Remittances Review
January, 2023
Volume: 8, No: 1, pp. 509-524
ISSN: 2059-6588 (Print) | ISSN 2059-6596 (Online)
remittancesreview.com

Received: 11 November 2022 Accepted: 15 March, 2023
DOI: https://doi.org/10.33182/rr.v8i1.32

A study of cross-linguistic sequential type features based on the decision tree model

Zijun Wang*[1]

Abstract

At present, with the development of information technology and computer science, decision trees, random forest quantitative research methods and multidimensional research perspectives all play an increasingly important role in linguistic typology research. In this paper, we propose four methods to calculate the classification weights of decision trees, including OBB evaluation, sample data correlation coefficient evaluation, chi-square evaluation and mutual information evaluation, through which the computational classification of a single decision tree is achieved. The final results are then obtained by involving all decision trees in the classification, which effectively avoids the problem of overfitting and the relative independence of constructing decision trees is suitable for parallel computation to improve the classification efficiency of the model. Based on the decision tree model, the cross-linguistic sequential type features within the Indo-European language family are classified. The results show that the common correlation coefficient of the decision tree model is 0.85, and the dominant sequential information of the random forest model based on the weighted decision tree is exactly the same as that of the dominant sequences in WALS, with an accuracy rate of 100%, and can distinguish the languages of each language family within the Indo-European family well. This study is well applied to the study of sequential typology and can accurately capture cross-linguistic sequential features.

Keywords: *decision tree model, OBB evaluation, common correlation coefficient, quantifiers of language order, cross-lingual, classification efficiency*

Introduction

The development of Internet technology has changed the way humans communicate, and with the successful application of monolingual word vectors in many NLP tasks, the potential of word vectors in cross-lingual natural language processing has attracted a lot of attention (Bouraoui, Jamoussi, & Hamadou, 2022; Davis & Aid, 2022). The ultimate goal of cross-lingual sentiment classification is to classify target language text datasets using sentiment classifiers trained on English text datasets and sentiment labels (He et al., 2008; Qiang et al., 2014). The feature migration method based on spatial mapping is currently the mainstream approach to solve the cross-lingual sentiment classification problem, which trains word vectors on two monolingual

[1] Art Management, Department of Performance, Film, Animation, Sejong University, Seoul, 05006, South Korea
Corresponding author: Zijun Wang (Clivia996@163.com)

datasets separately and uses word alignment to map monolingual word vectors to a shared space, thereby generating bilingual text vectors for sentiment classification, and thus requires high structural similarity between the two monolingual vector spaces (Dassa, 2012; Li, Cao, & Min, 2018). However, the performance of the same cross-lingual sentiment classifier in different target languages varies greatly due to the structural and syntactic differences between languages (Basu, 2018; Chen, 2022). And as one of the text pairs with the greatest feature variability, the shared space learned using only word alignment does not represent the bilingual text vector well, and whether it can accurately reflect the approximation of bilingual text will affect the accuracy of cross-lingual sentiment classification (Barbosa et al., 2020; Wang P 2017).

The literature (A, 2016) explores the multilingual motivational components of Chinese university students' choice to learn six non-English second and third public foreign languages and makes cross-linguistic comparisons. Literature ("Research on Cross-cultural Communication and Language Awareness," 2013) based on artificial intelligence for multimodal cultural communication and language communication with more intimacy and storytelling, aiming to build a bridge for culture in cross-language communication and pave the way for language in cross-cultural communication, so that language and culture can work together to build a civilizational chain of international communication. The literature (Wu, Wang, & Wang, 2020) proposes a cross-language retrieval method based on multi-task learning, using a text classification task as a secondary task, using a shared text feature extraction layer to capture feature information of 2 tasks simultaneously so that it learns the feature patterns of different tasks, and then inputting the feature vectors into a neural retrieval model and a text classification model to complete the 2 tasks respectively. The literature (Zou, Wang, & Zuo, 2010) constructs a cross-language information retrieval model based on multilingual ontology, which helps users to access information resources in different languages using their familiar languages through this model. The literature (Jia et al., 2019) improves the accuracy of cross-lingual sentiment classification by narrowing the distribution of bilingual text pairs in the shared space.

In this paper, we address algorithmic shortcomings such as overfitting easily in most machine learning algorithms through a random forest model, using two random processes to construct individual decision trees by sampling the training data set and the feature set respectively, which makes a difference in the classification ability of the constructed obtained decision trees. The shortcomings of decision trees in the classification process are investigated, and the classification process of the original decision tree model is studied, and a random forest model with weighted decision trees is proposed. Examining the problem of sequence typology from a discrete perspective is important for our intuitive understanding of the commonality and types of sequence typology, but it also leads to inaccurate descriptions of the unique features of each language and ignores the probabilistic nature of sequence features. Examining the quantitative features of sequences through a continuous perspective can provide new possibilities for sequential typology. Therefore, using both discrete and continuous perspectives can capture the features of each level of the sequence more comprehensively and accurately, and reveal linguistic commonalities and differences.

Random forest based on weighted decision tree

Decision tree model

Figure 1 shows the diagram of the binary decision tree algorithm model. There are many algorithms for building decision trees, but basically they all use a top-down greedy algorithm, which eventually forms a tree model containing multiple child nodes with two types of child nodes: non-leaf nodes and leaf nodes.

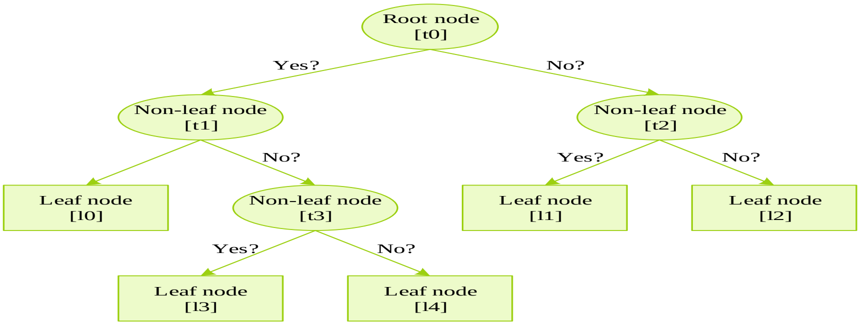

Figure 1: Two-class decision tree algorithm model diagram

Random Forest Model

Figure 2 shows the flow chart of the random forest model construction. In order to avoid overfitting, the decision tree needs to be pruned, and excessive pruning will reduce the prediction ability of the decision tree.

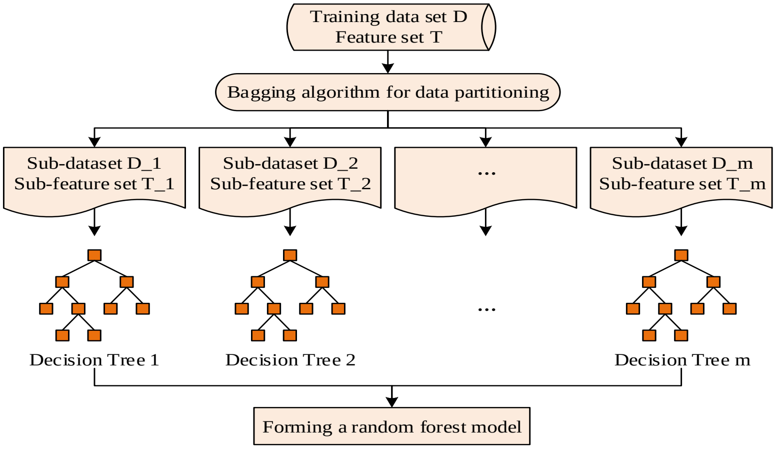

Figure 2: Flow chart of random forest model construction

Meta-classifier evaluation basis

Figure 3 shows the meta-classifier evaluation process. In conducting the assessment of meta-classifier predictive ability, a variety of assessment bases are used, including out-of-bag prediction accuracy, correlation coefficient, cardinality, and mutual information, and these assessment criteria can be broadly classified into two types.

(1) Using out-of-bag data to assess the predictive ability of decision trees, this assessment method is for the predictive ability of decision trees as a whole, without caring about the process of decision tree construction.

(2) The assessment of the predictive ability of the decision tree is based on the sample features. We know that the original classifier of the random forest model is the CART decision tree, and the node selection criterion of the CART decision tree is the Gini index, and in each creation of non-leaf nodes, the feature with the smallest Gini index before and after the classification of the data set is selected as the node feature attribute to create a new node. feature can obtain the largest amount of information when the data is classified, indicating that the feature contributes the most to the model prediction. Based on this reason, we first evaluate the importance of each feature using the training sample data, and then superimpose the feature importance on the decision trees containing different features, and use the superimposed structure as the weight of a single decision tree, while the analysis of feature importance can be performed using coefficients, cardinality, and mutual information.

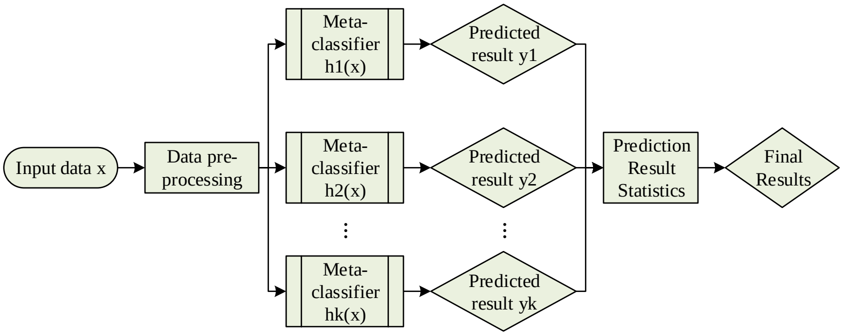

Figure 3: Meta-classifier evaluation process

OOB estimation

Figure 4 shows the OOB estimation process. In the composition of the random forest model construction, the decision tree construction and decision tree OOB estimation can be done serially, and each decision tree gets a corresponding OBB evaluation value at the same time the random forest model is finished construction, so as to assign the weights of the corresponding decision trees, for decision tree $h(x)$, define its weighted value as $Poob$, denoted as:

$$P_{OOB} = \alpha \times \frac{S_+}{S} \qquad (1)$$

Where S_+ is the number of samples correctly predicted by the decision tree using OOB data for data prediction, S denotes the total number of samples involved in the decision tree for OOB evaluation, and α is the tuning factor.

The forest model $\{h_1(x), h_2(x), \cdots, h_k(x)\}$ is trained and the OOB prediction accuracy $\{p_1, p_2, \ldots p_k\}$ of each decision tree is obtained at the same time, then the prediction result of the final model can be expressed as:

$$\max\left\{c \,|\, c_i = \sum_{k}^{j=1} P_{OOB\_j} I\left(h_j(x) = I_i\right), I_j \in C, j = 1, 2, 3\ldots, k\right\} \qquad (2)$$

C is the set of all category labels, I_i is the i th category label in the set C, $I\left(h(x) = I_i\right)$ is the indicative function, when the prediction result of decision tree $h(x)$ is category label I_i, the indicative function is equal to 1, otherwise it is equal to 0, p_i is the weighted value of the j th decision tree in the training process, c_i is the weighted result of the i th category label, and the final prediction result of the model is which category label is the largest among the total weighting obtained by each category label.

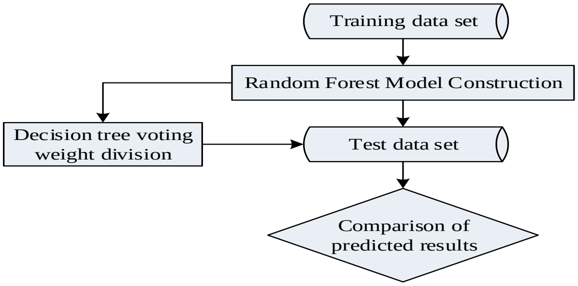
Training data set
Random Forest Model Construction
Decision tree voting weight division
Test data set
Comparison of predicted results

Figure 4: OOB estimation process

Correlation coefficient

When calculating the correlation coefficient between two variables, the difference between the individual variables and their respective means is first calculated, using the product of the two differences as the basis for the calculation of the correlation coefficient. Two variables, x and y, have statistical data sets $x\{x_1, x_2, \cdots, x_k\}$ and $y\{y_1, y_2, \cdots, y_k\}$, respectively, then the correlation coefficient r of variables x and y can be expressed as:

$$r = \frac{\sum_{i=1}^{k}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{k}\left(y_i - \bar{y}\right)^2}} \qquad (3)$$

$\bar{x}$ and $\bar{y}$ in the above equation are the means of the statistics for variables x and y, respectively.

For the set of n-dimensional features $\{t_1, t_2, \ldots, t_a\}$, use the training sample dataset to calculate the corresponding set of correlation coefficients $\{r_1, r_2, \ldots r_n\}$, use this feature set and this training dataset to construct the random forest model $\{h_1(x), h_2(x), \ldots h_k(x)\}$, and for the decision tree $h(x)$ define the weighted values:

$$p_r = \alpha \sum_{j=1}^{m} \left|r_j\right| \qquad (4)$$

The above equation sums the absolute values of the correlation coefficients of all the features used to construct the decision tree $h(x)$, where α is the conditioning factor and m is the number of features used to construct the decision tree, because the correlation coefficients themselves are relatively small, the weighted values of the decision tree are adjusted appropriately by adding the conditioning factor.

The random forest model $\{h_1(x), h_2(x), \ldots, h_k(x)\}$ has weighted weights $\{p_1, p_2, \ldots, p_k\}$, and for any input vector x, the structure of the model prediction is represented as:

$$\max\left\{c \mid c_l = \sum_{k}^{j=1} p_{r\_i} I\left(h_j(x) = I\right), I \in C\right\} \qquad (5)$$

Where C is the set of all classification labels, p_{r_i} is the weight of the i rd decision tree classifier, $I(*)$ is the schematic function, the number of weighting of each classification label is counted in the above equation, and the classification label with the largest number of weighting is used as the final output of the prediction structure.

Cardinality

The main function of the chi-square is to compare the correlation between two or more sample sizes, giving the correlation between the variables by comparing the agreement between the theoretical and actual frequencies of the sample to be analyzed.

The calculation of the chi-square statistic can be explained by the following formula:

$$\chi^2 = \sum \frac{(A-T)^2}{T} \qquad (6)$$

In the above expression, A denotes the actual distribution frequency of the variable and T denotes the theoretical distribution frequency of the variable. The training sample has n as feature $\{t_1, t_2, \mathrm{K}\, t_n\}$, and the cardinality verification is first performed on the n-dimensional features, and the cardinality value of each feature is calculated $\{\chi_1^2, \chi_2^2, \mathrm{K}\,, \chi_n^2\}$. The random forest model constructed using this training sample set can be expressed as $\{h_1(x), h_2(x), \mathrm{K}\, h_k(x)\}$, and the classification weights of the decision tree $h(x)$ can be defined as:

$$p_c = \alpha \sum_{j=1}^{m} \chi_j^2 \qquad (7)$$

The above equation is used as the classification weight of the decision tree by summing the cardinality values of all the features involved in the construction of the decision tree $h(x)$, χ_j is the cardinality value of the features used in the construction of the decision tree and the j feature, and α is the tuning factor of the classification parameters. The classification results of the data x are predicted using the constructed random forest model $\{h_1(x), h_2(x), \mathrm{K}\,, h_k(x)\}$ as:

$$\max\left\{ c \mid c_i = \sum_{k}^{j=1} p_{c\_i} I\left(h_j(x) = l\right), l \in C \right\} \qquad (8)$$

Where C is the set of all classification labels, $I(*)$ is the indicative function, $I\left(h(x)=1\right)$ takes the value of 1 when the prediction result of the decision tree is I, and the opposite takes the value of 0. The above expression counts the weighted results of each classification label during the training process, and the classification label with the most weighted votes is the final prediction result output.

Mutual Information

Mutual information is a measure used in information theory to indicate the strength of correlation between two variables. Mutual information reflects the degree of reduction in uncertainty of a random variable y after the appearance of a random variable x, or the increase in the amount of information brought about by the appearance of a random variable x. The minimum value of mutual information is 0, which means that the appearance of random variable x does not bring any information to variable y, and there is no relationship between random variable x and random variable y, and there is no mutual influence between the two variables. The maximum value of mutual information is the entropy of random variable y, which means that the appearance of random variable x can completely eliminate the uncertainty of variable y .

Two random variables x and y are known, then the mutual information of these two variables is defined as:

$$I(x,y)=\int_x\int_y P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \qquad (9)$$

where $P(x,y)$ denotes the joint distribution of variable x and variable y, and $P(x)$ and $P(y)$ are the marginal distributions of variable x and variable y, respectively. The uncertainty of variable y is described by the entropy of variable y, and the entropy $H(y)$ of variable y can be described as:

$$H(y)=-\int_y P(y)\log P(y) \qquad (10)$$

There is an integral theorem that rewrites the mutual information $I(x,y)$ as:

$$I(x,y)=H(y)-H(y|x) \qquad (11)$$

In the above expression, $H(y)$ represents the entropy of variable y, and the higher the dispersion of the distribution of variable y, the higher the entropy $H(y)$ of y. $H(x|y)$ represents the magnitude of the entropy of y in the case of known variable x. This is known from the above expression for the mutual information of random variable x and random variable y.

The mutual information reflects the strength of the correlation between two variables and can be used as a basis for determining the correlation assessment between a feature and a classification label. In practice, there is a training sample set $\{x_1, x_2, \mathrm{K}, x_n \mid y\}$ calculated to get the mutual information as $\{I \mid I = I(x_i, y), i = 1, 2, \mathrm{K}, n\}$, and using this training sample for the construction of the random forest model to get the classifier as $\{h_1(x), h_2(x), \mathrm{K}, h_k(x)\}$, for the decision tree $h(x)$ calculated to get the weights:

$$p_I = \alpha \sum_{j=1}^{m} I(x_i, y) \qquad (12)$$

In the above expression, the mutual information of all the features involved in the construction of the decision tree h(x) is summed by means of summation as a method of calculating the voting weight of the decision tree, I(xj,y) is the mutual information of the *j* rd feature in the feature set of this decision tree construction, and α is the tuning factor.

$$\max\left\{ c \mid c_i = \sum_{k}^{j=1} p_{I\_i} I\left(h_j(x) = l\right), l \in C \right\} \qquad (13)$$

Where *C* is the set of all classification labels and *P*$_I$ is the mutual information classification weight of a single decision tree, the above expression counts the weighted training volume obtained from all classification labels in this data prediction.

Typological classification of cross-linguistic sequences within the same language family

Cross-linguistic classification of quantitative features of language order in a discrete perspective

Cross-linguistic classification based on subject-predicate dichotomous order

Table 1 shows the results of the dominant order of subject and predicate for the 11 Indo-European languages according to the WALS criterion and their dominant order results among the WALS. From the matching of the dominant order extracted from the decision tree model with the dominant order of each language in WALS, according to the proportion of SV order relations between the subject and the predicate in the treebank, we can find that the proportion of SV order relations of the 11 Indo-European interlingual languages varies and is distributed in the interval [79.28%, 98.94%], while the proportion of their corresponding VS order ranges from [1.06%, 20.72%]. Even the language with the lowest proportion of SV order (Czech) has a proportion of SV order (79.28%) that is much more than twice the proportion of its VS order

(20.72%). In other words, these 11 languages all have SV order ratios well above 66.67%. Therefore, all 11 Indo-European languages can be identified as SV-dominant languages based on the cross-linguistic parallel treebank. This result is highly consistent with the dominance order information of these languages in WALS, where all 8 languages corresponding to the UD treebank are SV dominant languages. The remaining three languages, namely Italian, Spanish and Polish, are lacking dominance order in WALS.

Table 1: Subject-predicate word order relation based on UD data set in cross-language languages

language	SV word order proportion in tree base (%)	UD preponderant word order	WALS Advantage Word Order	family of languages
Hindi	98.93	SV	SV	Indo-Iranian group
English	96.13	SV	SV	Germanic language family
French	97.56	SV	SV	Roman language family
Portuguese	95.49	SV	SV	Roman language family
Italian	90.84	SV	SV	Roman language family
Spanish	90.23	SV	Lack of advantageous word order	Roman language family
Russian	84.96	SV	Lack of advantageous word order	Slavic language family
Swedish	82.87	SV	SV	Germanic language family
German	81.58	SV	SV	Germanic language family
Polish	79.84	SV	Lack of advantageous word order	Slavic language family
Czech	79.27	SV	SV	Slavic language family

Although the cross-linguistic scale and the corpora used are uneven, the results of the decision-tree-based dominant order for Italian and Spanish are consistent with the results of this study. These results show that the high proportion of SV order in the real corpus of these three languages reflects the real use of subject and predicate order. of the dominant inflectional information is generally consistent (8/11 = 72.73%). In addition, the chi-square test results (p=0.22>0.05) based on 9999 replications of the approximate substitution test indicated that the null hypothesis of no significant difference between the two could not be rejected. Thus, the above results indicate that the results of dominant order categorization based on the order relations of subjects and predicates in the random forest model of weighted decision trees under a discrete perspective are highly similar to the information provided in traditional typological databases.

Cross-linguistic classification based on predicate-object binary order

Next, we pay attention to the discrete perspective on the order relations between predicate and object. 11 cross-linguistic languages based on the dominant order of UD and WALS, as shown in Table 2. In this study, we focus on the relationship between predicate and object order in a discrete perspective. 11 cross-linguistic languages have a dominant order situation based on UD and WALS, for example, the proportion of OV order in Hindi (99.31%) is much higher than twice the proportion of VO order (0.69%), so it is classified as an OV dominant language. German, on the other hand, has a higher percentage of OV order (58.88%) than VO order (41.12%), but not twice as much, so it is a language lacking dominant order. The remaining nine languages all have more than twice the proportion of VO order than OV order, and are all classified as VO dominant languages. The dominance order information of the above random forest model based on weighted decision trees is fully consistent with the dominance order in WALS, with 100% accuracy.

In other words, the results of language categorization based on the dominance order information of predicate and object in the decision tree in the discrete perspective are no different from those in the traditional database.

Table 2: Predicate-object order relation based on UD data set in cross-language languages

language	Proportion of OV word order in tree database (%)	UD preponderant word order	WALS Advantage Word Order	family of languages
Hindi	99.32	OV	OV	Indo-Iranian group
German	58.87	Lack of advantageous word order	Lack of advantageous word order	Germanic language family
Czech	21.14	VO	VO	Slavic language family
French	19.18	VO	VO	Roman language family
Spanish	9.76	VO	VO	Roman language family
Russian	9.22	VO	VO	Slavic language family
Portuguese	8.06	VO	VO	Roman language family
Polish	7.34	VO	VO	Slavic language family
Italian	6.37	VO	VO	Roman language family
Swedish	2.68	VO	VO	Germanic language family
English	2.27	VO	VO	Germanic language family

Cross-linguistic classification of quantitative features of sequences in a continuous perspective

Characterization of the frequency of major binary sequential relations

The random forest model approach of weighted decision trees to calculate the corresponding optimal number of classifications is shown in Figure 5. In addition to counting the frequencies of the five sets of major binary sequences in the decision trees of the 11 Indo-European cross-linguistic languages, we are also concerned with whether we can conduct a feature analysis of the Indo-European cross-linguistic languages based on the frequency information of the relationships of these five sets of major binary sequences. Before performing the feature analysis, an important issue is: We need to find a suitable number of classifications, i.e., the optimal number of classifications, so that the whole dataset is classified into n class to achieve the smallest possible variation within the dataset, and the optimal number of classifications based on the frequencies of the major ordinal relations of the Indo-European cross-linguistic languages is 4.

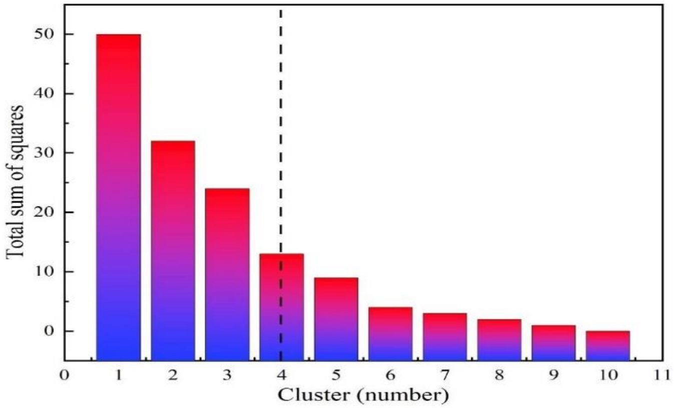

Figure 5: Calculate the corresponding optimal number of clusters

Figure 6 shows the results of language feature classification based on the frequency of the main sequential relations of the cross-linguistic languages. In this study, 11 decision trees of 11 cross-linguistic languages were classified into 4 categories, and the results are shown in Figure 6.

The classification results for the cross-linguistic languages of the Indo-European family are very good: Slavic languages (Russian, Czech and Polish), Romance languages (Spanish, Italian, Portuguese and French), Germanic languages (Swedish, German and English) and Indo-Iranian languages (Hindi) can be well classified into one category and are in full agreement with the traditional linguistic genealogy classification results. In addition, the common correlation coefficient of the decision tree model was 0.85, indicating good classification results. In summary, the frequency information of the main sequential relations can well distinguish the languages of the various language families within the Indo-European family.

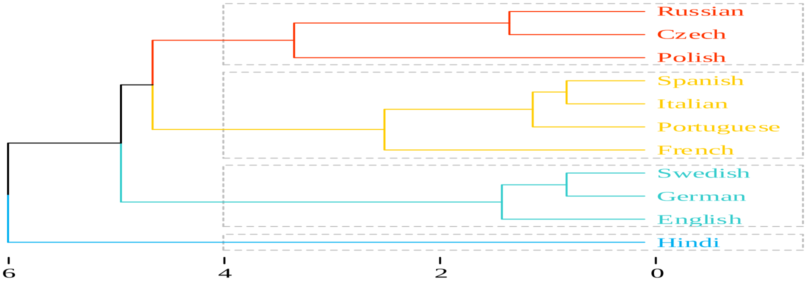

Figure 6: Classification results based on cross-linguistic features

Principal component analysis of the frequency of major binary sequential relations

Table 3 shows the eigenvalues of the major components of the frequency of major ordinal relations across languages. We used principal component analysis to reduce the dimensionality of the data to determine the similarity between languages by reducing the five sets of major sequences of each language to a few major components. Based on the frequency information of the relationship of each major sequence, the contribution value of each major component to the language for type differentiation. The number of major components is determined by the magnitude of the eigenvalue of each major component, so that the major component with the largest eigenvalue is retained. According to the Kaiser criterion, the component with an eigenvalue greater than 1 is the principal component. The first, second, and third components have eigenvalues greater than 1. Therefore, these three components are the principal components, and they explain 87.46% of the variance.

Table 3: Main eigenvalues of cross-language families

index	characteristic value	Variance percentage (%)	Proportion of cumulative variance (%)
PC1	2.04	40.98	40.98
PC2	1.35	27.15	68.14
PC3	0.94	18.52	86.67
PC4	0.62	12.27	98.94
PC5	0.05	1.05	100

We plotted the principal component analysis based on the first two major components, see Figure 7. We can classify each principal component by the magnitude and direction of the coefficients corresponding to each inflectional order combination degree of freedom in the horizontal and vertical coordinates. The larger the absolute value of each coefficient, the more important the corresponding inflectional order The first principal component has a strong positive association with the inflectional order freedom of the preposition-center noun inflectional relationship and a strong negative association with the inflectional order freedom of the predicate-object inflectional relationship and the subject-predicate inflectional relationship.

Similarly, for the second main constituent, the degree of inflectional freedom of the subject-predicate inflectional relationship and the collar-center noun inflectional relationship are strongly negatively associated with it, while the degree of inflectional freedom of the adjective-center noun inflectional relationship is strongly positively associated with it. Thus, for the third main component, the Romance languages in the blue circle show more freedom and flexibility in the adjective-center noun inflectional order, a result that is consistent with the results in Figure 6 (French, Italian and Spanish rank in the top four in terms of cosine similarity values), which is an important reason for the strong differentiation of their typological features.

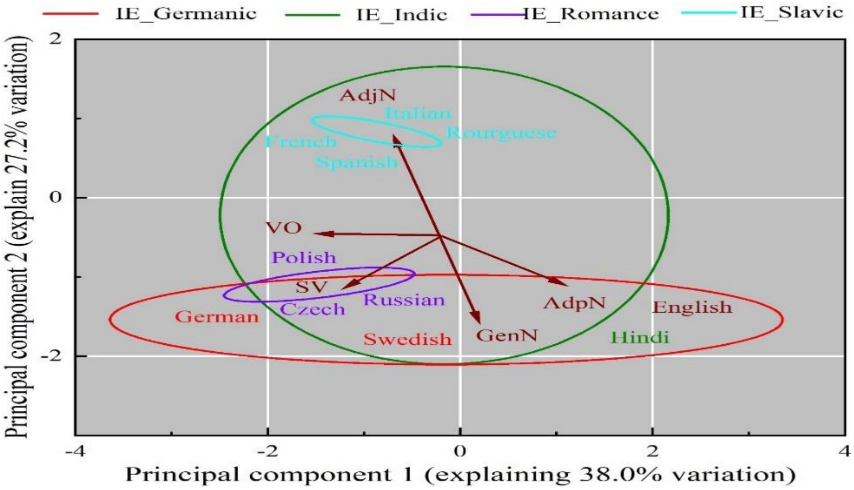

Figure 7: Principal component analysis based on the degree of freedom of cross-language families

Conclusion

Based on big data technology and artificial intelligence, a random forest model with weighted decision trees is proposed to study and analyze the morphological and syntactic features of language and to realize the classification of type features across language sequences. The following conclusions can be drawn:

(1) The match between the dominant order extracted from the decision tree model and the dominant order of each language in WALS, according to the proportion of SV order relations between the subject and the predicate in the treebank, we can find that the proportion of SV order relations of 11 Indo-European cross-linguistic languages varies and is distributed in the range of [79.28%, 98.93%], and the proportion range of their corresponding VS order The corresponding proportion of VS sequences was [1.06%, 20.72%]. Even the language with the lowest proportion of SV order (Czech) has a proportion of SV order (79.27%) that is much more than twice the proportion of VS order (20.72%). The results of dominant inflectional order categorization based on the inflectional order relations of subjects and predicates in the random

forest model of weighted decision trees are highly similar to the information provided in traditional typological databases.

(2) Each principal component is classified by the magnitude and direction of the coefficients corresponding to each degree of freedom of the combination of inflectional order in the horizontal and vertical coordinates. The more important the corresponding order the first principal component has a strong positive correlation with the order freedom of prepositional-central noun order relations, for the second principal component the order freedom of subject-predicate order relations and collateral-central noun order relations have a strong negative correlation with it, and for the third principal component the Romance languages in the blue circle show more free and flexible.

References

A, D. (2016). Motivation Types of Chinese Language Learners in China. *Overseas Chinese Education*.

Barbosa, G., Camelo, R., Cavalcanti, A. P., Miranda, P., Mello, R. F., Kovanović, V., & Gašević, D. (2020). *Towards automatic cross-language classification of cognitive presence in online discussions.* Paper presented at the Proceedings of the tenth international conference on learning analytics & knowledge.

Basu, R. (2018). The Problematic Influences Of First Language Culture On Esl And Modes To Overcome The Same. *International Journal of English Learning & Teaching Skills, 1*(1), 62-65.

Bouraoui, A., Jamoussi, S., & Hamadou, A. B. (2022). A comprehensive review of deep learning for natural language processing. *International Journal of Data Mining, Modelling and Management, 14*(2), 149-182.

Chen, Y. (2022). Study on the evolutionary game theory of the psychological choice for online purchase of fresh produce under replicator dynamics formula. *Applied Mathematics and Nonlinear Sciences, 7*(1), 641-652.

Dassa, L. (2012). Book Review: Literacy Language & Culture: Methods and Strategies for Mainstream Teachers with Not-So-Mainstream Learners. *ABAC Journal, 32*(3).

Davis, C., & Aid, G. (2022). Machine learning-assisted industrial symbiosis: Testing the ability of word vectors to estimate similarity for material substitutions. *Journal of Industrial Ecology, 26*(1), 27-43.

He, H., Jin, J., Xiong, Y., Chen, B., Sun, W., & Zhao, L. (2008). *Language feature mining for music emotion classification via supervised learning from lyrics.* Paper presented at the Advances in Computation and Intelligence: Third International Symposium, ISICA 2008 Wuhan, China, December 19-21, 2008 Proceedings 3.

Jia, X., Tai, J., Huang, Q., Li, Y., Zhang, W., & Du, H. (2019). EG-GAN: Crosslanguage emotion gain synthesis based on cycle-consistent adversarial networks. *arXiv preprint arXiv:1905.11173*.

Li, S., Cao, Y., & Min, X. (2018). Research on the text and emotion classification of particle swarm optimization in natural language processing. *Mach Tool Hydraulics, 46*(4), 150-155.

Qiang, C., Yanxiang, H., Xule, L., Songtao, S., Min, P., & Fei, L. (2014). Cross-Language Sentiment Analysis Based on Parser. *Acta Scientiarum Naturalium Universitatis Pekinensis, 50*(1), 55.

Research on Cross-cultural Communication and Language Awareness. (2013). *Foreign Language Research in Northeast Asia*.

Wang P , Z. X., Li N , et al. (2017). Research of cross-language sentiment classification based on structural correspondence learning. *Journal of Nanjing University(Natural Science)*.

Wu, X., Wang, T., & Wang, S. (2020). Cross-modal learning based on semantic correlation and multi-task learning for text-video retrieval. *Electronics, 9*(12), 2125.

Zou, X., Wang, M., & Zuo, J. (2010). New Multilingual Information Retrieval Model Based On Latent Inter lingua Semantics. *Journal of Chinese Computer Systems, 31*(04), 696-701.