Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques

Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee
Department of Information Technology, Faculty of Computer Science and Information Technology
Rambhai Barni Rajabhat University, Thailand
Email: chanusorn@gmail.com, thipwan.f@rbru.ac.th, tniyomwong@yahoo.com, wandeep@yahoo.com, sathit.s@rbru.ac.th, nitat.n@rbru.ac.th

Abstract— **Predictive analytics for healthcare using machine learning is a challenged task to help doctors decide the exact treatments for saving lives. In this paper, we present machine learning techniques for predicting the chronic kidney disease using clinical data. Four machine learning methods are explored including K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and decision tree classifiers. These predictive models are constructed from chronic kidney disease dataset and the performance of these models are compared together in order to select the best classifier for predicting the chronic kidney disease.**

Keywords-classification; chronic kidney disease (CKD); machine learning

I. INTRODUCTION

Currently, kidney disease is a major problem. Because there are so many people with this disease. Kidney disease is very dangerous if not immediately treated on time, and may be fatal. If the doctors have a good tool that can identify patients who are likely to have kidney disease in advance, they can heal the patients in time. Ho, Pai, Pheng, Lee and Chen [1] presented a computer-aided diagnosis implement based on analyzing images. This system used for detecting and classifying different stages of CKD. The K-means clustering was used for detecting after image preprocessing step. Miguel Estudillo-Valderrama, Alejandro Talaminos-Barroso, Laura Roa;David Naranjo-Hernández, Javier Reina-Tosina, Nuria Aresté-Fosalba, José Milán-Martín [2] suggested the feasibility study of using distributed approach for management of alarms from chronic kidney disease patients. They managed alarms related to the monitoring of CKD patients within the eNefro project. Rosmani, Mazlan, Ibrahim and Zakaria [3] developed CKD patient self-care guidelines for Chronic Kidney Disease using Adobe Flash CS5.5. This website of CKD patient self-care guidelines has been developed using Adobe Dreamweaver. The CKD patient self-care creating more effective information channel is developed for them. This web site has been helping them in their daily self-care management. Hsieh, Lee, Chen, Lee and Chiang, [4] suggested that a real time system to analyze chronic kidney disease can be developed by using only ultrasound images. The ensemble learning have also used for classifying chronic kidney disease by building a strong classifier using SVM to predict and classify CKD stage form ultrasound images. Singh, Nadkarni, Guttag and Bottinger [5] showed different methods to leverage the hierarchical structure in ICD-9 codes to improve the performance of predictive models. This research proposed and evaluated a novel feature engineering approach to leverage this hierarchy, while simultaneously reducing feature dimensionality. Chiu, Chen, Wang, Jian [6] presented intelligent model for detecting chronic kidney disease for evaluating the severity of a patient. This intelligent model used artificial neural networks. Three types of artificial neural networks were used in this model including back-propagation network (BPN), generalized feed forward neural networks (GRNN) and modular neural network (MNN).

In this paper, we present predictive models using machine learning techniques that can predict chronic kidney disease including decision tree, logistic regression, K-nearest neighbors and support vector machine models.

In part 2, we present the detail of these models. Part 3 will present the proposed method. In part 4, the results of experiments are described. The final section describes conclusions for this work.

II. MACHINE LEARNING TECHNIQUES

A. K-nearest neighbors

K-nearest neighbors (KNN) is the classification method for classifying unknown examples by searching the closest data in pattern space [10]. KNN predicts the class by using the Euclidean distance defined as follows:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (1)$$

The Euclidean distance $d(\mathbf{x}, \mathbf{y})$ is used to measure the distance for finding the k closest examples in the pattern space. The class of the unknown example is identified by a majority voting from its neighbors.

B. Support Vector Machine

For the classification problem, the support vector machine (SVM) is the popular data mining method used to predict the category of data [9]. The main idea of SVM is to find the optimal

978-1-5090-4105-3/16/$31.00 ©2016 IEEE

hyperplane between data of two classes in the training data [2]. SVM finds the hyperplane by solving optimization problem:

$$\max Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j \quad (2)$$

where $0 \leq \alpha_i \leq C$ for $i = 1, 2, \ldots, n$

SVM uses the decision function $f(x)$ defined in form of the kernel function for calculating the output as

$$f(x) = \text{sign}\left[\sum_{i=1}^{l} \alpha_i d_i K(x, x_i) + b\right] \quad (3)$$

where $K(x, x_i)$ is the kernel function.

C. Decision Tree

Decision tree is the classification method frequently used in data mining task [8]. A decision tree is a structure that includes a root node, branches, and leaf nodes. It divides the data into classes based on the attribute value found in training sample. The details of the decision tree are described in Algorithm 1.

Algorithm 1: Decision Tree

1. create a node N;
2. **if** tuples in D are all of the same class, C **then**

   return N as a leaf node labeled with the class C;
3. **if** $attribute\_listis$ empty **then**

   return N as a leaf node labeled with the majority class in D;
4. apply **Attribute_selection_method**$(D, attribute\_list)$ to find the "best" $splitting\_criterion$;
5. label node N with $splitting\_criterion$;
6. **if** $splitting\_attributeis$ discrete-valued and multiwaysplits allowed **then**

   $attribute\_list \leftarrow attribute\_list - splitting\_attribute$;
7. **for each** outcome j of $splitting\_criterion$

   let D_j be the set of data tuples in D satisfying outcome j;

   if D_j is empty **then**

   Attach a leaf labeled with the majority class in D to node N;
8. **else** attach the node returned by **Generate_decision_tree** $(Dj, attribute\_list)$ to node N;
9. return N;

D. Logistic Regression

Logistic Regression (LR) is the linear regression model [11]. LR computes the distribution between the example X and boolean class label Y by $P(X|Y)$. Logistic regression classifies boolean class label Y as follows:

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)} \quad (4)$$

$$P(Y = 0|X) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)} \quad (5)$$

III. The Propose Method

In this paper, four types of machine learning techniques are used to predict the case of chronic kidney disease. The proposed method compares classification performance of K-nearest neighbors, support vector machine, decision tree and logistic regression. The proposed processes of constructing the predictive models are shown in the figure 1.

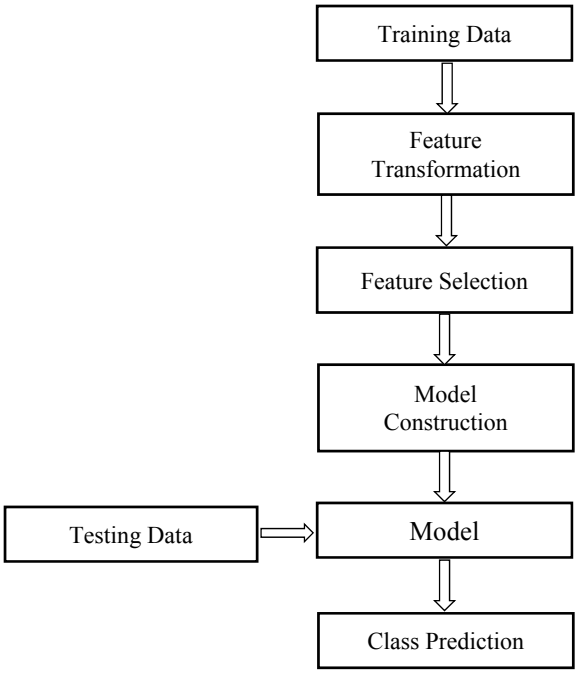

Figure 1. The process of creating the model to predict the case of chronic kidney disease.

In the first step, we transform nominal attributes to binary attribute in the training data. In the second step, the BestFirst features selection method is used to select subset of features to reduce the number of attributes and the training time. The BestFirst searches the space of feature subsets by greedy hillclimbing augmented with a backtracking facility [12]. In the third step, the classifier model is trained for creating the predictive model to predict unseen data. In the final step, the chronic kidney class is predicted by testing data.

IV. EXPERIMETAL RESULTS

In this experiment, the machine learning methods described in section 3 are trained to predict the chronic kidney disease. Four classifier methods are used in this experiment consisting of K-nearest neighbors (KNN), support vector machine (SVM) with Gaussian kernel, logistic regression (LR) and decision tree. The experiments are constructed on Matlab and Weka data mining tool.

A. Dataset

The Indians Chronic Kidney Disease (CKD) dataset consists of 400 instances and 24 attributes with 2 classes collected from UCI machine learning repository [7]. The Attribute of this dataset consists of two types of attributes which are numeric and nominal attributes divided into 11 numeric attributes and 14 nominal attributes. This dataset is collected from the patients in Apollo Hospitals Indians. The dataset is divided into two groups, one for training and another for testing. The ratio of traing and testing data is 70% and 30% respectively.

B. Performance Measurement

In this paper, the performance of the proposed method is measured by sensitivity, specificity and accuracy described as follows.

- Accuracy (ACC) is the overall success rate of the classifier defined as

$$ACC = (TP+TN) / (P + N) \quad (6)$$

where TP is the true value of positive rate, P is the positive class or yes class, N is the negative class or no class.

- Sensitivity or the true positive rate (TPR) which is defined as the fraction of positive instances predicted correctly by the model defined as

$$Sensitivity = TP / (TP+FN). \quad (7)$$

- Specificity is the true negative rate (TNR) which is defined as the fraction of negative instances predicted correctly by the model defined as

$$Specificity = TN / (FP+TN). \quad (8)$$

C. Results

The results of four machine learning techniques are compared in the experiments. All machine learning techniques are trained and tested by the proposed method. In this experiment, the 5-fold cross validation is used to train and test the machine learning models and the average results are shown in fig. 2

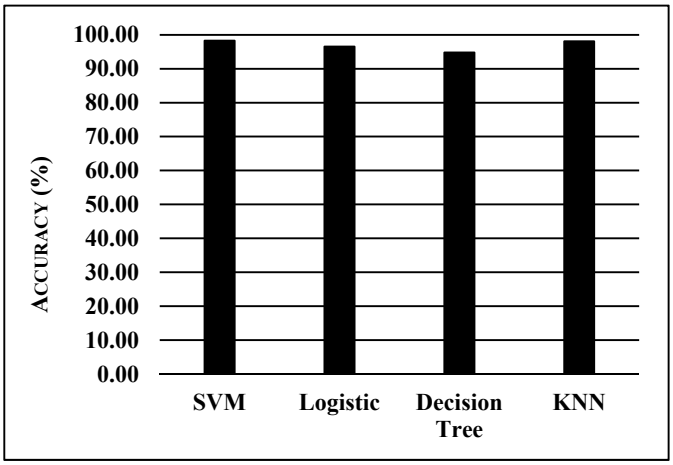

Figure 2. The average accuracy of four classifiers

Figure 2 shows the average accuracy of four classifiers conducted on five times. From the experimental results, it can be seen that the SVM classifier gives the highest accuracy than the others with 98.3% while Logistic, Decision Tree and KNN can produce the average accuracy of 96.55%, 94.8% and 98.1% respectively.

The accuracy of each class is also important because if the classifier predicts incorrectly, it may be a detriment to the patient. Therefore, the sensitivity and specificity value is used in the experiments for evaluating the performance of the proposed methods.

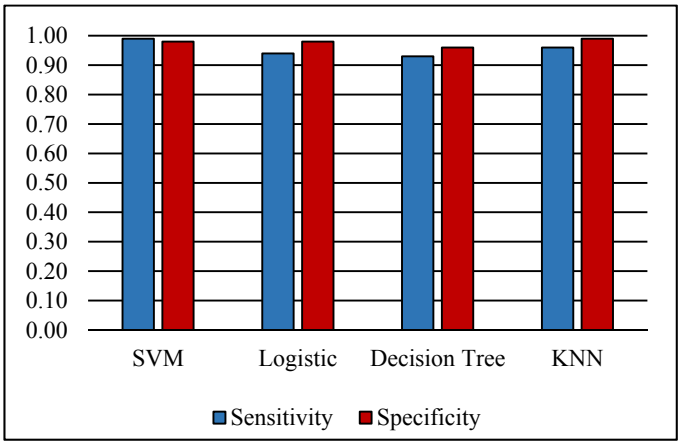

Figure 3. The comparisons of sensitivity and specificity.

Fig 3 illustrates the averages of sensitivity and specificity run on five time. The sensitivity of SVM has slightly higher than other methods at 0.99 and the sensitivity of Logistic, Decision Tree and KNN are 0.94, 0.93 and 0.96 respectively. For specificity value, the specificity of KNN is slightly higher than other methods at 0.99 and the specificity of SVM, Logistic and Decision Tree are 0.98, 0.98 and 0.96 respectively. Form the experimental results, it can be concluded that SVM is appropriated for predicting the chronic kidney disease.

V. CONCLUSION

In this paper, we present the predictive models by using machine learning methods including K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and decision tree classifiers to predict chronic kidney disease. From the experimental results, it can be seen that SVM classifier gives the highest accuracy. In addition, SVM has highest sensitivity after training and testing by the proposed method. Therefore, it can be concluded that SVM classifier is appropriated for predicting the chronic kidney disease.

REFERENCES

[1] C. Ho, T. Pai, Y. Peng, C. Lee, Y. Chen, Y. Chen, "Ultrasonography Image Analysis for Detection and Classification of Chronic Kidney Disease," IEEE Complex, Intelligent and Software Intensive Systems, pp. 624 – 629, July 2012.

[2] Miguel A. Estudillo-Valderrama; Alejandro Talaminos-Barroso; Laura M. Roa;David Naranjo-Hernández; Javier Reina-Tosina; Nuria Aresté-Fosalba; José A. Milán-Martín, "A Distributed Approach to Alarm Management in Chronic Kidney Disease ," IEEE Transl. Biomedical and Health Informatics , vol. 18, pp. 1796 – 1803, November 2014.

[3] A. Rosmani, U. Mazlan, A.Ibrahim, D. Zakaria, " i-KS: Composition of Chronic Kidney Disease (CKD) Online Informational Self-Care Tool," Computer, Communication, and Control Technology, IEEE , April 2015, pp. 379 – 383.

[4] Jun-Wei Hsieh,C.-Hung Lee,Y.-Chih Chen,W.-Shan Lee, H.-Fen Chiang, "Stage Classification in Chronic Kidney Disease by Ultrasound Image," International Conference on Image and Vision Computing New Zealand, ACM, pp. 271-276, 2014.

[5] A. Singh, G. Nadkarni, J. Guttag and E. Bottinger, "Leveraging hierarchy in medical codes for predictive modeling," Bioinformatics, Computational Biology, and Health Informatics, ACM, pp. 96-103, 2014.

[6] R. Kei Chiu, R. Y. Chen, S. Wang, S. Jian, "Intelligent systems on the cloud for the early detection of chronic kidney disease," Machine Learning and Cybernetics, IEEE, pp. 1737 – 1742, July 2012.

[7] A. Asuncion and D. J. Newman. (2007). UCI Machine Learning Repository [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[8] J. R. Quinlan, C4.5: programs for machine learning: Morgan Kaufmann Publishers Inc., 1993.

[9] R. G. Brereton, and G. R. Lloyd, "Support Vector Machines for classification and regression," Analyst, vol. 135, no. 2, pp. 230-267, 2010.

[10] S. Galit, R. P. Nitin, and C. B. Peter, Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner: Wiley Publishing, 2010.

[11] R. Xi, N. Lin and Y. Chen, "Compression and Aggregation for Logistic Regression Analysis in Data Cubes," IEEE Transl. Knowledge and Data Engineering, vol. 21, pp. 479 - 492, April 2009.

[12] Pearl, J. Heuristics: Intelligent Search Strategies for Computer Problem Solving. Addison-Wesley, pp. 48 1994.