Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/54/20241498

Decision tree based prediction system for word difficulty classification

Xinyi Jiang

Liaoning University, No.66 Chongshan Middle Road, Huanggu District, Shenyang, Liaoning Province 110036, China

18068313875@163.com

Abstract. In the era of information technology in education, personalised learning is becoming increasingly important, especially in the English learning process. An important aspect of facilitating personalised learning in English reading is the use of a reliable objective word difficulty classification system to quickly capture and understand reading difficulties and core concepts. The aim of this study was to use decision tree modelling to predict the general proficiency of the public in words with different attributes. In addition, a K-Means clustering algorithm was used to categorise words into five classes based on their level of difficulty. By adopting this approach, the prediction of word difficulty becomes both fast and objective, and by testing the accuracy of the model, it was found that our model achieved an accuracy of 0.95. Accurate classification of word difficulty will play an important role in facilitating personalised learning in English reading and improve the efficiency of English reading.

Keywords: Word Difficulty Classification, Decision Tree Model, K-Means Clustering, Personalized Annotations.

1. Introduction

1.1. Research Background

With the development of globalization and the rapid spread of Internet technology, education informatization has become an inevitable trend [1]. When people no longer worry about obtaining a high-quality learning material, the ensuing problem is how to be able to form a personalized learning report. Especially when learning English, facing the same English reading, different readers have different levels of mastery, professional orientation, learning ability and so on [2], which leads to different difficulties and focuses when reading. Some of the word annotation software available on the market roughly lists a few difficult words based on subjective judgement, which does not allow all readers to read without barriers; while others are too general, translating almost all the words, which does not allow readers to quickly grasp the key points [3]. The most important indicator for objectively analyzing the suitability of a reading material for a reader's individual ability to annotate words is the degree of difficulty of the words in the reading material.

DOI: 10.54254/2755-2721/54/20241498

1.2. Research Status

The study of English word classification originated from semantic research in the field of linguistics. In the early days, researchers mainly classified words based on their lexical meaning and grammatical function. With the development of computer science, methods based on statistics and machine learning have also been applied to word classification research.

In 1972, a study based on the "American Heritage Book of Word Frequency" compared and discussed the utility of various word frequency counts and frequency measures for assigning frequency values to words in dictionaries [4]. The report analysed word frequency data from the English corpus and pointed out that there is a definite link between the frequency of word use and word difficulty(Manelis L, 1972).

Since then, more scholars have classified word difficulty by analysing the lexical properties and roots of words as well as the word frequencies in the above studies.

In 2000, scholars in the book "Vocabulary in Language Teaching-Cambridge University Press", which contains practical information about the classification of word difficulty, analysed word difficulty from the factors of frequency of use, lexical complexity and linguistic knowledge [5], and so on. Factors such as frequency of use, lexical complexity and linguistic knowledge were analysed to determine word difficulty(Schmitt N, 2000).

In 2011, a study examined the effects of word frequency and working memory on word recall during reading across different task types [6]. The study found that the task type effect was superior to the word occurrence effect in recall only, confirming the efficiency of word-centred learning during reading(Laufer B, 2011).

Today, with the support of big data algorithms, a number of researchers are using machine learning techniques for word classification based on corpus analysis.

In 2022, a study was conducted to identify the difficulty of words based on a phonological and LSTM approach [7]. The study trained and tested 1800 words for complexity prediction in terms of both qualitative dimensions: linguistic regularity, clarity and quantitative dimensions: word frequency, word length(Shivam Parihar, 2020).

1.3. Purpose of the study

This study innovatively proposes to use the decision tree algorithm to classify words according to their attribute values, and to classify the classification results of the decision tree algorithm into difficulty levels by means of the K-Means algorithm, which objectively classifies the words of any piece of English reading, which can facilitate different readers to match the difficulty level of the article, achieve the effect of personalised annotation, and improve the efficiency of reading.

2. Methodology

2.1. Data Collection and Analysis

The data for this article comes from wordle, an anagram game presented by the New York Times. The player needs to guess a puzzle consisting of five letters within six chances, and after each guess, the player can learn part of the puzzle based on the change of the color of the ceramic blocks, until success or exhaustion.

Wordle is the first place in Google's 2022 hot search list, at a relatively fast pace popular around the world, this study used 2022 daily data. The names of the seven primitive variables were collected and their definitions were shown in the table below:

Table 1. seven types of raw data.

Variables	Definition of Variable
n try(tries)	Percentage of people who solved the puzzle in n chance(chances)
X tries	Percentage of people with unsolved the puzzle

DOI: 10.54254/2755-2721/54/20241498

In addition, some computational values were needed for this study. The sum of the percentages of the number of people with different number of attempts to solve the puzzle each day is always 100%, which is not a good trade-off to characterize the properties of the crossword puzzle for that day. Therefore, we introduce a weighted average number of attempts (try average), at the time of calculation, approximating X attempts as the percentage of people who had 7 chances to solve the puzzle.

$$try\ average = \sum_{n=1}^{7} n \times n\ tries(try) \tag{1}$$

2.2. Research Steps

Firstly, different attribute values of words were selected as quantitative indexes to classify them using a decision tree model. The classification was based on the quantitative values of the attributes of Wordle daily puzzles, aiming to predict the relationship between the percentage of people attempting each word and their attributes. Next, the sample words were clustered into five classes, ranging from easy to hard, using the K-Means clustering algorithm and try average as a basis. Finally, the try average in each subclass module of the decision tree was calculated to establish a correspondence between the quantitative value of a word's attributes and its difficulty level. This approach allowed an objective description of the difficulty level of a word based on its attributes.

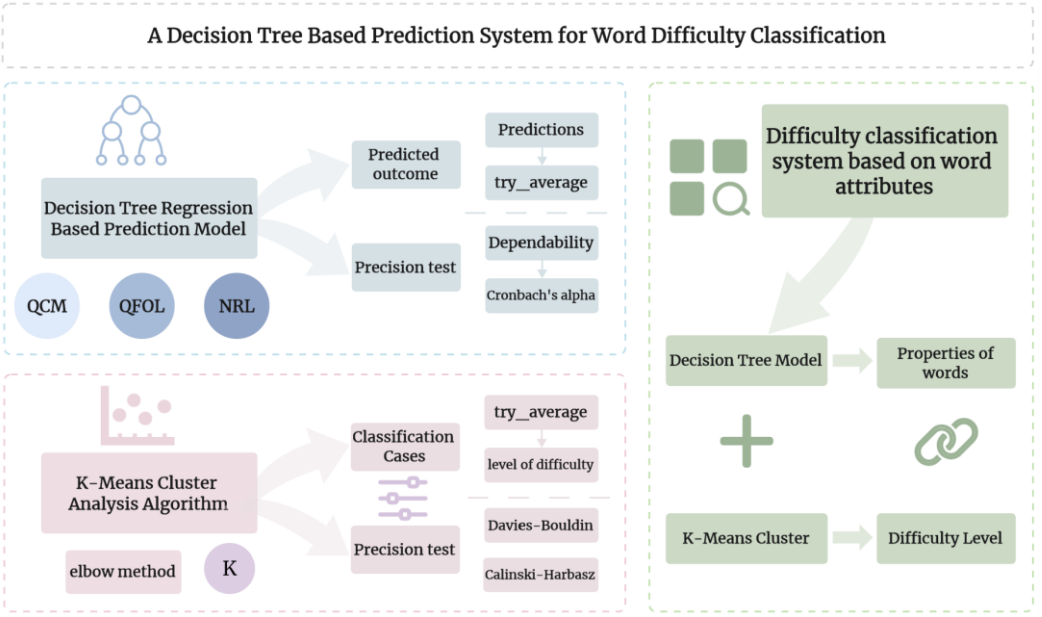

Figure 1. model flow chart.

3. Modeling

3.1. Quantitative Indicators

In order to reflect the characteristics of the words themselves, this study defines three attribute values of the words and quantifies them.

Quantitative commonness score (QCM): Quantifies how common the word is according to Google Ngram Viewer.

Quantitative score for frequency of occurrence of letters (QFOL): A quantitative sum of the frequency of individual letter occurrences in the word [8].

Number of Repeated Letters (NRL): The number of repeated letters in each word.

DOI: 10.54254/2755-2721/54/20241498

3.2. Decision Tree Model

The decision tree model is a supervised learning algorithm and an important classification and regression method in data mining techniques. It is a predictive analytics model represented as a tree structure (binary and multinomial trees). Each internal node represents an attribute test, each branch represents a test output, and each leaf node represents a category. Since each node has both "yes" and "no" judgements [9], the boundaries of the division are parallel to the coordinate axes.

The feature selection of CART algorithm for continuous type metrics is implemented based on mean square error (MSE) [10]. In this study, the quantitative value of the attributes of the word was used as the basis for each node division and selected the optimal division value. The formula for calculating MSE is as follows:

$$MSE(D, A = a_i) = \sum_{x_i \in D_1} (y_i - c_1)^2 + \sum_{x_i \in D_2} (y_i - c_2)^2 \tag{2}$$

$$c_1 = \frac{1}{N_1} \sum_{x_i \in D_1} (y_i)^2 , c_2 = \frac{1}{N_2} \sum_{x_i \in D_2} (y_i)^2 \tag{3}$$

Where c1 denotes the predicted value of sample subset D1 and c2 denotes the predicted value of sample subset D2. N1,N2is the number of samples in region D1,D2.

For all values of all features, the feature with the smallest squared error is chosen as the cut-off point:

$$\min_{A,a} \left[\min_{c_1} \sum_{x_i \in D_1} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in D_2} (y_i - c_2)^2 \right] \tag{4}$$

After selecting the optimal features, the percentage of the number of people in each number of attempts is divided to get the sub dataset. This study selected 70% of the samples as the training set, and 30% of the samples as the test set to present the results of prediction. In order to prevent overfitting, the decision tree model cut off some leaf nodes by pruning to improve the judgement efficiency and got the optimal subtree.

3.3. K-Means Clustering Algorithm

K-Means clustering algorithm is an unsupervised classification learning algorithm. In this study, the similarity between data objects was measured using the Euclidean distance, with smaller distances indicating higher similarity.

Chose the elbow method to determine the number of clustering centers by calculating the degree of distortion of the categories. Randomly selected k initial clustering centers, calculated the distance of each sample data to the k clustering centers and assigned it to the nearest clustering center[11].Repeat the calculation of new clustering centres until the clustering centre no longer changes or has reached the maximum number of iterations. The Euclidean distance formula was used to calculate the distance between data points and clustering centers [12]:

$$d(t, C_i) = \sqrt{\sum_{j=1}^{m} (t_j - {C_{ij}}^2)} \tag{5}$$

Where t represents the data object, Ci represents the i-th clustering center, m represents the dimension of the data object, and tj and Cij represent the j-th attribute values of t and Ci.

In this study, the clustering error of all sample data was evaluated using the sum of squared errors (SSE), which represents the overall clustering effectiveness. The SSE for the entire dataset was calculated as:

DOI: 10.54254/2755-2721/54/20241498

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} |d(x, C_i)|^2 \tag{6}$$

4. Results

4.1. Prediction of Results Based on Decision Tree

The results of using decision trees to predict the percentage of people per number of attempts for words with different attributes are shown in the figure below:

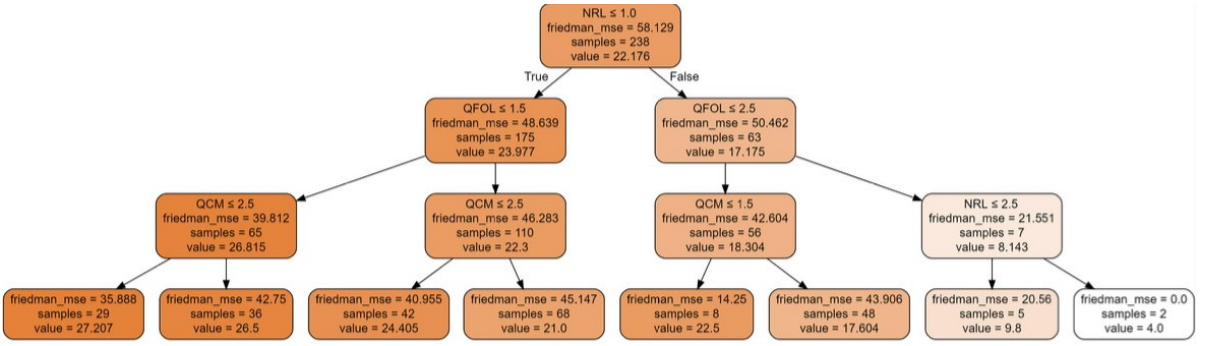

Figure 2. Decision tree predictions for 3 tries.

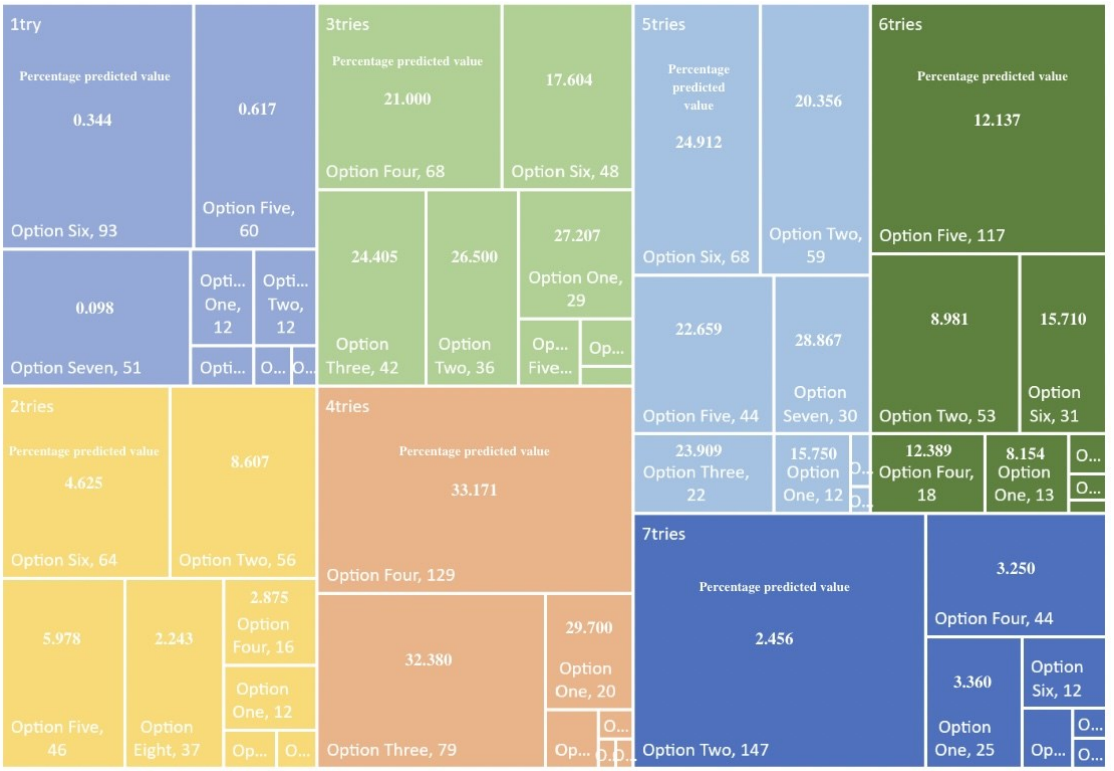

Figure 3. Rectangular dendrogram of decision tree.

The visualization charts reflect more intuitively the classification of words based on the quantitative values of their attributes as features and the prediction of their corresponding percentage of people per number of attempts, thus calculating the average number of attempts for any word.

DOI: 10.54254/2755-2721/54/20241498

Use Cronbach's alpha coefficient [13] for credibility analysis, calculated using the following formula:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum S_i^2}{S_x^2}\right) \tag{7}$$

Generally 0.6 or above is considered fair, 0.7 or above is good, 0.8 or above is very good, and 0.9 or above is ideal.

Table 2. Cronbach's alpha coefficient for reliability statistics.

Cronbach's alpha coefficient	Based on normalized terms	Number of terms
0.942	0.953	8

The value of Cronbach's coefficient α is 0.942, which indicates that the decision tree model is highly reliable.

4.2. K-Means based Word Classification

The clustering identified 5 centers by elbow method and the results of clustering words using K-Means clustering algorithm are shown below:

Table 3. K-Means clustering features table.

K-Means	Clustering categories (mean ± standard deviation)					F	P
	Easy	Fair	Moderate	Difficult	Extremely difficult		
Average	3.62±0.16	3.98±0.19	4.29±0.11	4.57±0.16	4.97±0.23	556	0.000

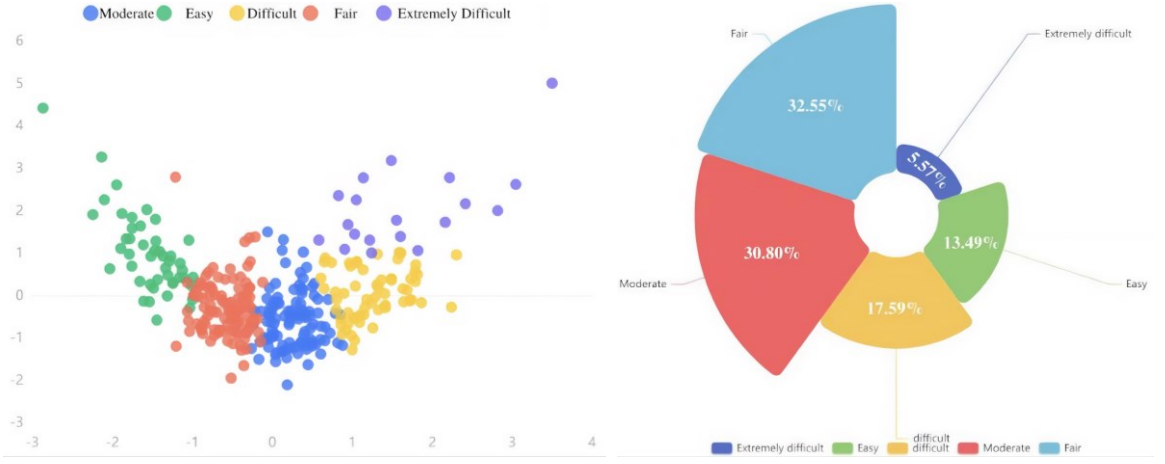

Figure 4. Scatterplots and Rosettes for K-Means Clustering.

In total, the words were classified into 5 categories according to their difficulty level, which are easy, fair, medium, difficult, difficult and extremely difficult.

In this study, two metrics, Davies-Bouldin (DBI) and Calinski-Harbasz Score (CH), were used to test the accuracy of the K-Means clustering model. The formula is as follows:

$$DBI = \frac{1}{N}\sum_{i=1}^{N} \max_{j \neq i} \frac{\overline{S_i} + \overline{S_j}}{||w_i - w_j||_2}, \qquad CH = \frac{SS_B}{K-1} \Big/ \frac{SS_w}{N-K} \tag{8}$$

DOI: 10.54254/2755-2721/54/20241498

Table 4. Evaluation indicators for clustering effects.

DBI	CH
0.983	265.132

The low DBI and high CH values suggest that the K-Means clustering model performs well.

4.3. Difficulty Level Classification of Words

The average number of attempts of any word calculated according to the decision tree was matched with the results of K-Means clustering to get the results of difficulty level classification based on the attributes of the word as shown in the figure below:

The Rising Sun diagram has three layers [14]. The first layer shows different color blocks that represent the difficulty level. The second layer depicts the decision tree classification cells, with a total of 17 cells divided as follows: 1 cell under the easy category, 5 cells under the average category, 4 cells under the medium category, 5 cells under the difficult category, and 2 cells under the very difficult category. These cells display predicted values for the average number of attempts, which serve as the main criteria for categorization. The third layer is a subset of the second layer and displays metric values for word attributes in each cell, such as QCM, QFOL, and NRL.

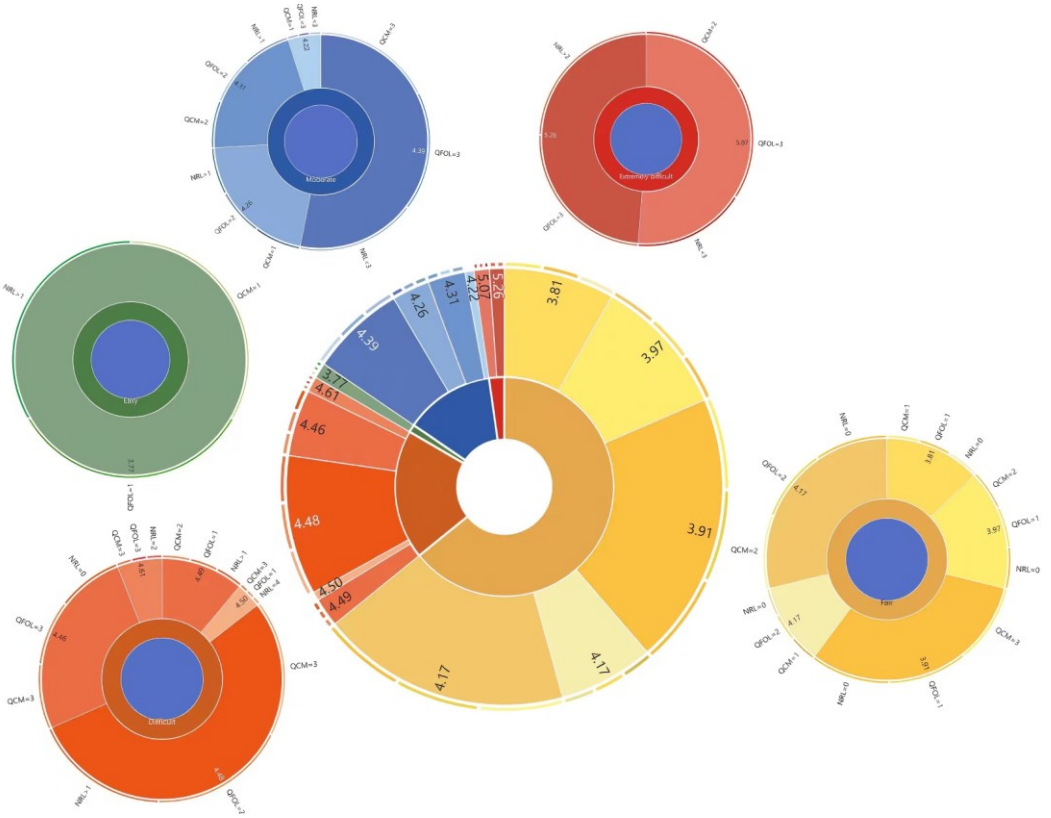

Figure 5. Difficulty Category Rising Sun Chart.

A more intuitive description of condition correspondence and word difficulty can also be based on a visual diagram, as shown below:

DOI: 10.54254/2755-2721/54/20241498

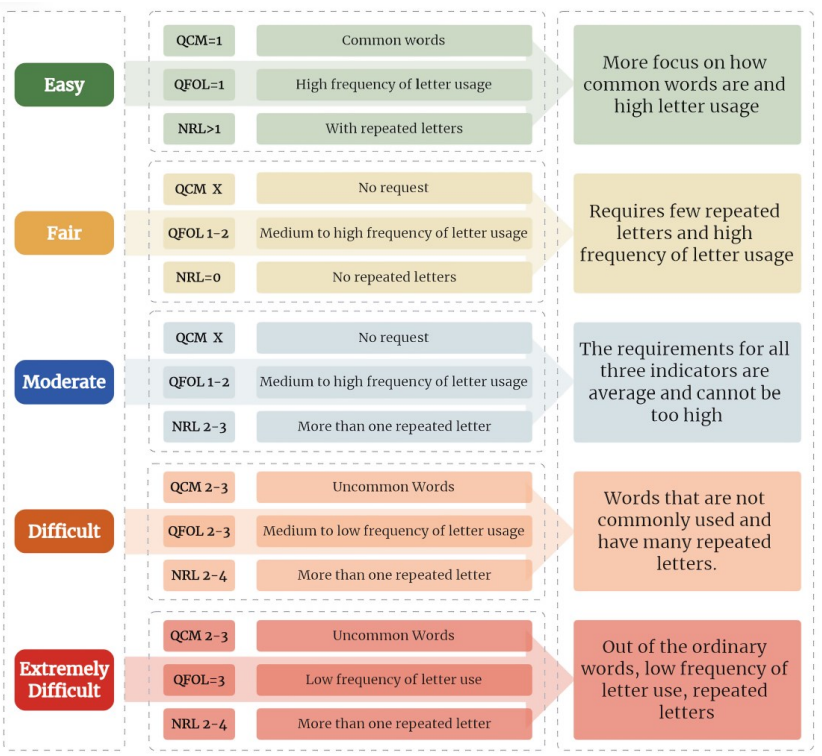

Figure 6. Relationship between attributes and clusters.

5. Conclusion

Through the fusion of the decision tree model and the K-Means clustering algorithm, this study have developed an objective English word difficulty classification system. Then tested this model extensively and it has demonstrated a high degree of credibility. With this classification system, English words in a reading can be quickly categorized based on their difficulty. Subsequent research can explore the integration of an intelligent word extraction system that can extract words with varying levels of difficulty in advance. This way, the system can better discern the readers' abilities, professional orientations, etc., thereby creating a personalized English word annotation system. In turn, this will facilitate more efficient and targeted reading and learning experiences.

References

[1] Ma L, Li J 2022 Influence of Educational Informatization Based on Machine Learning on Teaching Mode. *J.International Transactions on Electrical Energy Systems.* **2022**:7 page

[2] Liu X H, Dong M X 2023 Exploring the relative contributions of learning motivations and test perceptions to autonomous English as a foreign language learning and achievement. *J.Frontiers in Psychology.* **14**:1059375-1059375

[3] Tam Y M, Wu K 2014 A lexical annotation algorithm for English articles. Journal of Beijing University of Posts and Telecommunications. 37:120-124

[4] Manelis L 1972 "the american heritage word frequency book" and its relation to the communication skills lexicon. technical note no. 2-72-38. *J.Communication Skills.* **22**:no.2-72-38

[5] Schmitt N 2000 *Vocabulary in Language Teaching*. Cambridge University Press, Cambridge.

[6] Laufer B, Rozovski-Roitblat B 2011 Incidental vocabulary acquisition: the effects of task type, word occurrence and their combination. *J. Language Teaching Research.* **15**: 391-411

[7] Shivam P, Shubhi M, Abhinav P and Ruchi P 2020 Word Difficulty Level Predction System Using Deep Learning Approach. *J. Ethics And Information Technology.* **2**:109-112

DOI: 10.54254/2755-2721/54/20241498

[8] Wu J B, Wu S, Wu X J 2016 Alphabet frequency-based single table replacement password deciphering algorithm. *J.Computer and Digital Engineering.* **44**:583-585+634

[9] Dong H Z, Xu H P, Lu B & Yang Q 2019 A CART regression tree prediction study of NOx concentration on urban traffic roads. *J.Journal of Environmental Science.* **39**:1086-1094

[10] Alghamdi A 2022 A Hybrid Method for Big Data Analysis Using Fuzzy Clustering, Feature Selection and Adaptive Neuro-Fuzzy Inferences System Techniques: Case of Mecca and Medina Hotels in Saudi Arabia. *J.Arabian Journal for Science and Engineering*. **48**:1693-1714

[11] Chen X M, Su H 2023 Cluster analysis of sea state based on kernel KMeans and SOM neural network algorithm. *J.Journal of Shaanxi University of Science and Technology.* **45**:208-214

[12] Dong H R, Fu Y J, Zhang S A, Yu Y J, Chen J and Xi D H 2023 An adaptive oversampling method based on Euclidean distance clustering. *J.Research on Printing and Digital Media Technology.* **2023**:26-41

[13] Karl S 2011 On the changing role of Cronbach's α in the evaluation of the quality of a measure. *J.European Journal of Psychological Assessment.* **27**:143-144

[14] Yi X Q, Li T R and Chen C 2019 Sunburst graph visualisation for review text data. *J.Computer Scienc.* **46**:14-18