論文の内容の要旨

論文題目　**Ensemble feature selection approaches for discovery of candidate cancer biomarkers**

（癌バイオマーカー候補発見のためのアンサンブル特徴選択技法）

氏　　名　　文　銘塡

Introduction

Lately, biomarker discovery has become one of the most significant research issues in the biomedical field. Owing to the presence of high-throughput technologies, genomic data such as microarray data and RNA-seq have become widely available. As biomarker discovery is typically modeled to determine the most discriminating features from the datasets, it can be described as a feature selection problem regarding class. Many kinds of feature selection techniques have been applied to retrieve significant biomarkers from these kinds of data. However, they tend to contain high-dimensional features with only a small number of samples; thus, conventional feature selection approaches may be problematic in terms of performance and reproducibility. In addition, conventional studies mostly focus on differentially expressed genes only; however, noise in gene expression datasets and insufficient information in limited datasets impede precise analysis of novel candidate biomarkers. Thus, integrative analysis using different kind of datasets is required.

In this thesis, we propose ensemble feature selection approaches for discovery of candidate cancer biomarkers. Here, ensemble feature selection implies integration of different kinds of feature selectors to yield more robust results through instance perturbation or merging of multiple datasets. Section 1 describes a novel ensemble feature selection method based on the l_1-norm support vector machine (SVM). To be specific, l_1-norm SVM, which efficiently reduces the number of irrelevant or redundant features and produces sparse feature sets, is applied over bootstrap samples produced by random sampling of the original dataset. Through this process, high stability as well as high classification performance can be achieved. Section 2 focuses on integrative analysis of gene expression and DNA methylation. It is known that DNA methylation is a key regulator of gene expression; thus, it is expected that better understanding of gene-regulatory mechanisms can be acquired by integrating DNA methylation and gene expression datasets. We use normalization and unsupervised feature extraction methods to integrate DNA methylation and gene expression datasets into a one-dimensional dataset containing their main characteristics. Integrative analysis shows generally superior performance compared to use of the gene expression or DNA methylation dataset only. We believe that the proposed methods can be applied efficiently to several kinds of biomarker discovery task.

Section 1: Stable Feature Selection Based on the Ensemble l_1-norm Support Vector Machine for Biomarker Discovery

From RNA-seq dataset of renal clear cell carcinoma acquired from the cancer genome atlas (TCGA), we took into account two stage information, i.e., stage I and IV, as stage-I renal clear cell carcinoma involves local tumors that only exist in the kidney, whereas tumors at stage-IV have grown into other tissues outside the kidney or have spread widely to other lymph nodes. Thus, the use of these stages could provide significant clues regarding tumor advancement and tumor metastasis. SVM has been a popular method for classification of the biomedical problems, and SVM-RFE enabled SVM for feature selection processes. Applying l_1-norm which tends to produce sparse solution which makes it possible to reduce the number of features of a large feature set, but applying to a single dataset might produce a result that is excessively dependent on the sample set, which even causes uncertainty of reproducibility on datasets that are only slightly different. This problem can be solved by perturbed datasets produced by bootstrap aggregation. The overall process of l_1-norm SVM-RFE is described as below.

(1) Generate n random bootstrap samples, X_1, X_2, …, X_n, containing i% of the data of the whole training dataset X.
(2) Perform a cross-validation test on each bootstrap sample to set regularization parameter C.
(3) Apply the l_1-norm SVM to each bootstrap sample. Then, the weight vector w is calculated for each feature.
(4) Eliminate features for which the coefficient $w = 0$ in each bootstrap sample.
(5) Record the cross-validation score for each bootstrap sample for step (7).
(6) Repeat steps (2) ~ (5) until no more features with $w = 0$ are available for any bootstrap.
(7) Select optimal feature subsets for each bootstrap sample, which maximize the cross-validation score recorded in step (5).
(8) Produce the integrated feature set of size k by aggregating all the remaining features in the bootstrap samples.
(9) Convert X to the reduced dataset X' that consists of features in k. Here, the number of bootstrap samples that contain a given feature is considered the "stability score" S for that feature ($1 \leq S \leq n$).

The performance of the proposed model was compared with three well-known feature selection methods, i.e., fast correlation-based filter (FCBF), random forest, and an ensemble version of l_2-norm SVM-RFE. The stability of the proposed method was tested by using the Tanimoto distance, while the classification performance was measured by area under the curve (AUC) score of 10-fold cross-validation and independent data tests. First, we carried out the stability test by creating 20 random subsamples from the original dataset, which is constructed with 80% of the data, and the proposed method generally demonstrates high stability among all the methods. Then the classification performance was tested using four popular classifiers, i.e., adaptive boosting (AdaBoost), logistic regression, random forest, and SVM with an RBF kernel. Figure 1 demonstrates the results of the performance test using each feature selection method. The proposed method shows the overall best performance both in the cross-validation test and independent data test. Also, we conducted an additional classification performance test using the top 20 genes selected by each feature selection method. The results show that the proposed method demonstrated the best performance even with the limited number of features. We took into account those top 20 genes, and have found that 14 of 20 genes have known relation with tumor, by searching

literature, databases and gene annotations. Also, 8 genes among 14 genes, have known relationship with tumor progression and metastasis which are directed related to the cancer stage. Thus, other 6 genes which have no relation with tumor might also be considered as candidate genes for tumor or tumor progression and metastasis. In addition, the proposed method showed the best performance in classification test using those top 20 genes compared to other feature selection methods.

Figure 1. The result of classification performance test

Section 2: Integrative Analysis of Gene Expression and DNA Methylation using Unsupervised Feature Extraction for Detecting Candidate Cancer Biomarkers

Conventionally, gene expression analysis using microarray or RNA-seq data has been an effective method for cancer biomarker discovery. However, gene expression profiles do not contain a sufficient number of samples, and large quantities of noise exist in the datasets. To deal with this problem, we integrated gene expression and DNA methylation datasets to into a one-dimensional dataset which retains the characteristics of the original datasets. As the units denoting gene expression and DNA methylation levels differed, i.e. transcripts per million (TPM) and β-value, respectively, box-cox normalization was applied to both datasets. Additionally, the datasets were rescaled into the same range of [0, 1]. After data normalization and rescaling, unsupervised feature extraction feature extraction, i.e. principal component analysis (PCA) and autoencoder were applied to the gene expression and DNA methylation datasets to produce the integrated dataset. Dimension reduction to one-dimensional data with PCA is achieved by deriving a 1st principal component, which describes 71.3% of the data. For autoencoder, the one-dimensional integrated dataset can be acquired similar to that in PCA by setting the number of the intermediate node to 1. Welch's t-test was utilized for each integrated dataset to calculate the feature importance, and performed 10-fold cross-validation test using forward feature selection for performance evaluation. As described in Figure 2, the results of cross-validation test demonstrate that integrative analysis

using feature extraction generally outperforms compared to the use of gene expression or DNA methylation alone. In addition, the integrative analysis also shows the better performance when only top 20 genes are selected. For the further analysis, we focused on the top genes from the integrated datasets with ranks in both the gene expression and DNA methylation datasets at 200+, finding that their ranks improved in the integrated datasets. Among those 11 genes, 6 genes have been previously reported as cancer-related candidate biomarkers. These results imply that differentially expressed genes found in the integrated dataset and not previously reported as being related to cancer progression, may constitute strong potential candidates as cancer-progression biomarkers.

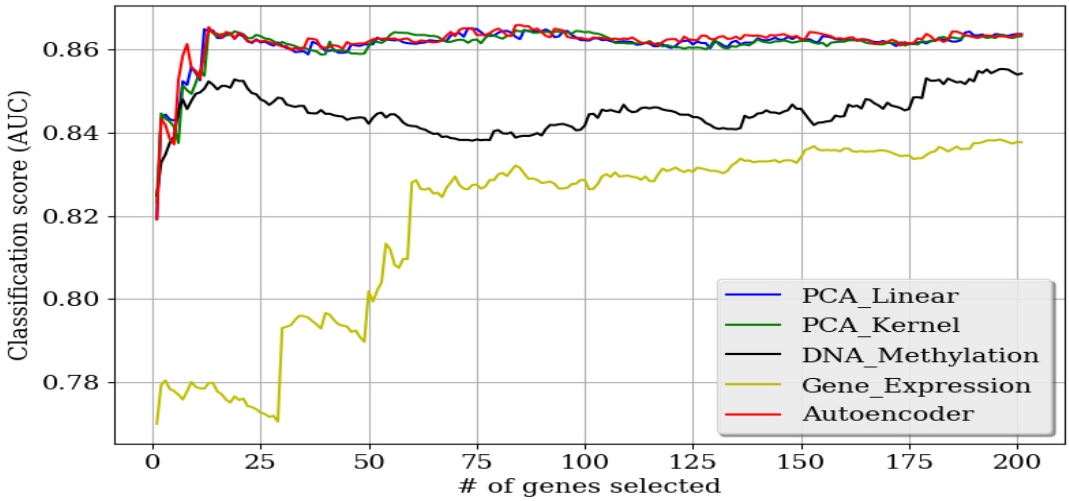

Figure 2. The result of cross-validation tests

Conclusion

In this thesis, we presented ensemble feature selection approaches for discovery of candidate cancer biomarkers. We first suggested a novel feature selection method based on application of the l_1-norm SVM over data perturbation, which was shown to be efficient for biomarker discovery. We applied our approach to RNA-seq data of renal clear cell carcinoma to find candidate biomarkers related to stage progress, which may be closely associated with tumor advancement and the metastasis issue. Through comparison with established feature selection methods, the good performance of our algorithm in terms of classification performance and stability was established. In addition, we proposed an integrative analysis method combining gene expression and DNA methylation data using unsupervised feature extraction for cancer biomarker discovery. The simple integrative analysis with an additional DNA methylation dataset, along with implementation of normalization and unsupervised feature extraction, yielded improvement over the results obtained using a single gene expression dataset. Additionally, we showed that candidate biomarkers related to cancer progression can be acquired by analysis of integrated datasets. The proposed approaches are expected to be applied to various research studies aimed at discovery of candidate cancer biomarkers.