2019 6th International Conference on Dependable Systems and Their Applications (DSA)

A Double Weighted KNN Algorithm and Its Application in the Music Genre Classification

Meimei Wu
School of Data Science and Media Intelligence
Communication University of China
Beijing, China
Email: wumm@cuc.edu.cn

Xingli Liu
School of Computer and Information Engineering
Heilongjiang University of Science and Technology
Harbin, China
Email: 570407553@qq.com

***Abstract*—This paper proposes a double weighted KNN algorithm, and applies it to the research of music genre automatic classification. This algorithm makes improvements in two aspects in the traditional KNN algorithm, which can effectively solve the problem that traditional KNN algorithm ignores the degree of correlation between attributes and categories in the classification process, and the problem that it only considers the number of the nearest samples and ignores the existence of similarity differences between the nearest samples and the samples to be classified in the process of category judgment, thus can effectively improve the classification accuracy. In this paper, this algorithm is applied to the music genre classification, and experiments prove that the algorithm can achieve higher classification accuracy in terms of music genre classification, and even has a better classification performance especially where there is no obvious difference between some of the categories and in the situation of cross-category, and this algorithm is simple and symmetrical, with no complex dependency, high calculation efficiency, and adaptable to the demand of mass music data classification.**

***Index Terms*—KNN; Double weighted KNN; Music genre classification; Attributes dependency degree;**

I. Introduction

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]. Mladeni et al. [4] combined classifiers of SVM, machine perception etc. to weight, and when classifiers are distinguishing samples of positive and counter examples, score the documents, and this score is the basis of weighting. Keller et al. [5] put forward a fuzzy KNN method that in the category judgment, gives full play to the effect of nearest neighboring samples and reduces the error due to the uneven distribution of training samples, therefore to improve the classification accuracy, and can to a certain extent solve the issue that traditional KNN algorithm is sensitive to the value of k.

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 to-be-classified 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 DW-KNN algorithm has better classification accuracy for music samples with cross-genre.

978-1-7281-6057-3/19/$31.00 ©2019 IEEE
DOI 10.1109/DSA.2019.00051

II. Basis of relevant theories

A. 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. 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 $\{a_1(x), a_2(x), \ldots, a_r(x)\}$, 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.

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

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 [7] [8] .

Rough set theory proposed by Professor Pawlak of Poland, is a method of portraying uncertain or imprecise unknown knowledge by utilizing the approximate knowledge in knowledge base. Rough set theory is a way to deal with uncertain information, and one of its important applications is to eliminate redundant attribute and redundant data [2].

The concept of attribute dependency degree in rough set theory: Let $K = (U, R)$ be knowledge base, and $P, Q \subseteq R$; when $k = \gamma_p\left(Q\right) = card\left(pos_p\left(Q\right)\right)/card\left(Q\right)$,$pos_p(Q) = \bigcup \underline{R}(x)$, $x \in U/ind(P)$, knowledge Q is dependent on P at degree of $k(0 \leq k \leq 1)$, referred to as $P \Rightarrow_k Q$, 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.

III. Double weighted KNN (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 $X, X = (x_1, x_2, \ldots, x_n)$, and training sample set as $Y, Y = (y_1, y_2, \ldots, y_n)$.

In the calculation of the nearest neighbors, first calculate the attribute dependency of decision attribute on each condition attribute, as shown in Formula 2, and, after removing the characteristic with the attribute dependency degree equal to zero, 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 3 to formula 4. 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.

$$k = \gamma_p(Q) = \frac{card(pos_p(Q))}{card(U)} \tag{2}$$

$$dist(x_i, y_j) = \sqrt{\sum_{r=1}^{n} \left(a_r\left(x_i\right) - a_r\left(y_j\right)\right)^2} \tag{3}$$

$$dist(x_i, y_j) = \sqrt{\sum_{r=1}^{n} k_r \left(a_r\left(x_i\right) - a_r\left(y_j\right)\right)^2} \tag{4}$$

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 $y_1, y_2, \ldots, y_k$ as the k nearest neighbors found for the unknown sample x; according to the distance between x and $y_1, y_2, \ldots, y_k$ 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 5.

$$W_j = \frac{1}{dist(x, y_j)} \tag{5}$$

Finally, according to the weight-sum size of samples involved in category judgement in each category, make the judgement which category does the unknown sample belong to. This method can well solve the problem that the strategy of category judgement of traditional KNN algorithm only considers the number of the nearest neighbors but ignores the similarity differences between the nearest neighbors and the samples to be classified. On top of the idea above, the algorithm procedures of the improved double weighted KNN algorithm (DW-KNN) are as follows:
Step 1: Build characteristic matrix
Step 2: Use rough set theory to calculate the attribute dependency degree $k = \gamma_p(Q)$ of decision attribute on condition attribute, and remove the attribute with $k = 0$, then generate the characteristic vector matrix after characteristic selection.
Step 3: Calculate the distance between the unknown sample $x(i)$ and each training sample y_j is shown as in formula 4.
Step 4: Select in y_j the sample k with the smallest distance from x.
Step 5: In category judgement, firstly calculate the occurrence times of the k nearest samples in each category, if in the category of the biggest occurrence times the number of the nearest neighbors exceeds 60% the value of k,then directly do the category judgement according to the traditional KNN category judgement method; if it does not exceed 60%, then do the category judgement in all categories where k occurs according to the sum of weighted distance of samples. Set k nearest neighbors occur in category C_1,C_2,...,C_p,. The category C_p includes samples y_{p1},y_{p2},...,y_{pj}. The sum of weighted distance is caculated as shown in formula 6 and 7.

$$W_{pj} = \frac{1}{dist(x, y_{pj})} \tag{6}$$

$$C_p = \sum_{j=1}^{n_p} W_{pj} \tag{7}$$

Step6: Select the class label of C_p maximum as the class label of the unknown sample.

IV. The application of DW-KNN in music genre classification

A. Categorical dataset

Experiments in this paper is based on genres dataset; this dataset contains 10 categories of European and American songs, respectively, 1 blues, 2 classic, 3 country, 4 disco, 5 rap, 6 jazz, 7 heavy metal, 8 popular, 9 reggae, 10 rock. Every category contains 100 pieces of songs, 1000 pieces in all. In order to highlight the genre characteristic of song and reduce the amount of calculation, only 30 seconds of chorus in each song is captured.

B. Data preprocessing

1) Generate characteristic matrix: for 1000 songs in the dataset, extract the 59-dimensional characteristic of each song, then obtain a characteristic matrix with 59 columns and 1000 rows.

2) 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;

3) 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]. The normalization uses the normalization functions mapminmax in MATLAB to normalize training set and test set 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.

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. The experimental results show that however many categories are considered, when the value of k is 13, basically the maximum of accuracy rate is reached, as shown in Figure 1 and Figure 2.

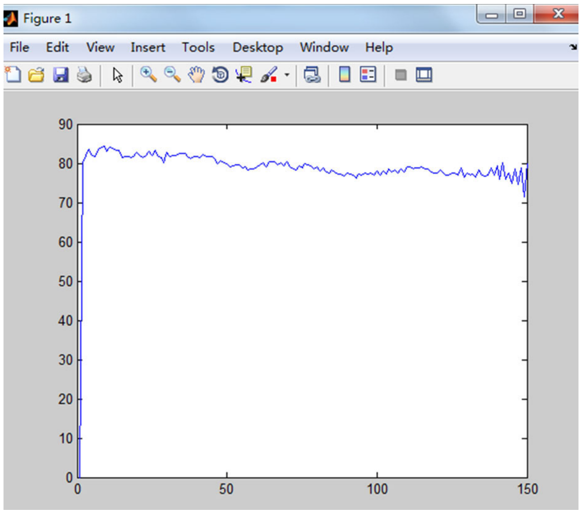

Fig. 1. k value in 2 categories classification

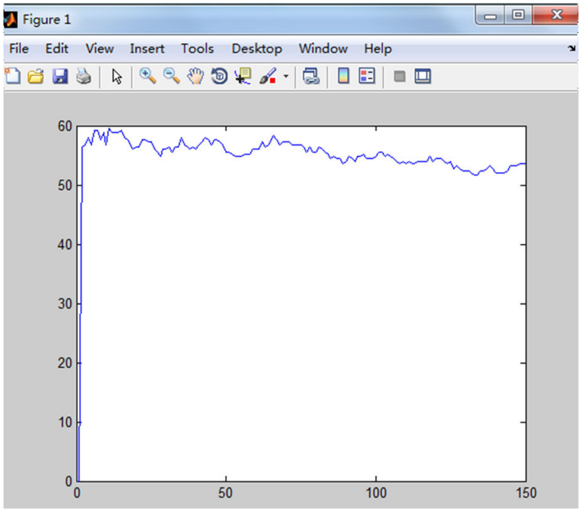

Fig. 2. k value in 10 categories classification

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.

(1) 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 94%;the lowest accuracy rate is 60%; average accuracy rate is 82.49%. Accuracy rate decreases as the number of categories increases, and, when the number of categories reach 10, the accuracy rate is 46%.

Table 1 Classification Accuracy rate of traditional KNN algorithm

Categories	Highest accuracy rate (%)	Lowest accuracy rate (%)	**Average accuracy rate (%)**
2	94	60	82.49
3	89.33	52	73.13
4	89	33	66.41
5	81.6	28.8	60.64
6	76.67	28.67	56.74
7	74.86	28.57	53.67
8	68	32	51.17
9	55.56	37.33	48.71
10	46	46	46

(2) 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 100%; the lowest accuracy rate is 58%; average accuracy rate is 87.02%. Accuracy rate decreases as the number of categories increases, and, when the number of categories reach 10, the accuracy rate is 58%.

Table 2 Classification accuracy rate of Weighted (W-KNN) algorithm

Categories	Highest accuracy rate (%)	Lowest accuracy rate (%)	**Average accuracy rate (%)**
2	100	58	87.02
3	90.67	46.67	73.39
4	83	47	68.06
5	79.2	45.6	64.80
6	77.33	46	62.37
7	73.71	47.43	60.47
8	67	49.5	59.19
9	64	52.44	58.31
10	58	58	58

(3) 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 59%; average accuracy rate is 87.15%. Accuracy rate decreases as the number of categories increases, and, when the number of categories reach 10, the accuracy rate is 59.16%.

Table 3 Classification accuracy rate of double weighted (DW-KNN) algorithm

Categories	Highest accuracy rate (%)	Lowest accuracy rate (%)	**Average accuracy rate (%)**
2	100	59	87.15
3	91.58	47.14	73.37
4	84.66	47.94	68.17
5	80.78	46.51	64.76
6	78.88	46.92	63.62
7	75.18	48.38	61.28
8	68.34	50.49	60.38
9	65.28	53.49	59.48
10	59.16	59.16	59.16

(4) When traditional KNN, weighted KNN (W-KNN) and double weighted KNN (DW-KNN) algorithm classify music genres, they classify data in sample set into 2-10 categories; respective classification accuracy rate of each algorithm is shown as in Figure 3. It can be inferred that, when the number of categories is 2-5, the accuracy rate of DW-KNN and W-KNN is very close, but, as the number of categories increases, compared with W-KNN the accuracy rate of DW-KNN improves gradually, which demonstrates that DW-KNN has better classification accuracy rate in the condition that the number of categories is increasing and no insignificant difference is between categories.

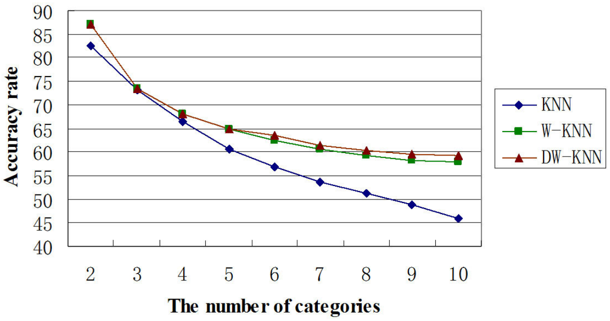

Fig. 3. Comparison of classification accuracy rate of Traditional KNN, W-KNN and DW-KNN algorithm

3) Analysis of time complexity: Use respectively the traditional KNN and DW-KNN algorithm to classify music genres, the calculation efficiency of each algorithm is shown as in Table 4. Because DW-KNN first needs to use ReliefF algorithm to calculate each characteristic weight and then when deciding category needs to calculate distance, thus the calculation time is longer than KNN. Traditional KNN algorithm takes up most of its calculation time calculating the distance between each sample and in this process finding the k nearest neighbors, thus, in the condition that the number of samples is n, this algorithm time complexity is $O(n)$. The time complexity of DW-KNN is divide into two parts, the time complexity of weight calculation and the time complexity of classification. Weight calculation is only calculated once in the whole classification process, and can be finished by offline calculation, thus its time complexity can be omitted. The process of classification compared with traditional KNN only adds weights in the calculation of distance, thus its time complexity is still $O(n)$. The time cost of the calculation of distance between samples to be classified and 1000 pieces of songs in music library is 0.0573 second, and, as the number of songs in music library increases, according to prediction of linear increase, the time cost of calculation of 1 million samples is approximately 53.7 seconds; the time cost of calculation of 10 million samples is approximately 9 minutes. The classification work is finished mainly by offline calculation and calculation time can be accepted by the system.

Table 4 Comparison of time complexity of DW-KNN and traditional KNN algorithm

	DW-KNN	**KNN**
Time Complexity	O(n)	O(n)

E. Comparison of DW-KNN and Support Vector Machines (SVM) classification algorithm

Support Vector Machines, SVM, is a very effective method of dealing with problems of supervised learning in current machine learning field; this method is based on the VC-dimension theory and structural risk minimization principle in statistical learning theory, and has solid basis of mathematics. It has comprehensive application in solving classification problems of all kinds, especially in music classification problems, and can achieve higher accuracy rate in many conditions compared with other classification methods [9] [10] [11]. However, experiments find that the situation of SVM with lower classification accuracy rate happens in where there is no obvious difference between categories and there is cross-class field. This paper used DW-KNN method to classify categories where SVM got lower classification accuracy rate, and it was found that DW-KNN achieves higher classification accuracy rate than SVM in some of the 2 categories and 3 categories, such as: jazz and reggae, reggae and rock, blues and rap, reggae and rock and jazz, as shown in Figure 4. In fact, many music websites indeed have cross in classification of those categories.

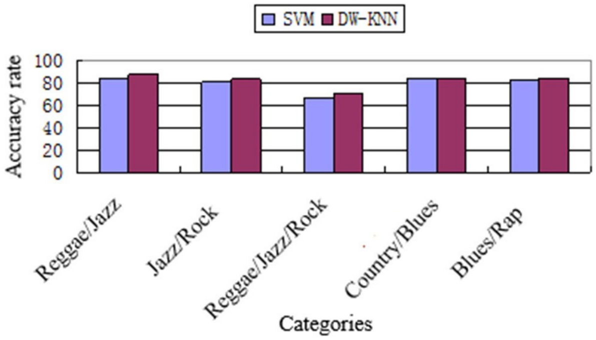

Fig. 4. Comparison of accuracy rate between SVM and DW-KNN

As can be seen from the results, in the condition where there is prone to have cross-class field among several categories, DW-KNN indeed has a little higher classification accuracy rate than SVM.

V. Conclusion

This paper innovatively proposes a DW-KNN algorithm for music genre automatic classification. This algorithm makes improvements in two aspects based on traditional KNN algorithm, and the advantages are manifested as follows:

1. This algorithm makes improvements of two problems in the traditional KNN algorithm. 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. 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.
2. Improved DW-KNN classification algorithm has higher classification accuracy rate in music genres classification than traditional KNN algorithm.
3. This algorithm has better classification accuracy rate in the condition where there is prone to have cross-class field.
4. The calculation of this algorithm is simple and symmetrical, with no complex dependency relation, high calculation efficiency, and adaptable to the demand of mass music data classification.

Acknowledgement

This paper is supported by the subject of Based on Deep Learning Voice Evaluation Model Research (3132018XNG1813) and by the subject of Based on Geographic Themes and User Preferences Personalized Mobile News Recommendation (CUC18QB44).

References

[1] W. Q. Shang, H. K. Huang, Y. L. Liu, Y. M. Lin, Y. L. Qiu, and H. B. Dong, "Research on the algorithm of feature selection based on gini index for text categorization," *Journal of Computer Research and Development*, vol. 43, no. 10, pp. 1688–1694, 2006.

[2] P. J. Wang, Y. L. Zhao, and J. F. Lv, "New method of attribute reduction based on rough set," *Computer Engineering and Applications*, vol. 48, no. 2, pp. 113–115, 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," *Journal of Beijing University of Technology*, vol. 39, no. 6, pp. 828–834, 2013.

[4] D. Mladenic, J. Brank, M. Grobelnik, and N. Milicfrayling, "Feature selection using linear classifier weights: interaction with classification models," pp. 234–241, 2004.

[5] J. M. Keller, M. R. Gray, and J. 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 S. U. Zhan, "An improved knn text categorization method based on data uneven," *Microelectronics and Computer*, vol. 27, no. 3, p. 51, 2010.

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

[8] Q. Hu and D. Yu, *Applied Rough Set*. Science Press, 2012.

[9] K. Aryafar and A. Shokoufandeh, "Multimodal sparsity-eager support vector machines for music classification," pp. 405–408, 2014.

[10] S. H. Chen, S. H. Chen, and R. C. Guido, "Music genre classification algorithm based on dynamic frame analysis and support vector machine," *international symposium on multimedia*, pp. 357–361, 2010.

[11] P. L. Deepa and K. Suresh, "An optimized feature set for music genre classification based on support vector machine," in *Recent Advances in Intelligent Computational Systems*, pp. 610–614, 2011.