Journal of Kerman University of Medical Sciences

JKMU

Journal of Kerman University of Medical Sciences, 2019; 26 (3): 207-213

Classification of Chronic Kidney Disease Patients via k-important Neighbors in High Dimensional Metabolomics Dataset

Hadi Raeisi Shahraki, Ph.D.[1], Shiva Kalantari, Ph.D.[2], Mohsen Nafar, Ph.D.[3]

1- Assistant Professor, Department of Biostatistics and Epidemiology, Faculty of Health, Shahrekord University of Medical Sciences, Shahrekord, Iran (Corresponding author; E-mail: raeisi.shahraki_hadi@yahoo.com)

2- Assistant Professor, Chronic Kidney Disease Research Center, Labbafinejad Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran

3- Professor, Urology and Nephrology Research Center, Labafinejad Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Received: 22 April, 2019 Accepted: 21 May, 2019

ARTICLE INFO

Article type:
Original Article

Keywords:
Chronic kidney disease
Classification
High dimensional data
KNN
SCAD

Abstract

Background: Chronic kidney disease (CKD), characterized by progressive loss of renal function, is becoming a growing problem in the general population. New analytical technologies such as "omics"-based approaches, including metabolomics, provide a useful platform for biomarker discovery and improvement of CKD management. In metabolomics studies, not only prediction accuracy is attractive, but also variable importance is critical because the identified biomarkers reveal pathogenic metabolic processes underlying the progression of chronic kidney disease. We aimed to use k-important neighbors (KIN), for the analysis of a high dimensional metabolomics dataset to classify patients into mild or advanced progression of CKD.

Methods: Urine samples were collected from CKD patients (n=73). The patients were classified based on metabolite biomarkers into the two groups: mild CKD (glomerular filtration rate (GFR)> 60 mL/min per 1·73 m^2) and advanced CKD (GFR<60 mL/min per 1·73 m^2). Accordingly, 48 and 25 patients were in mild (class 1) and advanced (class 2) groups respectively. Recently, KIN was proposed as a novel approach to high dimensional binary classification settings. Through employing a hybrid dissimilarity measure in KIN, it is possible to incorporate information of variables and distances simultaneously.

Results: The proposed KIN not only selected a few number of biomarkers, it also reached a higher accuracy compared to traditional k-nearest neighbors (61.2% versus 60.4%) and random forest (61.2% versus 58.5%) which are currently known as the best classifieres.

Conclusion: Real metabolomics dataset demonstrate the superiority of proposed KIN versus KNN in terms of both classification accuracy and variable importance.

Citation: Raeisi Shahraki H, Kalantari SH, Nafar M. Classification of Chronic Kidney Disease Patients via k-important Neighbors in High Dimensional Metabolomics Dataset. *Journal of Kerman University of Medical Sciences*, 2019; 26 (3): 207-213.

Introduction

Chronic kidney disease (CKD), characterized by progressive loss of renal function, is becoming a growing problem in the general population (1). CKD is associated with high mortality and increased risk of several diseases including cardiovascular diseases (CVD), infectious diseases and acute kidney injury (AKI) (1). Identification of biomarkers could improve our insight regarding the diagnosis, progression process and pathogenic mechanism. New analytical technologies such as "omics"-based approaches, including

metabolomics, provide a useful platform for biomarker discovery and improvement of CKD management (2). In metabolomics studies, not only prediction accuracy is attractive, but also variable importance is critical because the identified biomarkers reveal pathogenic metabolic processes underlying the progression of chronic kidney disease. Moreover, these biomarkers illustrate the impaired pathways that could be used as the target for therapeutic agents and consequently better management of these patients.

Classification, as one of the oldest statistical issues, was considered by Fisher in 1936 for the first time. Although up to now about 200 different classifiers have been developed, there has not been a method as the best one in all situations (3). K-nearest neighbors [KNN] classifier is known as a very popular approach, due to its simplicity and accuracy in practical problems (4). In KNN, new observation was assigned to the class of most of their k-neighbors. But, high dimensional scenarios pose some challenges for KNN (5, 6). High dimensional settings refer to situations in which the number of predictors is large relative to the sample size and its related challenges are called "*curse of dimensionality*" (7-9). There is a great deal of literature on the deleterious effects of curse of dimensionality in KNN. For example, Lu et al. showed that KNN in high dimensional datasets represent unstable results (8). Pal et al. have noted that KNN is affected by nuisance (irrelevant with outcome) variables in high dimensional problems and nearly half of the observations may be misclassified (6). As a consequence of curse of dimensionality, many studies have argued that nearest neighbor can become ill posed due to a phenomenon called distance concentration (6, 7, 10). Beyer et al. demonstrated that distance concentration occurs when all pairwise distances concentrate around a single value and inferred that, in this situation, nearest neighbor is not meaningful (10).

In order to handle the aforementioned problems facing high dimensional setting in KNN, both dimension reduction and dimension extraction techniques have been proposed including random projection in Fern et al. study (11), projection based on principal components in Deegalla et al. study (12) and robust nearest neighbor in Chan et al. study (13). In 2016, Pal et Al. defined mean absolute difference of distances [MADD] as a novel dissimilarity measure to avoid curse of dimensionality but like the other proposed methods, MADD does not take into account the variables importance (6). For high dimensional datasets, due to the nature of their sparsity, lack of imposing variable importance in the classifier causes hard identification of real patterns and decreased accuracy of classification.

In the last decades penalized models were developed to avoid curse of dimensionality in some statistical methods like regression (14, 15), discriminant analysis (16, 17), principal component (18) and clustering (19). Simultaneous estimation and variable selection in penalized methods lead to stable results and higher accuracy even when the number of variables is higher than the sample size (15, 17). *Smoothly clipped absolute deviation* [SCAD] regression as one of the well-known penalized regressions was proposed by Fan et al. in 2001. SCAD is able to estimate coefficients of all the informative variables as non- zero and all non-informative variables equal to zero with a probability very close to one (oracle property) (20).

Recently, Raeisi et al. incorporated the importance of each variable using a function of SCAD logistic regression in construction of dissimilarity measure. Using this hybrid dissimilarity, which combines information of variables and

distances, leads to considering k-important neighbors [KIN] instead of k-nearest neighbors in the assignment procedure (21). In the present study, we attempted to classify the patients based on metabolite biomarkers into the mild and advanced CKD and also to identify the biomarkers associated with progression of CKD that are excreted in the urine using metabolomics tools and via KIN. The value of these potential biomarkers was assessed in a cross sectional design, and their performance in the prediction of the renal function decline was evaluated.

Materials and methods

Urine samples were collected from 73 CKD patients who had membranous glomerulonephritis (MGN, n=29), focal segmental glomerulosclerosis (FSGS, n=29) and IgA nephropathy (IgAN, n=15). According to KDIGO guideline, CKD is defined based on the presence of either kidney damage or decreased kidney function for three or more months, irrespective of cause (22). Hence, glomerular diseases are considered as kidney damage and categorized as CKD. We classified the patients based on metabolite biomarkers into two groups: mild CKD (glomerular filtration rate (GFR)> 60 mL/min per 1·73 m2) and advanced CKD (GFR<60 mL/min per 1·73 m2). Accordingly, 48 and 25 patients were in mild (class 1) and advanced (class 2) groups respectively. The urine samples were analyzed using H1-NMR technique and the spectrum was subdivided into 205 regions, having an equal bin size of 0.04 ppm over a chemical shift range of 0.2–10.0 ppm. The bins/chemical shifts were considered as variables consistent with urinary metabolites.

If $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ be a vector of variables for ith observation, $y_i \in \{0, 1\}$ represents class membership and training dataset is defined as $L = \{(y_i, x_i),\ i = 1, \dots, n_L\}$, weight of each variable can be calculated as follows:

$$w_j = \frac{|\beta_j|}{\sum_{j=1}^{p}|\beta_j|} \quad j=1,2,...,p$$

Where, β_j is estimated coefficient of jth variable through fitting SCAD regression as follow:

$$L(\beta;\lambda) = l_n(\beta) + \lambda \sum p\ (\beta)$$

$$p\ (\beta) = I(|\beta| \le \lambda) + \frac{(3.7\lambda - |\beta|)_+}{2.7\lambda} I(|\beta| > \lambda)$$

Where, $l_n(\beta)$ is traditional MLE and penalty function was denoted as $p\ (\beta)$ in which λ regulates amount of penalty and can be estimated using cross validation technique (20). In the next step, instead of using traditional Euclidean distance, a novel dissimilarity measure is used to take into account both distances and variable importance. If we denote the distance of two points like a and b with $d(x_a, x_b)$, our dissimilarity measure is defined as follows:

$$d(x_a, x_b) = (\sum_{j=1}^{p} w_j (x_{aj} - x_{bj})^2)^{\frac{1}{2}}.$$

An observation is assigned to class 1 (y=0) when k_1 (the number of class 1 votes among k-important neighbors) is higher than k_2 (the number of class 2 votes among k-important neighbors) and is assigned to class 2 (y=1) when $k_2 > k_1$. Decision rule in the tie occurrence ($k_1=k_2$) is as follows:

$$\begin{cases} y = 0 & \text{if } \sum_{k_1} (\sum_{j=1}^{p} w_j (x_{aj} - x_{bj})^2)^{\frac{1}{2}} < \sum_{k_2} (\sum_{j=1}^{p} w_j (x_{aj} - x_{bj})^2)^{\frac{1}{2}} \\ y = 1 & \text{if } \sum_{k_1} (\sum_{j=1}^{p} w_j (x_{aj} - x_{bj})^2)^{\frac{1}{2}} > \sum_{k_2} (\sum_{j=1}^{p} w_j (x_{aj} - x_{bj})^2)^{\frac{1}{2}} \end{cases}$$

After estimation of the best value of neighbor, we constructed dissimilarity measure for testing set and assigning each observation into a class (21).

To perform classification task, we randomly split our dataset to training (n=58) and testing (n=15) with the ratio of 80/20 and replicated it 600 times to achieve stable results in all of the classifiers. In addition to KNN and KIN, we implemented random forest [RF] classifier because this method

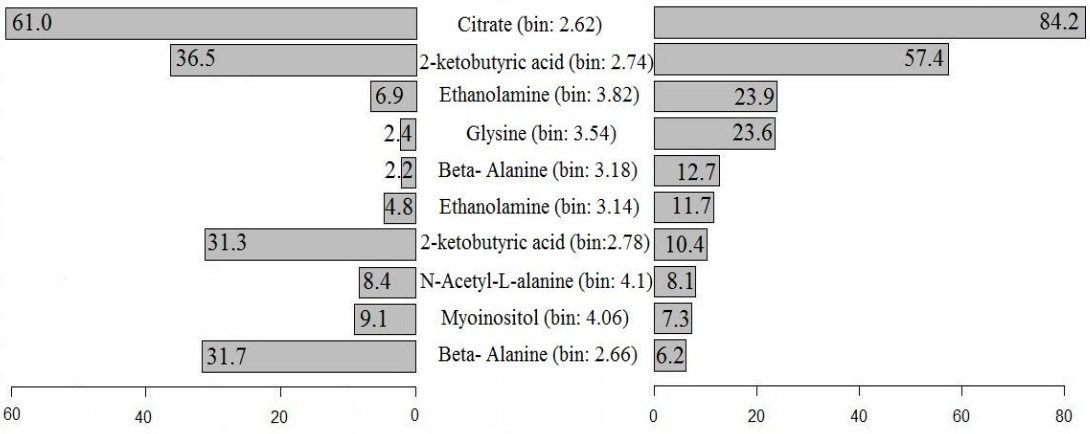
is the best classifier as mentioned in Delgado et al. (3). We used *random Forest* package for fitting RF, *caret* and *class* packages for KNN and *ncvreg* package for fitting KIN in R 3.5.0 software.

Results

Descriptive statistics about the accuracy of the classifiers have been summarized in Table 1. First of all, it should be noted that although KNN performed classification task, there was no information about the metabolite biomarkers with a key role in this classification. Moreover, there was no cut off point in random forest method to distinguish its importance from other biomarkers in relative importance index represented by RF. KNN and RF, respectively, reached 60.4% and 58.5% accuracies by incorporating almost all of the biomarkers, but mean number of considered biomarkers for the proposed KIN was only three and the number of biomarkers that were selected more than 5% of the replications, was not greater than 10 biomarkers. In Fig. 1, we displayed these biomarkers in terms of both percentage of selection (right) and mean contribution (left).

Table 1. Accuracy of different classifiers on metabolomics dataset

Method	Mean (SD) accuracy	Median accuracy	Number of used variables
Random forest	58.5 (11.5)	59.6	Almost all of the 205
K-nearest neighbors	60.4 (11.8)	59.8	All of the 205
K-important neighbors	61.2 (10.8)	61.5	3 out of the 205 in average



Proposed KIN not only selected a few number of biomarkers, it also reached a higher accuracy (61.2%) compared to traditional KNN and random forest which is known as the best classifier.

Discussion

Identified biomarkers in classification of CKD progression have been given more attention in several investigations and were consistent with other studies on kidney diseases

categorized as CKD. Citrate was earlier reported as the differential metabolite in diabetic patients with and without CKD (23). Furthermore, decreased level of myoinositol was suggested by Zhao et al. as a biomarker for diabetic nephropathy (24). Recently, Sekula et al. suggested the association of several metabolites, including N-acetylalanine, with eGFR (25). Ethanolamine is a known polar head of glycerolipids that might be a biomarker for CKD (1). Detection of this molecule as a differential marker in advanced CKD confirms previous data on the diagnostic value of this metabolite.

In line with our results on urine sample, several studies revealed the altered level of amino acids in plasma of CKD patients (26, 27). The impaired pathways in CKD patients were ”transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds” and ”urea cycle and metabolism of arginine, proline, glutamate, aspartate and asparagine” based on the analysis of the results in IMPaLA tool.

This study tried to improve the accuracy of k-nearest neighbors as a classifier in high dimensional setting which has become widespread in the recent decades. To recognize more complex patterns in high dimensional datasets, a hybrid dissimilarity measure was presented which imposed variable importance into distance calculation. Real metabolomics dataset demonstrate the superiority of proposed k-important neighbors versus KNN in terms of both classification accuracy and variable importance. KIN is also capable of managing all of the curse of dimensionality challenges and enjoys oracle property.

In the current study, to estimate variable estimation, we only considered SCAD logistic regression, however, one can use the other penalized logistic regressions. As another limitation of this study, we just considered binary classification. Although classification for more than two groups is more complex than binary classification, finding a way to determine the importance of each variable and tie management in KIN with more than two groups are topics that require further studies.

Ethics Statement

The current study was approved by the Ethics Committee of the Urology-Nephrology Research Center in Shahid Beheshti University of Medical Sciences (UNRC.930726/17).

Author Disclosure Statement

The authors declare no potential conflicts of interest with respect to the research, authorship, financial interests and/or publication of this article.

References

1. Hocher B, Adamski J. Metabolomics for clinical use and research in chronic kidney disease. Nat Rev Nephrol 2017; 13(5):269-84.
2. Nkuipou-Kenfack E, Duranton F, Gayrard N, Argilés À, Lundin U, Weinberger KM, et al. Assessment of metabolomic and proteomic biomarkers in detection and prognosis of progression of renal function in chronic kidney disease. PLoS One 2014; 9(5):e96955.
3. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research 2014; 15:3133-81.

4. Lin Y, Jeon Y. Random forests and adaptive nearest neighbors. J Am Stat Assoc 2006; 101(474):578-90.

5. Lantz B. Machine Learning With R. Birmingham, Mumbai: Packt Publishing; 2013.

6. 6. Pal AK, Mondal PK, Ghosh AK. High dimensional nearest neighbor classification based on mean absolute differences of inter-point distances. Pattern Recognition Letters 2016; 74:1-8.

7. Aggarwal CC, Hinneburg A, Keim DA. On the Surprising Behavior of Distance Metrics in High Dimensional Space. Berlin: Springer; 2001. p. 420-34.

8. Lu CY, Min H, Gui J, Zhu L, Lei YK. Face recognition via weighted sparse representation. J Vis Commun Image Represent 2013; 24(2):111-6.

9. Radovanović M, Nanopoulos A, Ivanović M. Hubs in space: popular nearest neighbors in high-dimensional data. The Journal of Machine Learning Research 2010; 11:2487-531.

10. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is "nearest neighbor" meaningful? London: Springer-Verlag; 1999. p. 217-35.

11. Fern XZ, Brodley CE. Random projection for high dimensional data clustering: a cluster ensemble approach. Twentieth International Conference on Machine Learning; 2003 Aug 21-24; Washington, DC: AAAI Press; 2013.

12. Deegalla S, Boström H, editors. Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. 5th International Conference on Machine Learning and Applications; 2006 Dec 14-16; Orlando, FL, USA: IEEE; 2006.

13. Chan Yb, Hall P. Robust nearest-neighbor methods for classifying high-dimensional data. Ann Stat 2009; 37(6A):3186-203.

14. Raeisi Shahraki H, Pourahmad S, Ayatollahi SM. Identifying the prognosis factors in death after liver transplantation via adaptive LASSO in Iran. J Environ Public Health 2016; 2016(1):1-6.

15. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996; 58(1):267-88.

16. Raeisi Shahraki H, Bemani P, Jalali M. Classification of bladder cancer patients via penalized linear discriminant analysis. Asian Pac J Cancer Prev 2017; 18(5):1453-7.

17. Witten DM, Tibshirani R. Penalized classification using fisher's linear discriminant. J R Stat Soc Series B Stat Methodol 2011; 73(5):753-72.

18. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. J Comput Graph Stat 2006; 15(2):265-86.

19. Witten DM, Tibshirani R. A framework for feature selection in clustering. J Am Stat Assoc 2010; 105(490):713-26.

20. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 2001; 96(456):1348-60.

21. Raeisi Shahraki H, Pourahmad S, Zare N. K important neighbors: a novel approach to binary classification in high dimensional data. Biomed Res Int 2017; 2017:7560807.

22. Levey AS, Coresh J. Chronic kidney disease. Lancet 2012; 379(9811):165-80.

23. Sharma K, Karl B, Mathew AV, Gangoiti JA, Wassel CL, Saito R, et al. Metabolomics reveals signature of mitochondrial dysfunction in diabetic kidney disease. J Am Soc Nephrol 2013; 24(11):1901-12.

24. Zhao L, Gao H, Lian F, Liu X, Zhao Y, Lin D. 1H-NMR-based metabonomic analysis of metabolic profiling in diabetic nephropathy rats induced by streptozotocin. Am J Physiol Renal Physiol 2011; 300(4):F947-56.

25. Sekula P, Goek ON, Quaye L, Barrios C, Levey AS, Römisch-Margl W, et al. A metabolome-wide association study of kidney function and disease in the general population. J Am Soc Nephrol 2016; 27(4):1175-88.

26. Qi S, Ouyang X, Wang L, Peng W, Wen J, Dai Y. A pilot metabolic profiling study in serum of patients with chronic kidney disease based on 1H-NMR-Spectroscopy. Clin Transl Sci 2012; 5(5):379-85.

27. Rhee EP, Clish CB, Ghorbani A, Larson MG, Elmariah S, McCabe E, et al. A combined epidemiologic and metabolomic approach improves CKD prediction. J Am Soc Nephrol 2013; 24(8):1330-8.