Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 127 (2018) 293–299

Procedia
Computer Science

www.elsevier.com/locate/procedia

The First International Conference On Intelligent Computing in Data Sciences

Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis

Walid CHERIF[a,*]

[a]*Laboratory SI2M, Department of Computer Science,*
National Institute of Statistics and Applied Economics,
B.P. 6217, Rabat, Morocco

Abstract

There is a growing trend towards data mining applications in medicine. Different algorithms have been explored by medical practitioners in an attempt to assist their work; the diagnosis of breast cancer is one of those applications. Machine learning algorithms are of vital importance to many medical problems, they can help to diagnose a disease, to detect its causes, to predict the outcome of a treatment, etc. K-Nearest Neighbors algorithm (KNN) is one of the simplest algorithms; it is widely used in predictive analysis. To optimize its performance and to accelerate its process, this paper proposes a new solution to speed up KNN algorithm based on clustering and attributes filtering. It also includes another improvement based on reliability coefficients which insures a more accurate classification. Thus, the contributions of this paper are three-fold: (i) the clustering of class instances, (ii) the selection of most significant attributes, and (iii) the ponderation of similarities by reliability coefficients. Results of the proposed approach exceeded most known classification techniques with an average f-measure exceeding 94% on the considered breast-cancer Dataset.

Keywords: data mining; cancer diagnosis; supervised classification; unsupervised classification; k-nearest neighbors; k-means; similarity measurement.

* Corresponding author. Tel.: +212-672-277-806.
E-mail address: w.cherif@insea.ac.ma.

10.1016/j.procs.2018.01.125

1. Introduction

In medical domains, the volume and complexity of collected data is growing at a rapid pace. This includes besides the information coming from clinical studies, other data on patients [1]. The analysis of such data makes it possible to come up with new medical hypotheses or to confirm existing hypotheses. This partially overcomes the limitations of traditional medical studies which were restricted to small numbers of parameters and small numbers of instances.

In particular, breast-cancer is the most common type of cancerous disease among women of the western world. Approximately every tenth woman suffers from it in her lifetime, half of whom does not survive. Even though breast-cancer is such a severe disease, effective treatment is possible if it is detected at an early stage [2].

The social and economic values of breast-cancer diagnosis are very high [3]. As a result, the problem has attracted many researchers in the area of data mining recently [4, 5]. Their efforts generated various approaches for automatic diagnosis.

As breast-cancer diagnosis is an important and complicated task that needs to be extremely accurate and efficient. Its automation would be very advantageous. It may probably exceed traditional approaches [6]. However, standard algorithms are still limited and no algorithm has proved perfect diagnosis.

Technically, the problem of diagnosis belongs to binary classifications. It stays at the cross junction of statistics and artificial intelligence [7], it aims to classify instances into two groups on the basis of classification criteria.

Binary classification includes two types of models: predictive (supervised) such as KNN algorithm [8], in which the class of each instance is known, and exploratory (unsupervised) such as k-means algorithm whose task is the creation of clusters (classes of instances) [9]. Supervised classification models consist of two major steps: training and testing [10].

Therefore, this paper proposes a novel approach that extends the KNN algorithm by a normalization stage and by creating clusters inside each class. The second improvement is based on the elimination of insignificant attributes. Finally, obtained clusters are then compared to each new instance in terms of a weighted similarity measure. In experiments, each one of these improvements proved quite interesting and contributed into the overall performance.

The rest of this paper is structured as follows: Section 2 summarizes main approaches applied to breast-cancer datasets. Section 3 meticulously details the proposed approach; and in section 4, the obtained results are compared to those of most known classification techniques. Finally, the last section concludes this work.

2. Background

Several researchers in the literature have measured their performances on recognizing benign from malignant breast-tumors.

Sarkar and Leong [11] treated the breast-cancer diagnosis as a pattern classification problem. They used KNN algorithm as a nonparametric classifier to predict malignant samples. Their study included another enhancement: the fuzzy KNN.

Setanio proposed [12] a rule extraction algorithm based on artificial neural networks. The rules were extracted from the network. The pruning algorithm was used to remove redundant connections, and a clustering step was used to discretize the activation values of the input pattern. A similar rule extraction model was presented by Taha et al. [13]. It included three rule extraction algorithms based on artificial neural networks.

In the work of Reyes and Sipper [14], fuzzy logic and genetic algorithm were combined into a same classifier system which outperformed other artificial neural networks approaches.

Abbass [3] proposed an evolutionary artificial neural network approach based on the pareto-differential evolution algorithm augmented with local search for the prediction of breast-cancer.

Kuo at el. [15] opted for decision tree technique to classify breast cancers. Their work aimed to reduce the number of unnecessary biopsies and to increase the diagnosis confidence. 24 features were used to create a decision tree with the ability of recognizing malignant breast-cancers.

Sawarkar et al. [16] have used, in addition to artificial neural networks, support vector machines for breast cancers diagnosis. The implemented algorithm maps the input data into a high-dimensional space. Further, it associates the instances into their respective classes by separating formed hyperplanes.

In many cases, nonparametric classification techniques such as KNN algorithm have become an attractive approach [17, 18].

3. The proposed approach

The proposed approach is an improvement of KNN algorithm. This section will thus briefly explain the principle of the algorithm, highlights its limits, and introduces the clustering stage and the reliability computations to improve it:

3.1. Theory of K-nearest neighbors

The k-nearest neighbors (KNN) algorithm is one of the simplest similarity-based artificial learning algorithms, offering interesting performance in some contexts [19].

When classifying a given instance, the basic idea is to make the nearest neighbor instances, in the sense of a predefined distance, vote on. The class of the new instance is then determined according to the most frequent class among the k nearest neighbors. The choice of the value of k must be chosen a priori; various techniques have been proposed to select it such as cross-validation and heuristics. This value should not be a multiple of the number of classes to avoid tie votes. Thus, in the case of a binary classification, it is necessary to take a value of k odd so that a majority necessarily emerges. The performance of KNN also depends largely on the measure used to calculate the distances between the instances [20].

The KNN method is non-parametric; this means that the algorithm makes it possible to classify without making any assumptions about the function $Y = f(x_1, x_2, \ldots x_n)$ which associates the class Y to the attributes x_j $(1 \le j \le n)$.

In general, large values of k reduce the effect of noise on classification but make class boundaries less distinct [21].

3.2. Limits of K-nearest neighbors

The limits of KNN algorithm are as follows:

- The KNN algorithm is slow since it reviews all the instances each time.
- The algorithm is vulnerable to dimensionality.
- The algorithm is sensitive to irrelevant and correlated attributes.
- A wrong choice of the distance or the value of k degrades the performance.

3.3. Optimizations of K-nearest neighbors

The KNN algorithm has seen several improvements in an attempt to overcome its limitations [22]:

Bailey introduced weights to classical KNN to present weighted K-nearest neighbors (WKNN) [23]. In WKNN, the weights are assigned to each calculated value, then the nearest neighbors are computed, and finally, the class is assigned to the processed instance.

The Condensed Nearest Neighbor algorithm (CNN) eliminates duplicate data, removes irrelevant instances which do not give additional information, and shows similarity with other training datasets. The Reduced Nearest Neighbor algorithm (RNN) meanwhile includes an additional step: it eliminates patterns which don't affect the result.

WKNN algorithm extends classical KNN in two ways [21]:

- A weighting scheme for nearest neighbors.
- A standardization of distances.

Indeed, in classical KNN, the k nearest neighbors influence the prediction in an identical way, regardless of their degree of similarity with the new instance. This extension is based on the idea that samples that are particularly close to the new instance must have a higher weight in the decision than those that are farther apart.

Thus, the distances of the k nearest neighbors are standardized by the $(k+1)^{th}$ neighbor, and all obtained standardized distances take values in the interval [0, 1].

The weights are calculated according to a kernel function K. The purpose of such functions is to weight the observations relative to a reference point so that the closer an instance is to the reference, the greater its weight will be.

Yong, et al. [24] applied a new model of KNN, improved by clustering, in text classification. The training instances of each class are clustered by k-means algorithm, and resulting cluster centers are taken as the new training dataset. Each training instance is weighted by a value representing its importance.

Su [25] proposed a method to identify flooding attacks in real-time, based on anomaly detection by genetic weighted KNN. A genetic algorithm is used to train an optimal weight vector for features; meanwhile, an unsupervised clustering algorithm is applied to reduce the number of instances in the sampling dataset, in order to shorten training and execution time, as well as to promote the system's overall accuracy.

3.4. The proposed model

3.4.1. Normalization

The main drawback of similarity-based approaches for many attributes is the normalization:

Let's consider the binary classification problem in table 1:

The instance: E_1 which belongs to the first class a, and E_2 which belongs to the second class b.

The new instance E_3 has equal values to E_2 on three attributes: x_2, x_3 and x_4; but the order of magnitude of x_1 (10^3) makes the Euclidean distance very far ($d(E_3,E_2)=950; d(E_3,E_1)=50.87$), favoring thus the instance of the class a which has a relatively close value only for x_1.

Table 1. Data Normalization

Instance	x_1	x_2	x_3	x_4	Y
E_1	3000	35	20	15	a
E_2	1200	50	30	20	b
E_3	3100	50	30	20	

A normalization step reduces the values e_{ij} of the attribute x_j for the example E_i to the interval [0,1] by the formula:

$$e_{ij} \leftarrow \frac{e_{ij} - \min_j(e_{ij})}{\max_j(e_{ij}) - \min_j(e_{ij})}$$

The previous distances become: $d(E_3,E_2)=0.5$ and $d(E_3,E_1)=0.87$

3.4.2. Reliability measure

The second contribution of this paper is reliability coefficients. They enable the selection of significant attributes on the one hand, and the reduction of the dimensionality on the second hand.

Table 2. Reliability Coefficients

Instance	x_1	x_2	x_3	Y
E_1	0.1	0	0.3	a
E_2	0	0.2	0	a
E_3	0.2	1	0.7	a
E_4	0.5	0	1	b
E_5	1	0.5	0.5	b
E_6	0.5	1	0.6	b

From the abovementioned table, the classification of a new instance (x_{71}, x_{72}, x_{73}) distinguishes three types of attributes:

If $x_{71} \leq 0.2$, then E_7 has a big chance to belong to the class a.

If $x_{71} \geq 0.5$, then E_7 has a big chance to belong to the class b.

The attribute x_1 is decisive in the classification.

3 values of each class belong to the interval of the class. The reliability coefficient is therefore maximal: 1. No similar deduction is possible for the attribute x_2 since both classes share values in the interval [0, 1]

The reliability coefficient is therefore minimal for x_2: 0. x_2 is insignificant and can be ignored.

The attribute x_3 has reliable intervals: [0, 0.5[for the class a, and]0.7, 1] for the class b, and intermediate values are common. The reliability coefficient is therefore 0.5: only 3 of the 6 values taken by x_3 belong to reliable intervals.

3.4.3. Clustering

Instances of the dataset are replaced by less, but more significant, centers of clusters. K-means algorithm is used to form clusters, and the classification will be based on the centers of this new set of clusters.

Thus, classifying a new instance into one of the k clusters instead of comparing it to the initial n instances divides the computation time of the algorithm by $\frac{n}{k}$.

Finally, the distance between a given instance and the center of each cluster is restricted to significant attributes, and weighted by their reliability coefficients.

4. Results and Analysis

4.1. Dataset

The considered breast-cancer dataset contains 355 benign and 210 malignant cancers [26].

The attributes are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Each instance contains the ID, the diagnosis (malignant or benign) and 30 attributes such as radius, texture, fractal dimension, etc.

4.2. Computation

The dataset was divided into 5 subsets of 113 instances each (71 benign and 42 malignant).

At each iteration, 4 subsets are considered for training and the fifth for the test, the process is thus repeated five times, and the average f-measure is retained.

The F-Measure used for evaluation is written as [27]:

$$precision = \frac{a}{a+b} \quad ; \quad recall = \frac{a}{a+c}; \quad F-measure = \frac{2 \times precision \times recall}{precision + recall}$$

where a: is the number of instances correctly classified,
b : the number of false positives
and c : is the number of false negatives.

4.3. Results

Figure 1 summarizes the results of classification of the proposed approach compared to four other algorithms, namely SVM [28] with linear kernel, ANN [29], NB [30] and KNN with k=3 [31]. The proposed approach yielded an average f-measure: 94.1%, exceeding: SVM (89.7%), NB (92.2%) and KNN (91.1). the only algorithm which returned slightly better performance is ANN with an average f-measure: 95.6%. but this latter is the slowest one among this list of algorithms. It consumed 2.2 times more computation time than the proposed algorithm.

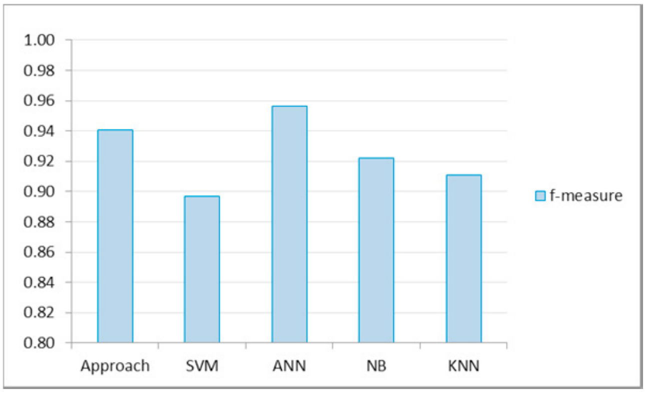

Fig. 1. F-measure of the considered algorithms

Finally, on both accuracy and classification time, the proposed approach has undoubtedly yielded the best performance. The extension of the proposed approach to larger and more complex datasets is the focus of our current research.

5. Conclusion and perspectives

The advent of high-performance computing has benefited various disciplines in finding practical solutions to their problems, and breast-cancer diagnosis is no exception to this. Such problems when mined properly can lead to better diagnosis.

This paper investigates a novel approach for binary classification of breast-cancers. It selects most reliable attributes, and then weights them according to their level of reliability. To classify a given instance into one of the two classes (benign or malignant), the algorithm starts by forming clusters inside each class in order to speed up the KNN process. This process is of major interest especially for large and complex datasets since it reduces both the number of computed distances (k classes instead of n instances), and the number of attributes figuring in these distances since it eliminates insignificant ones.

The results of classification indicate that the proposed algorithm outperforms KNN, NB, SVM with an f-measure slightly exceeding 94% on the considered dataset.

References

[1] Mueller, M. L. (2012). Data Mining Methods for Medical Diagnosis. Technical University of Munich.

[2] Oskouei, R. J., Kor, N. M., & Maleki, S. A. (2017). Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges. American journal of cancer research, 7(3), 610.

[3] Abbass, H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. Artificial intelligence in Medicine, 25(3), 265-281.

[4] Diz, J., Marreiros, G., & Freitas, A. (2016). Applying Data Mining Techniques to Improve Breast Cancer Diagnosis. Journal of medical systems, 40(9), 203.

[5] Chaurasia, V., & Pal, S. (2017). Performance Analysis of Data Mining Algorithms for Diagnosis and Prediction of Heart and Breast Cancer Disease.

[6] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

[7] Tung, H. H., Cheng, C. C., Chen, Y. Y., Chen, Y. F., Huang, S. H., & Chen, A. P. (2016, November). Binary Classification and Data Analysis for Modeling Calendar Anomalies in Financial Markets. In Cloud Computing and Big Data (CCBD), 2016 7th International Conference on (pp. 116-121). IEEE.

[8] Hwang, W. J., & Wen, K. W. (1998). Fast KNN classification algorithm based on partial distance search. Electronics letters, 34(21), 2062-2063.

[9] Mirošević, I. (2017). k-means Algorithm. KoG, 20(20), 91-98.

[10] Lad, H., & Mehta, M. A. (2017). Feature Based Object Mining and Tagging Algorithm for Digital Images. In Proceedings of International Conference on Communication and Networks (pp. 345-352). Springer, Singapore.

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

[12] Setiono, R. (1996). Extracting rules from pruned neural networks for breast cancer diagnosis. Artificial Intelligence in Medicine, 8(1), 37-51.

[13] Taha, I., & Ghosh, J. (1997, June). Evaluation and ordering of rules extracted from feedforward networks. In Neural Networks, 1997., International Conference on (Vol. 1, pp. 408-413). IEEE.

[14] Pena-Reyes, C. A., & Sipper, M. (1999). A fuzzy-genetic approach to breast cancer diagnosis. Artificial intelligence in medicine, 17(2), 131-155.

[15] Kuo, W. J., Chang, R. F., Chen, D. R., & Lee, C. C. (2001). Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. Breast cancer research and treatment, 66(1), 51-57.

[16] Sawarkar, S. D., Ghatol, A. A., & Pande, A. P. (2006, June). Neural network aided breast cancer detection and diagnosis using support vector machine. In Proceedings of the 7th WSEAS International Conference on Neural Networks (pp. 158-163).

[17] Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3), 238-247.

[18] Bueno, G., Vállez, N., Déniz, O., Esteve, P., Rienda, M. A., Arias, M., & Pastor, C. (2011). Automatic breast parenchymal density classification integrated into a CADe system. International journal of computer assisted radiology and surgery, 6(3), 309-318.

[19] Diz, J., Marreiros, G., & Freitas, A. (2016). Applying Data Mining Techniques to Improve Breast Cancer Diagnosis. Journal of medical systems, 40(9), 203.

[20] Goldberger, J., Hinton, G. E., Roweis, S. T., & Salakhutdinov, R. R. (2005). Neighbourhood components analysis. In Advances in neural information processing systems (pp. 513-520).

[21] Mathieu-Dupas, E. (2010). weighted KNN Algorithm and application in diagnosis Algorithme des k plus proches voisins pondérés et application en diagnostic. In 42nd Days of Statistics..

[22] Bhatia, N. (2010). Survey of nearest neighbor techniques. arXiv preprint arXiv:1007.0085.

[23] Bailey, T., & Jain, A. K. (1978). A note on distance-weighted $ k $-nearest neighbor rules. IEEE Transactions on Systems, Man, and Cybernetics, (4), 311-313.

[24] Yong, Z., Youwen, L., & Shixiong, X. (2009). An improved KNN text classification algorithm based on clustering. Journal of computers, 4(3), 230-237.

[25] Su, M. Y. (2011). Using clustering to improve the KNN-based classifiers for online anomaly network traffic identification. Journal of Network and Computer Applications, 34(2), 722-730.

[26] Bennett, K. P. (1992). Decision tree construction via linear programming (pp. 97-101). Center for Parallel Optimization, Computer Sciences Department, University of Wisconsin.

[27] Powers, D.M.W., 2011. Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

[28] Vapnik V (1995) The nature of statistical learning theory. Springer, New York.

[29] Wasserman, P. D. (1993). Advanced methods in neural computing. John Wiley & Sons, Inc.

[30] Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications, 6(2), 256-261.

[31] Hwang, W. J., & Wen, K. W. (1998). Fast KNN classification algorithm based on partial distance search. Electronics letters, 34(21), 2062-2063.