Development of cancer classifier based on copy number variation in urinary cell-free DNA

Sohyun Im[1], Seungmin Wi[1], Minji Kim[1], Kwang Hyun Kim[2], Donghwan Lee[1]

[1]Department of Statistics, Ewha Womans University, Republic of Korea

[2]Department of Urology, Ewha Womans University College of Medicine, Republic of Korea

Abstract

Copy number variation (CNV) is one of structural variation in the human genome, and it plays an important role in development of malignant disease. Through the next generation sequencing (NGS), CNV information can easily be collected for each individual and various machine learning methods have been developed to analyze genetic information for cancer diagnosis. Bladder carcinoma is characterized by a large number of genetic alterations. DNA from urine is a promising source for liquid biopsy in bladder cancer diagnosis. In this study, we propose a pipeline for predicting cancer type using CNV information of urinary cell free DNA. We applied machine learning algorithms to classify multiple cancer types based on CNV features. TCGA dataset is big data including genetic information of multiple cancer types and used as training dataset. We show the proposed methods perform well in terms of high accuracy in the internal validation. Finally, we validated the prediction performances of proposed methods with CNV features of urinary cell-free DNA from independent samples from patients with or without malignancies. Specifically, we examined the performance of this model for diagnosis bladder cancer using urinary cell free DNA.

Data and Methods

Training Data

- Copy number variation (CNV) : variation of genomic sequence ranging from a kilobase to multiple mega base pairs in length.
- Obtain samples from TCGA website.
- Number of cancer types: 26 classes (incl. Normal, bladder cancer, etc)

Validation Data

- Urine from 116 patients were obtained in Ewha Womans University Medical Center with informed consent.
- Patient samples were categorized as bladder cancer (BC, n=42), prostate cancer (PC, n=23), renal cell cancer (RC, n=23), and normal (NL, n=28).
- Shallow whole genome sequencing was performed for obtaining CNV data of urinary cell free DNA. In prior study, we demonstrated the similarity of CNV features between cancer tissues and urinary cell free DNA in bladder cancer (Lee et al. 2018).

Feature extraction

- Cytoband matching: Following UCSC, the Chromosome 1 to 22 is partitioned into about 800 regions. For each cytoband, compute mean of CNV values in overlapped positions.
  - ⇒ Size of training sample: X: 20715*761, Y: 20715*1
  - ⇒ Size of validation sample: X: 116*761, Y: 116*1

Machine learning methods : Classical classification methods and Boosting

- XGBoost(Tianqi Chen et al.2016) is a new gradient boosting decision tree implementation with *Weighted Quantile Sketch* and *Sparsity-aware algorithm*.
  - *Weighted Quantile Sketch* : an approximation algorithm for determining how to make splits in a decision tree.
  - *Sparsity-aware algorithm* : It takes advantage of the sparsity in the dataset, reducing the computation to a linear search on only the non-missing entries.
- LightGBM(Ke et al.2017) uses XGBoost as a baseline and outperforms with *Gradient-based One-Side Sampling(GOSS)* and *Exclusive Feature Bundling(EFB)* to save the time and memory.
  - *GOSS* : It inspects the most informative samples while skipping the less informative samples
  - *Exclusive Feature Bundling* : It takes advantage of sparse datasets by grouping features in a near lossless way.

Results 1: Learning performance in TCGA

Accuracy (%) of internal validation in TCGA

# class	Logistic	SVM	RandomForest	LGBM	XGBoost
26	79.95	70.55	78.29	80.18	**80.27**
4	91.94	91.86	91.77	93.32	**93.55**
2 (NL vs BC)	97.87	96.45	**98.44**	98.16	98.25

Table 1. 5-fold CV accuracy in TCGA. Bladder cancer patients and normal are classified with binary class-model. 4 class-model includes 3 cancers (Bladder, Kidney, Prostate) and a normal. 26 class-model classifies 25 cancers including all of above and a normal

Results 2: Cancer prediction in urine data

Performance of 2-class model

We validated urinary cell-free DNA collected from independent bladder cancer patient and normal samples based on the best-fitting models.

	Logistic	SVM	RandomForest	LGBM	XGBoost
Accuracy(%)	50.63	62.03	67.09	**69.62**	64.56
Sensitivity	0.431	0.725	0.588	0.714	0.686
Specificity	0.857	0.714	0.679	0.643	0.607

Table 2. Prediction accuracy, sensitivity, specificity of binary classification model (NL vs BC)

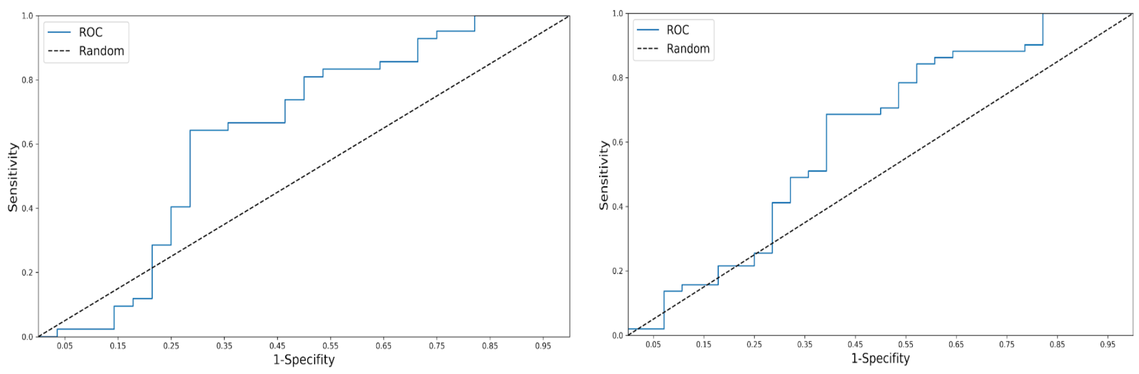

Figure 1. ROC curves of bladder cancer classification models on urinary cell-free DNA by using XGBoost on the left, Light GBM on the right

Performance of multi-class model

Focusing on how accurately proposed models detect bladder cancer patient and normal, we visualized each percentage of patient labels assigned by predicted values on all true bladder patients or all true normal.

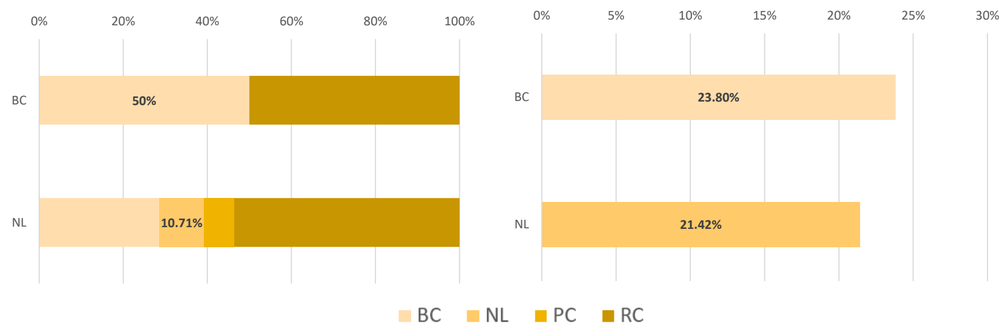

Figure 2. Proportions of predicted labels of true Normal and true bladder patients from 4-class classification model (left) ; 26-class classification model (right).

Comparison of 2, 4, 26 classification model in urine data

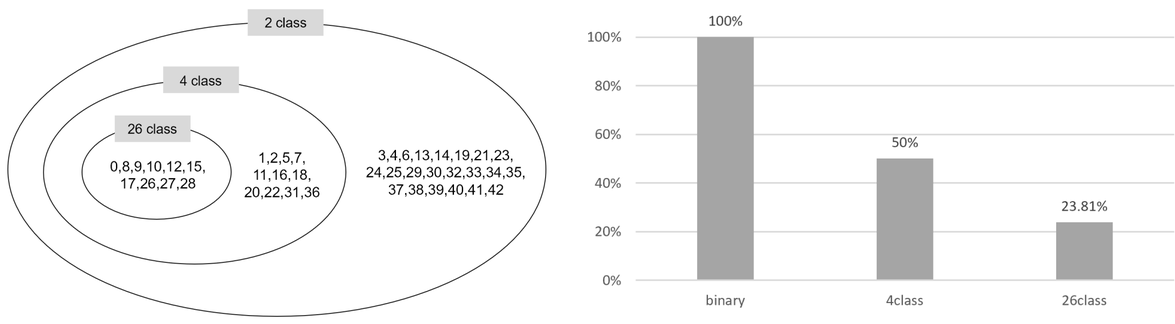

Figure 3. Venn diagram of bladder cancer patient index predicted correctly (left). Bar plot of bladder cancer sensitivity when the cutoff is 0.5 (right).

Conclusions

- We consider various machine learning methods to identify the cancer types based on CNV collected from NGS.
- Cytoband-based CNV features are useful to identify the cancer types.
- In TCGA data learning, all classification methods works well. Especially, the boosting methods such as LGBM and XGBoost outperform.
- In the external validation (urine data), although predictive accuracy decreases, the proposed method showed acceptable performance and can be utilized in urine based liquid biopsy for diagnosis of bladder cancer.

References

• Itsara et al., Population analysis of large copy number variants and hotspots of human genetic disease (2009)

• Lee et al., Urinary Exosomal and cell-free DNA Detects Somatic Mutation and Copy Number Alteration in Urothelial Carcinoma of Bladder, Scientifc Reports, 8: 14707 (2018)

• Zhang et al., Classification of cancers based on copy number variation landscapes, Biochimica et Biophysica Acta (2016)

• Chen and Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,(2016)

• Ke et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, NIPS (2017)