International Journal of Computer Engineering and Information Technology

VOL. 8, NO. 6, June 2016, 90–95
Available online at: www.ijceit.org
E-ISSN 2412-8856 (Online)

A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm

Maryam Kuhkan

Department Of Computer Engineering, Malayer Branch, Islamic Azad University, Malayer, Iran

m.276k9@gmail.com

ABSTRACT

K-Nearest Neighbor Algorithm (Knn) is one of the best and most widely used classification algorithms with a variety of applications. One of the most important challenges to the application of the algorithm is the same impact of all characteristics in doing the classification, while some of the characteristics are less important for classification; this may deviate the process of classification and reduce the accuracy of the Knn algorithm.

Using a new method in this study, a certain weight is allocated to various features based on their importance, so that the same effect of all features is avoided in doing the classification and deviation of classification process, thereby increasing the accuracy of Knn algorithm classification. The comparison of the results of the proposed algorithm implementation and five other classification algorithms on 10 datasets selected from UCI repository indicated the considerable improvement of the classification by the algorithm.

Keywords: *Knn Algorithm, Accuracy Improvement, Data Mining, Classification.*

1. INTRODUCTION

Nowadays, with advances in technologies especially those related to information and communication, huge amounts of data are created in communication and information networks. One of the requirements of success in different businesses is the possibility of using the information and data. Different business managers know that data and information collection from customers is one of the factors needed for the growth and development of the companies, but it is only half of the path; if data are just collected and remained unusable, practically the main goal of collecting the information has not been accomplished. The crucial step that must be taken after data collection is to extract knowledge and to make the collected data more perceptible. Data mining is one of the most important methods for extracting knowledge from a large amount of data.

Data mining means to find hidden patterns available in data set. It makes use of analysis models, classification and estimation of information [1]. In summary, data mining functions in order to discover hidden information and to predict unknown or unseen cases.

This study examines one of the data mining algorithms known as Knn algorithm, The Knn algorithm is used for classification; the aim of data classification is to place the data in distinct categories or classes. One of the problems in using this algorithm is the similar effect of all the features on classification while some features are less important to the classification. The minor features cause two records that are close to each other to be recognized far from one another. This misleads the classification process and reduces the accuracy of Knn algorithm. In this paper a new method is presented to prevent the diversion of the classification process and to increase the accuracy of the algorithm.

In what follows in the second section, an account of some basic concepts in the field of data mining and classification is provided, and in the third section the studies conducted by other researchers are presented; in the fourth section the proposed method of the article will be explained, and in the fifth part the obtained results will be presented. Finally, in the sixth part the conclusion will be stated.

2. DATA MINING

Different definitions are proposed in various scientific literatures for data mining. Some of the most common definitions are as follows:

- Data mining includes the detection of valid, new, and understandable patterns in data sets; in other words, it is a process that extracts knowledge from data sets by using smart techniques [2].
- Data mining is an interdisciplinary field that has integrated different fields such as database, statistics, machine learning and other related fields, so that invaluable information and

knowledge hidden in large amounts of data can be extracted [3].

2.1 Data Mining Methods

The main methods of data mining are two categories; predictive and descriptive methods.

2.1.1 Predictive Methods

The predictive methods use some characteristics in order to predict the value of a specific characteristic. In different scientific texts, different predictive methods are known as the methods with supervisor. Classification, regression, and deviation detection methods are three learning methods of the model in data mining with predictive nature [4]. The three methods are briefly explained here:

- Classification: Classification issues lead to the identification of characteristics that specify to which class each record belongs. In classification algorithms the primary datasets are divided into training datasets and experimental datasets. The model is made by using training datasets and the experimental dataset is used for evaluation and computation of the model accuracy [5]. Some common classification methods are support vector machine (SVM), K - nearest neighbor (KNN), genetic algorithm, neural network, Bayesian classification.
- Regression: Prediction of the value of a continuous value based on the values of other variables and on the basis of a linear or non-linear dependence model is called regression such as the prediction of wind speed as a function of air pressure, humidity, and temperature [4].
- Anomaly detection: This application is used just when the samples with similar labels which usually show normal situation are available. Therefore, as only the samples of normal class are available, the algorithm makes model for the normal situation and with regard to a specified threshold, and considers any violation of that threshold as the abnormal situation such as detection of credit card fraud [4].

2.1.2 Descriptive Methods

The descriptive methods find describable patterns that describe the relationships governing the data regardless of any label or output variable. In different scientific literature, the descriptive methods are known as unsupervised methods. Some examples of the descriptive methods are clustering, exploring association rules, and discovering sequential patterns [5].

2.2 K-Nearest Neighbor Algorithm (Knn)

The knn algorithm is one of the most famous classification algorithms used for predicting the class of a record or (sample) with unspecified class based on the class of its neighbor records. The algorithm is made of three steps as follows [5]:

1. Calculating the distance of input record from all training records
2. Arranging training records based on the distance and selection of K-nearest neighbor
3. Using the class which owns the majority among the k-nearest neighbors (this method considers the class as the class of input record which is observed more than all the other classes among the K-nearest neighbors).

In general, for predicting a new record class the algorithm looks for similar records among the set of training records, so that if the records have n attributes, then it will consider them as a vector in n-dimensional space and predict the class label of the new record based on a distance criterion in this space such as the Euclidean distance as well as the class label of the neighbors.

The classifier assumes the distance of records from each other as a criterion for their nearness and selects the most similar records. There are numerous methods to compute the distance such as the function of Euclidean distance, Manhattan, etc., among which the function of Euclidean distance is one of the most common ones defined as Equation 1.

$$x1 = (x_{11}, x_{12}, \dots, x_{1n})$$
$$x2 = (x_{21}, x_{22}, \dots, x_{2n})$$
$$dis(x1, x2) = \sqrt{\sum_{i=1}^{n} (x_{1i-} x_{2i})^2} \quad (1)$$

One of the most important parameters in the knn algorithm is K value; in fact, there is no accurate value for k and its proper amount depends on the data distribution and space of the problem.

3. RELATED STUDIES

Researchers have offered different methods in order to solve the problem of Knn algorithm and improve its accuracy; for instance in [6] the genetic algorithm is used for solving the problem, and the genetic algorithm is combined with knn. (Knn-Ga), so that every characteristic is given a weight by using genetic algorithm; in the genetic algorithm used in [6], the sum of weights is considered as a chromosome, and their fitness is evaluated through classification error; then, using the obtained weights, the knn algorithm is weighted and classification is performed.

In [7], a synthetic method consisting of knn algorithms and (SVM) is proposed to identify hand writing. Since one of the problems with identifying handwriting is the similarity between characters, which in turn leads to the confusion of knn algorithm, SVM algorithm is used in this paper to separate characters which cause confusion in knn. In this article, SVMs act as decision-maker classifiers when knn algorithm is plunged into confusion.

In [8], a specific method is used By Mr. David Aha to weight the characteristics and to solve knn problems, so that by using a proposed algorithm and based on the accuracy of classification and the difference between the characteristics in the new sample and training samples, the weight of each characteristic is changed according to Table 1 to achieve the right weight for each characteristic and then knn is weighted and classified using the obtained weights.

Table 1: Changing the weight of characteristics according to the accuracy of classification and its difference in input and training samples

Difference/ Classification	***Correct***	***Wrong***
Little	Much Increase	Much Decrease
Much	Little Increase	Little Decrease

In Aha method, for instance, if the classification is performed correctly and the difference between one feature and its corresponding feature in the nearest neighbor is little, this means that that characteristic is of paramount importance; therefore, the weight of that characteristic is increased more than the other states to enhance the effect of that characteristic on calculating the distance (In other cases, as displayed in Table 1, the weights are changed based on the same argument).

4. PROPOSED METHOD

Our proposed method for solving the problem of knn algorithm is to allocate weight to each characteristic; that is, for each characteristic like i we have defined a weight like wi (in order to define the weights we have used a specific rule that will be explained in the following). It enables us to separate the characteristics with different levels of importance and consequently to prevent the effect of less important characteristics on the classification process and classification deviation.

The main idea of the research on how to allocate weight to the features is inspired from the method developed by David Aha in [8] (Aha method is entirely described in Part III), so that by doing various studies and tests on the Aha method we came to the conclusion that the method has certain drawbacks as the following:

1. When classification is performed accurately, it is logical to assume that the value of characteristics weight is correct, so there is no reason to change the weights. On the other hand, when the classification is done correctly and the rate of the characteristics is the same, changing the weight of characteristics will have no effect on making decisions on classification.
2. If classification is performed correctly, and the value of characteristics is not the same, changing the weight of the characteristics will change distances, and consequently it may affect making decisions on classification.

Given the above mentioned problems, we have offered a method in this article which not only removes the problems of Aha method, but also significantly enhances the accuracy of Knn algorithm classification.

Our proposed algorithm acts in such a way that at first the weight algorithm of all the characteristics is considered to be the same and then in each step based on the fact that whether for the new sample (the input sample or the sample the class of which is to be predicted) classification has been done correctly or not as well as on the basis of the difference between the value of each characteristic in the new sample and its corresponding characteristic in the nearest neighbor, the weight of the characteristic is changed according to Table 2 until the ultimate weight for each characteristic is achieved.

Table 2: Changing the weight of characteristics according to the accuracy of classification and its difference in input and training samples

Difference/ Classification	***Correct***	***Wrong***
Little	Unchanged	Much Decrease
Much	Unchanged	Little Decrease

For example, suppose that the sample test of x is used to determine the class and y is selected as the nearest neighbor; in order to determine the feature like i, first the difference of this feature in X and Y is obtained and then with regard to the correctness or incorrectness of the classification, the weights will be changed according to Table 2.

In the proposed model, if the new sample is correctly classified, the weight of the characteristic will not be changed because changing their weight in this condition has no effect on the classification; however, if classification is wrong, the value of the characteristics will be reduced according to the rate of their differences, so that the differences will be more revealed. For instance, when the classification is done wrongly and the difference between one characteristic and its corresponding characteristic in the nearest neighbor is low, it means that the characteristic is of little importance, so the weight of that characteristic is reduced more than the other states (much decrease) to reduce the effect of that characteristic on computing the distance and in the other states (as displayed in Table 2), the weights will be changed based on the same argument.

Having obtained the weight of each characteristic, knn algorithm will be revised by making some changes in its distance function; then, the classification will be performed. Since the Euclidean distance function is used in this article to calculate the distance, the function is revised as Equation 2.

$$\mathrm{dis}(x1,x2) = \sqrt{\sum_{i=1}^{n} \left(w_i\left(x_{1i} - x_{2i}\right)\right)^2}$$

$$w = (w_1, w_2, \dots, w_n) \qquad (2)$$

Some notes taken into account in the implementation of the proposed algorithm are as follows:

1. 30% of the data set is taken into account as the testing sample and 70% as the training sample.
2. In the knn algorithm, the number of neighbors is considered to be equal to 7 (k=7).
3. The Euclidean distance function is used to calculate the distance.
4. Matlab R2014b programming context and Weka 3.6.11 software program are used to implement the proposed algorithm.
5. In order to achieve the increase or decrease step (the amount of increase or decrease of weight) of the characteristics in each phase, by calculating the accuracy of classification and the rate of error an average number is obtained in each step.

5. RESULTS

In this part, the suggested algorithm in this research is compared with five other classification algorithms on 10 data sets of the UCI repository [9] in terms of classification accuracy. The data sets are completely described in Table 3.

Table 3: Description of the data sets used in the research

Row	*Dataset*	*Size*	*Attribute*	*Class*
1	Iris	150	4	3
2	Diabetes	768	8	2
3	Credit	690	15	2
4	Labor	57	16	2
5	Haberman	306	3	2
6	Zoo	101	17	7
7	Vehicle	946	18	4
8	Hepatitis	155	19	2
9	Wine	178	13	3
10	Breast-cancer	699	10	2

The algorithms that are used for comparison are as follows:

1. K-nearest neighbor algorithm
2. The algorithms presented by David Aha in [8]
3. The Naïve Bayesian algorithm (NB) [10]
4. J48 algorithm [11]
5. LWL algorithm [12]

It should be noted that all of the classification algorithms were implemented in WEKA 3.6.11 and Matlab R2014b software. The results of implementing the proposed algorithm and the other five classification algorithms are shown in Table 4.

Table 4: Comparison of accuracy of the proposed classification method and five other classification algorithms on 10 UCI datasets

Row	*Dataset / Method*	*Proposed method*	*Knn*	*Aha Method*	*NB*	*J48*	*Lwl*
1	Iris	94.97	89.74	89.09	85.91	89.01	82.04
2	Diabetes	83.75	73.12	83.9	78.70	79.00	67.74
3	Credit	93.90	83.40	87.05	84.25	86.02	84.60
4	Labor	69.96	52.20	63.01	59.60	61.12	58.12

5	Haberman	89.75	72.42	82.16	73.12	75.11	73.10
6	Zoo	97.32	86.06	90.2	91.60	89.50	89.26
7	Vehicle	84.69	75.10	79.13	77.68	78.42	77.14
8	Hepatitis	87.81	72.37	80.23	78.20	79.90	74.93
9	Wine	78.26	64.08	71.12	62.50	73.46	65.85
10	Breast-cancer	89.69	75.04	82.08	81.35	84.00	87.30
	Average	87.01	74.35	80.79	77.29	79.55	76.00

According to the Table 4, it can be concluded that the mean of the classification accuracy of the proposed algorithm in this research showed remarkable improvement in all the datasets compared with Knn algorithm (13% in average). Its accuracy also has increased 6% to 11% in comparison to the other algorithms as follows:

1. Compared with Aha method, its accuracy has increases in most datasets (in average, it has an accuracy increase of about 6% compared to this algorithm), which indicated the better performance of our proposed method than the method offered by Aha in [8].
2. Compared with NB algorithm, it has an average accuracy increase of about 10%.
3. Compared with J48 algorithm in most data sets it has an increased accuracy (in average 7% increase in accuracy in comparison to this algorithm).
4. Compared with LWL algorithm, it has an average accuracy increase of about 11%.

6. CONCLUSION

In this paper, a new algorithm was introduced based on dynamic weighting to the characteristics in order to improve the classification accuracy of the Knn algorithm. It was observed that we can differentiate less important characteristics from other characteristics by using the method, in the sense that more important characteristics exert greater effect on the calculation of the distance between the records in Knn algorithm, and thus the impact of less important characteristics on the classification accuracy and on deviation of classification process will be prevented in this method.

According to the practical experiments, it is observed that the proposed algorithm increases the accuracy of knn algorithm and own higher accuracy than the other classification algorithms.

REFERENCES

[1] J. Han, M. Kamber and J. Pei, "Data Mining And Techniques", 3nd ed, Elsevier, 2006.

[2] S. Larose, T. Daniel, "Discovering Knowledge In Data An Introduction To Data Mining", John Wiley And Sons Press, 2005.

[3] Gp. Shapiro, "Knowledge Discovery In Databases: 10 Years After", ACM SIGKDD Explorations, Vol. 1¸ No. 2, 2000, pp. 59-61 .

[4] M. Kantardzic, "Data Mining: Concepts, Methods¸ Models And Algorithms", 2th ed, John Wiley And Sons Press, 2011.

[5] AH. Wahbeh¸ QA. Al-Radaideh¸ MN. Al-Kabiand¸ EM. Al-Shawakfa¸ "A Comparition Study Between Data Mining Tools Over Some Classification Methods", International Journal Of Advanced Computer Science And Applications. Vol. 35¸ 2011, pp. 18-26.

[6] N. Suguna, K. Thanushkodi¸ "An Improved K-Nearest Neighbor Classification Using Genetic Algorithm", International Journal Of Computer Science Issues, Vol. 7, No. 2, 2010, pp. 32-61.

[7] C. Zanchetitin¸ BL. Bezerra¸ W. Azevedo¸ "A Knn-Svm Hybrid Model For Cursive Hand writing Recognition"¸ The International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, June 10-15, 2012, pp. 1-8.

[8] DW. Aha. "Feature Weighting For Lazy Learning Algorithms", The Springer International Series In Engineering And Computer Science¸ Vol. 453, 2005, pp. 13-32.

[9] A. Asuncion, DJ. Newman¸ "UCI Machine Learning Repository", Irvine, CA: University of California, School of Information and Computer Science. 2007. [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html

[10] H. George, J. P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers", Proceedings of the Eleventh conference on Uncertainty in artificial

intelligence, Montreal, Canada, August 18-20, 1995, pp. 338-345.

[11] J.R. Quinlan¸ "C4.5: Programs for Machine Learning", San Francisco: Morgan Kaufmann Publishers, San Mateo, 1993.

[12] E. Frank, M. Hall, B.P. fahringer¸ "Locally Weighted Naive Bayes", 19th Conference On Uncertainty In Artificial Intelligence, Acapulco, Mexico, Aug 7-10, 2003, pp. 249-256.