On the Study of Small Class Sizes of Primary School Using Machine Learning

Tian Litong(✉)

Department of Mathematics, Wenzhou Kean University Yingkou, Yingkou, China
1098422@wku.edu.cn

Abstract. The purpose of this study is to explore the phenomenon of small classes size in elementary school education in recent years and the differences in the implementation of small-class size education policies in different regions in China. This essay also talked about the motivation, background information and advantages of small class size in Primary Education. Besides, the study use data visualization to analyze the number of primary schools and the number of teachers in 34 provinces in China from 1978 to 2020. Furthermore, the exploration is also focused on the analysis of full-time teachers, enrolments, and their relationship in 34 provinces in China. The last but not the least, this study uses machine learning and other skills to make a prediction model, such as K- NearestNeighbor(KNN). In conclusion, the government should take measures to narrow the gap of educational resources in China.

Keywords: small class size · primary school · data visualization · cluster · K-NearestNeighbor(KNN) prediction model

1 Introduction

With the negative growth of the birth population, the abundance of educational resources, and the increasing expectations of parents for their children's high-quality education, small-class education, known as "exquisite education", has gradually emerged in some large and medium- sized cities and economically developed regions in China. Since students in primary schools have poor concentration ability and self-management ability, its motivation is to pay attention to individual differences of students [2]. Small class teaching has become a new education reform, which is a hot spots of exploration. The purpose of this paper is to explore the implementation degree of small class education policy in defferent regions of China. The remainder of this paper will use data visualization, including line chart, area chart, etc.. And machine learning, such as Cluster and KNN prediction model to analyze the policy of small class size of primary school in China specifically.

Z. Zhan et al. (Eds.): ICBDIE 2022, AHCS 5, pp. 1148–1157, 2023.
https://doi.org/10.2991/978-94-6463-034-3_118

2 Analyze the Phenomenon of Small Class Siza in Elementary School Education in Recent Years Through Data Visualization

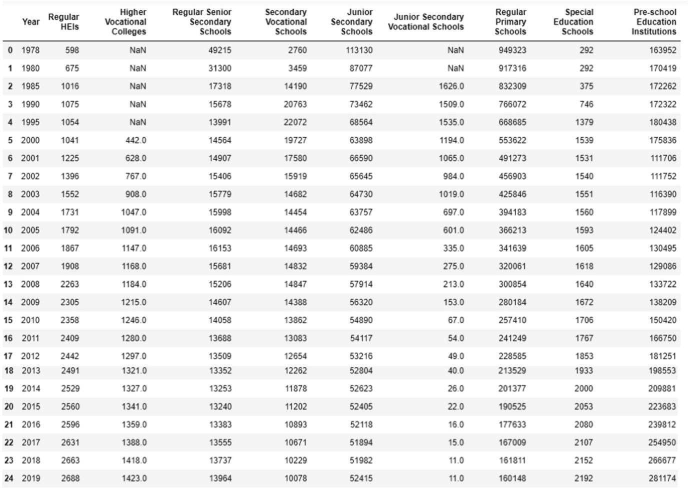
2.1 Sorting and Analyzing Data

We firstly import the excel file of "Number of Schools by Type and Level" and "Number of Full-time Teachers of Schools by Type and Level" into the Jupyter notebook and get the data summarization. After sorting the data, we get:

We have such large data; so we pay more attention to the number of teachers and students at the primary education stage, because primary school students have poor concentration ability, self-management ability, etc.. [1] And the teachers claimed that smaller classes increased the amount of individual attention that pupils received [4].

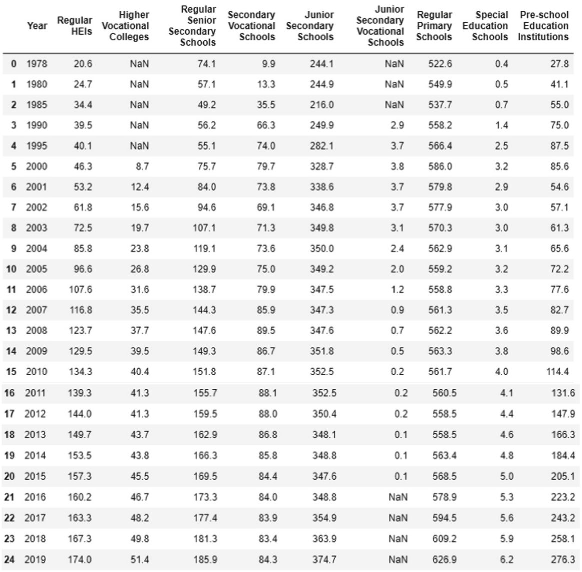
2.2 Data Prepossessing and Mapping

We use data visualization to analyze the phenomenon of small classes in elementary school education in recent years, including line chart, box diagram, bar chart, pie chart, and area chart. According to the line chart of schools and teachers from 1980 to 2020, it can be seen that the teacher-student ratio in elementary schools is increasing year by year. Similarly, the box diagram, bar chart, and pie chart also show that the number of teachers is increasing year by year, and the number of schools is decreasing year by year. Thus, from the data point of view, this matches the full implementation of the small class system in primary schools by the Ministry of Education. Besides, preschool education has the same situation as elementary education, which is that the number of schools for pre-school education is decreasing, but the number of teachers is increasing. It could be suggested by the area chart below:



	Year	Regular HEIs	Higher Vocational Colleges	Regular Senior Secondary Schools	Secondary Vocational Schools	Junior Secondary Schools	Junior Secondary Vocational Schools	Regular Primary Schools	Special Education Schools	Pre-school Education Institutions
0	1978	598	NaN	49215	2760	113130	NaN	949323	292	163952
1	1980	675	NaN	31300	3459	87077	NaN	917316	292	170419
2	1985	1016	NaN	17318	14190	77529	1626.0	832309	375	172262
3	1990	1075	NaN	15678	20763	73462	1509.0	766072	746	172322
4	1995	1054	NaN	13991	22072	68564	1535.0	668685	1379	180438
5	2000	1041	442.0	14564	19727	63898	1194.0	553622	1539	175836
6	2001	1225	628.0	14907	17580	66590	1065.0	491273	1531	111706
7	2002	1396	767.0	15406	15919	65645	984.0	456903	1540	111752
8	2003	1552	908.0	15779	14682	64730	1019.0	425846	1551	116390
9	2004	1731	1047.0	15998	14454	63757	697.0	394183	1560	117899
10	2005	1792	1091.0	16092	14466	62486	601.0	366213	1593	124402
11	2006	1867	1147.0	16153	14693	60885	335.0	341639	1605	130495
12	2007	1908	1168.0	15681	14832	59384	275.0	320061	1618	129086
13	2008	2263	1184.0	15206	14847	57914	213.0	300854	1640	133722
14	2009	2305	1215.0	14607	14388	56320	153.0	280184	1672	138209
15	2010	2358	1246.0	14058	13862	54890	67.0	257410	1706	150420
16	2011	2409	1280.0	13688	13083	54117	54.0	241249	1767	166750
17	2012	2442	1297.0	13509	12654	53216	49.0	228585	1853	181251
18	2013	2491	1321.0	13352	12262	52804	40.0	213529	1933	198553
19	2014	2529	1327.0	13253	11878	52623	26.0	201377	2000	209881
20	2015	2560	1341.0	13240	11202	52405	22.0	190525	2053	223683
21	2016	2596	1359.0	13383	10893	52118	16.0	177633	2080	239812
22	2017	2631	1388.0	13555	10671	51894	15.0	167009	2107	254950
23	2018	2663	1418.0	13737	10229	51982	11.0	161811	2152	266677
24	2019	2688	1423.0	13964	10078	52415	11.0	160148	2192	281174

Fig. 1. Snapshot of the data of "Number of Schools by Type and Level" (Original).



	Year	Regular HEIs	Higher Vocational Colleges	Regular Senior Secondary Schools	Secondary Vocational Schools	Junior Secondary Schools	Junior Secondary Vocational Schools	Regular Primary Schools	Special Education Schools	Pre-school Education Institutions
0	1978	20.6	NaN	74.1	9.9	244.1	NaN	522.6	0.4	27.8
1	1980	24.7	NaN	57.1	13.3	244.9	NaN	549.9	0.5	41.1
2	1985	34.4	NaN	49.2	35.5	216.0	NaN	537.7	0.7	55.0
3	1990	39.5	NaN	56.2	66.3	249.9	2.9	558.2	1.4	75.0
4	1995	40.1	NaN	55.1	74.0	282.1	3.7	566.4	2.5	87.5
5	2000	46.3	8.7	75.7	79.7	328.7	3.8	586.0	3.2	85.6
6	2001	53.2	12.4	84.0	73.8	338.6	3.7	579.8	2.9	54.6
7	2002	61.8	15.6	94.6	69.1	346.8	3.7	577.9	3.0	57.1
8	2003	72.5	19.7	107.1	71.3	349.8	3.1	570.3	3.0	61.3
9	2004	85.8	23.8	119.1	73.6	350.0	2.4	562.9	3.1	65.6
10	2005	96.6	26.8	129.9	75.0	349.2	2.0	559.2	3.2	72.2
11	2006	107.6	31.6	138.7	79.9	347.5	1.2	558.8	3.3	77.6
12	2007	116.8	35.5	144.3	85.9	347.3	0.9	561.3	3.5	82.7
13	2008	123.7	37.7	147.6	89.5	347.6	0.7	562.2	3.6	89.9
14	2009	129.5	39.5	149.3	86.7	351.8	0.5	563.3	3.8	98.6
15	2010	134.3	40.4	151.8	87.1	352.5	0.2	561.7	4.0	114.4
16	2011	139.3	41.3	155.7	88.1	352.5	0.2	560.5	4.1	131.6
17	2012	144.0	41.3	159.5	88.0	350.4	0.2	558.5	4.4	147.9
18	2013	149.7	43.7	162.9	86.8	348.1	0.1	558.5	4.6	166.3
19	2014	153.5	43.8	166.3	85.8	348.8	0.1	563.4	4.8	184.4
20	2015	157.3	45.5	169.5	84.4	347.6	0.1	568.5	5.0	205.1
21	2016	160.2	46.7	173.3	84.0	348.8	NaN	578.9	5.3	223.2
22	2017	163.3	48.2	177.4	83.9	354.9	NaN	594.5	5.6	243.2
23	2018	167.3	49.8	181.3	83.4	363.9	NaN	609.2	5.9	258.1
24	2019	174.0	51.4	185.9	84.3	374.7	NaN	626.9	6.2	276.3

Fig. 2. Snapshot of the data of "Number of Full-time Teachers of Schools by Type and Level" (Original).

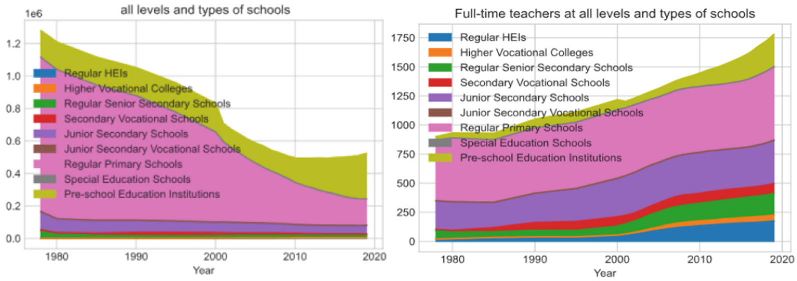


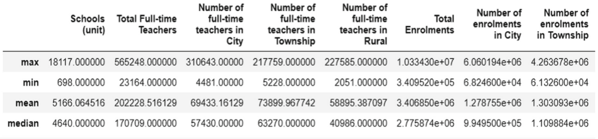
Fig. 3. Area chart of pre-school education institutions and full-time teachers from 1980 to 2020

Thus it could be seen that the small class of preschool education can lay a good foundation for the small class of primary education, and enable students to adapt to the policy of small class of primary education as soon as possible.

3 The Analysis of Full-Time Teachers, Enrolments, and Their Relationship in 34 Provinces in China by Data Visualization and Machine Learning

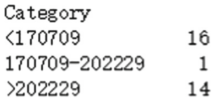
3.1 Data Collating and Category

We firstly import the excel file of "Statistics on Regular Primary Schools by Region (2019)" into the Jupyter notebook and get the data summarization. We have such large



	Schools (unit)	Total Full-time Teachers	Number of full-time teachers in City	Number of full-time teachers in Township	Number of full-time teachers in Rural	Total Enrolments	Number of enrolments in City	Number of enrolments in Township
max	18117.000000	565248.000000	310643.00000	217759.000000	227585.000000	1.033430e+07	6.060194e+06	4.263678e+06
min	698.000000	23164.000000	4481.00000	5228.000000	2051.000000	3.409520e+05	6.824600e+04	6.132600e+04
mean	5166.064516	202228.516129	69433.16129	73899.967742	58895.387097	3.406850e+06	1.278755e+06	1.303093e+06
median	4640.000000	170709.000000	57430.00000	63270.000000	40986.000000	2.775874e+06	9.949500e+05	1.109884e+06

Fig. 4. Summary, max, min, mean, median of the data (Original).



```
Category
<170709            16
170709-202229       1
>202229            14
```

Fig. 5. Category of Regular Primary Schools by Region (2019)

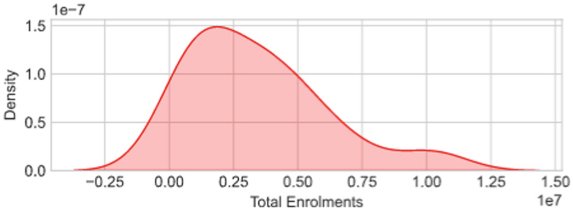


Fig. 6. Density chart of total full time teachers

data; we firstly filtrate which provinces have worse educational resources in primary schools and which provinces have better educational resources in primary schools. So we acquire the max, min, mean and median data of these provinces.

Then we pay more attention to the "Total Full-time Teachers" in 34 provinces in China. According to the mean value and median value, we do the category method. We categorize full-time teachers in 34 provinces in China based on median and average values:

It could be seen that in 52% of the provinces, the number of teachers is below the median, and in 47% of the provinces, the number of teachers is above the average. There is one province, Fujian, where the number of teachers is between the median and average values.

3.2 Data Prepossessing and Mapping

- Density chart of teacher resource analysis.
  In order to understand the distribution of the data density of the number of full-time teachers and students in school, use the density map to analyze:

  According to the chart, the number of teachers with less than 200000 has the highest probability density, and the number of pupils with less than 2.5 million has the highest probability density. This also proves that the number of full-time teachers in primary education in most provinces in China is below the average value, which is 202229. Therefore, the education department should narrow the gap in education resources as soon as possible.

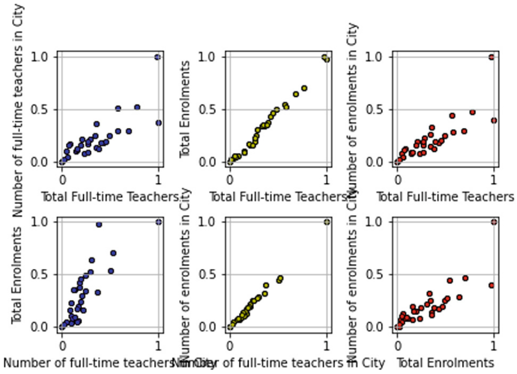


Fig. 7. Scatter plot of total full time teachers and total enrolments

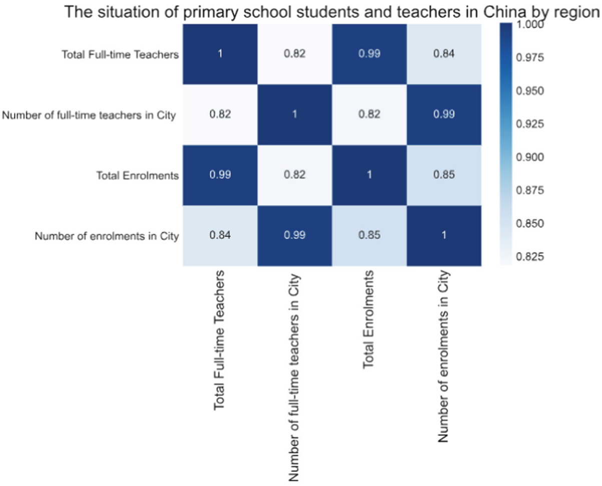


Fig. 8. Correlation coefficient thermal diagram

- Scatter plot

  We perform regression analysis on the number of full- time teachers and students. Observe the general trend of the number of full-time teachers with the number of enrolments and the general trend of the number of enrolments with the number of full-time teachers to judge the relationship between the two variables.

  From the data visualization, it could be seen that both the relationship between the total number of enrolments and total teachers in elementary schools and the relationship between the number of enrolments and teachers in the city are all proportional to each other.

- Heat map

  We also use the heat map to observe the similarity of multiple features in the data table:

It can be seen from the heat map that the correlation between the total number of teachers and the total number of students is the same as the correlation between the number of teachers in the city and the number of students in the city, and both are very high, which is 0.99.

3.3 Machine Learning - Cluster

We present a methodology, based on machine learning, which can break the trace down into clusters of traffic where each cluster has different traffic characteristics [5]. In order to eliminate the influence of different variable units on the clustering results, we should first standardize all data, using the formula:

$$X_{ij} = \frac{\mathrm{x}_{ij} - \overline{x_j}}{S_J} \tag{1}$$

Then calculate the "distance" between the objects to get the "similar relationship" matrix between the objects. And "Distance" has following expressions:

The square of the Euclidean distance:

$$\mathrm{r}_{ij}^2 = \frac{1}{p}\sum_{k=1}^{p}(x_{ik} - x_{jk})^2 \tag{2}$$

Deviation distance:

$$r_{ij} = \frac{1}{p}\sum_{k=1}^{p}|x_{ik} - x_{jk}| \tag{3}$$

Correlation coefficient

$$r_{ij} = \frac{\sum\limits_{k=1}^{p}(x_{ik} - \overline{x_i})(x_{jk} - x_j)}{\sqrt{\sum\limits_{k=1}^{p}(x_{ik} - \overline{x_i})^2}\sqrt{\sum\limits_{k=1}^{p}(x_{jk} - \overline{x_j})^2}} \tag{4}$$

The distance between classes can be represented by the distance between "representative points". Finally, clustering: treat each point or object as a category, and then find the d_{ij} with the smallest or largest distance, so as to obtain the two categories i and j with the closest or the farthest distance, and merge them into a higher category. This is repeated until all the points are merged into one category. And according to the four data characteristics, respecively, "Number of Full-time Teachers in rural", "Number of full-time teachers in City", "Number of Enrolments in rural", and "Number of enrolments in City", the educational resources of small class teaching in primary schools in 34 provinces are divide into three groups:

This shows that the level of small class teaching of primary education development varies greatly in China. The level of educational development in economically developed areas is relatively high.

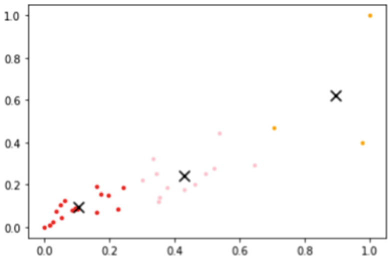


Fig. 9. Cluster

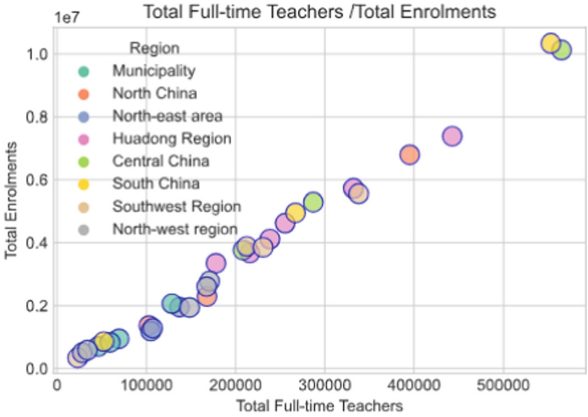


Fig. 10. Scatter plot of full-time teachers and enrolments of seven regions in China

4 Analyze the Implementation of Small Classes In Primary Education in China by Region

In order to better understand the implementation of the policy of small-class primary education in different regions of China, we expand and merge the data, and divide China's 34 provinces and cities into 7 regions, respectively, Northwest China, North China, Northeast China, Southwest China, East China, South Central China, and municipality.

In 7 regions of China, both the relationship between the total number of students and total teachers and the relationship between the number of students and teachers in the city are all positive correlation, which could be seen from the following picture:

The data also shows that the difference of educational development level among the provinces in the eastern region shows an inverted U-shaped trend, which is firstly rising and then declining, and the difference of educational development level between the provinces in the central and western regions also tends to narrow after a small fluctuation.

4.1 Machine learning - K- NearestNeighbor(KNN) prediction model

Because the data in machine learning needs to be sorted into the form of vectors. Therefore, in order to make prediction model, it must first be processed and transformed into a vector in order to meet the needs of prediction. The vector can be regarded as a sequence x, which is generally represented by $\{a_1\ (x), a_2\ (x), \ldots, a_n\ (x)\}$, where $a_i(x)$ represents the i-th component of the vector x. After the data is sorted into vector form, it generally has the following form: $\{a_1\ (x), a_2\ (x), \ldots, a_n\ (x); y\}$, where the first m items represent m different characteristics of the value, and the last one the item y represents the classification or target value of the data.

Educational resources are distinguished by different regions in China. We have data on the number of full-time teachers and students in each region, which is a data sequence. Combined with the prediction method, the value of educational resources can be expressed as $\{a_1\ (e), a_2\ (e), \ldots, a_n\ (e); y)$, where the first m items are the characteristics of teachers and students in a certain region. The last item y represents the region. In this way, if m characteristics of educational resources are obtained, it can be predicted which region of 7 regions in China belongs to.

Form a sequence $\{a_1, a_2, \ldots, a_n\}$ with the number of teachers and students in the known 7 regions, and use this set of data to predict which region it belongs to given the data about the number of teachers and students.

Use a vector of length m, which is $\beta_0 = \{a_{n\text{-}m+1}, a_{n\text{-}m+2}, \ldots, a_{n\text{-}1}, a_n\}$ to predict y_n. Since y_n is unknown, firstly find $\beta_0 = \{a_{n\text{-}m+1}, a_{n\text{-}m+2}, \ldots, a_{n\text{-}1}, a_n\}$'s K nearest neighbors. In $\{a_1, a_2, \ldots, a_n\}$, $\beta_0 = \{a_{n\text{-}m+1}, a_{n\text{-}m+2}, \ldots, a_{n\text{-}1}, a_n\}$ as the base, and sequentially take (n-m) sub-columns of length m: $\beta_1 = \{a_{n\text{-}m}, a_{n\text{-}m+1}, \ldots, a_{n\text{-}1}\}$, $\beta_2 = \{a_{n\text{-}m\text{-}1}, a_{n\text{-}m}, \ldots, a_{n\text{-}2}\}$, …, $\beta_{n\text{-}m\text{-}1} = \{a_2, a_3, \ldots, a_{m+1}\}$, $\beta_{n\text{-}m} = \{a_1, a_2, \ldots, a_m\}$, in these sub-columns, find the K nearest neighbors of $\beta_0 = \{a_{n\text{-}m+1}, a_{n\text{-}m+2}, \ldots, a_{n\text{-}1}, a_n\}$, and use the cosine of the angle between the two vectors to describe the proximity of the two vectors, that is,

$$\cos(\beta, \beta_i) = \frac{\beta\beta_i}{|\beta||\beta_i|} = \frac{\sum\limits_{j=1}^{m} \beta_{0j}\beta_{ij}}{\sum\limits_{j=1}^{m} \beta_{0j}^2 \sum\limits_{j=1}^{m} \beta_{ij}^2} \quad (i = 1,\ 2,\ \ldots \text{n - m}) \tag{5}$$

where β_{ij} represents the jth component of the vector β_i. Then, the greater the cosine value, the closer the two vectors are. Through calculation, we find $\beta_0 = \{a_{n\text{-}m+1}, a_{n\text{-}m+2}, \ldots, a_{n\text{-}1}, a_n\}$'s K nearest neighbors in $\beta_1, \beta_2, \ldots\beta_{n\text{-}m}$, denoted as $\alpha_1, \alpha_2, \ldots, \alpha_k$. Because $\{a_{n\text{-}m+1}, a_{n\text{-}m+2}, \ldots, a_{n\text{-}1}, a_n\}$ is used to predict yn, the element next to the last component of the K vectors is considered to be the nearest neighbor of y_n. For example, if $\alpha_1 = \{a_1, a_2, \ldots, a_m\}$, then a_{m+1} is taken as a nearest neighbor of y_n so that K nearest neighbors of y_n are obtained, which are $b_1, b_2, \ldots, b_k$, and then we can calculate y_n by weighted average of these K numbers, that is, $y_n = \frac{\sum_{i=1}^{k} b_i}{k}$. The key of K nearest neighbor algorithm lies in the definition of similarity function. As mentioned above, the similarity of two points is generally defined as the cosine of the angle between two points. This will lead to a key deficiency, that is, when searching for training samples which are similar to query samples from the sample library, if a small number attributes have a greater impact than other attributes on the classification results, the similarity between neighbors is dominated by a large number of unimportant attributes, which leads to misleading classification.

For the example of educational resources {a_1 (x), a_2 (x), …, a_n (x); y}, because the importance of the first m items will be different, we can try to assign weights the education resource information of the corresponding region. The more similar the data is to the area to be predicted, the greater the weighting. The K nearest neighbors of y_n, b_1, b_2,…, b_k, are sorted according to the similarity degree of data information. The data which is closest to the data of nth region is ranked first, denoted by c_1, followed by c_2, and so on. A sequence c_1, c_2, …, c_k is obtained. For each c_i, a weighted.

$$\mathrm{w_i} = 2^i / \sum_{j=1}^{k} 2^j \text{ and } \sum_{i=1}^{k} w_i = 1 \tag{6}$$

is given. In this way,

$$\mathrm{yn} = \sum_{i=1}^{k} w_i c_i \tag{7}$$

can effectively solve the previous problem. Thus, the improved KNN algorithm as a prediction model has excellent performance [3].

In order to process data more quickly and conveniently, we carried out data normalization and get the result. The accuracy of the prediction model can reach 97.78%, which is a very high accuracy. Therefore, according to the four data information, respectively, "Number of Full-time Teachers in rural", "Number of full-time teachers in City", "Number of Enrolments in rural", and "Number of enrolments in City", we can infer that it belongs to which region in China.

5 Conclusion

This paper firstly analyzes the phenomenon of small classes in elementary school education in recent years, from 1978 to 2020, through data visualization. And we concluded that the teacher-student ratio in elementary schools is increasing year by year in China. Thus small class teaching is the trend of teaching reform and development, and it is also one of the trends of contemporary Chinese social development. Besides, this essay is also focused on the exploration of the analysis of full-time teachers, enrolments, and their relationship in 34 provinces in China by data visualization and machine learning, such as Cluster. One point shows that there is a huge difference about the level of small class size of primary education development in China. Therefore, the government should take measures to narrow this gap as soon as possible. Furthermore, the implementation of small classes of primary education in China by region is explored. And we find that in 7 regions of China, the relationship between the number of students and teachers are all positive correlation. Apart from this, the KNN algorithm is also used to predict the region of educational resources. And the improvement of KNN algorithm is simple in calculation, low in complexity and high in accuracy, which is of great significance in practice.

References

1. Blatchford, P., Bassett, P., & Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. Learning and Instruction, 21(6), 715–730.
2. Brühwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. Learning and Instruction, 21(1), 95–108.
3. Chen, H.-L., Yang, B., Wang, G., Liu, J., Xu, X., Wang, S.-J., & Liu, D.-Y. (2011). A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method. Knowledge-Based Systems, 24(8), 1348–1359.
4. Galton, M., & Pell, T. (2012). Do class size reductions make a difference to classroom practice? The case of Hong Kong primary schools. International Journal of Educational Research, 53, 22–31.
5. McGregor, A., Hall, M., Lorier, P., & Brunskill, J. (2004). Flow Clustering Using Machine Learning Techniques. In C. Barakat & I. Pratt (Eds.), Passive and Active Network Measurement (pp. 205–214). Springer