An Improved Algorithm for School Bullying Based on K-means Clustering

Xinyi Gong

School of Information Science and Sechnology, Fudan University, Shanghai 200433, China.
gongxy14@fudan.edu.cn

Keywords: Component, formatting, Style, Styling, Insert.

Abstract. School bullying has become a very important problem in education circle. In order to discuss this problem in the rational and scientific eye, we can make use of big data analysis that can provide data and decision support. Aiming at the time-varying abstract characteristics, this paper proposes an improved algorithm on school bullying based on K-means clustering, establishes the whole data analysis structure of data collection, data processing and inferential decision, analyzes the data feature of school bullying. It also gains the time, geographic and age characteristic though chi square test, linear regression and logistic regression. On the other hand, behavior data is also processed with K-means clustering algorithm. Time varying characteristic and uncertain factors are greatly reduced through density, grid and models. In the end, data support for the solution of school bullying behavior is offered by means of inferential decision based on neural network.

1. Introduction

Big data analysis, which conducts frequency statistics among a large number of irregular events, is an important analysis method to obtain the trend and regularity of events [1]. With the help of computer, the generalization and analysis of mass data characteristics can be realized and provides effective conditions for human to explore the microcosmic and macro trends.

Being different from common probability statistics, big data analysis uses multi-scale data acquisition equipment, high-efficiency data processing unit and smart network, making the characteristic analysis faster [2], more accurate and pervasive

Therefore, a large number of international large enterprises and research institutions are now highly concerned about big data analysis. IBM, Microsoft, Alibaba, and Baidu, for example, have already launched researches to make data analysis the technical core of Big Data [3]. Also, techniques such as unstructured data processing, highly artificial intelligence and distributed processing architecture, become the trend of big data development gradually.

Bullying on campus refers to the mistreatment and oppression of the unequal rights between students. This phenomenon comes to appear with the development and the opening up of the country. Intensifying these years, it has become a very important problem in educational circle. With the help of Big Data analysis, data and decision support can be provided and help to view the problem in the eye of ration and science [4]. Since bullying on campus varies according to time, space and individuation, Big Data analysis takes uncertainty and redundancy into account.

Lin Zhou from The PLA Information Engineering University proposed a kind of clustering algorithm based on spectral clustering. Using the three-tuple algorithm to calculate similarity matrix, the similarity information between the data points is expanded to avoid the problem of selecting the scale parameters accurately. Xiaodong Yu from Air Force Engineering University put forward an intuitionistic fuzzy nuclear clustering algorithm based on particle swarm optimization. It takes advantage of global search capability and the fast convergence speed of particle swarm optimization algorithm so as to optimize the initial clustering center of the PSO-based intuitionistic fuzzy kernel clustering algorithm, enhance the clustering performance and improve the efficiency of the algorithm. Donghui, Chen advanced a fuzzy clustering algorithm based on target function that is suitable for

high-dimensional arbitrary distributed data sets, extending the application scope of clustering analysis.

To sum up, the clustering method mentioned above can easily ignore the time-varying characteristic and uncertainty of the feature, and therefore resulting in the decline of the clustering accuracy.

Aiming at the time-varying abstract characteristics, this paper proposes a study on school bullying with an improved algorithm based on K-means clustering algorithm, establishes the whole data analysis structure of data collection, data processing and inferential decision, analyzes the data feature of school bullying. It also gains the time, geographic and age characteristic though chi square test, linear regression and logistic regression. On the other hand, behavior data is also processed with K-means clustering algorithm. Time varying characteristic and uncertain factors are greatly reduced through density, grid and models. In the end, data support for the solution of school bullying behavior is offered by means of inferential decision based on neural network.

2. Data characteristics of bullying behavior on campus

2.1 The time characteristics of the chi-square testing of bullying behavior on campus

Time is a very important indicator of human action and has a great impact on the frequency of school bullying [5]. Here we use chi-square test method to analyze the probability of bullying behavior occurring at different periods. The chi-square test method is usually used for the correlation analysis of two or more samples and two classification variables. Its model is as follows.

$$\chi^2 = \Sigma \frac{(A-T)^2}{T} \quad (1)$$

Where A is the observation frequency of the sample, T is the theoretical frequency of the sample [6]. First we can establish the original hypothesis H_0:

（1） People of all age have same psychological condition.

（2） The number of people is same at all periods.

Then we classify the original time data into different periods and name them.

T1：6:00 ~ 8:00.；

T2：8:00 ~ 12:00；

T3：12:00 ~ 13:00；

T4：13:00 ~ 17:00；

T5：17:00 ~ 19:00；

T6：19:00 ~ 24:00；

Input these data into the chi-square test model, we can get the result (table 1) about the frequency of school bullying in different periods.

Table 1 the value of chi-square test at different time period

Period	High	Medium	Low
T_1	0.34	0.36	0.20
T_2	0.02	0.08	0.90
T_3	0.23	0.26	0.51
T_4	0.03	0.09	0.89
T_5	0.21	0.15	0.64
T_6	0.10	0.12	0.78

According to table 1, we can conclude that school bullying is more likely to happen when students are out of the classroom.

2.2 The linear regression of bullying behavior on campus

Although the correlation between school bullying and spatial factor is relatively lower, the distance between campuses and where bullying happens still matters. Linear-regression method is used to analyze the frequency of school bullying [7]. Linear regression model can help to analyze linear regression between continuous dependent variables and continuous independent variables. It

uses the best fitting line to establish relationship between the dependent variable (Y) and the independent variable (X）. The regression model is as follows.

$$Y = aX + b + e \quad (2)$$

Where, a is the rate of data change, b is the initial state of the bullying frequency, and e is the error value of the estimated data. After analyzing and processing the original data, we can gain the following conclusion.

（0~1）means bullying happens on campus

（1~3）means bullying happens around school

（3~5）means bullying happens far from the school

（5~∞）means bullying happens near home

After inputting these data into the linear regression model, the relationship between location and the occurrence of school bullying can be found in table 2.

Table 2 Linear regression values of different geographical intervals

Geo-denote	large	medium	small
（0~1）	0.04	0.06	0.90
（1~3）	0.13	0.19	0.79
（3~5）	0.30	0.24	0.46
（5~∞）	0.12	0.16	0.72

It is very clear that school bullying is more possible to happen where is far from school and home.

2.3 The logistic age characteristics of bullying behavior on campus

Student of different age have different behavior characteristics at school. Here we use age as characteristic quantity to analyze the difference between the bullying behaviors of different age groups through logistic models. Logistic model is mainly utilized to study the relationship between the binary variables or multicomponent dependent variables and a set of independent variables, which is shown by the value of P and OR. The regression model is as follows.

$$P = \frac{1}{1+\exp[-(\beta_0+\beta_1 X_1+\cdots+\beta_m X_m)]} \quad (3)$$

Where β_0 is a constant term, β_1 and β_m are regression coefficient.

The source of the data in this paper is the attachment of 2016 China mathematical modeling network challenge. First we can category the original data into several groups.

8~：students from 8 to 10 years old ；

11~：students of 11 and 12 years old；

13~：students of 13 and 14 years old；

15~：students of 15 and 16 years old；

17~：students from 17 to 20 years old.

In order to improve the accuracy of the data, following hypothesis are put forward.

（1） Data in the attachment is real and effective

（2） Every sample is not defective

（3） The samples are random

As table 3 shows, put the data into logistic model can we achieve a result about the possibility of school bullying among students from different age groups.

Table 3 Logistic regression values of different ages

Age	Strong	Medium	Weak
8~	0.12	0.15	0.73
11~	0.13	0.16	0.71
13~	0.12	0.18	0.70
15~	0.13	0.17	0.70
17~	0.12	0.15	0.73

We can see that the difference between the possibilities of school bullying among students from different age groups exists, however, it is still not clear that age is a very decisive factor.

3. Behavior data processing based on k-means algorithm

K-means is a clustering algorithm based on division, which is simple and fast. For numeric attributes, it also does a great job on data merging. Therefore, this paper will use k-means algorithm to process the data of school bullying so that the redundancy of the data can be reduced.

3.1 Multi-scale data standard conversion

Since the data of school bullying varies from time, space and personal characters, multiscale standard transformations are required when we have to consider it as a whole. Because of the different units of multiscale data, small-scale unit is usually used [8]. Thus, the range of properties can be better contractile.

If we transfer the original data into non-unit variable and a property a is given, through calculating the mean of absolute deviation we can get

$$S_a = \frac{1}{n}(\sum_{i=1}^{i=n}\sqrt{(x_{na} - m_a)^2}) \quad (4)$$

Where $m_a = 1/n(\sum_{i=1}^{i=n} x_{na})$, and the standard measure of calculation is

$$x_{na} = \frac{x'_{na}}{s_a} \quad (5)$$

3.2 Determine the original clustering center based on Density aggregation factor

In order to deal with the different features of samples, we can use the following formulation to calculate the local density of samples.

$$\rho_i = \sum_{j=0}^{\infty} X(d_{ij} - d_c) \quad (6)$$

Here

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \cdots (x_{in} - x_{jl})^2} \quad (7)$$

d_{ij} shows the distance between the ith sample and the jth sample. Each sample has l attributes. d_c is a truncation distance. Therefore, ρ_i means the number of points the distance between point i and which is less than d_c.

3.3 Data clustering by k-means

Suppose that the data features of school bullying can be abstracted to n data sets, then

$$x = \{x_1, x_2, \ldots, x_n\}. \quad (8)$$

Here each data can own q attributes:

$$x_j = \{x_{j1}, x_{j2}, \ldots x_{jq}\} \epsilon R_q \quad (9)$$

Using the k-means algorithm, the target data group can be divided into k clusters.

Step 1: Initially there are k' original cluster centers

$$C = \{c_1, c_2, \ldots, c_n\} \quad (10)$$

Decide the maximum number of iterations of clustering and the minimum target function when the iteration ends.

Step 2: According to Euclidean distance formula, the distance from every data point to the clusters can be worked out [9]. In order to distribute all the data points to the cluster which is the nearest, we can use the formulation

$$d(x_j, c_n) = \sqrt{\sum_{n=1}^{l}(x_{j1} - c_{n1})^2} \quad (11)$$

Where $d(x_j, c_n)$ is the distance between the jth data point and the nth cluster center.

Step 3: Calculate the center of k' clusters once again.

$$C = \{m_1, m_2, \ldots, m_n\} \quad (12)$$

The formulation is as follows.

$$m_j = \frac{1}{n}\sum_{x_j \epsilon c_n} x_j \quad (13)$$

Here $m_j$is the center of the jth cluster.

Step 4: After iterating for m times, clustering ends. Otherwise, we should judge whether the result of clustering is smaller than the given parameter T. Clustering ends as well if yes. Or else, step 1 and 2 should be repeated.

Picture 1 shows the clustering condition of this moment.

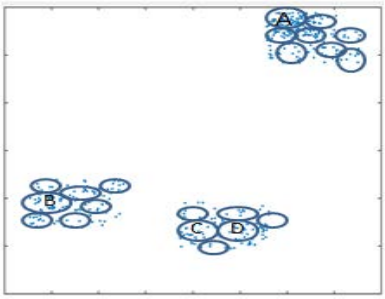

Figure 1 Clustering of bullying data on campus

4. Reasoning on campus bullying based on neural network

Campus bullying is a time-varying event, as well as personalized time-varying events. Through data collecting and clustering, we can get characteristic data behavior. Then, with the help of the neural network, the data is further analyzed in order to obtain the trend of bullying behavior.

Neural network is a kind of machine learning technology that simulates human brain to realize artificial intelligence to some extent. Figure 2 shows its basic structure.

4.1 Input layer design

The neural network changes the network's connection weight continuously under the stimulation of the external input sample, so that the output of the network gradually approaches the expected output. Therefore, it has a similar trajectory to the data on bullying behavior on campus.

Assume that the input layer has n neurons and create an associative matrix for different school bullying data. Also, give each connection weight a random value in the range of (-1, 1) and set the error function e. If given the calculation accuracy of ε and maximum learning times M, choose the kth input sample and its corresponding expected output, we can get

$$Y(k) = (y_1(k), y_2(k), \dots y_n(k)) \quad (14)$$

$$(k) = (d_1(k), d_2(k), \dots d_n(k)) \quad (15)$$

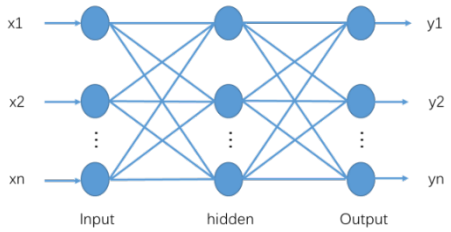

Figure 2 Basic structure of neural network

4.2 Hidden layer design

In order to calculate the input and output of each neuron in the hidden layer, first we assume the number of the hidden layer units of neural network is M and M equals to one. After several learning iterations, stop the iteration if the error condition is satisfied [10]. After it achieves the maximum number of learning times, if the error condition still cannot be satisfied, then the number of transformation units is increased by 1, and the above process is repeated until the error condition is satisfied. In this way, the number of transformation units is able to be determined automatically according to every specific problem. Finally using the network expectation output and actual output, we can calculate the partial derivative of each neuron in the output layer.

4.3 Output layer design

The partial derivatives of error function and each neuron in the hidden layer are calculated by using the connection weight from implicit layer to output layer and the output of these two layers. At the same time, the connection weights are corrected by referring to the output of each neuron in the output layer and the output of each neuron. With the result of global error, we can tell whether the error of network meets the requirement [11]. When the error reaches the preset accuracy or the maximum number of learning times is greater than the set, the algorithm should be ended. Otherwise,

select the next learning sample and the corresponding expected output, return to step 3, and enter the next round of learning. Therefore, to which degree the predicted probability of occurrence of different geographic ranges in campus bullying behavior matches the actual value can be obtained. The result is as figure 3 shows.

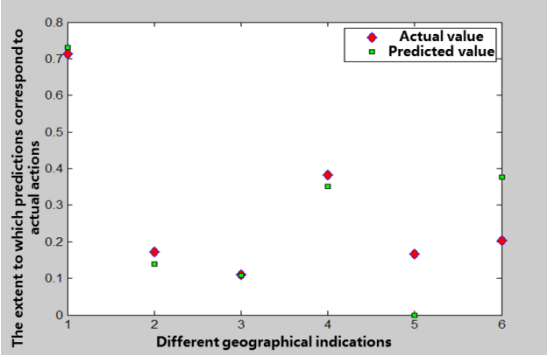

Figure 3 Degree of concept conformity in different geographical intervals

5. Conclusion

By studying the large data k-means clustering algorithm on campus bullying behavior, the data acquisition, data processing and inference decision-making structure of big data analysis can be obtained. This paper analyzes the data characteristics of bullying behavior on campus, and constructs the time characteristics of chi-square method, linear regression characteristics and logistic age characteristics of the card on campus bullying behavior. The behavior data processing based on k-means algorithm is proposed, which can effectively reduce the time-varying characteristics and uncertainties by density, grid and model.

Reference

[1] Jain, Anil K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 2010, 31(8):651-666.
[2] Hartigan J A, Wong M A. A K-means clustering algorithm. Applied Statistics, 1979, 28(1):100-108.
[3] Z. Deng, W. Lin, et al. The uncertainty entropy of low-rate speech quality evaluation and the analyses of the gray correlation. IEICE Electronics express. 2015, 12(3): 20141019.
[4] Ostrovsky R, Rabani Y, Schulman L J, et al. The Effectiveness of Lloyd-Type Methods for the k-Means Problem. Journal of the ACM (JACM), 2012, 59(6):28.
[5] Celebi M E, Kingravi H A, Vela P A. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications, 2013, 40(1):200-210.
[6] Liu X M, Lei D. An improved K-Means clustering algorithm. Microcomputer Information, 2014, 9(1):44-46.
[7] He K, Wen F, Sun J. K-Means Hashing: An Affinity-Preserving Quantization Method for Learning Binary Compact Codes. Computer Vision and Pattern Recognition. IEEE, 2013:2938-2945.
[8] Coates A, Ng A Y. Learning Feature Representations with K-Means. Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, 2012:561-580.
[9] Kulis B, Jordan M I. Revisiting k-means: New Algorithms via Bayesian Nonparametrics. 2011.
[10] W. Lin, Z. Deng. Dimensional functional differential convergence for Cramer-Rao lower bound. Journal of difference equations and applications. 2017, 23(1-2):249-257.
[11] Wang J, Su X. An improved K-Means clustering algorithm// IEEE, International Conference on Communication Software and Networks. IEEE, 2011:39-43.