University of Nottingham
UK | CHINA | MALAYSIA

Final Year Project Proposal
COMP3050 Individual Dissertation Single Honours 2019/20

Machine Learning tools for prediction of cell subtypes from single cell expression (Mouse data)

Supervised by Prof. Vladimir Brusic

Bangrui Zheng(16522091)
zy22091@nottingham.edu.cn
BSc Computer Science

School of Computer Science, University of Nottingham Ningbo China

Motivation and Background

Biology and biotechnology.

Single cell expression is one of the cutting-edge biotechnology areas enabling a more precise and detailed analysis of gene expression pattern than bulk transcriptomics. This technology can be applied in multiple fields, including diagnosis of cancer, multiple sclerosis, diabetes, and infection [1-5]. Highly accurate detection of cell subtype and disease with machine learning tools can promote the study of developmental biology and cell biology in life science research field and provide early diagnosis, prognosis and treatment. The single cell transcriptome data has been collected from qualified sources (e.g. GEO database, 10X GemCode platform support, etc.), but there is a lack of optimized computational tools to analyze the data. Machine learning tools are well needed to analyze and classify SCT data, distinct and refined cell subtypes can be identified despite the problem of missing data and noise from sparse count matrix. Existing technology mainly focus on using unsupervised machine learning methods to analyze SCT data to predict cell cycle stage and cell life status [6]. However, supervised machine learning tools with reference genome assembly as annotation labelling and the optimizations of model structure and parameters are much required in this area. The data sets used in this project are mouse SCT data, which contains less noise than human SCT data. Furthermore, mouse data sets have more cell subtypes in comparison with human datasets, the control experiments of SCT process are easier to be conducted on mouse. ***This project focuses on standardizing the mouse data, developing and implementing supervised machine learning methods to classify single cells by their gene expression profile and simply optimize different machine learning tools to provide the most proper method in different situations.***

Computer science/Machine learning

Numerous mouse datasets are available in Professor Brusic's research group. These datasets are ethic and collected from different database and essays. The project will be divided in following two steps.

Data preparing:

We already have numerous datasets, so we don't need to focus on data collecting. We will clean and standardize the mouse data in order to apply the machine learning method and validate the practicability of this method. Most of datasets are collected from different sources whose formats are totally different. Therefore, we must define the standard data format and define the symbol to represent missing data or error data. Then we must do the data cleaning on the datasets. Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records datasets and refers to identifying incomplete, incorrect, inaccurate parts of the data and then replacing, modifying, or deleting the dirty or coarse data.[7] Following are some problems occurring in datasets and related data cleaning methods.

- **Missing data**

  In most situation, missing data should get from corresponding experiments. However, some missing data could be derived from source datasets. Missing values can be deduced from statistical methods by plausible values. Sometimes, we also can use "unknown" to represent them to meet the demand for data cleaning.

- **Error/Noise detection**

  Applying statistical methods to detect possible error value or exception value, such as deviation analysis, non-compliance distribution, value from regression equation. In addition, some rule bases (common sense rules, etc.) are available to clean noise and error value.

- **Detecting duplicate records**

  Some records which have same attributes value in the datasets are considered as duplicate records. Most of time, duplicated records will be combined as the one record. A determination algorithm is required to define whether the datasets contain duplicate records.

- **Inconsistency**

  Datasets from different sources may have semantic conflict. For example, two different experiments may have same sample name. We must define integrity constraint for the whole datasets.

Machine Learning:

We will establish a system based on different machine learning tools (SVM, ANN, etc) to distinguish the subtype of cells with multiple statistical methods and identified critical features. The main machine learning methods used are mainly from the scikit-learn package in python. Scikit-learn is one of the most common package in Python used to do machine learning programming, it combines many advanced machine learning algorithms.[8] In this package, we can find numerous classification methods that are suitable for our works.

Following lists the classification methods in scikit-learn package and the analyzing of whether they are suitable for our project from my point of view.

- **Support Vector Machine**

Support vector machine is one of common classification methods. It builds a model to separate different categories of training samples and classify new sample to specific category.[9] In our project, the data are multiple dimensions, and SVM can map the eigenvector to multiple dimension through kernel function. So, I think they are suitable in our project.

- **Neural network models**
  Neural network can simulate the human brain to process the classification work, and it build many layers of interconnected processing units to represent neurons in human brains.[10] In my opinion, it is a useful method to do classification. But, the method in scikit-learn package doesn't support large-scale project which requires GPU-based implementation.

- **Other methods**
  In the Python scikit-learn library, there are also many other classification methods. Many methods are based on mathematical model. They divide space in small zones and each zone represents one specific class. In addition, Nearest neighbors method which is quite simple, it saves all data in the model. When the new data comes, it finds closest data groups, and classify it. Naive Bayes is a form of supervised machine learning algorithm based on Bayes' theorem and it requires the data in datasets are independent [11]. If we have enough time, all of these methods will be implemented and validated.

Aims and Objectives

The project concentrates on standardizing and reproducing mouse datasets, utilizing machine learning methods to identify cell subtypes that can be used for early disease detection in medicine field [12]. The practical objective of this project is to develop and implement a platform which can deal with the existing single cell datasets, optimizing machine learning tools and test the performance of machine learning models. The project will include following steps: standardizing datasets, training of classification system, testing learning model using new datasets, optimizing the most suitable model. Specific objectives are:

1. Standardize and clean the mouse datasets from different literatures and sources.
2. Develop and implement machine learning algorithm for classification of SCT data.
   a. ANN (artificial neural network)
   b. SVM (support vector machine)
   c. Explore other machine learning methods
3. Optimize different machine learning tools and provide the suitable one for each case.
4. Prepare an article for publication
5. Complete and submit the final year dissertation.

Project Plan

Software development will utilize Agile development [13]. Agile development is suitable because this project is small-scale and it's convenient to communicate with customers directly and modify current software. Machine learning tools and statistic methods will be performed in Python environment and the system will be established on webpage.

Specific tasks are:

Preparatory

1.1 Complete and submit supervisor project proposal, detailed project proposal, revised project proposal, and preliminary research ethics checklist

1.2 Review literature, study the methods for distinguish cell subtypes, select the suitable machine learning methods.

1.3 Complete the project poster based on project proposal and prepare for transferring to essay.

System development

2.1 Standardize and reproduce the mouse data sets.

2.2 Develop and implement system for distinguish cell subtypes.

2.3 Evaluate the performance of machine learning method and optimize.

2.4 Establish online platform to manipulate new datasets.

Reporting and publication

4.1 Provide weekly incremental progress reports and short monthly written reports

4.2 Complete and submit interim report (deadline December 16, 2019)

4.3 Develop a plan and schedule for preparing the final dissertation, preliminary and revised.

4.4 Write and submit the final dissertation (deadline April 20, 2020).

4.5 Prepare and submit an article for publication (desired but not compulsory)

References

1. Powell AA, Talasaz AH, Zhang H, Coram MA, Reddy A, Deng G, Telli ML, Advani RH, Carlson RW, Mollick JA, Sheth S. Single cell profiling of circulating tumor cells: transcriptional heterogeneity and diversity from breast cancer cell lines. PloS one. 2012 May 7;7(5):e33788.
2. Babbe H, Roers A, Waisman A, Lassmann H, Goebels N, Hohlfeld R, Friese M, Schröder R, Deckert M, Schmidt S, Ravid R. Clonal expansions of CD8+ T cells dominate the T cell infiltrate in active multiple sclerosis lesions as shown by micromanipulation and single cell polymerase chain reaction. Journal of Experimental Medicine. 2000 Aug 7;192(3):393-404.
3. Segerstolpe Å, Palasantza A, Eliasson P, Andersson EM, Andréasson AC, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, Smith DM. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell metabolism. 2016 Oct 11;24(4):593-607. A
4. Baxter AE, Niessl J, Fromentin R, Richard J, Porichis F, Charlebois R, Massanella M, Brassard N, Alsahafi N, Delgado GG, Routy JP. Single-cell characterization of viral translation-competent reservoirs in HIV-infected individuals. Cell host & microbe. 2016 Sep 14;20(3):368-80.
5. Heldt FS, Kupke SY, Dorl S, Reichl U, Frensing T. Single-cell analysis and stochastic modelling unveil large cell-to-cell variability in influenza A virus infection. Nature communications. 2015 Nov 20;6:8938. A
6. Scialdone A, Natarajan K N, Saraiva L R, et al. Computational assignment of cell-cycle stage from single-cell transcriptome data[J]. Methods, 2015, 85: 54-61.Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python[J]. Journal of machine learning research, 2011, 12(Oct): 2825-2830.[8]
7. Wu S. A review on coarse warranty data and analysis[J]. Reliability Engineering & System Safety, 2013, 114: 1-11.
8. Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.
9. Venkatalakshmi B, Shivsankar M V. Heart disease diagnosis using predictive data mining[J]. International Journal of Innovative Research in Science, Engineering and Technology, 2014, 3(3): 1873-7.
10. Kalogirou S A. Artificial neural networks in renewable energy systems applications: a review[J]. Renewable and sustainable energy reviews, 2001, 5(4): 373-401.
11. Taheri S, Mammadov M, Bagirov A M. Improving naive Bayes classifier using conditional probabilities[C]//Proceedings of the Ninth Australasian Data Mining Conference-Volume 121. Australian Computer Society, Inc., 2011: 63-68.
12. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. The New England journal of medicine. 2017 Jun 29;376(26):2507.
13. Paetsch F, Eberlein A, Maurer F. Requirements engineering and agile software development[C]//WET ICE 2003. Proceedings. Twelfth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2003. IEEE, 2003: 308-313.