
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1588–1602, DOI 10.12694/scpe.v25i3.2778

PREDICTION OF ENGLISH TEACHER CAREER DEVELOPMENT BASED ON DATA MINING AND TIME SERIES MODEL

LIPING FAN*

Abstract. With the gradual growth of the teaching profession, the teaching profession is facing new trends in reform and development, and the same dilemma exists for English teachers' career development and planning. To this end, the study first uses a modified K-means clustering method to cluster and analyse the factors affecting English teachers' professional development, forming a system of indicators on English teachers' professional development. The Long Short-Term Memory (LSTM) network employed the time-series features to create a time-series model, and the Support Vector Machine (SVM) was used to forecast the course of English teachers' career development. To assess the current career status of English teachers and their impact on people and organizations, this study proposes a career prediction model for English teachers. This model utilizes data mining and time series modeling to provide accurate predictions. In accordance with the experimental findings, the precision of the upgraded K-Means model was 98.58%, and the error between the projected sample data and the actual sample data for the training of the trend prediction model for English teachers' career growth was 0.032. It was able to accurately predict teachers' career development and explore the specific factors affecting English teachers' career development, so as to solve the problems in teachers' career development.

Key words: Data mining, Time Series, English Teacher, Career development, Predictive models

1. Introduction. As modern science and technology develop rapidly, the Internet and artificial intelligence and other science and technology are reshaping educational forms such as teacher education development, and the development of network technology also provides data support for English teachers' professional development planning [1]. Based on data mining technology applied to teachers' professional education, through the collection of statistical teaching big data, deep learning and intelligent analysis of teachers' characteristics, it can provide personalized guidance solutions for teachers' quality improvement [2]. The career development stages of English teachers are generally divided into career exploration stage, career establishment stage, mid-career stage and late career stage [3]. Factors influencing English teachers' career development are generally categorized as social, family, personal and organizational [4, 5]. The Department of Teacher Education of the Ministry of Education in 2001 suggested that the main influencing factors for in-service teachers include school environment, life environment, teachers' social status, students, and teachers' peer groups [6]. There are serious issues with the professional development of English teachers, and the number of people participating in team building for teacher development is growing. In addition, teacher professional education is confronting new challenges related to reform and development [7]. The study proposes to use K-Means algorithm as a cluster analysis algorithm for data to predict teachers' career development, I use a time series prediction model and a Support Vector Machine (SVM) to categorize, analyze, and determine the stages of English teachers' promotion opportunities and the direction of their career development. It also explores the specific factors that affect English teachers' career development so as to solve the problems in teachers' career development.

2. Related works. Addresses the difficulties of choice and career planning faced by English teachers in their professional development. With the rapid development of information technology, English teachers can efficiently and accurately choose the right career path for their development by using technologies such as data mining and artificial intelligence [8]. Tim A et al. addressed data mining for electrochemistry, with a general discussion from information to knowledge, describing the location of the nanochannels themselves by performing species transport on them [9]. Zou C et al. discovered high-strength ductile titanium alloys based on the integration of data mining and machine learning. And the integration allowed for more efficient and

*College of Humanities and Law, Gannan University of Science and Technology, Ganzhou, 341000, China (flp080881@163.com)

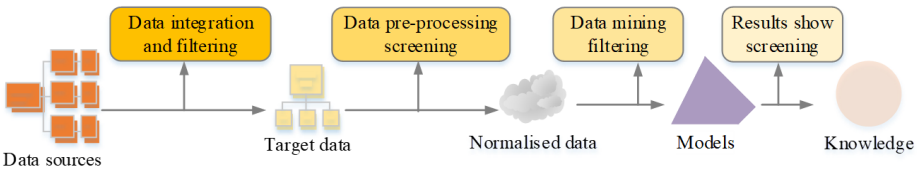


Fig. 3.1: Data mining specific flow chart

cost-effective design of high-strength and ductile titanium alloys [10]. By using data mining approaches, Feng Z et al. suggested identifying additional suspected harmful viruses in pangolins. They found two genomes of the genus Gemykibivirus and nine types of bat-associated circovirus, respectively [11]. With the implementation of five distinct text different classifiers that perform well in tweet categorization, Rahman R presented a real-time Twitter data mining method for inferring user mobile perceptions [12]. He Y et al. used data mining techniques with statistical metrics to analyse strategies for database-based energy-efficient building design, building a database of near-zero energy-consuming buildings and developing a customised data mining [13]. Parashkooh H I proposed a data mining technique based on oil-in-water droplet aggregation to guide the study of molecules and performed a series of molecular dynamics models. The findings exhibit a comprehension of the joint conduct of all species in multiphase systems, which can be applied to numerous emulsion formation fields [14]. Chen H et al. investigated an adaptive recommendation method based on online learning styles that can personalise learning resources according to users' pedagogical needs and personal preferences. Experimental results showed that the model facilitated data mining for learners and that the accuracy of recommendations was higher than other traditional recommendation models [15].

To process this temporal data, many researchers have created both short- and long-term memory artificially learning modules for temporal prediction models [16]. Wang H et al. proposed a more efficient fusion algorithm by combining BP neural networks with genetic algorithms and particle swarm optimization algorithms in data mining techniques [17]. The fusion algorithm was found to consume less energy and run more stably after experimental comparison [18]. To address the problem of predicting the career development of English teachers (CDET), this study proposes a model for predicting the CDET based on data mining and time series models. The model classifies and analyses the stages of teachers' career development and the direction of English teachers' career development based on the time series prediction model and SVM to explore the specific factors affecting English teachers' career development. The model uses an improved K-Means algorithm as a clustering analysis algorithm for data to predict teachers' career development.

3. English Teacher Career Development Analysis and Forecasting Design.

3.1. Construction of a Career Development Indicator System Based on K-means Clustering.

The study uses techniques such as data analysis to predict the CDETs in relation to their current situation. Among the factors that affect English teachers' career development are the imbalances in gender, education, age, title and professionalism in the English teaching force, which in turn affect the development of the teaching force. Individual English teachers' professional development is influenced by both personal and organizational factors. Therefore, when studying the CDETs, the data can be aggregated to analyse the factors that influence their career development, and then to predict the CDETs. The study first uses data mining techniques to analyse the data and further optimize the predicted data in English teachers' career development. The specific flow chart is shown in Figure 3.1.

Finding information data or specific correlations between data that fulfill English teachers' objectives of professional growth from a huge amount of information data is the main goal of the data mining technique shown in Figure 1. The main steps are data integration and screening of data sources to form target data, followed by pre-processing screening operations on the target data. After the data has been normalized, the

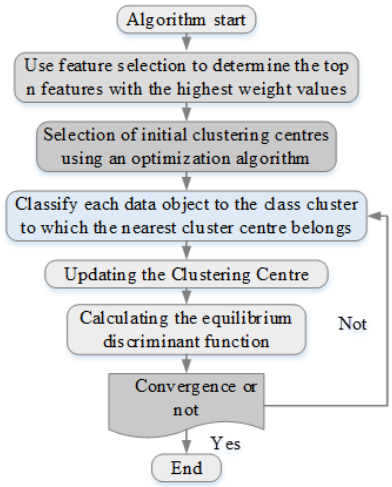


Fig. 3.2: Flow chart of improved K-Means feature selection algorithm

corresponding data model is formed by mining and filtering, and finally the results are presented. The application areas of data mining include clustering analysis, association rule analysis, feature algorithm analysis and classification analysis. The study uses the K-Means algorithm and improves it according to the problem of predicting the CDETs. The dataset chosen for this study originates from English professors in a university. It encompasses teachers' personal information, teaching accomplishments, research achievements, career plans, and other relevant details. The K-means algorithm is applied for cluster analysis due to its simplicity and ease of implementation. It is optimal for handling large-scale datasets. When selecting the basic principle behind K-means algorithm, the research mainly considers the following aspects: The K-means algorithm demonstrates a strong clustering effect. It offers high processing efficiency, enabling fast cluster analysis of large-scale data sets. Moreover, K-means has good interpretability and can directly reflect the distribution of data. Additionally, data mining technology is applicable to various data types. To better adapt K-means algorithm to the needs of English teacher career development prediction in cluster analysis, the initial cluster center is optimized. In the initialization phase of the algorithm, a more effective method for selecting the initial cluster centers is employed to circumvent local optimal solutions. Following this, the number of clusters is adjusted dynamically. During algorithm operation, the number of clusters gets dynamically adjusted based on evaluation index of clustering results for optimal clustering effect. Subsequently, to accurately represent the CDET, teacher's personal information, teaching achievements, scientific research achievements, and other relevant data characteristics are considered. Through the aforementioned enhancements, this study employs the K-means algorithm to extract and group the data. A comprehensive flowchart of the revised method is presented in Figure 3.2.

The detailed steps of the enhanced K-means feature selection algorithm are shown in Figure 3.2, starting with the initialization of the weights for each feature property. Secondly, the weight vector of feature attributes is calculated by the feature selection algorithm and reassigned. Third, the obtained feature attribute weight vector is sorted according to the weight value from largest to smallest, and then the top n elements are selected from all features. Fourth, n clustering centers with optimal contribution are selected among the initial ones. Fifth, the Euclidean distance calculation between data objects is performed, and based on the result of the calculation, the division is made into the class clusters with the smallest Euclidean distance value from the current data object, respectively. Sixth, a reassignment and update is performed at the centre of each class cluster. Seventh, a balanced discriminant function is used to calculate and then determine whether the final value gradually converges to 0. Eighth, if the function value is getting closer to 0, then the algorithm run ends, otherwise go to step 5. In terms of feature selection, the selected dataset D dataset is divided into

partitions of $D = D_1, D_2, .., D_n$. The equation for calculating the weights of specific feature attributes is given in equation 3.1.

$$W_t^{i+1} = W_t^i - \sum_{x \epsilon T(c)} \frac{diff(t, D_i, x)}{n * d} + \sum_{x \epsilon S(S_i)} \left[\frac{q_c}{1 - q(D(D_i))} \sum_{x \epsilon G(c))} diff(t, D_i, x) \right] /(n * d) \tag{3.1}$$

In equation 3.1, each partition contains q attributes. where $D = D_{i1}, D_{i2}, \ldots, D_{in}$ and D_i attributes have the category $C_i \in C$, $C = C_1, C_2, \ldots C_k$ is a different set of categories for k, and the centre of mass is D_i. the data set will then be partitioned into categories, and subsequently will then go on to select d data objects from each category of data samples at a distance of D_i, where the updated weight vector for the feature attributes is $W = W_1, W_2, \ldots W_k$. n is the number of times the sample data is extracted. Equation 3.2 is used to calculate the diff(t, S_i, x) function, which indicates the differential function of the data objects.

$$\text{diff}(t, S_i, x) = \left| \frac{D_{it} - D_{jt}}{max_t - min_t} \right| \tag{3.2}$$

Equation 3.2 equalizes D_i distance to be more comparable to the d sample data by using the maximum and lowest values on the particular desired. For the optimal selection of the initialized clustering centers, assume that the data set, $X = x_1, x_2, \ldots, x_n$ has n different data objects, each containing p features, i.e. $x_i = x_{i1}, x_{i2}, \ldots x_{ip}$. Equation 3.3 defines the Euclidean distance between any two data objects, x_i and $1 < i < j < n$.

$$d(x_i, x_j) = \sqrt{\sum_{a=1}^{p} (x_{ia} - x_{ja})^2} \tag{3.3}$$

Equation 3.3 defines the distance density function for a sample data set X that corresponds to a data object named $x_i (1 < i < n)$ in equation 3.4.

$$\text{density}(x_i) = \sum_{j=1}^{n} \frac{d(x_i, x_j)}{\sum_{i=1}^{n} d(x_i, x_j)} \tag{3.4}$$

In equation 3.4, it is assumed that the neighbourhood radius R_i of data object $x_i (1 \leq i \leq n)$ in the dataset can be defined in equation 3.5, where $cR (0 \leq cR \leq 1)$ is used as a moderating factor for the neighbourhood radius. The clustering effect is expected to be more favorable if the value of cR is assumed to be 0.1, based on prior experience. The details are shown in equation 3.5.

$$R_i = n^{cR} * \frac{1}{n} \sum_{i=1}^{n} e^{-\mathbf{density}(x_i)} \tag{3.5}$$

The data set X's data object is supposed to be $X_i (1 \leq i \leq n)$, with R_i centre and radius neighbourhood radius of the spherical region contains many data pairs, i.e. corresponding point density X_i , denoted as $S(X_i)$. In the spherical region where this data object is placed, there are more points per unit area the greater the value of $S(X_i)$, as shown in equation 3.6.

$$S(x_i) == |p| d(x_i, p) \leq R_i, p \in X| \tag{3.6}$$

Assume that $MS(x)$ is the average density of data objects in dataset x , see equation 3.7.

$$MS(x) = \frac{1}{n} \sum x \in XS(x) \tag{3.7}$$

The citation criterion function, also known as the objective function, is the criteria specified to decide data clustering. The objective function is used to calculate whether the similarity of the data objects in the same category meets the requirements and whether the differences between the different categories are close

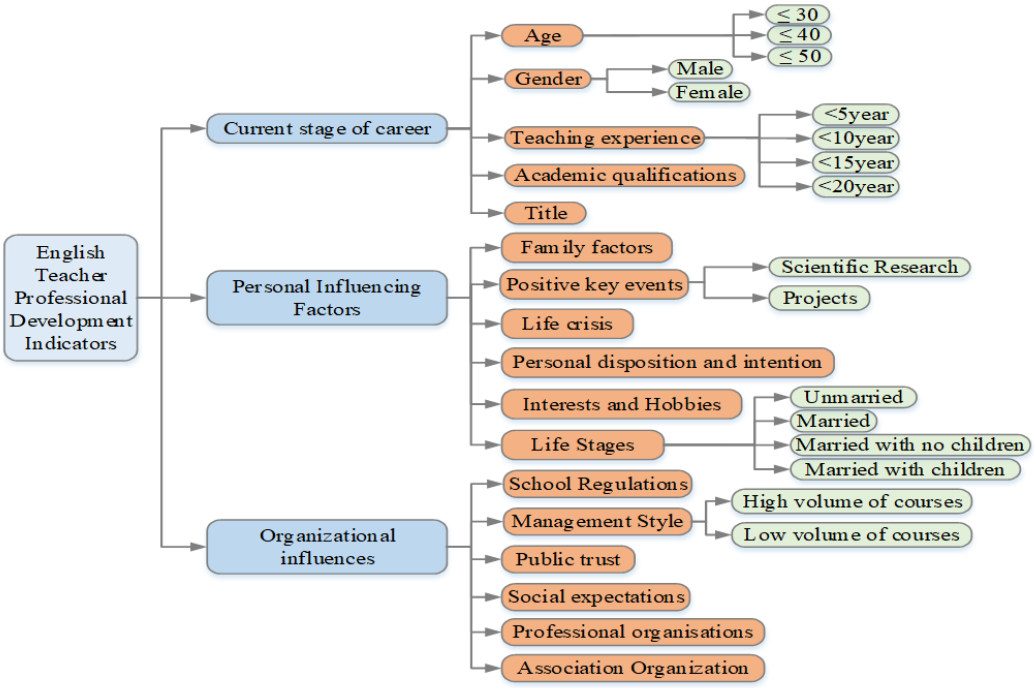


Fig. 3.3: Model diagram of English teacher career development indicator system

to the maximum. Based on this discriminative clustering, the optimal number of clusters can be obtained. The squared distance between the cluster's data objects and its center must be determined when intra-cluster variance is used to assess how compactly the clusters inside a cluster are organized, as shown in equation 3.8.

$$w(c) = \sum_{i=1}^{k} w(c_i) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x_i, c_i)^2 \tag{3.8}$$

The computing the Euclidean distance from the cluster centers and subsequently the difference between the clusters, the difference between the clusters can be determined. Assuming that c_i and c_j are the centers of the ith and jth class clusters respectively, then the difference between the two class clusters $b(c)$. The detail is shown in equation 3.9.

$$b(c) = \sum_{1 \le j \le k} d(c_j, c_i)^2 \tag{3.9}$$

The equilibrium discriminant function is then introduced in equation 3.10, where k is the number of clusters, and the differences between class clusters $b(c)$ and within class clusters $w(c)$ need to be first normalized.

$$W(c, k) = \frac{1}{1 + e^{b(c) - w(c)}} \tag{3.10}$$

The factors that affect English teacher's professional development were analyzed through data clustering and aggregation. This process led to the identification of different factors that influence their development and resulted in the creation of an indicator system. The specific English teacher career development indicator system model is shown in Figure 3.3.

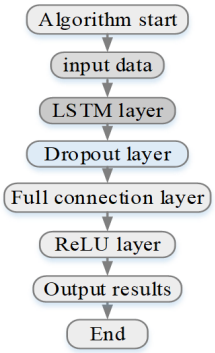


Fig. 3.4: LSTM model structure diagram

3.2. Time Series Model-based Design for Predicting English Teachers' Career Development. The theory of teacher professional development is divided into two areas: stages of professional development and influencing factors. Data mining techniques are used to determine and analyse the influencing factors in teacher professional development, while English teacher professional development requires a time series model of the stages of teacher professional development. The three main categories, from point to point and then comprehensive, are usually period theory, stage theory and cycle theory. Due to recent advancements in the realm of data science, LSTM has emerged as a highly efficacious solution for all time series prediction issues. As a recurrent neural network architecture, it can operate on varied interval values, rendering it a perfect fit for classifying, processing, and prognosticating time series with unknown durations or lags. The specific LSTM model structure is illustrated in Figure 3.4.

In Figure 3.4, a hierarchical analysis of the specific LSTM model structure is presented. In order to build the time-series features, the data was first pre-processed and then the factors influencing English instructors' professional development were identified. In constructing time-series features, objective information is melded to unify data traces from multiple behaviors in a time series. English teachers are then structurally entered into the LSTM network chronologically, based on age and teaching experience. Subsequently, an LSTM with a Dense function is assembled, holding an input layer with two fully connected layers. Finally 50 more features were extracted separately as time-series features to represent the dynamic changes of the above indicators. To further improve the interpretability of the behavioral features. The weights of each linear indicator for each teacher are calculated in equation 3.11.

$$\begin{cases} (N - \mathrm{Rank}(x_n))/N, \mathrm{Corr}(X_k) > 0 \\ \mathrm{Rank}(x_n)/N, \mathrm{Corr}(X_k) < 0 \end{cases} \tag{3.11}$$

In equation 3.11, N denotes N English teachers and K features extracted. $\mathrm{Corr}(X_k)$ is the Pearson correlation coefficient between the kth feature X_k and the individual influences on English teachers' professional development, where $k \leq K$. $\mathrm{Rank}(x_n)$ denotes the ranking of the Nth English teacher's professional development feature (denoted as $u_n$and$n \leq N$) among all English teachers' professional development factors. For example, there are four English teachers (u_1, u_2, u_3, u_4), if their kth characteristic (e.g., English teachers under thirty) is (0.8, 0.5, 0.7, 0.6), then $\mathrm{Score}_{k1} = 0, \mathrm{Score}_{k2} = 0.65, \mathrm{Score}_{k3} = 0.25 and \mathrm{Score}_{k4} = 0.15 \mathrm{Corr}(X_k) > 0$. The weighted average of the characteristics is obtained by further substituting the following equation.

$$\sum_{k=1}^{k} = (|\mathrm{Corr}(X_k)|^* \mathrm{Score}_k^n) \tag{3.12}$$

Specifically, the average value of impacts on English teacher professional development is determined by a weighted average of all the influences with the respective weights defined by correlation coefficients. Following

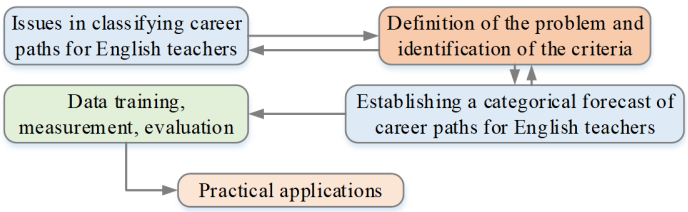


Fig. 3.5: English Teacher Professional Development Forecast Implementation Flow Chart

the above steps, a weighted average of English teacher career development and a weighted average of all characteristics can be calculated into the weighted average of English teacher career development factors and all characteristics, respectively. The data is then normalized. The specific normalization equation is shown below.

$$x_{scale} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.13}$$

In equation 3.13, x denotes the normalized variable object, x_{min} and x_{max} denote the minimum and maximum values of the variable object respectively, and x_{scale} is the normalized value, specifically ranging from 0 to 1. The practical linear classifier is then constructed, and the regression algorithm is later implemented by SVM for outlier detection to do the classification work. The specific English teacher career development prediction implementation process is shown in Figure 3.5. The SVM model was applied to the directional classification problem of English teacher career development. The goal of the model is to maximize the classification interval. The separation distance between two hyperplanes is the classification interval of the SVM, and the hyperplane should be oriented in such a way that it is as far away from the nearest data point of each class as possible. A decision boundary with a larger interval means that the model has a smaller generalization error. Whereas a smaller interval means that the model may overfit, the training samples that are closest to the hyperplane interval are the support vectors. With w serving as the weight vector and serving as the bias, the ideal hyperplane can be defined as $wx^T + b = 0\dot{w}$ and b will meet the requirements in the equation for each component of the training set.

$$\begin{cases} wx_i^T + b \geq 1, \text{if} = 1 \\ wx_i^T + b \geq -1, \text{if} = \text{-1} \end{cases} \tag{3.14}$$

Identifying w and b is the key to training the SVM model so that the hyperplane is as distant from the various classes of points as is feasible. Figure 3.6 depicts the SVM model's underlying principles.

500 linear features, 120 non-linear features, 60 LSTM-based features, and 30 basic features were among the 780 distinct types of characteristics that were retrieved for the study (including gender, age and teaching age). Firstly, pre-processing was carried out to zero-fill the behavioral features that were missing. Also, to eliminate order-of-magnitude differences between features, all features were scaled to between 0 and 1 using the maximum-minimum normalization method. The strategy for dealing with data gaps in the sample was to select the feature to be omitted when more than half of the gaps were present, and its gaps were zero-filled.

$$x' = \frac{x - min}{max - min} \tag{3.15}$$

Next, in the selection of features, the SelectKBest function in the scikit-learn library was used as the feature selection function and f-classif was used as the evaluation index in the python 3.8 environment to screen the set of features that have a more significant impact on the professional development of English teachers. Indeed, data usage and privacy protection are crucial ethical concerns for data-driven research. In this study pertaining to the prediction model for CDET, the following ethical implications must be addressed:

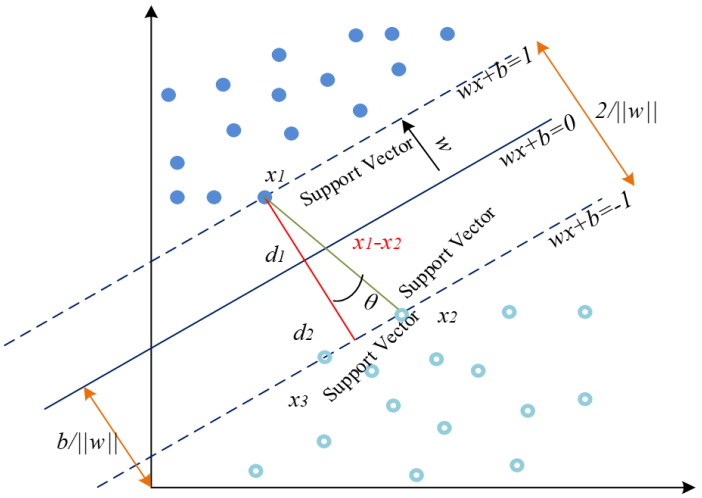


Fig. 3.6: Classification and prediction principle of SVM model

1. Data Use and Privacy protection: When collecting and utilizing teacher data, strict adherence to data protection principles and privacy policies is imperative. The collected data must be utilized exclusively for research purposes while ensuring necessary security measures are taken to safeguard against unauthorized access and use.
2. Transparency and research purposes: The utilization of teacher data in research should be transparent and overt. It is incumbent on researchers to explicitly articulate to faculty and other stakeholders the employment and objectives of the data while ensuring ethical standards are upheld.
3. Respect the wishes of teachers: Teachers' will and rights must be respected when collecting and using their data. Researchers should respect teachers' decisions not to participate in research or provide personal information and ensure that their choice does not hinder their professional growth.
4. Fairness and harmlessness: Research utilizing teacher data must abide by the principles of impartiality and non-harmfulness. The study must guarantee that it does not have a negative impact on the educators' professional growth, and that the complete variety of potential outcomes and impacts are adequately considered.
5. Sustainability and accountability: Research utilizing teacher data should take into account sustainability and accountability. Researchers must ensure that the data is sustainable and accessible, and implement measures to protect against unauthorized access and usage. In conclusion, the ethical considerations surrounding the use of teacher data for research are crucial. It is imperative that researchers abide by strict ethical guidelines, laws and regulations to safeguard against privacy breaches while ensuring transparency, respect for faculty wishes, fairness, harmlessness, sustainability, and accountability.

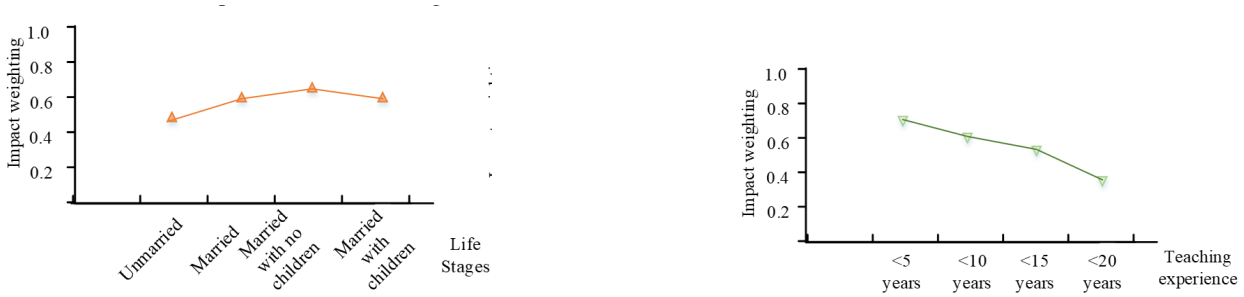
4. Analysis of the Effects of Predicted Professional Development of English Teachers. The system software environment based on the experiment was a Windows 10 system, an Intel(R) Core(TM) i7-6700 processor with 4G installed memory. The existing English teachers' career development database was imported into the prediction software and predicted online, and the existing data samples were compared and analysed according to the prediction results of the prediction model. A total of 1,000 English teacher career development data were processed to obtain 880 valid data, of which the results of the weighting of influencing factors in some English teacher career development predictions are shown in Table 4.1. The research findings are based on a substantial amount of experimental data and repeated tests. Senior university faculty were employed to analyze the test results, providing a high level of credibility for the study.

Table 4.1: Forecast results table

Tier 1 indicators	Weighting	Secondary indicators	Weighting
Career Status	0.332	Age	0.312
		Gender	0.251
		Teaching experience	0.354
		Title	0.345
Personal Influences	0.335	Family factors	0.325
		Positive key events	0.321
		Life crisis	0.254
		Personal disposition and intention	0.326
		Interests or hobbies	0.214
		Life Stages	0.462
Organizational influences	0.333	School Regulations	0.123
		Management Style	0.231
		Public Trust	0.341
		Social Expectations	0.325
		Professional Organization	0.252
		Association Organization	0.125

As the English teacher career development predictors are positive indicators. After calculating the weighting of the indicators in the table above, the results show that the most influential factor in the CDETs is life stage, with a weighting ratio of 0.462, which is one of the more important indicators of personal influence. This further indicates that each English teacher's career development focus varies at certain stages of life, either by improving their theoretical knowledge of English or pursuing further studies. Alternatively, they may choose to evaluate their titles and pursue additional education or concentrate on teaching and prioritize their families. The next factor is the number of years teaching English, which is weighted at 0.354, with different goals for English teachers in their career development. For new teachers, the immediate goal is to continue to hone their teaching and gain experience in teaching English. For teachers with some years of experience, their career development may be focused on refining their professionalism. The less influential factor in English teachers' career development is the organizational factor of school regulations, with a specific weighting of 0.123. The main reason for this is that school regulations are generally humane and therefore less influential in English teachers' career development. Life stage refers to the life cycle of teachers and is categorized into different periods: exploration, establishment, stability, maintenance, plateau, and retirement. The practical implications of life stage and school regulations are as follows. The impact of life stage varies during different stages of career development and presents English teachers with varying tasks and challenges. For instance, during the exploratory stage, teachers must adjust to the teaching setting, acquire teaching expertise, and establish their own educational concepts. In the stable stage, they face the challenge of career plateau and must sustain their enthusiasm and teaching drive. Hence, differences in life stages exert a significant influence on teacher professional growth. The Effects of School Regulations: School regulations are the norms of teaching and management that English teachers must adhere to. Reasonable regulations can safeguard the rights and interests of teachers while providing an optimal teaching and development environment. However, unreasonable regulations may impede teachers' instructional autonomy and professional growth. Consequently, school regulations are crucial factors that impact the CDET. Based on the analysis, it is evident that numerous factors impact the career growth of English teachers, with life stage and school regulations being the most significant. Thus, when developing the career plan for English teachers, one should consider the full impact of these factors. This consideration will provide better career advancement prospects and support for teachers. A visualization was used to show the variability of life stage and teaching age in English teachers' career development, as shown in Figure 4.1.

In the visual comparison in Figure 4.1, life stage is a personal influence on the CDETs, with a high weighting in the influence indicators. The English teacher career development model was completed and the data set was



(a) The impact of life stages on the professional development of English teachers

(b) The impact of teaching age on the career development of English teachers

Fig. 4.1: The impact of different life stages and years of teaching on the professional development of English teachers

(a) Comparison chart of the results of the first set of comparison experiments

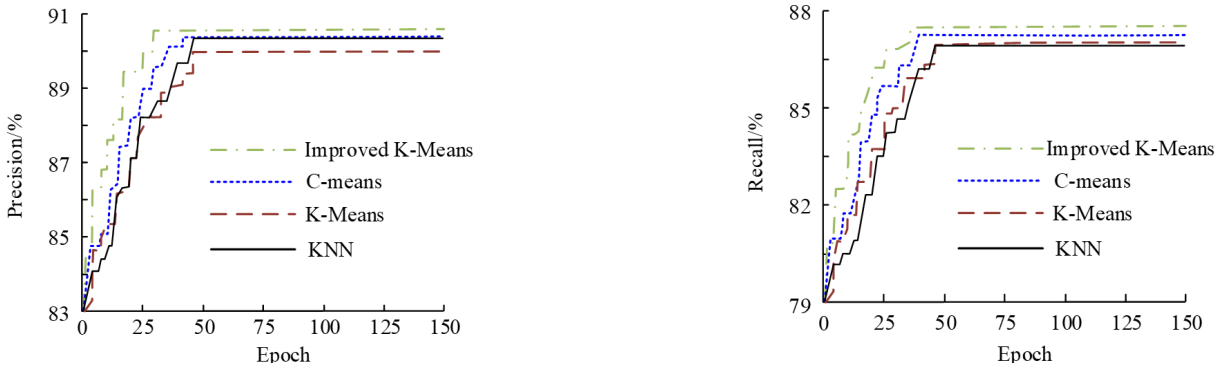
(b) Comparison chart of the results of the comparison experiment for group 2

Fig. 4.2: Validation of the graph of the correctness of the boys' data set with change

analyzed using data from the previous research. Experiments were also conducted on the relevant data to verify the performance and accuracy of the English teacher career development prediction model. The 2 sets of experimental results are shown in Figure 4.2.

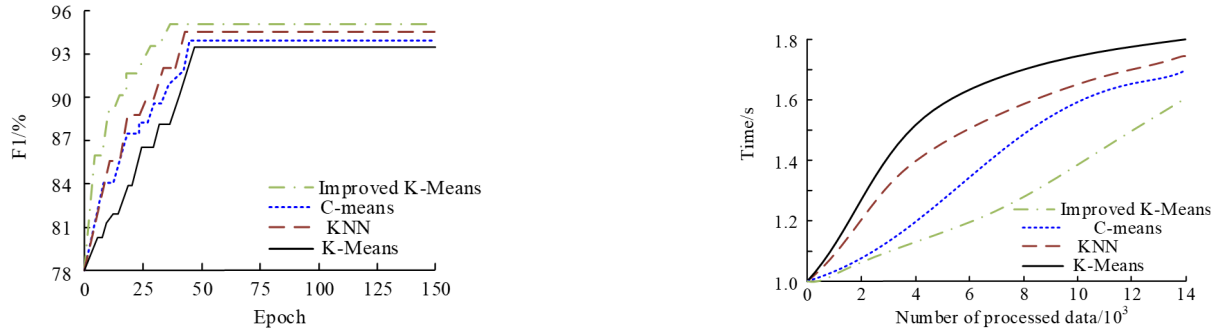
Figure 4.2 demonstrates that the accuracy of the CDET prediction model following the aforementioned training model has met the accuracy criterion between the absolute error limit of 0.1. The two sets of sample data were compared, where the error between the actual sample data and the predicted sample data was 0.032. The results indicated that the trend prediction model for English teachers' career development had a high accuracy. The performance of the constructed indicators to evaluate the improved K-Means algorithm was tested to achieve the matching of the actual influencing factors of English teachers' career development. The data were divided into two datasets, one of which contained 480 records, 33 attributes and 323 entities. The other dataset contained 520 records, 33 attributes and 432 entities. The improved K-Means algorithm is compared with the three current state-of-the-art models. The Precision and Recall of the above three models are shown in Figure 4.3.

In Figure 4.3a, the highest Precision of the study model is achieved at 32 iterations, which is 10 and 16 times less than the C-means model and the KNN model respectively. At this stage, the precision of the improved K-Means model stands at 98.58%, surpassing the C-means model, KNN model, and traditional K-Means model by 0.12%, 0.36%, and 28%, respectively. In Figure 4.3b, the improved K-Means model attains the maximum recall of 44 iterations, which is significantly lower than the C-means model, KNN model, and traditional K-



(a) Precision (b) Recall

Fig. 4.3: Precision and Recall of the model



(a) Epoch (b) Time

Fig. 4.4: F1 and running time of the model

Means model, by 2, 5, and 7 times, respectively. The recall rate for the improved K-Means model was 86.62%, surpassing the C-means, KNN, and traditional K-Means models by 0.13%, 0.29%, and 0.32%, respectively. The above results demonstrated that the performance of the English teacher career development model constructed by the study based on the improved K-Means was superior. F1 and running time of the C-means model, KNN model and traditional K-Means model are shown in Figure 4.4.

In Figure 4.4a, the F1 of the C-means model, KNN model and traditional K-Means model are roughly positively correlated with the number of iterations. However, after a certain number of iterations, F1 stops growing and stabilizes. The F1 score for the improved K-Means model peaked at 31 iterations and then stabilized at 95.67%. In comparison, the C-means model only achieved its maximum F1 score after 46 iterations and remained stable thereafter. The F1 score for the traditional K-Means model eventually stabilized at 93.12%, which is 2.55% lower than that of the improved K-Means. The final F1 for the KNN model stabilized at 94.20%, which is 1.27% lower than that of the improved K-Means. When the sample size reaches 14000, the improved K-Means model requires 1.51 seconds in Figure 4.4b, which is 0.13 seconds, 0.19 seconds, and 0.21 seconds less than the C-means model, KNN model, and traditional K-Means model, respectively. It displays better performance than the other three models and proves to be superior. The improved K-Means model proposed in the study has a much better performance in entity matching. The performance of the four algorithms in

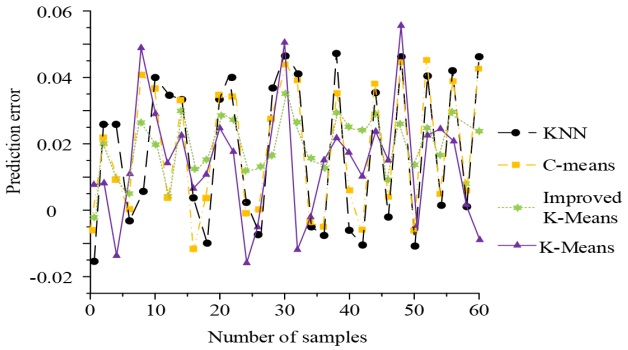


Fig. 4.5: Evaluation error results of different algorithms in training samples

Table 4.2: Comparison of the prediction results of several models under different data sets

Project	Data set 1		Data set 2		Data set 3	
	Forecast accuracy (%)	Prediction time (s)	Forecast accuracy (%)	Prediction time (s)	Forecast accuracy (%)	Prediction time (s)
Model 1	94.81	1.25	93.66	1.08	94.05	1.11
Model 2	90.44	1.88	90.13	1.96	90.28	1.97
Model 3	85.12	2.53	86.07	2.52	85.49	2.50
Model 4	89.47	2.15	89.18	2.22	89.72	2.10
Model 5	80.74	3.00	80.28	2.97	80.15	2.91

terms of evaluation error in the training sample is shown in Figure 4.5. As can be observed from Figure 4.5, all four algorithms are able to evaluate the given training samples, but the Improved K-Means algorithm is much better than the other three models. The prediction error of the improved K-Means algorithm can be controlled to within 0.002-0.032, possessing a better prediction accuracy compared to the remaining three algorithms. The traditional K-Means algorithm has a maximum prediction error of 0.055, with the worst prediction effect. kNN has a prediction error within -0.016-0.046, with a maximum prediction error value of 0.046. the C-means algorithm is second only to KNN, with a prediction error that can be controlled within -0.016-0.045, with a maximum prediction error value of 0.045. To assess the predictive capability of the model (Model 1) built by the Institute, Model 1 was compared with the more advanced career development prediction models in existing studies in three different data sets. Comparison models consist of a career development prediction model based on graph convolutional network (Model 2), a career development prediction model based on data mining (Model 3), a career development prediction model based on multi-attribute important weighted K-nearest neighbor algorithm (Model 4), and a career development prediction model based on extreme learning machine (Model 5). Table 4.2 displays the comparison results.

As Table 4.2 illustrates, model 1 shows an average prediction accuracy of 94.17% and an average prediction time of 1.15 seconds. In comparison to model 2, the average accuracy of its predictions is improved by 3.89%, 8.61% compared to model 3, 4.51% compared to model 4, and 13.78% compared to model 5. Furthermore, the average prediction time is reduced by 0.79 seconds compared to model 2, 1.37 seconds compared to model 3, 1.01 seconds compared to model 4, and 1.81 seconds compared to model 5. The model constructed by the research institute exhibits excellent predictive performance and achieves efficient and accurate career development predictions.

Precision, recall, and F1 values are widely employed in machine learning to assess classification model performance. These measures are interdependent. The accuracy rate represents the fraction of correctly predicted positive samples in the total predicted positive sample. The recall rate is the percentage of true positive cases

correctly identified as positive. F1 score represents the balance between precision rate and recall rate, providing a comprehensive evaluation of the model's performance. In machine learning, it is crucial to consider both accuracy and recall for optimal classification performance. Focusing on one of these metrics exclusively may result in underperformance of the model in certain scenarios. For instance, if emphasis is solely placed on accuracy, it can lead to overlooking several true positive samples. Similarly, if only recall rates are given importance, there may be incorrect reporting of multiple negative samples. Thus, it is imperative to weigh the accuracy rate and recall rate to obtain an improved F1 value. Based on an analysis of various dimensions, the results demonstrate that the model significantly enhances the matching of entities in the K-Means algorithm. Compared to the other three models, the enhanced K-Means model exhibits superior performance in accuracy rate, recall rate, F1 value, and running time. Therefore, this model demonstrates stronger stability and higher efficiency when handling large-scale datasets. Comparative analysis reveals that the traditional K-Means algorithm has the least satisfactory performance in entity matching. The traditional K-Means algorithm is often impacted by noisy data and outliers, leading to subpar clustering outcomes, particularly when working with large data sets. By incorporating the weight mechanism, the enhanced K-Means algorithm accounts for the similarity among samples, leading to improved accuracy and stability of clustering. Moreover, despite the satisfactory accuracy and recall rates demonstrated by the KNN algorithm, its F1 value remains relatively low. This suggests that the KNN model may suffer from overfitting or underfitting issues when dealing with large-scale data sets. The C-means algorithm generally performs well in terms of accuracy and recall rates, but the F1 value is relatively high. The C-means model demonstrates stability and accuracy when handling large-scale data sets. The improved K-Means algorithm offers significant advantages in entity matching. In comparison to the other three models, the improved K-Means algorithm exhibits superior performance in accuracy rate, recall rate, F1 value, and running time. Therefore, in practical applications, the enhanced K-Means algorithm could enhance entity matching and improve the efficiency and accuracy of data processing.

5. Conclusion. In the field of English language teacher career development, this paper presents a data mining and time series model-based predictive approach for English teacher career growth. This study aims to thoroughly examine the current career status of English teachers, including individual and organizational factors that influence it. The K-means algorithm has been enhanced and utilized to implement cluster analysis. Next, time series features are constructed and inputted into an LSTM to develop a time series model. Finally, SVM is employed to predict the career progression direction of English teachers. The study revealed that life stage and teaching age were the most important factors influencing English teachers' career development, and their weight ratios were 0.462 and 0.354, respectively. These results showed that English teachers' career development had different characteristics and needs at different stages of their life cycle. Organizational factors like school rules and regulations had little influence on the CDET, with a weight ratio of 0.123. This could be attributed to the fact that the policies and regulations implemented in most schools tended to be more humane and had relatively little impact on the CDET. During the training of the prediction model, the error between the actual sample data and the predicted sample data was 0.032. When compared to the conventional K-means model, the enhanced K-means model demonstrated a 28% surge in precision, achieving 98.58%. This suggested that the model had a high degree of accuracy. These results indicated that the enhanced K-means model performs better in forecasting the career progression of English teachers. The study presents a fresh view and technique for promoting the career development of English instructors. It aids in a thorough comprehension of the requirements and traits for career progression and offers tailored policies and proposals for educational institutions and departments. Nevertheless, limitations exist, including scarce data sources and insufficient consideration of other factors that may influence English teacher career development. Future research can expand data sources and consider additional influencing factors to enhance the model's accuracy and applicability.

Although the research model demonstrates high predictive ability for the CDET, certain limitations persist. These limitations are as follows:

1. Limitation of data sources: The study data was obtained from a specific English teacher database, which may not represent all English teachers. Additionally, this study did not consider certain individual and organizational factors that could have a more significant effect on English teacher career development.
2. Limitations of model application: Although the model constructed in the study performs well in training sets, it may encounter uncertainties in forecasting novel situations or instructors. This is due to the

fact that models are trained using existing data and may lack sufficient information to make predictions about new scenarios or instructors.

3. Limitations of the improved K-Means algorithm: Although the enhanced K-Means algorithm performs well in cluster analysis, it exhibits weak outlier processing capabilities. The presence of outliers can significantly impact the clustering outcomes, ultimately affecting the model's overall performance.
4. Failure to consider individual differences of teachers: The study did not give full consideration to the individual differences among teachers, including their personal background, educational philosophy, and teaching style. These differences could potentially impact the career development of teachers in important ways.
5. Lack of long-term observation: The brief observation period of this study limits its ability to capture the long-term changes and trends in English teacher career development. A longer observation period could reveal more about the rules and factors influencing teacher career development. To address these limitations, future research can refine data collection methods and broaden data sources to enhance the model's accuracy in capturing the real-world conditions of diverse English teachers. Additionally, alternative clustering algorithms or ensemble learning techniques may be employed to upgrade the overall efficacy of the model. In addition, the influence of individual teacher differences on career development should be further explored in order to provide teachers with more targeted career development suggestions and planning.

Fundings. The research is supported by: Provincial Project, Teaching Design and Practice of Ideological and Political Case English Writing Based on POA Concept, (NO., JXJG-22-36-2).

REFERENCES

[1] Dey, L. & Mukhopadhyay, A. Biclustering-based association rule mining approach for predicting cancer-associated protein interactions. *IET Systems Biology*. **13**, 234-242 (2019)

[2] Eriya, K., Nugrahani, F. & Ghosh, A. Recommendation system using hybrid collaborative filtering methods for community searching. *Journal Of Physics: Conference Series*. **17**, 27-35 (2019)

[3] Jiang, W., Liu, P. & Wen, F. Speech Magnitude Spectrum Reconstruction from MFCCs Using Deep Neural Network. *Chinese Journal Of Electronics*. **3**, 42-47 (2018)

[4] Hai-Tao, L. & Yuan, S. Corrosion prediction of marine engineering materials based on genetic algorithm and BP neural network. *Marine Sciences*. **44**, 33-38 (2021)

[5] Zhou, K., Lin, W., Sun, J., Zhang, J., Zhang, D. & Feng, X. Prediction model of end-point phosphorus content for BOF based on monotone-constrained BP neural network. *Journal Of Iron And Steel Research International*. **29**, 751-760 (2022)

[6] Liu, M., Yao, D., Guo, J. & Chen, J. An Optimized Neural Network Prediction Model for Reservoir Porosity Based on Improved Shuffled Frog Leaping Algorithm. *International Journal Of Computational Intelligence Systems*. **15**, 11-19 (2022)

[7] Solodovnik, D., Tatonova, Y., Urabe, M., Besprozvannykh, V. & Inoue, K. Three species of Exorchis Kobayashi, 1921 (Digenea: Cryptogonimidae) in the East-Asian region: Morphological and molecular data. *Parasitology*. **148**, 1578-1587 (2021)

[8] Liu, M., Zhang, B., Li, X., Tang, W. & GQ., Z. An Optimized k-means Algorithm Based on Information Entropy. *The Computer Journal*. **64**, 1130-1143 (2021)

[9] Tim, A., Cao, X., Chen, D., Manuel, C., Edwards, M., Andrew, E., Stefano, F., Justin, G., Luke, G. & Mining, A. from information to knowledge: general discussion. *Faraday Discussions*. **233** pp. 58-76 (2022)

[10] Zou, C., Li, J., Wang, W., Zhang, Y. & Xu, D. Integrating data mining and machine learning to discover high-strength ductile titanium alloys. *Acta Materialia*. **202** pp. 211-221 (2021)

[11] Feng, Z., Dai, Z., Zhao, C., Jin, K., Shen, Q., Sun, R., Zhang, W., Yang, S., Wang, X. & Ning, S. Novel putative pathogenic viruses identified in pangolins by mining metagenomic data. *Journal Of Medical Virology*. **94**, 2500-2509 (2022)

[12] Rahman, R., Shabab, K., Roy, K., Zaki, M. & Hasan, S. Real-Time Twitter Data Mining Approach to Infer User Perception Toward Active Mobility. *Transportation Research Record*. **2675**, 947-960 (2021)

[13] He, Y., Chu, Y., Song, Y., Liu, M., Shi, S. & Chen, X. Analysis of design strategy of energy efficient buildings based on databases by using data mining and statistical metrics approach. *Energy And Buildings*. **258**, 1-11181 (2022)

[14] Parashkooh, H. & Jian, C. Data Mining Guided Molecular Investigations on the Coalescence of Water-in-Oil Droplets. 2022. (0)

[15] Chen, H., Yin, C., Li, R., Rong, W., Xiong, Z. & David, B. Enhanced learning resource recommendation based on online learning style model. *Tsinghua Science And Technology*. **25**, 348-356 (2020)

[16] Ta, X., Liu, Z., Hu, X., Yu, L., Sun, L. & Du, B. Adaptive Spatio-temporal Graph Neural Network for traffic forecasting. *Knowledge-based Systems*. **3**, 242-251 (2022)

[17] Wang, H., Song, L., Liu, J. & Xiang, T. An efficient intelligent data fusion algorithm for wireless sensor network. *Procedia Computer Science*. **183**, 418-424 (2021)
[18] Bollé, D. & Blanco, J. The Blume-Emery-Griffiths neural network with synchronous updating and variable dilution. *European Physical Journal, B*. **47**, 281-290 (2021)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Nov 1, 2023
Accepted: Dec 12, 2023