Application of an Optimized KNN Algorithm in Music Classification

Yunjie Ji
School of Computer Science
Nanjing University of Posts and Telecommunications
Nanjing, China

***Abstract*—This study introduces an enhanced KNN algorithm tailored for music genre classification. It addresses two key limitations in the traditional KNN approach, effectively resolving issues of disregarding attribute-category correlation and overlooking similarity disparities between nearest samples and those to be classified. By doing so, it significantly enhances classification accuracy. The algorithm's application in music genre classification demonstrates superior performance, particularly in scenarios where category distinctions are subtle or in cross-category situations. Notably, this algorithm boasts simplicity, symmetry, minimal dependencies, high computational efficiency, and scalability for large-scale music data classification needs.**

***Index Terms*—KNN, Music genre classification, Attributes dependency degree**

I. Introduction

With the growth of online music databases and easy access to music content, people find it increasing hard to manage the songs that they listen to. One way to categorize and organize songs is based on the genre, which is identified by some characteristics of the music such as rhythmic structure, harmonic content and instrumentation (Tzanetakis and Cook, 2002). Being able to automatically classify and provide tags to the music present in a user's library, based on genre, would be beneficial for audio streaming services such as Spotify and iTunes. K-NearestNeighbor (hereinafter referred to as KNN) classification algorithm is a typical non-parametric, effective and popular lazy learning method, has always been the hot topics in research on data mining, machine learning and statistical pattern recognition. Its advantages are mainly manifested in its simple principle, convenient implementation, support of incremental learning, ability to build model for ultra-polygon complex decision space, and better classification performance in the situation of cross-class field. But its problems are mainly reflected in two aspects, one is that, when the traditional KNN algorithm measures the similarity, it supposes that the effect of each attribute in the process of distance calculation is the same, and ignores the problem of degree of correlation between attributes and its categories, thus affects the classification accuracy. The second is that, in the process of category judgement, it only considers the number of nearest neighbors in each category, and ignores the similarity differences between the nearest neighbors and the samples to be classified. Many domestic and overseas scholars brought up improvements for these two problems. Teacher Qian Shangwen [1] proposed a weighting function based on the Gini coefficient. Wang Peiji et al., who proposed an attribute reduction algorithm based on attribute dependency degree, it can simplify attributes in system containing uncertain information and data noise, and delete redundant rules, and keep the system functionality and performance unchanged [2]. Wang Shiqiang and others used fuzzy-rough set as model, and proposed a two-step simplification method to extend the concept of fuzzy dependency degree used to describe the dependency of condition attribute and decision attribute, and it can be used to measure the dependencies between the condition attributes [3].

This paper proposes a double weighted KNN (DW-KNN) classification algorithm. This algorithm is double weighted based on the traditional KNN algorithm, and can solve the problem that the traditional KNN algorithm considers that the effect of each attribute is the same and ignores the degree of correlation between attribute and its category, and the problem that the strategy of category judgement only considers the number of nearest samples in each category, and ignores the similarity differences between the nearest neighbors and the samples to be classified in different categories, when judging the categories of samples to be classified. KNN classifier is mainly used for text classification, and few researchers used KNN for music classification. However, studies have found that the KNN is more suitable than other methods for tobeclassified samples with more cross or overlap in class field [6]. This article applies DW-KNN classification algorithm to the music genre classification, and experiments prove that DWKNN algorithm has better classification accuracy for music samples with cross-genre.

In summary, the paper's main contributions are:

1. This algorithm addresses two key issues found in the traditional KNN approach. Firstly, it rectifies the oversight in similarity measurement where the traditional method assumes equal impact of each attribute during distance calculation, disregarding the correlation between attributes and their respective categories. Secondly, it improves the category judgment process by not solely focusing on the number of nearest neighbors within each category but also considering the dissimilarities in similarity between these neighbors and the samples undergoing classification.

Identify applicable funding agency here. If none, delete this.

2. The enhanced KNN classification algorithm demonstrates superior classification accuracy in music genre classification compared to the conventional KNN algorithm.

II. Background

A. Traditional KNN Algorithm

KNN algorithm is a lazy learning algorithm, with significant differences from other classification algorithms, which, such as SVM, HMM and others, firstly do machine learning for data in the training set, and build classification model, and then do the classification work supported by the classification model. KNN is a passive classification process and it does the testing while training and while building classification model, which is based on statistical methods that firstly find the number of k nearest samples in the characteristic space for testing samples, and, in the principle of minority obeying the majority, determine the category of test samples according to the categories of majority samples among the nearest k samples [4] . The basic method is as follows:

Assuming that all of the samples are in N-dimension space, each sample x is represented in the form of characteristic vector as $a_1(x), a_2(x), ..., a_r(x)$, and $a_i(x)$ represents the **NO.** i attribute value of the sample x. The similarity between the two samples x_i and x_j is generally calculated through the Euclidean distance between two vectors, as shown in formula 1.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \quad (1)$$

The advantages of KNN algorithm are mainly manifested in its simple principle, convenient implementation, support of incremental learning, ability to build model for ultra-polygon complex decision space, and better classification performance in the situation of cross-class field. But its problems are mainly reflected in two aspects, one is that, when the traditional KNN algorithm measures the similarity, it supposes that the effect of each attribute in the process of distance calculation is the same, and ignores the problem of degree of correlation between attributes and its categories, thus affects the classification accuracy. The second is that, in the process of category judgement, it only considers the number of the nearest neighbors in each category, and ignores the similarity differences between the nearest neighbors and the samples to be classified. These two problems above can be improved by the following methods.

B. Attributes dependency degree

Rough set theory has developed rapidly since proposed in 1982, due to its unique advantages in dealing with large data sample sets, eliminating redundant information and other issues, and for many years has been widely applied in attribute selection, rule learning, classifier design and other fields [5]-[6]

The concept of attribute dependency degree in rough set theory: Let $K = (U, R)$ be knowledge base, and $P, Q \subseteq R$; when $k = \gamma_p(Q) = card\left(pos_p(Q)\right)/card(Q), pos_p(Q) = \bigcup \underline{R}(x), x \in U/ind(P)$, knowledge Q is dependent on P at degree of $k(0 \leq k \leq 1)$, referred to as $P \implies {}_k$Q , where $card(pos_p(Q))$ represents the number of elements that can certainly be classified among Q according to the attributes P, U. When k = 1, Q is fully dependent on P; when $0 \leq k \leq 1$, Q is roughly (partially) dependent on P; when $k = 0$, Q is fully independent of P. Attribute dependency degree can be understood as the ability to classify objects. When $k = 1$, all elements in the domain of discourse can be included by P in the elementary category of U/Q; when $k \neq 1$, only elements in positive region can only be included by P in the category of knowledge Q; when $k = 0$, the no elements in the domain of discourse can be included by P in the elementary category of Q. The attribute dependency degree introduced by this algorithm into rough set theory mainly plays two roles. It uses attribute reduction in rough set theory to select characteristic, and then introduces attribute dependency degrees as weights into the calculation formula of the distance between KNN algorithm samples.

In the calculation of the nearest neighbors, first calculate the attribute dependency of decision attribute on each condition attribute, as shown in Formula 2.

$$k = \gamma_p(Q) = \frac{card(pos(Q))}{card(U)} \quad (2)$$

III. DW-KNN Classification Algorithm

This paper proposes a DW-KNN algorithm, which makes improvements both in distance calculation and category judgement of traditional KNN algorithm. Set unknown sample set as X,$X = (x_1, x_2, ..., x_n)$, and training sample set as Y,$Y = (y_1, y_2, ..., y_n)$.

Introduce attribute dependency degree as the weight of each characteristic into the neighbor distance calculation formula of KNN algorithm, thus, in this way, the distance calculation formula of the distance between two vectors x_i and y_j is transformed from formula 1 to formula 3. The first weighting can effectively solve the problem that the traditional KNN algorithm considers that the effect of each attribute is the same and ignores the degree of correlation between attribute and its category.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} k_r (a_r(x_i) - a_r(x_j))^2} \quad (3)$$

In the process of category judgement, if the numbers of the nearest neighbors found for the unknown sample x_i in several categories are relatively close, then during classification not only the number of the nearest neighbors but the factor of distance between unknown samples and training samples as well should be considered. Set the number of the nearest neighbors as k, and $y_1, y_2, ..., y_n$ as the k nearest neighbors found for the unknown sample x; according to the distance between x and $y_1, y_2, ..., y_n$ in increasing sequence, take turns to assign weights in decreasing sequence for samples involved in category judgement, and calculation of each weight W_j is shown as in formula 4.

$$W_j = \frac{1}{dist(x, y_j)} \tag{4}$$

IV. EXPERIMENTS

A. Categorical dataset

The experiments in this paper utilize a genres dataset comprising 10 categories each of European and American songs. These categories encompass various genres such as blues (1), classical (2), country (3), disco (4), rap (5), jazz (6), heavy metal (7), pop (8), reggae (9), and rock (10). Each category comprises 100 songs, totaling 1000 songs overall. To emphasize the genre-specific attributes of songs while minimizing computational load, only the 30-second chorus segment of each song is captured. The dataset can be downloaded from *https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification/data*.

B. Data preprocessing

- Create a characteristic matrix by capturing the 50-dimensional traits of each of the 1000 songs in the dataset. Condense these characteristics into a matrix format with 1000 rows and 50 columns for further analysis and representation.
- Form the training set and test set: apply the quartered crossover trial method. Divide the data in sample set into four equal parts; take out three parts as the training set and 1 part as the test set; take turns to do loop test; test the algorithm accuracy rate; finally take the average of accuracy rate as the test result;
- Data normalization: due to the inconsistency of data units used to represent each characteristic, the data must be normalized, data in sample set were normalized uniformly into the range of [-1,1].

C. Evaluation criteria and validation methods

1) Evaluation criteria: The evaluation criteria of classification problem prediction takes the prediction accuracy rate to measure. For scientific experiments, the accuracy rate refers to that for several measured values under certain experimental conditions the proportion of the measured values meeting the thresholds, commonly represented as coincidence rate. Namely accuracy rate = the number of eligible measured values / the number of total measured values * 100% . So in the music classification, the accuracy rate = the number on record of the correctly classified / the total number on record in test set * 100% . Accuracy rate is the most widely used evaluation criteria in classification prediction, the greater the value of the accuracy rate, the better classification effect.

2) Validation method: The validation method in use is the cross-validation method; the sample set is randomly divided into k collections; select $k-1$ collections as the training set and the one remaining rest as test set; use data in training set for training to obtain a classification model, and then use that classification model to do testing for data in test set. This procedure was repeated for k times, the average accuracy rate calculated during the k times is the experiment final accuracy rate.

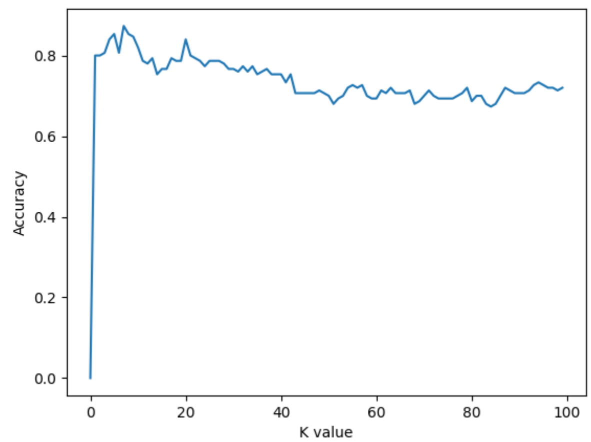

Fig. 1. categories classification

D. Experimental results and analysis

1) *Selection of the value of k*: Through the cross-validation method, judge the value of k in traditional KNN algorithm and double weighted KNN algorithm, and find the value of k with the highest classification accuracy rate; if the value of k is too small, it represents that the number of nearest neighbors is too small, which results in the decrease of classification accuracy rate; if the value of k is too great, it is prone to generate more noise data thus decrease the classification accuracy rate. [7] The experimental results show that however many categories are considered, when the value of k is 10, basically the maximum of accuracy rate is reached, as shown in Figure 1.

2) Classification accuracy rate: Use respectively the traditional KNN, weighted KNN and double weighted KNN algorithm to classify music genres. Classify respectively the data in sample set into 2-10 categories. Calculate the accuracy rate of each algorithm.

- Use the traditional KNN algorithm to classify music genres, classify respectively the data in sample set into 2-10 categories, accuracy rate as shown in the Table 1 below. Wherein, the 2 categories accuracy rate can reach up to 93.6%;the lowest accuracy rate is 70.3%; average accuracy rate is 81.95%. Accuracy rate decreases as the number of categories increases, and, when the number of categories reach 10, the accuracy rate is 34.7%.
- On top of the traditional KNN algorithm, weighting is only done in the calculation of distance, classification accuracy rate as shown in Table 2. Wherein, the 2 categories accuracy rate can be up to 99.6%; the lowest accuracy rate is 80.6%; average accuracy rate is 90.1%. Accuracy

TABLE I
ACCURACY RATE OF TRADITIONAL KNN

Categories	Highest(%)	Lowest(%)	***Average(%)***
2	93.6	70.3	81.95
3	89.5	64.7	77.1
4	81.6	56.3	68.95
5	72.2	48.2	60.2
6	63.4	40.1	51.75
7	52.9	35.4	44.15
8	44.3	34.4	39.35
9	43.6	30.6	37.1
10	40.5	28.9	34.7

rate decreases as the number of categories increases, and, when the number of categories reach 10, the accuracy rate is 34.4%. Table 2 Classification accuracy rate of Weighted (W-KNN) algorithm

TABLE II
ACCURACY RATE OF W-KNN

Categories	Highest(%)	Lowest(%)	***Average(%)***
2	99.6	80.6	90.1
3	90.5	74.5	82.5
4	86.2	69.3	77.75
5	74.2	60.2	67.2
6	65.4	50.1	57.75
7	54.9	45.8	50.35
8	43.3	35.6	39.45
9	45.6	29.8	37.7
10	40.5	28.3	34.4

- On top of the traditional KNN algorithm, weighting is done both in distance calculation and statistical results, classification accuracy rate as shown in Table 3. Wherein, the 2 categories accuracy rate can be up to 100%; the lowest accuracy rate is 85.9%; average accuracy rate is 92.95%. Accuracy rate decreases as the number of categories increases, and, when the number of categories reach 10, the accuracy rate is 49.85%.

TABLE III
ACCURACY RATE OF DW-KNN

Categories	Highest(%)	Lowest(%)	***Average(%)***
2	100	85.9	92.95
3	93.6	78.7	86.15
4	88.5	73.4	80.95
5	80.3	65.9	73.1
6	76.5	57.6	67.05
7	73.6	50.6	62.1
8	70.4	44.8	57.6
9	65.3	43.6	54.45
10	59.8	39.9	49.85

- When traditional KNN, weighted KNN (W-KNN), and double weighted KNN (DW-KNN) algorithms are applied to classify music genres into 2-10 categories within the sample set, the respective classification accuracy rates of each algorithm are illustrated in Figure 2. The analysis reveals that, for categories ranging from 2 to 5, the accuracy rates of DW-KNN and W-KNN exhibit close proximity. However, as the number of categories expands, the accuracy rate of DW-KNN steadily outperforms W-KNN. This observation signifies that as the number of categories increases, DW-KNN consistently demonstrates superior classification accuracy compared to W-KNN, especially in scenarios where no significant difference exists between categories.

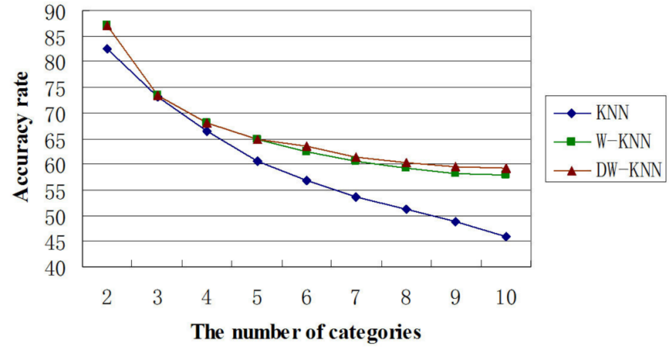

Fig. 2. Comparison of classification accuracy rate of Traditional KNN, WKNN and DW-KNN algorithm

V. CONCLUSION

This research introduces an innovative DW-KNN algorithm designed specifically for automatic music genre classification. Addressing limitations in the traditional KNN method, this algorithm brings advancements in two key areas. Firstly, it rectifies the assumption made by traditional KNN regarding equal attribute impact during similarity measurement, disregarding attribute-category correlation. Secondly, it enhances category judgment by considering not just the count of nearest neighbors within each category but also the dissimilarities in similarity between these neighbors and the samples undergoing classification. The improved DW-KNN classification algorithm demonstrates superior performance in music genre classification compared to the traditional KNN approach, particularly in achieving higher classification accuracy rates.

REFERENCES

[1] W. Shang, H. Huang, Y. Liu, Y. Lin, Y. Qu, and H. Dong, "Research on the algorithm of feature selection based on gini index for text categorization.," *Jisuanji Yanjiu yu Fazhan(Computer Research and Development)*, vol. 43, no. 10, pp. 1688–1694, 2006.

[2] P. Wang, Y. Zhao, and J. Lv, "New method of attribute reduction based on rough set," *Jisuanji Gongcheng yu Yingyong(Computer Engineering and Applications)*, vol. 48, no. 2, 2012.

[3] S.-Q. Wang, D.-F. Zhang, D.-Y. Bi, and L.-D. Zhang, "Two-step attribute reduction method based on fuzzy rough sets dependency," *Beijing Gongye Daxue Xuebao(Journal of Beijing University of Technology)*, vol. 39, no. 6, pp. 828–834, 2013.

[4] D. Mladenić, J. Brank, M. Grobelnik, and N. Milic-Frayling, "Feature selection using linear classifier weights: interaction with classification models," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 234–241, 2004.

[5] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE transactions on systems, man, and cybernetics*, no. 4, pp. 580–585, 1985.

[6] H.-f. LIU, Q. CHEN, S.-s. LIU, and Z. SU, "An improved knn text categorization method based on data uneven," *Microelectronics & Computer*, vol. 27, no. 3, pp. 51–53, 2010.

[7] J. Duan, Q. Hu, L. Zhang, Y. Qian, and D. Li, "Feature selection for multi-label classification based on neighborhood rough sets," *Journal of computer research and development*, vol. 52, no. 1, pp. 56–65, 2015.