defined as below:

The first row is the header it defines the list of attribute and class label for each tuple. The first column defines the unique identifier for the record. Each attribute and class label is separated by semicolon(;) and the data tuple separated by a newline.



Methodology



Scrum Model. Adapted from "Nutcache",

retrieved from https://www.nutcache.com/blog/leverage-scrum-to-manage-your-projects/

A scrum model (agile) is implemented in this project. Agile is an iterative approach for software development. The software is developed and delivered to customers in increments. Agile has the flexibility to accommodate frequent changes in the design. Scrum is one of the agile process frameworks which include product owner, scrum master, and development team. Scrum breaks the task into goals that can be completed within the timeboxed iteration, which call sprint. This is a lightweight. Iterative and incremental approach. The sprints should not longer than one month. The development team is self-organized, and responsible convert the backlog into an actual system. Eight members of the development team are required in this project. Product owner representing stakeholders and the voice of the customer. Only 1 product owner is required in this project to maximizing the value delivered by the development team. Scrum Master is responsible for ensuring the Scrum framework is followed and acts as a buffer between the team and any distracting influences. Each team required a scrum master, in this project, there is two development team which are Team A, and Team B. Scrum also included sprint planning, daily scrum, sprint review, sprint retrospective, backlog refinement, cancelling a sprint. These will be implemented in this project. (Fernandes, 2015) (scrum.org, n.d.)

Definition

Scrum events are defined as the following:

**Sprint:**A time-boxed mini project which less than 4 weeks, at the end for each sprint should deliver a releasable product or feature, Sprints include the planning, design, development and testing phase. Each sprint can assign to a synchronized team.

**Sprint planning:** A planning stage before the sprint began. Each backlog is prioritizing and review the requirement (product backlog) and created an order list for a particular sprint. Analyze the feasibility for each requirement and features finish at a particular time.

**Daily Scrum:**A meeting that held every day morning and takes less than 5 minutes. The scrum master will coordinate the team and discuss their daily goal and achievement. The obstacle also will be discussed in the meeting to seek help from another team. An unclear goal can make the team focus on their daily tasks and increase productivity.

**Sprint review:**An informal meeting establishes at the end of the sprint. The increment (product backlog) will be demonstrating to the end-user if any improvement or changes will execute in the next sprint.

**Sprint retrospective:**A formal meeting that gathers all the scrum and reviews the sprint. Each sprint will be review in this stage, which included the factor that makes the sprint or goal fail, way to improve the sprint, and etc. Then continue next sprint.

Scrum artifacts are defined as the following:

**Story:**Describe what users need to solve their problems. It describes the functionality and the features of the system which is also known as user stories. For example, login, pay and update profile.

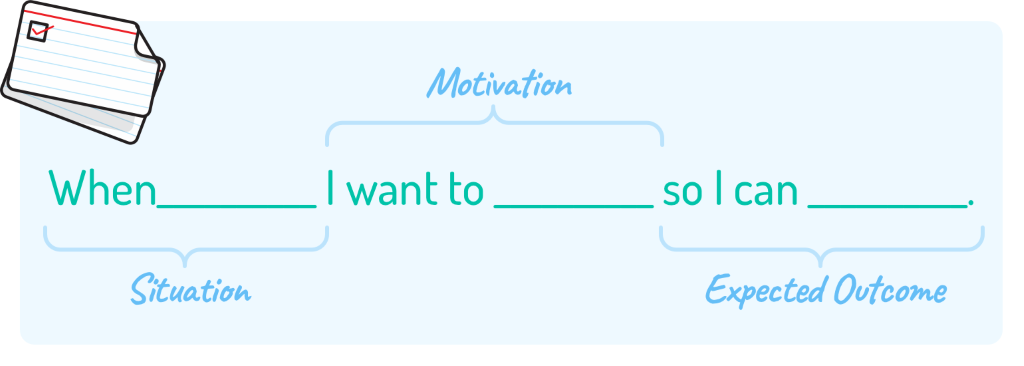


Figure 3.4 User Story. Adapted from "Mountain Goat Software", by Mike Cohn,

retrieved from <http://www.mountaingoatsoftware.com/blog/job-stories-offer-a-viable-alternative-to-user-stori>

es/feed

**Product Backlog**: An ordered list of the requirement for the product or features of the system. Product backlog includes all the requirements of the systems or features such as users can pay via credit card. A product backlog is never complete because product backlog will evolve throughout the entire development process. A product backlog is dynamic and frequently changes to fulfil the requirement of the product and what the product needed to be competitive. Adding detail, estimate, and order to items in the product backlog call product backlog refinement.  (Fernandes, 2015)

**Tasks:**A decomposed of a product backlog. Task refine the product backlog and the requirement of the product or features of the system.

User Story

User story able to help programmer to analyze the user requirement and turn it into module in programming easily.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| User ID | As a …. | I need … | So I can… | Priority |
| U001 | User | A program that able to read and input training file | Use to create a classification model based on training set | 5 |
| U002 | User | A program that able to read and input testing file | Classify the object by using classification model | 5 |
| U004 | User | Display performance metrics such as accuracy, error rate, sensitivity, specificity, precision and F-score | Evaluate the performance of classification model | 3 |
| U005 | User | Include C4.5 and Naïve Bayesian classification algorithm | Create classification model | 5 |
| U006 | User | Laplacian correction for Naïve Bayesian algorithm | Avoid any computing probability of values of zero | 4 |
| U007 | User | Display all measure value for c4.5 such as Info(D), InfoA(D), Gain(A), SplitInfoA(D) and GainRatio(A) | Evaluate the features | 3 |
| U008 | User | Display classification result of each testing data | Evaluate the features | 3 |

Deliverables

|  |  |  |
| --- | --- | --- |
| **Deliverables** | **Phase** | **Date** |
| Documentation (Chapter 1 to 3) | Analysis Phase (Planning – Define Scope) | 15/1/2021 |
| -Read from csv module  -C4.5 Module | Sprint 1 | 16/1/2021 |
| Performance Metrics Module  Naïve Bayesian Module | Sprint 2 | 17/1/2021 |
| User Interface (View)  Compile to executable file (console application) | Sprint 3 | 18/1/2021 |
| Source Code  Deployment and user manual  Final documentation  Presentation  Presentation video | Sprint 4 | 19/1/2021 |

Chapter 4: Implementation

Detailed description and explanation for each of the stages in the classification operations.

Detailed presentation, analysis and discussion of the evaluation of the trained classification models.

Chapter 5: Conclusion

Conclusion

Strengths

Weaknesses