The Performance of K-Nearest Neighbors on Malignant and Benign Classes: Sensitivity, Specificity, and Accuracy Analysis for Breast Cancer Diagnosis

Arash Roshanpoor
Department of Health Information Management
Tehran University of Medical Sciences, Tehran, Iran

Marjan Ghazisaeidi
Department of Health Information Management
Tehran University of Medical Sciences, Tehran, Iran

Sharareh R. Niakan Kalhor
Department of Health Information Management
Tehran University of Medical Sciences, Tehran, Iran

Keyvan Maghooli
Department of Biomedical Engineering
Science and Research Branch, Islamic Azad University, Tehran, Iran

Reza Safdari
Department of Health Information Management
Tehran University of Medical Sciences, Tehran, Iran

ABSTRACT
Breast cancer is one of the major threats to women nowadays. Early detection of breast cancer decreases mortality rate. Machine learning algorithms are used for this purpose. Accuracy is the most popular measure for evaluating machine learning algorithms for breast cancer diagnosis. However, it does not make a distinction between the performance of the classifier on malignant and benign test cases. This paper studies sensitivity and specificity along with accuracy to differentiate between KNN performance on malignant and benign classes for the different number of neighbors. Additionally, the standard deviations of sensitivity and specificity are studied to show KNN stability in malignant and benign classes. This study is critical because the cost of false negative is more than the cost of false positive in breast cancer detection. This study is conducted on Wisconsin breast cancer dataset (WBCD) from UCI repository. Stratified 10-fold cross-validation is used in this paper. Additionally, in order to increase the correctness of outcome, validation method repeated 100 times by considering that the samples are randomly reassigned to the folds again. The results show that KNN does not work well on malignant samples compared to the benign test cases, especially for higher values of neighbors. Also, the results for malignant samples are less reliable than benign ones. Furthermore, accuracy is more representative of specificity than sensitivity. It seems that the imbalance distributions of malignant and benign classes make difference between KNN performance on malignant and benign samples. It is recommended that a new study to be conducted to show the effect of imbalance numbers of positive and negative samples and also the difference between standard deviations of positive and negative classes on KNN performance.

General Terms
Data Mining, Classification, Pattern Recognition

Keywords
Breast Cancer Diagnosis, K-Nearest Neighbors, Imbalance Dataset, Sensitivity, Specificity

1. INTRODUCTION
Concerning the report of World Health Organization (WHO), breast cancer is leading women cancer in both developing and developed countries [1]. If breast cancer is detected in early stages, the survival rate will be increased [1–3]. The significance of breast cancer detection in early stages led to huge amount of research in the field of machine learning [4].

Based on the literature review [1,3–12], it is obvious that accuracy is the most popular metric for evaluating the performance of the classifier in breast cancer detection. Although the performance of classifier could be different on positive (malignant) and negative (benign) classes, the accuracy cannot make a distinction between false positives and false negatives, and so it does not show the performance of the classifier on positive and negative classes, separately [13].

Having access to the performance of the classifier on individual positive and negative classes is essential in healthcare since positive class indicates conditions that are more serious and so decreasing false negative is more valuable than false positive [13]**.** In breast cancer, malignant class indicates more serious conditions than benign class. As a result, the study of classifier performance on individual positive and negative classes is highly beneficial as the cost of false negative is higher than false positive for breast cancer detection [14].
K-Nearest Neighbors (KNN) is one of the most prominent classification algorithms because it is simple, effective, and more accurate than many other classification algorithms [2,8,15]. This algorithm does not require any assumption for data distributions as it is a non-parametric algorithm [16]. According to these reasons, KNN is one of the most interesting algorithms in machine learning [3,5].

Katsuyoshi Odajima et al. [2] studied the effect of the sample size and the number of neighbors (K) on accuracy in breast cancer diagnosis. They demonstrated that accuracy is reduced by increasing the number of neighbors as well as decreasing the sample size. In addition, they showed that the standard deviation of accuracy is almost fixed for every value of K.

In this paper, sensitivity and specificity along with accuracy are studied to show KNN performance on individual malignant and benign classes for different values of

neighbors. Also, this paper examines the standard deviations of sensitivity and specificity to show how much classifier is reliable in each positive and negative class. Moreover, the minimum and maximum values of sensitivity and specificity are investigated in this research. Katsuyoshi Odajima et al. studied the maximum and minimum values of accuracy [2], but they did not report any results for minimum and maximum values of sensitivity or specificity.

The rest of this paper is organized as follows. At first, some details about KNN algorithm are provided. Then, material and method section gives some pieces of information about the description of the breast cancer dataset and the methodology of this study. After that, the results are provided for comparing sensitivity, specificity, and accuracy as well as their standard deviations. In the discussion, we interpret results to find out how much classifier works well in positive and negative classes. The last section is the conclusion that presents some suggestions for future work.

2. KNN

KNN is one of the most prominent algorithms in data mining. This algorithm is used vastly in machine learning [3]. It classifies a test case directly by the samples in the dataset, and so it does not require train phase [1–3].

KNN uses majority vote approach [1,2]. First, it searches for K-most nearest samples from classification-set (training-set) to a particular test sample, where K shows the number of neighbors. After that, KNN counts the number of neighbors belonging to each class. Finally, it assigns test case to the class with the maximum number of samples in the neighborhood.

This algorithm highly depends on the definition of similarity[3]. The distance between two pieces of data shows how much they are different from each other and so the inverse distance represents the similarity between them.

Euclidean distance is one of the most appropriate distance metric[3]. Equation 1 shows the Euclidean distance between two n-dimensional samples, where x_i and y_i are features of first and second data-points, respectively. Medjahed et al. [3] have reported that the Euclidean distance gives the best result on breast cancer dataset. Additionally, Euclidean distance is noted that works well on high-dimensional data points [17]. Hence, Euclidean distance is used as a measure of similarity in the current study.

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Additionally, the number of nearest neighbors (K) mostly affects KNN performance[3]. Decreasing the number of neighbors generates a more complex classifier that decides by fewer neighbors. Although this classifier could have less bias, its variation is high, and so it is not stable enough to be generalized easily in new cases. On the other hand, the increase in the number of neighbors makes more robust classifier. This classifier has much less variation, and so it is more stable. This classifier can be generalized with less difficulty on unseen data points. Unfortunately, increase in the number of neighbors selects more samples from the other classes as the neighbor. It makes more undesirable Bias[17]. Additionally, this classifier is sensitive to outliers[15].

In addition, if the value of K is less than a threshold, the classifier follows every single sample in the dataset. This classifier is sensitive to noise, and so many data points are mislabeled in this case[15,17]. Bias and variation are in conflict with each other. Thus making the compromise between them is essential in classification.

3. METHOD AND MATERIALS

3.1 Dataset Description

Studies in this paper are conducted on Wisconsin breast cancer dataset (WBCD) from UCI repository [18]. This dataset has 699 clinical cases, each one labeled as malignant (cancerous) or benign (non-cancerous). The number of malignant and benign cases are 241(24.5%) and 458 (65.5%), respectively. This dataset has 16 samples (cases) with some missing values. Removing these samples from dataset decreases the sample size to 683. Every sample has 11 features (Table 1). The first feature is sample id, and the last one is a class label that keeps two values: 2 for benign and 4 for malignant. The classifier uses other nine features as the predictor.

3.2 Performance Measures

Equations 2, 3, and 4 show how accuracy, sensitivity, and specificity are calculated based on the values of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Concerning these equations, it is obvious that accuracy depends on both false positive and false negative while sensitivity and specificity depend only on false negative and false positive, respectively.

Equation 5 shows the effectiveness of sensitivity and specificity on accuracy. Accuracy is calculated as the weighted sum of sensitivity and specificity in the way that sensitivity and specificity are multiplied in "Prevalence" and "1-Prevalence", respectively. Prevalence is the number of positive samples to all samples, and "1-Prevalence" is the number of negative samples to all samples as shown in Eq. 6 and Eq. 7.

$$\text{Accuracy} = \frac{\text{TP+ TN}}{\text{TP+ TN+ FP+ FN}} \quad (2)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+ FN}} \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+ FP}} \quad (4)$$

$$\text{Accuracy=Prevalence}\times\text{Sensitivity+(1-Prevalence)}\times\text{Specificity} \quad (5)$$

$$\text{Prevalence} = \frac{\text{Positives}}{\text{Positives+ Negatives}} \quad (6)$$

$$\text{"1-Prevalence"} = \frac{\text{Negatives}}{\text{Positives+Negatives}} \quad (7)$$

3.3 Validation Method

In this paper, stratified 10-fold cross-validation is used for measuring the sensitivity, specificity, and accuracy. 10-Fold cross-validation splits the dataset into ten non-overlapping folds. Each fold contains some randomly selected samples. The number of samples in every fold is roughly equal. Additionally, stratification keeps the same ratio between the numbers of positive and negative samples in every fold.

Practically is proved that stratified 10-fold cross validation is one of the best methods due to low bias and variance[19].

After dividing the dataset into ten folds, first fold is selected for testing and the combination of the other nine folds for training. The numbers of test and train samples are equal to 69 and 614 in each run. The numbers of positive (malignant) and negative (benign) train-samples are equal to 215 and 399. The standard deviations of positive and negative classes are 8.269 and 3.143, respectively.

After that, every sample in the test fold is classified by finding K nearest samples from the training set. Now, the values of accuracy, sensitivity, and specificity are measured for the selected test fold. This process is repeated ten times by selecting each fold exactly once for testing. At this point, we have ten values for accuracy, sensitivity, and specificity.

In order to increase the correctness of outcome, these steps are repeated 100 times by considering that the samples are randomly reassigned to the folds again. Finally, we have 1000 values for each one of accuracy, sensitivity, and specificity. The average of accuracy, sensitivity, and specificity are calculated by getting averages over these 1000 fold results. The similar steps can be followed for calculating minimum values, maximum values, and standard deviations of accuracy, sensitivity, and specificity over 1000 folds.

Table 1. Description of WBCD Features

#	Feature	Domain
1	Sample Code Number	Id Number
2	Clump Thickness	1-10
3	Uniformity of Cell Size	1-10
4	Uniformity of Cell Shape	1-10
5	Marginal Adhesion	1-10
6	Single Epithelial Cell Size	1-10
7	Bare Nuclei	1-10
8	Bland Chromatin	1-10
9	Normal Nucleoli	1-10
10	Mitoses	1-10
11	Class Label	2 for Benign 4 for Malignant

4. RESULTS

The result section is organized as follows. At first, the values of sensitivity, specificity, and accuracy for different values of K between 1 and 614 is reported to show the individual performance of the classifier on positive and negative classes. After that, the maximum values, minimum values, and standard deviations of positive and negative classes are examined to show the stability of classifier over positive and negative classes.

Averages of sensitivity, specificity, and accuracy for different values of neighbors are compared in Figure 1. It shows that sensitivity is always less than specificity for every value of neighbors. This figure illustrates that sensitivity, specificity, and accuracy are unstable for K<5, regarding that their variation is high in this area. The maximum values for sensitivity and accuracy obtained for K=5. After that, sensitivity decreases sharply and specificity rises a bit for the values of K from 5 to 430. Finally, sensitivity and specificity reach to 0% and 100%, respectively, for every value of K>430. In addition, this figure shows that the accuracy is always biased toward specificity, and yet sensitivity strongly affects it.

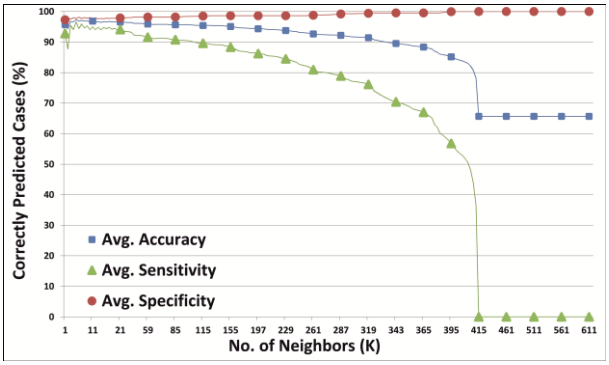

Fig.1 Averages of Sensitivity, Specificity, and Accuracy

Figure 2 shows the maximum values for accuracy, sensitivity, and specificity. This figure illustrates how much classifier works well on positive and negative classes in the best case. The maximum value of specificity is always 100% for every value of K. This figure clarifies that the maximum value of sensitivity can reach up to 100% in the best case for smaller values of neighbors, but by increasing the number of neighbors the maximum value of sensitivity starts to decrease.

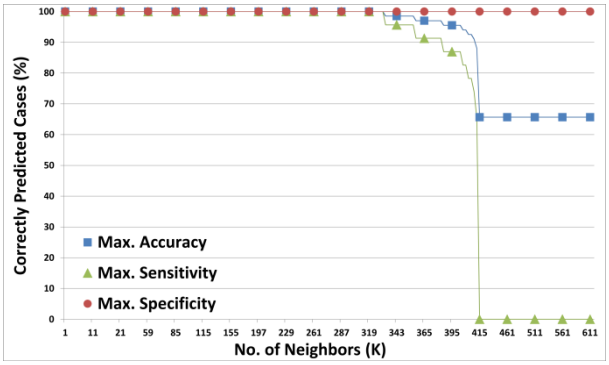

Fig.2 Maximum values of Sensitivity, Specificity, and Accuracy

Sometimes, the classifier provides good results in ideal conditions, but it does not make a proper decision in a critical situation [2]. Accordingly, the study of the minimum values for sensitivity, specificity, and accuracy is important in breast cancer detection. Minimum values of sensitivity, specificity, and accuracy show how many times the classifier fails to predict positive samples, negative samples, and total (positives plus negatives) samples in the worst case, respectively.

Figure 3 shows the minimum values for accuracy, specificity, and sensitivity. In the worst case, sensitivity drops more than specificity and also accuracy is more representative of specificity than sensitivity. It means that the condition of the positive class is more critical than negative class in the worst case. Additionally, by an increase in the number of neighbors, specificity goes better while sensitivity drops extremely.

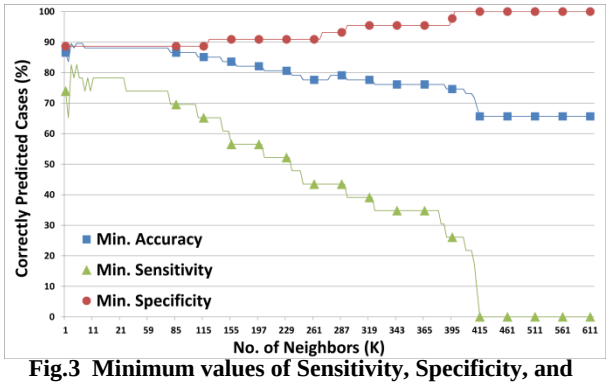

Fig.3 Minimum values of Sensitivity, Specificity, and Accuracy

Figure 4 represents the standard deviations of accuracy, sensitivity, and specificity. It is obvious that the standard deviation of sensitivity always is more than the standard deviation of specificity. Additionally, it is remarkable that the standard deviation of accuracy always is close to the standard deviation of specificity.

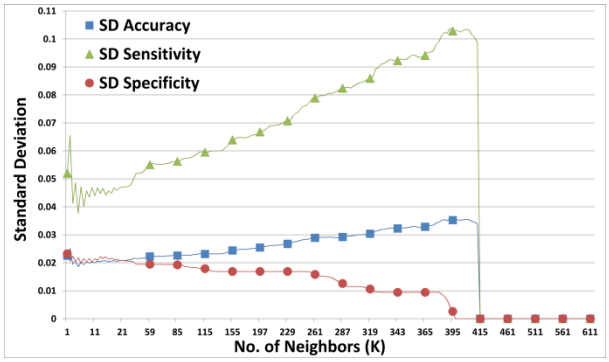

Fig.4 Standard Deviations of Sensitivity, Specificity, and Accuracy

Figure 5 shows minimum, average, and maximum values for sensitivity. These values are depicted in Figure 6 for specificity. Comparing Figures 5 and 6 reveals that the difference between maximum and minimum values of sensitivity is more than the difference between maximum and minimum values of specificity. In other words, under some conditions, the classifier provides proper results for the positive class, but it fails many times to predict positive samples correctly in the worst case.

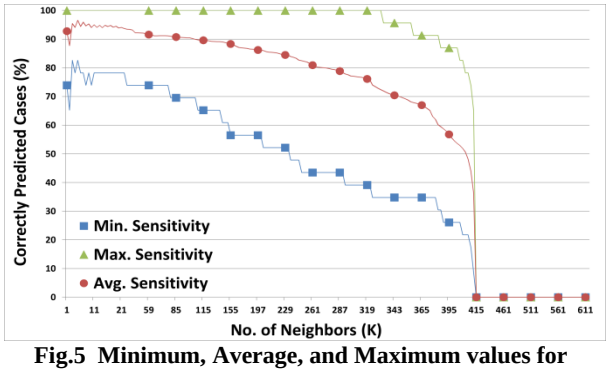

Fig.5 Minimum, Average, and Maximum values for Sensitivity

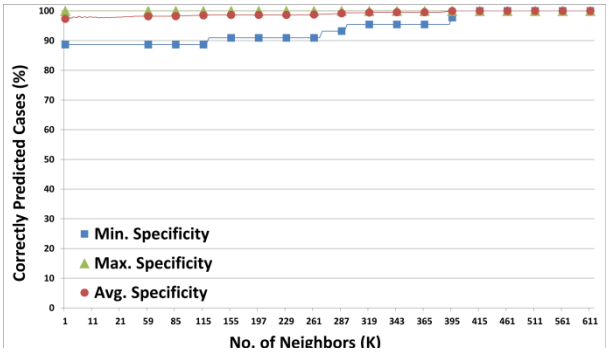

Fig.6 Minimum, Average, And Maximum values for Specificity

5. DISCUSSION

The first study answers to this question that how much KNN works well on malignant and benign test samples. Figure 1 shows that sensitivity is less than specificity for every value of neighbors. It means that the classifier does not work well on samples with malignancy in comparison with samples belong to the benign class. Additionally, it reveals that increase in the number of neighbors (K) reduces KNN predictive power on positive samples remarkably, while the rise in KNN performance on negative samples is inconsiderable. Regarding that the cost of false negative is more than false positive in breast cancer detection, deciding under this condition is risky.

Besides, by making the comparison between Figures 5 and 6, it is obvious that sensitivity drops more than specificity in the worst case. It means that the number of false negatives is increased more than the number of false positives in the worst case and so the performance of the classifier on positive test samples drops more than negative test samples in this situation. In other words, although the averages of accuracy, sensitivity, and specificity could be acceptable, KNN classifier does not have enough power to decide on positive test cases, in the worst case, especially for bigger values of neighbors.

The change in accuracy for different values of neighbors is depicted in Figure 1 that is similar to the results reported by Katsuyoshi Odajima et al. in [2]. However, they did not provide any evidence to show the reason for the drop in accuracy for bigger values of neighbors. Figure 1 clarifies that specificity is almost stable for different values of neighbors while sensitivity drops extremely for greater values of neighbors. Hence, it can be declared that the drop of sensitivity has an impact on the decrease in accuracy for bigger values of neighbors (Eq. 5). In other words, by increasing the number of neighbors, false negative is increased and false positive is consistent. It causes the drop of sensitivity. As a result, the main reason for the drop of accuracy for bigger values of neighbors is the increase in false negative.

Concerning that the number of negative samples is more than the number of positives, the dataset is imbalanced. As a result, finding positive neighbors is more probable than negative neighbors around test cases, especially for bigger values of neighbors. Consequently, it seems that the imbalance dataset causes the increase in false negative for bigger values of neighbors. This is the reason for the drop in sensitivity and also accuracy for greater values of neighbors.

The second study deals with this question that how much the results belong to malignant and benign classes are reliable. Figure 2 shows the standard deviation of sensitivity is more than the standard deviation of specificity. It means that KNN is less reliable in malignant test cases than benign ones while the stability in malignant class is more important than benign class in breast cancer detection.

Additionally, Figure 2 reveals that the standard deviation of accuracy is almost steady for different values of neighbors. This result is fitted into the results provided in [2]. Although the standard deviation of accuracy is almost fixed for the different values of neighbors, Figure 2 makes it obvious that the standard deviation of sensitivity rises and the standard deviation of specificity drops for greater values of neighbors. It means that by an increase in the number of neighbors, sensitivity will be less reliable while specificity is more stable. It seems that the difference between numbers of positive and negative samples gives us an explanation for this behavior of KNN classifier.

6. CONCLUSION

The performance of KNN classifier on individual malignant and benign classes has been studied in this paper. The results have shown that the performance of the KNN on malignant class is less than the benign class. Also, the result in malignant class is less reliable than benign class. This problem is intensified especially for higher values of K. By considering that the cost of false negative is more than false positive for breast cancer detection, It is strongly recommended that sensitivity and its standard deviations should be investigated cautiously, before any software implementation.

Additionally, In order to robustly understand the changes in accuracy, sensitivity, and specificity for different values of neighbors, it is recommended these measures to be examined under the condition that the numbers of samples in positive and negative classes are balanced.

Moreover, it seems that the difference between standard deviations of positive and negative classes could affect the decrease in sensitivity compared with specificity. The standard deviation of the positive class is 2.63 times the standard deviation of the negative class. It means that positive samples are far from each other compared to the negative class samples. As a result, finding neighbors in positive class could be harder than negative class. It seems that it makes sensitivity to drop more than specificity. It is recommended that the effect of standard deviations of positive and negative classes on the performance of classifier to be studied in the next research.

7. REFERENCES

[1] Senturk ZK, Kara R. Breast Cancer Diagnosis via Data Mining: Performance Analysis of Seven different algorithms. Comput Sci Eng. 2014;4(1):35.

[2] Odajima K, Pawlovsky AP. A detailed description of the use of the kNN method for breast cancer diagnosis. In: Biomedical Engineering and Informatics (BMEI), 2014 7th International Conference. IEEE; 2014 May 27. p. 688–692.

[3] Medjahed SA, Saadi TA, Benyettou A. Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. Int J Comput Appl. 2013;62(1).

[4] Gayathri BM, Sumathi CP, Santhanam T. Breast Cancer Diagnosis using Machine Learning Algorithms-A Survey. Int J Distrib Parallel Syst. 2013;4(3):105.

[5] Salama GI, Abdelhalim M, Zeid MA. Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer WDBC. 2012;32(569):2.

[6] You H, Rumbe G. Comparative study of classification techniques on breast cancer FNA biopsy data. IJIMAI. 2010;1(3):6–13.

[7] Gupta S, Kumar D, Sharma A. Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis.

[8] Sarkar M, Leong T-Y. Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 2000. p. 759.

[9] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2006;2.

[10] Pena-Reyes CA, Sipper M. A fuzzy-genetic approach to breast cancer diagnosis. Artif Intell Med. 1999;17(2):131–155.

[11] Karabatak M, Ince MC. An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl. 2009;36(2):3465–3469.

[12] Jhajharia S, Verma S, Kumar R. Predictive analytics for breast cancer survivability: A comparison of five predictive models. In: Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. ACM; 2016. p. 26.

[13] Rahman MM, Davis DN. Addressing the Class Imbalance Problem in Medical Datasets. Int J Mach Learn Comput. 2013;224–8.

[14] Das B, Krishnan NC, Cook DJ. Handling class overlap and imbalance to detect prompt situations in smart homes. In: 2013 IEEE 13th International Conference on Data Mining Workshops. IEEE; 2013. p. 266–273.

[15] Gou J, Du L, Zhang Y, Xiong T. A New Distance-weighted k-nearest Neighbor Classifier. J Inf Comput Sci. 9(6):1429–1436.

[16] Jain R, Mazumdar J. A genetic algorithm based nearest neighbor classification to breast cancer diagnosis. Australas Phys Eng Sci Med. 2003;26(1):6–11.

[17] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inf Syst. 2008;14(1):1–37.

[18] UCI Machine Learning Repository: Data Sets [Internet]. [cited 2016 Sep 13]. Available from: https://archive.ics.uci.edu/ml/datasets.html

[19] Han and Kamber: Data Mining---Concepts and Techniques, 2nd ed., Morgan Kaufmann, 2006.