International Journal of Mathematics and
Computer Science, **15**(2020), no. 4, 1337–1343

KNN Classification in Chronic Kidney Disease Dataset

V. Manoranjithem[1], **M. Venkatesulu**[2]

[1]Department of Computer Science Engineering
Kalasalingam Academy of Research and Education
Srivilliputtur, India

[2]Department of Information Technology
Kalasalingam Academy of Research and Education,
Srivilliputtur, India

email: mano.ranjithem@gmail.com, venkatesulum2000@gmail.com

(Received July 11, 2020, Accepted September 16, 2020)

Abstract

Nowadays Data mining has been broadly used in medical databases. The k-nearest neighbors (knn) algorithm is a more popular and efficient algorithm for pattern recognition. Nowadays a lot of people are sick with Chronic Kidney Disease (CKD). In this paper, data analysis on CKD Dataset is done using k-nearest neighbours algorithm.

1 Introduction

Data Mining is a collection of techniques which applies to highly complex database. To eliminate the uncertainty and find out the unseen pattern data mining tools and methods are used for enlightening patterns in data. The basic functions of data mining involve Classification, Neural Networks, Association Rules, Clustering, Visualization, and Decision Tree. Classification identifies the set of sub-populations, a new study belongs to the origin of a training set of data containing the studies and whose class membership

Key words and phrases: KNN, chronic kidney disease, data mining.
AMS (MOS) Subject Classifications: 68P01, 68P20.
ISSN 1814-0432, 2020, http://ijmcs.future-in-tech.net

is known. Training and Testing are the two phases used for classification. These partitions are calculated through labeled training data which classifies unlabeled testing data.

In the next section theoretical and practical view of knn algorithms and the CKD dataset description is fully explained along with the analysis. In the third chapter ten distance measures formulae has been shown. In section 3, error rates that are relevant to this particular study are summarized and the solutions to the difficulties encountered during classification are explained using R programming language which helps further related studies greatly. In the last section, we conclude our paper with the classification results.

2 KNN Classification

Knn Classification uses the Euclidean distance for classification. It calculates the distance between the new element and other elements classes which are known. In this paper, Chronic Kidney Disease dataset is taken from UCI database which consists of 25 variables with 400 instances. In that we have continuous, nominal and binary variables. Hence nominal variables attributes such as specific gravity, albumin and sugar are taken. We convert all the nominal variables to binary and we use knn classification. k values are chosen.

2.1 Training phase and test phase

In the training phase, a KNN algorithm is applied and in the test phase results are displayed. Hence the dataset partitioned into 2 phases in the ratio of 80: 20. R is a free software used by data miners which develops statistical data analysis. In R, a knn (train, test, cl) function is used for classification, where cl is the class label. The main goal of this paper is to evaluate the performance of ten distance formulae when KNN is used for binary data and also to find the best value of k. Here, we assign for k values ranging from 175 to 190 and find out the resulting error rates.

3 Distance Measures

The following 10 distance measures are binary similarity and distance measures, where S and D are similarity and distance measures, respectively.

$$(1) Sjaccard = \frac{a}{a+b+c}$$

$$(2)S3wjaccard = \frac{3a}{3a + b + c}$$

$$(3)Sczekanowski = \frac{2a}{2a + b + c}$$

$$(4)Srogertanimoto = \frac{a + d}{a + 2(b + c) + d}$$

$$(5)Ssokalmichener = \frac{a + d}{a + b + c + d}$$

$$(6)Srussellrao = \frac{a}{a + b + c + d}$$

$$(7)DEuclid = sqrt(b + c)$$

$$(8)DsquaredEuclid = sqrt(b + c)^2$$

$$(9)Dmeanmanhattan = \frac{b + c}{a + b + c + d}$$

$$(10)Dvari = \frac{b + c}{4(a + b + c + d)}$$

4 Result

The table below indicates the error rates for ten different binary formulae corresponding to their k values. Figures 1 and 2 indicate the error rates for the different values of k

From the above tables, we conclude that for the ten formulae, the error rate is minimum for different k values. For k=175,176,177 the error rates are found to be minimum.

5 Conclusion

The KNN Classification in Chronic Kidney Disease Dataset was analyzed. We have discovered that choosing the best value of k implies minimizing error rate. Data analysis done in CKD dataset using kNN for binary data implied that the distance measures 1,2,3,4,5,6,8 are considered to be the best among the ten binary distance measures considered in this work. Under these ten measures the error rate was minimum for most of the k values 175,176,177. However, finally the best k values turned out to be 175 and the distance

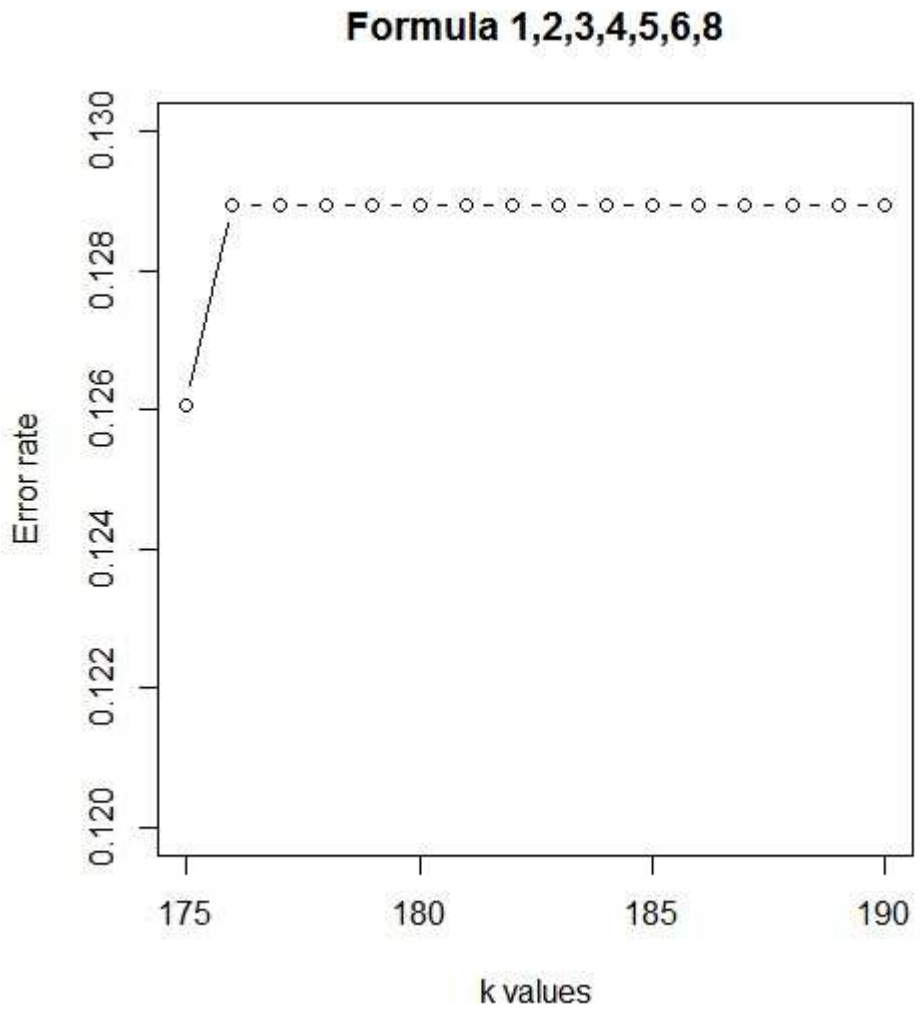

Figure 1: Formula 1,2,3,4,5,6,8

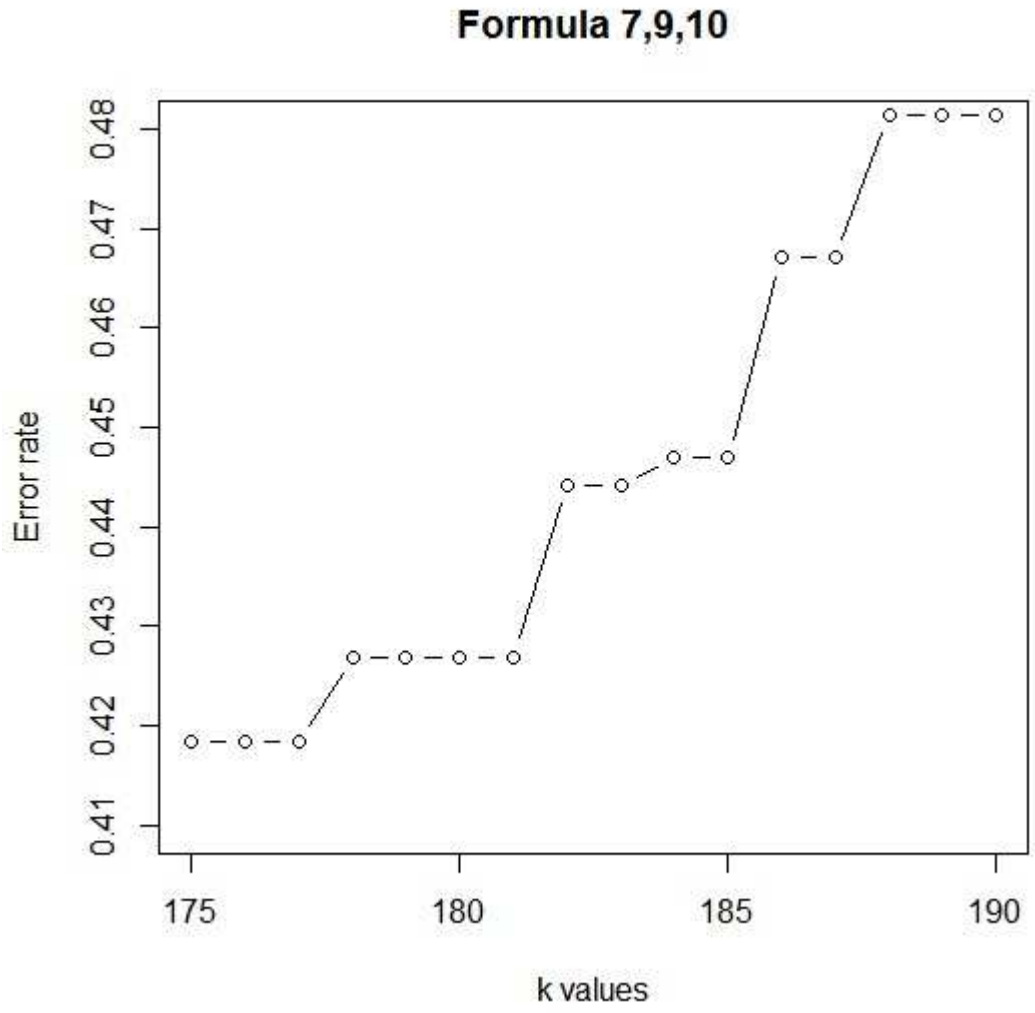

Figure 2: Formula7,9,10

k	er.k1	er.k2	er.k3	er.k4	er.k5	er.k6	er.k7	er.k8	er.k9	er.k10
175	0.1260	0.1260	0.1260	0.1260	0.1260	0.1260	0.4183	0.1260	0.4183	0.4183
176	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4183	0.1289	0.4183	0.4183
177	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4183	0.1289	0.4183	0.4183
178	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4269	0.1289	0.4269	0.4269
179	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4269	0.1289	0.4269	0.4269
180	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4269	0.1289	0.4269	0.4269
181	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4269	0.1289	0.4269	0.4269
182	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4269	0.1289	0.4269	0.4269
183	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4469	0.1289	0.4469	0.4469
184	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4469	0.1289	0.4469	0.4469
185	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4469	0.1289	0.4469	0.4469
186	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4670	0.1289	0.4670	0.4670
187	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4670	0.1289	0.4670	0.4670
188	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4813	0.1289	0.4813	0.4813
189	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4813	0.1289	0.4813	0.4813
190	0.1289	0.1289	0.1289	0.1289	0.1289	0.1289	0.4813	0.1289	0.4813	0.4813

Table 1: Error rates for different k values

measures calculated 1,2,3,4,5,6,8 are considered as the best among the ten binary distance formulae.

References

[1] V. Kunwar, K. Chandel, A. S. Sabitha, A. Bansal, Chronic Kidney Disease analysis using data mining classification techniques, 6th International Conference-Cloud System and Big Data Engineering, IEEE, (2016), 300–305.

[2] N. Tazin, S. A. Sabab, M. T. Chowdhury, Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique, International Conference on Medical Engineering, Health Informatics and Technology, IEEE, (2016), 1–6.

[3] M. Elhoseny, K. Shankar, J. Uthayakumar, Intelligent diagnostic prediction and classification system for chronic kidney disease, Scientific report, **9,** no. 1, (2019), 1–14.

[4] H. Polat, H. D. Mehr, A. Cetin, Diagnosis of chronic kidney disease based on support vector machine by feature selection methods, Journal of medical systems, **41,** no. 4, (2017), 55.

[5] T. Saidi, O. Zaim, M. Moufid, N. El Bari, R. Ionescu, B. Bouchikhi, Exhaled breath analysis using electronic nose and gas chromatography-mass spectrometry for non-invasive diagnosis of chronic kidney disease, diabetes mellitus and healthy subjects. Sensors and Actuators Chemical,**257,** (2018), 178–188.

[6] Z. Saringat, A. Mustapha, R. R. Saedudin, N. A. Samsudin, Comparative analysis of classification algorithms for chronic kidney disease diagnosis, Bulletin of Electrical Engineering and Informatics, **8,** no. 4, (2019), 1496–1501.

[7] W. H. S. D. Gunarathne, K. D. M. Perer, K. A. D. C. P. Kahandawaarachchi, Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD), 17th International Conference on Bioinformatics and Bioengineering, IEEE, (2017), 291–296.

[8] U. N. Dulhare, M. Ayesha, Extraction of action rules for chronic kidney disease using Nave Bayes classifier in 2016 IEEE International Conference on Computational Intelligence and Computing Research, IEEE, (2016), 1–5.

[9] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyaf, S. O. Olatunji, Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Computers in biology and medicine, **109,** (2019), 101–111.

[10] P. Yildirim, Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction in 2017 IEEE 41st Annual Computer Software and Applications Conference, **2,** IEEE, (2017), 193–198.