MDFP - Project plan

Giménez Gredilla, Daniel

9th October 2017

Contents

1 - Project context and justification

1.1 - General description

The current project is framed in the context of lymphocyte classification. Lymphocyte classification is achieved through the evaluation of morphologic, geometric and colorimetric features. Current classification workflows (*Puigví et al.* [1]) make use of dimension reduction techniques to avoid the biostatistical problems derived from an overabundance of variables. This project aims to make a comparison between different dimension reduction techniques and their outputs.

1.2 - MDFP justification

Current lymphocyte classification workflows, in an oncologycal background, aim to classify normal, abnormal and reactive lymphocytes, being neoplasic lymphoid cells the most difficult to be recognized by only qualitative morphologic features (*Alférez Baquero et al.* [2]). The medical problem is apparent in this stage: lymphocite features have been, up to this point, evaluated by experts in this area; this introduces a measure of intraobserver and interobserver variation (*Puigví et al.* [1]).

In these articles, quantitative, machine-learning oriented classification workflows have been described that take morphological and geometric-colorimetric features and, through the use of dimension reduction techniques, construct a limited set of significant features explaining the most variance; this enhances the accuracy of the classification while keeping the processing requirements to a reasonable amount.

Correct lymphocyte classification improves the diagnosis of lymphoma and thus increases the survival of individuals. The optimization of every step of this diagnosis should be a goal of research. A comparison of the behaviour of dimension reduction techniques is a logical step of this improvement, leading to a better understanding of the optimal parameters of classification, and potentially improving its accuracy.

2 - Goals

2.1 - General goals

1 - To design a **comparison protocol** for different dimension reduction techniques.

2 - To apply this protocol to each technique within the frame of lymphocyte classification, achieving an **objectively quantifiable scoring system**.

2.2 - Specific goals

Specific goals for Phase 1 (17-10-2017 through 20-11-2017)

1.1 - To choose an array of dimension reduction techniques for comparison.

1.2 - To choose a programming environment to work with (languages, frameworks...)

1.3 - To extract a subset of functions from the appointed languages and frameworks to apply to the test data.

Specific goals for Phase 2 (21-11-2017 through 18-12-2017)

2.1 - To set a scoring system to satisfy the need for an objective measure of accuracy.

2.2 - To apply each of the selected dimension reduction techniques, under equivalent parameters, to the test data.

2.3 - To classify the behaviour of the referred techniques, based on the selected scoring system, as applied to the stated problem (lymphocyte classification)

3 - Focus and methods

Given the specific goal of this project (the comparison of dimension reduction techniques), the focus to accomplish it will be directed to the evaluation of the accuracy of classification tasks implementing each of these techniques. Feature construction aims to explain the most variance through the less possible, most explicative, features. For a constant given amount of variance explained through variables in a classification task, the accuracy of the predicted classes improves while the number of dimensions decreases.

The methods to accomplish this will be a selection of dimension reduction techniques, including **PCA**, **ICA** and **Factor Analysis** amongst others, applied to the problem dataset and used as input for the same machine learning classification algorithm (SVM with an RBG kernel).

This method has been evaluated as the most appropriate, as it makes it possible to subject the objects of evaluation to an equal environment, under equal conditions, and give out a numerical, objective measure of correlation with observed results.

4 - Planning, milestones and timetables

4.1 - Tasks

Tasks for Phase 1:

1.1.1 - Choose a subset from the most used and widely applied dimension reduction techniques appliable to the present topic, including **PCA**, **ICA** and **Factor Analysis**. (1 week, 21 hours equivalent)

1.2.1 - Elaborate a list of widely bioinformatics-applied languages (and frameworks, if used within one). (7 days, 21 hours equivalent)

1.2.2 - Choose a subset from those languages and frameworks and elaborate a briefing of characteristics and examples of application. (3 days, 9 hours equivalent)

1.3.1 - Elaborate a list of dimension reduction packages and functions from chosen languages. (7 days, 21 hours equivalent)

1.3.2 - Choose a subset and elaborate a briefing of package traits: optimal application, parameters, example workflows it has actually been used for, etc. (4 days, 9 hours equivalent)

1.3.3 - Elaborate monitoring report for **Phase 1**. (7 days, 21 hours equivalent)

Tasks for Phase 2:

2.1.1 - Elaborate a short briefing on the value of prediction accuracy as output by this packages. (4 days, 12 hours equivalent)

2.1.2 - Assess the validity of it for all the packages selected, and, if it is not valid for all of them, extrapolate a valid, normalized scoring system. (6 days, 18 hours equivalent)

2.2.1 - Apply each and every package's or function's workflow to the supplied lymphocyte data. (4 days, 12 hours equivalent)

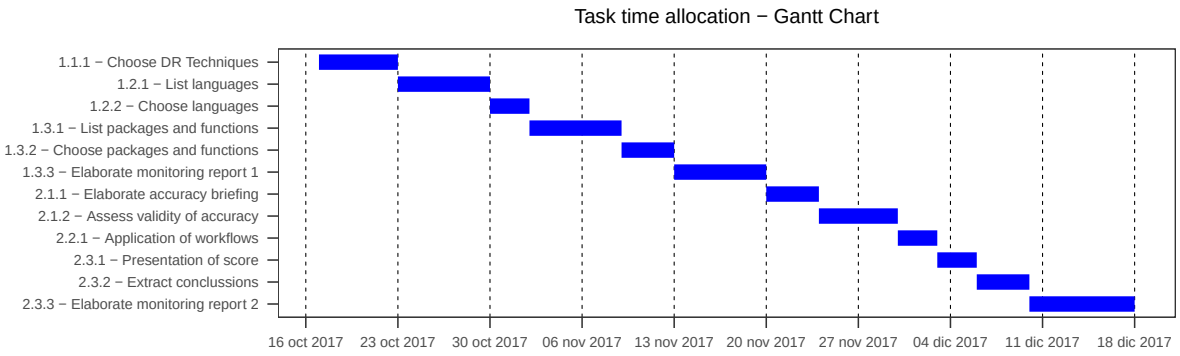
2.3.1 - Present the score output of each of the applications in a user-friendly manner. (3 days, 9 hours equivalent)

2.3.2 - Extract behaviour/comparison conclussions from this score output. (4 days, 12 hours equivalent)

2.3.3 - Elaborate monitoring report for **Phase 2**. (8 days, 24 hours equivalent)

4.2 - Calendar

The following Gantt diagram represents the division of time through the proccess of this project:



4.3 - Milestones

The following tables represent the milestones for each development phase.

Table 1: Phase 1 milestones

Deadline	Milestone
01-NOV-2017	Array of candidate bioinformatics languages, tools and protocols assessed
20-NOV-2017	Definitive subset of bioinformatics languages, tools and protocols selected
20-NOV-2017	Monitoring report for Phase 1

Table 2: Phase 2 milestones

Deadline	Milestone
28-NOV-2017	Scoring system completely defined
08-DEC-2017	Complete set of workflows applied
18-DEC-2017	Behaviour/comparison conclussions from output elaborated
18-DEC-2017	Monitoring report for Phase 2

Table 3: Post-production milestones

Deadline	Milestone
02-JAN-2018	Final report produced and delivered
10-JAN-2018	Virtual presentation produced and delivered

4.4 - Risk analysis

Some of the factors that could hinder the proposed work frames are the following:

1. Technical problems: a short buffer of time must be allocated for unexpected technical problems stemming from equipment malfunction, infrastructure breakdown, etc. Measures covering these problems include a recurrent backup system, cloud storage and accessibility to the project and its resources from several, if controlled, workstations.
2. Goal overextension: an incorrect or exaggerated choice of dimension reduction techniques or an overambitious reach could mean an ineffective use of time. This is controlled by allocating an initial time for a detailed judgement and selection of techniques to include in this project's comparison goal.
3. Incompatibilities: accuracy measurements between packages or functions in different languages or frameworks could demonstrate to be incompatible between them, or not fit to compare; this is avoided through both the allocation of time for a strict selection of these languages and frameworks, and for the production of a normalized scoring system.

Although there are many other factors that could mean an obstacle for the correct development of each phase, they are not foreseeable and, thus, to be assessed on an occurrence basis.

5 - Expected results

5.1 - Work plan

A document pertaining the project's planification will be delivered by October the 16^{th}, this being it. This document's aim is to reflect the project's expected goals and tasks to accomplish and the time frames in which to fit them. Pragmatism is expected in this planning, meaning the ability to fit realistic goals and tasks in realistic timeframes, acknowledge possible hindrances and obstacles, and establishing procedures to avoid or sort them out.

This project's work plan is been rendered via R, using packages **Rmarkdown**, **ggplot2**, **knitr**, and **reshape2**. The embedded Gantt graph is produced via **ggplot2** and **reshape2**, from an input of tasks in data frame format and a series of graphic parameters.

This document will also establish the products that will stem from the project, any additional outputs, and the monitoring and evaluation thereof.

5.2 - Report

Three reports will be made through this project's duration, structures as follows:

The first one will be a monitoring report, due November the 20^{th}, in which the project's ongoing evolution will be described. This will be composed of the description itself, a complete relation of overtaken activities, both foreseen and unforeseen, a relation of hindrances and obstacles and the measures taken to buffer them, complete with an update of time frames, a list of delivered partial results and any particular comment by the project's tutor.

Another monitoring report, due December the 18^{th}, will be generated with contents similar to the first one, this time with a focus on the completed second phase of the project and the degree of accomplishment of the planned goals for it.

The final report, due January the 2^{nd}, with a maximum length of 90 pages, will present the output of the project, with a justification of its interest, goals, methodology and materials, and results obtained.

5.3 - Product

In the course of this project, an automated comparison report script will be produced. The code used will be added as an addendum to the final report, along with code comments and protocols of use.

5.4 - Virtual presentation

The virtual presentation for this project will be carried out through **Present@**, a presentation tool offered by **Universitat Oberta de Catalunya** for the display of project results. This presentation will be comprised of approximately 20 slides with an oral presentation for a maximum of 20 minutes. This presentation's aim is to be as concise and informational as possible, while delivering the results and conclussions of the project in a clean, outreaching way.
The presentation will be produced between the 3^{rd} and 10^{th} of January 2018, January the 10^{th} being the deadline. Of special importance is the content, synthetic ability and clarity of purpose and expression. Evaluation criteria have been provided by the project's tutor.

5.5 - Project self-evaluation

This project's self-evaluation will confront it from two angles: first, a side-by-side comparison of initial goals and time schedules and final, actual results and time schedules, and second, a thorough analysis of style, clarity and informative value. Being this:

Goals and schedules:
1 - Correct assertion of techniques to compare: the techniques assessed are widely used, available to the general research personnel, and suited for the task at hand. Also, the number of techniques is decided pragmatically, avoiding overextension and, thus, decrease in effective time. 2 - Validity of scoring system and conclussions: the scoring system is, by itself or through normalisation, fit to give an objective, comparable value. The conclussions that follow are in agreement with this scoring system. 3 - Adecuation of assigned times: the assigned times corresponded to the times actually employed for each task, and so, milestones are accomplished within the expected period.

Style and structure:
1 - Style: the project is easily readable, is expressed in a correct way, follows correct style guidelines, quotes and references are strictly marked.
2 - Structure: the project follows the structure established by the documentation provided through the subject. Contents are correctly divided in sections. The project as a whole presents a semantic flow without logical leaps that may hinder the reader's comprehension.

6 - Project structure

The final project will fit the following structure.

1. Introduction

- 1.1 Context and justification for the project
- 1.2 Project goals
- 1.3 Focus and followed method
- 1.4 Project plan

- 1.5 Brief summary of products obtained
- 1.6 Brief description of other chapters

2. All other chapters
3. Conclussions
4. Glossary
5. Bibliography
6. Annexes

This structure will be part of the evaluation requirements.

Annex I - References

[1] Puigví, L., Merino, A., Alférez, S., Acevedo, A., & Rodellar, J. (2017). New quantitative features for the morphological differentiation of abnormal lymphoid cell images from peripheral blood. Journal of Clinical Pathology, jclinpath-2017.

[2] Alférez Baquero, Edwin Santiago. "Methodology for automatic classification of atypical lymphoid cells from peripheral blood cell images." (2015).