University of Nottingham Ningbo China

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

Bangrui Zheng, Vladimir BRUSIC

BACKGROUND & MOTIVATION

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. Most of mouse data sets have not been standardized. Defining standard data formant and reducing noise in datasets are essential for further machine learning.
Based on existing data sets and previous studying on human datasets, data processing is mainly about following steps:

- Experiments which contains incomplete or partial result (incomplete gene expression) will be abandoned.
- Missing data will be replaced by zero because most of genes in cell won't express. If sample contains too many missing data, the sample will be abandoned.
- Sample which didn't express any gene means experiments occurring errors, and the sample data will be abandoned.

This is the preliminary data processing plan, it will modify with the actual needs. Methods including non-negative matrix factorization, principal component analysis, and so on will be used to reduce noises in datasets and improve machine learning performance.

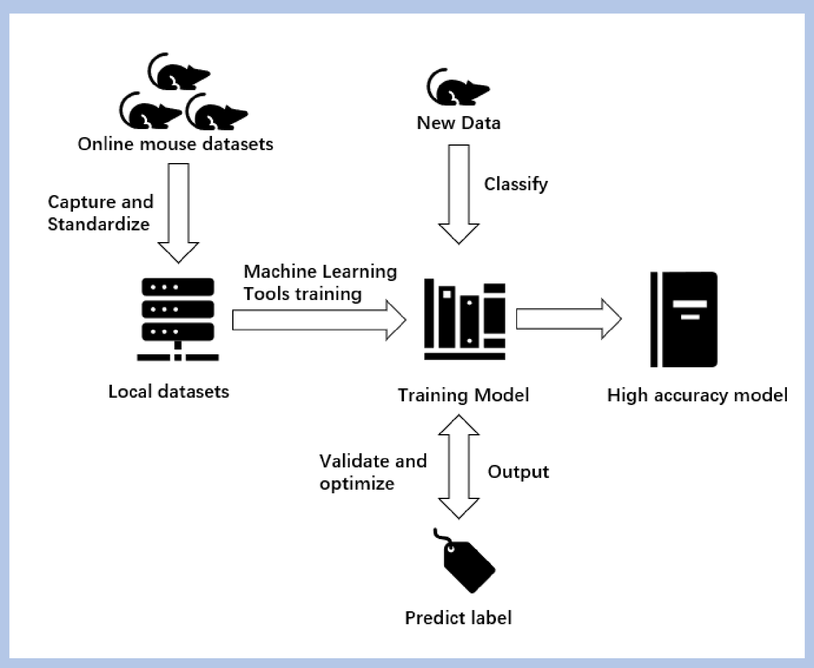
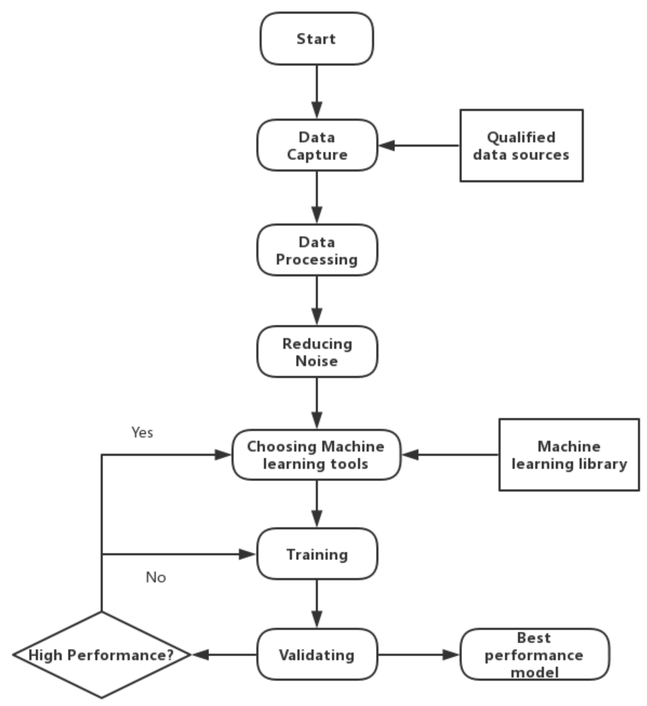


Figure 1: Project processes, data will be collected from dependable qualified sources (e.g. GEO database, 10X GemCode platform support, etc.) and store in local database after standardizing . Then we will do a iteration step to get a high accuracy model preparing for further studying.

Figure 2: Workflow diagram that represents high performance model based on validation.



CHALLENGES

- **Lack of standards**

  Not all data sets collected from qualified sources are the same standard.

- **Noise**

  Many experiments data may contain noises which affect the prediction accuracy

- **Heterologous source**

  Data sets are not collected from same sources, which leads low performance of machine learning prediction.

- **Prediction accuracy**

  Machine learning tools may not guarantee high prediction accuracy

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]

Contacts: Bangrui Zheng zy22091@nottingham.edu.cn
Vladimir BRUSIC vladimir.brusic@Nottingham.edu.cn