Research on the Application of Big Data Technology in Information Statistics Research System

Yuhong Hao a, Ying Huang b, * and Yu Feng c

School of Xi'an Traffic Engineering Institute, Xi'an, Shanxi 710300, China;

a 156472975@qq.com, *, b Corresponding author Email: Yuhong@xjgyedu.edu.cn,
c 1602350301@qq.com

Abstract. With the continuous development of computing science and technology, the application of big data technology is very popular, especially in the field of scientific research. The content of mathematical information science fully reflects that mathematics is an important part of scientific research. This article adopts big data technology, which has certain educational significance for training scientific research staff to analyze sports technology of mathematics. The application of big data teaching has enabled scientific researchers to have a new understanding of mathematics, stimulated their interest in mathematics research, and gradually realized that mathematics information science is the basic skill of modern scientists, thus inspiring the continuous deepening of research mathematics Scientific research work.

Keywords. Big data applications, computing science, Mathematical Statistics, Scientific Method.

1. Introduction
With the continuous development of big data technology, the division of disciplines is very obvious. It has always been considered that literature and science are relative, and the temperaments of the people they cultivated and nurtured are quite different. Everyone thinks that Neo-Confucianism is rigorous and it is difficult to express beautiful things, let alone any connection with literature. In the process of entering Neo-Confucianism, I have always thought that people who are infiltrated with numbers should have a more rigid and rough understanding of literature; when words encounter numbers, it should be a difficult encounter. But when I listened to Professor Chen Dakang’s anecdote about the Dream of Red Mansions, I suddenly felt that the inspiration from mathematics could be so interesting. When words encounter numbers, it may not be an encounter, but a beautiful encounter [1]. Everyone thinks it cannot be quantified. In fact, it can also be quantified through statistics to be qualitative.

2. Mathematical Statistics Methodology

2.1. Adjust the statistical correlation of mutual information random variables
Through the application of big data technology, Mutual information (MI) is used to express the relationship between information. It is a measure of statistical correlation between random variables. Its definition is similar to cross entropy. It is used to measure the degree of agreement between the two data
distributions and is used as the result of data clustering [2]. Figure 1 shows the process of adjusting the mutual information algorithm. Assuming that and are the distributions of samples, the entropy of the two distributions is:

\[ H(U) = \sum_{i=1}^{n} P(i) \log(P(i)) \]  
\[ H(V) = \sum_{j=1}^{n} P'(j) \log(P'(j)) \]

Figure 1. Adjust the mutual information algorithm flow

Among them, \( P(i) = \frac{|U_i|}{N} \), \( P'(j) = \frac{|V_j|}{N} \). Then the mutual information between U and V can be expressed as:

\[ MI(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{n} P(i,j) \log \left( \frac{P(i,j)}{P(i)P'(j)} \right) \]

In the formula, \( P(i,j) = \frac{|U_i\cap V_j|}{N} \) represents the probability that a random sample belongs to two classes at the same time. In order to improve MI's distinguishing degree of different categories and further adjust it, use Equation 4 to limit its value range within the interval \([-1,1]\] to obtain adjusted mutual information (AMI). The closer the AMI value is to 1, the more consistent the sample point is with the actual category. In the formula, \( E(MI) \) represents the expected value of mutual information.

\[ AMI = \frac{MI - E(MI)}{\text{mean}(H(U), H(V)) - E(MI)} \]

2.2 Principal component analysis and statistical correlation

Principal component analysis (PCA) is usually used as a multivariate statistical analysis method to reduce variables and control the complexity of research problems in the early stages of development. The main idea is to map the dimensional features to the dimensional orthogonal feature space through the maximum variance strategy, and use the size of the variance included in different dimensions as the standard to judge, ignore the feature dimensions whose variance is lower than the set standard, so as to retain most of the effective information. In the case of information, the multi-dimensional variables are converted into a few main indicators to realize the dimensionality reduction operation of the original feature data [3]. The main component analysis algorithm is shown in Figure 2.
Figure 2. Principal component analysis algorithm flow

Assuming that the original feature set contains $n$ sets of samples, each set of samples has $p$ features, then the feature set is transformed into a matrix form and the data is normalized to obtain:

$$X = \begin{pmatrix}
    x_{11} & x_{12} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2p} \\
    \vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix}$$  \hfill (5)

In the formula, $x_{np}$ represents the eigenvalue of the group of samples. Calculate the covariance matrix based on the sample matrix. In order to eliminate the influence of eigenvalue dimensions, first standardize the sample matrix $X$:

$$Z = \begin{pmatrix}
    z_1 \\
z_2 \\
    \vdots \\
z_m
\end{pmatrix} = \begin{pmatrix}
    z_{11} & z_{12} & \cdots & z_{1n} \\
z_{21} & z_{22} & \cdots & z_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
z_{m1} & z_{m2} & \cdots & z_{mn}
\end{pmatrix}$$  \hfill (6)

In the formula, $z_{ij} = \frac{q_{ij} - \bar{q}_j}{s_j}$, $\bar{q}_j$, $s_j$ are the average value and variance of column $j$ in matrix $Z$, respectively. Then the covariance matrix can be expressed as:

$$Y = \frac{Z^T Z}{m-1}$$  \hfill (7)

Then calculate the eigenvalue $\lambda$ of the covariance matrix and the corresponding eigenvector $u$ according to the eigen decomposition formula [4]. The characteristic decomposition formula is shown in Equation 8.

$$\sum_{i=1}^{m} -\lambda_i I = 0$$  \hfill (8)

In the formula, $I$ is the identity matrix. Since the eigenvectors are arranged from large to small according to the corresponding eigenvalues, for a given principal component score, the principal
component information contribution rate is calculated by the ratio of the sum of the first k eigenvalues to the sum of all eigenvalues, and the largest of them is obtained the first k eigenvectors form the principal component matrix U.

\[
U = \begin{bmatrix}
    u_1 & u_2 & \cdots & u_k \\
    u_2 & u_2 & \cdots & u_k \\
    \vdots & \vdots & \ddots & \vdots \\
    u_m & u_m & \cdots & u_m
\end{bmatrix}
\]  \hspace{1cm} (9)

2.3. Linear discriminant analysis
Linear Discriminant Analysis (LDA) Fisher proposed a statistical analysis method in 1963, which was initially only applied to binary classification problems. In order to make the method more widely applicable, in 1948, CRRao made a further summary, summary and promotion of it, and successfully applied it to the analysis of multi-category problems [5]. The biggest difference from the PCA algorithm is that the LDA algorithm fully considers the class label of the sample when processing data, and is a popular supervised dimensionality reduction algorithm. The main idea can be expressed as mapping the original sample data to the low-dimensional feature space in the best projection direction, so as to maximize the interval between sample points of different categories, and make the samples of the same category cluster as much as possible.

2.3.1. Two classification problems. For the binary classification problem, suppose the given data set is

\[
D = \{(x_i, y_i)\}_{i=1}^{n}, \quad y_i \in \{0, 1\},
\]

and the set of samples, the mean vector, and the covariance matrix are \( X_i, \mu_i, \) and \( \Sigma_i \) respectively. Since the data samples are projected in a certain direction, the resulting mapping range is a straight line \( w \). If any two samples of different types are projected onto this straight line, their projections on \( w \) can be set to \( \omega^T \mu_0 \) and \( \omega^T \mu_1 \) respectively. When all sample data are projected onto \( w \), the covariances of the two types of samples are \( \omega^T \Sigma_0 \omega \) and \( \omega^T \Sigma_1 \omega \) respectively [6]. In order to minimize the intra-class distance, make the distance between projections of the same class as small as possible, that is, minimize \( \omega^T \Sigma_0 \omega + \omega^T \Sigma_1 \omega \); in order to maximize the inter-class distance, make the distance between the centers of different classes as large as possible, that is, maximize \( \| \omega^T \mu_0 - \omega^T \mu_1 \|_2^2 \). Then the optimization goal of LDA can be expressed as:

\[
\arg \max_{\omega} \frac{\omega^T \mu_0 - \omega^T \mu_1}{\omega^T \Sigma_0 \omega + \omega^T \Sigma_1 \omega} = \frac{\omega^T \mu_0 - \omega^T \mu_1}{\omega^T \Sigma_0 \omega + \omega^T \Sigma_1 \omega}
\]  \hspace{1cm} (10)

Define the intra-class divergence matrix \( S_0 \) and the inter-class divergence matrix \( S_b \) as:

\[
S_0 = \Sigma_0 + \sum_{i=1}^{n} (x - \mu_0) (x - \mu_0)^T + \sum_{i=1}^{n} (x - \mu_1) (x - \mu_1)^T
\]  \hspace{1cm} (11)

\[
S_b = (\mu_0 - \mu_1) (\mu_0 - \mu_1)^T
\]  \hspace{1cm} (12)

Then the objective function J can be redefined as:

\[
J = \frac{\omega^T S_0 \omega}{\omega^T S_b \omega}
\]  \hspace{1cm} (13)

This formula is the generalized Rayleigh quotient of \( S_0 \) and \( S_b \). Since the numerator and denominator of the above formula are both quadratic terms with respect to \( \omega \), the solution of the above formula is only related to its direction, and the direction of \( S_b \omega \) is always \( \mu_0 - \mu_1 \), then \( S_b \omega = \lambda (\mu_0 - \mu_1) \) can be set and brought into \( S_0 \omega = \lambda S_0 \omega \) to obtain \( \omega = S_0^{-1} (\mu_0 - \mu_1) \). To
determine the value of $\omega$. In summary, for binary classification problems, only the mean and covariance of the original sample data can be solved to determine the best projection direction $w$.

2.3.2. Multi-classification problem. For the multi-classification problem, the given data set is $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where any sample $x_i$ is an $n$ dimensional vector, $y_i \in \{C_1, C_2, \ldots, C_k\}$. Let the number, set, mean vector, and covariance matrix of the $j$ sample be $N_j$, $X_j$, $\mu_j$, $\sum_j$ respectively. Different from the binary classification problem, the projection space of the multi-class sample is a hyperplane. Assuming that the dimension of the projection space is $d$ and the corresponding basis vector is $(w_1, w_2, \ldots, w_d)$, an $n \times d$ dimensional basis vector matrix is obtained [7]. Then the optimization goal of multi-class linear discriminant analysis can be expressed as:

$$J(W) = \frac{W^T S_b W}{W^T S_w W}$$  \hspace{1cm} (14)

And define the intra-class divergence matrix $S_b$ and the inter-class divergence matrix $S_w$ as:

$$S_b = \sum_{j=1}^{k} S_{\mu j} = \sum_{j=1}^{k} \sum_{x \in x_j} (x - \mu_j)(x - \mu_j)^T$$  \hspace{1cm} (15)

$$S_w = \sum_{j=1}^{k} N_j (\mu_j - \mu)(\mu_j - \mu)^T$$  \hspace{1cm} (16)

Since $W^T S_b W$ and $W^T S_w W$ in the above formula are both matrices, they cannot be optimized as scalar functions. Therefore, before obtaining the multi-category optimization objective function that is easy to optimize, it needs to be transformed first, and finally the optimization objective function shown in equation 17 is obtained.

$$\arg \max_{W} J(W) = \prod_{w} \frac{W^T S_b W}{W^T S_w W}$$  \hspace{1cm} (17)

In the formula, $\prod_{w} A$ is the product of diagonal elements of $A$, and $W$ is an $n \times d$ dimensional matrix. Then the optimization process of $J(W)$ can be transformed into:

$$J(W) = \prod_{w} \frac{\omega_f^T S_b \omega_f}{\omega_f^T S_w \omega_f} = \prod_{w} \frac{\omega_f^T S_b \omega_f}{\omega_f^T S_w \omega_f}$$  \hspace{1cm} (18)

The rightmost of the above formula is the generalized Rayleigh quotient, and the matrix $W$ is the matrix formed by the largest $d$ eigenvalues corresponding to the eigenvectors of $S_w^{-1} S_b$.

3. Examples of the application of mathematical statistical evaluation in literature teaching

"Quiet River Don" is a majestic, exquisite and profound Soviet novel. After liberation, the novel and the film based on it were popular in my country. The author of the novel Sholokhov won the 1965 Nobel Prize in Literature for this book. However, soon after the novel was published, someone suggested that Sholokhov had copied the work of another writer named Kryukov. The reason is that Sholokhov was young when he wrote the book and could not have such rich life experience; in addition, after the book was published by Sholokhov, he has never written a work of such literary value. In order to clarify this problem, scholars such as Jieze used computational stylistic methods to conduct research. They randomly selected 2000 sentences from "The Quiet Don River", and then randomly selected 500 sentences from each of Sholokhov's and Kryukov's novels. A total of 3 groups of samples, 3000 sentences, Enter the computer for processing. First calculate the length of the sentence. The 3 sets of
samples are very close. So, we subdivided it into several groups according to different lengths, and compared the corresponding sentence groups in the three groups of samples. It was found that Sholokhov’s novels are more consistent with "Quiet Don", while Kryukov’s novels are similar to "Quiet Don River is very different! They conducted part-of-speech statistical analysis and took out each of the three samples. With 10,000 words, it was found that, except for pronouns, there are six types of words appearing at the same frequency in Sholokhov’s novels, but Kryukov’s novels are not consistent with it. The different positions of the parts of speech in the sentence can well express the style characteristics of the style, especially the two words at the beginning of the sentence and the three words at the end, which can often play a role in distinguishing the style. Through statistics, Jieze and others found that Sholokhov's novel is very close to "The Quiet Don", while Kryukov's novel is quite far from it. They also counted the structures of the 15 opening sentences with the highest frequency among the 3 samples, and found that 14 of Sholokhov’s novels are consistent with "Quiet Don", while only 15 of Kryukov’s novels are in Appeared in "Quiet Don River". Finally, they also counted the structures of the 15 ending sentences with the highest frequency among the three samples, and found that Sholokhov’s novels are completely consistent with "Quiet Don", while the structure of the ending sentences of Kryukov’s novels is the same as "The Quiet Don River" is totally inconsistent.

4. Conclusion

Mathematics is not only the basis of natural sciences, it also permeates many aspects of literature, art, sports, health, etc., and becomes an indispensable tool for the development of humanities. Therefore, if we learn mathematics well today, it is of great practical significance for participating in any work in the future.

References

[1] Sharma, S. Definitions and models of statistical literacy: a literature review. Open Review of Educational Research, 4(1) (2017) 118-133.
[2] Gluzman, N. A., Sibgatullina, T. V., Galushkin, A. A., & Sharonov, I. A. Forming the basics of future mathematics teachers’ professionalism by means of multimedia technologies. EURASIA Journal of Mathematics, Science and Technology Education, 14(5) (2018) 1621-1633.
[3] Perry, J., Lundie, D., & Golder, G. Metacognition in schools: what does the literature suggest about the effectiveness of teaching metacognition in schools? Educational Review, 71(4) (2019) 483-500.
[4] Pulham, E., & Graham, C. R. Comparing K-12 online and blended teaching competencies: a literature review. Distance Education, 39(3) (2018) 411-432.
[5] Phuong, T. T., Cole, S. C., & Zarestky, J. A systematic literature review of faculty development for teacher educators. Higher Education Research & Development, 37(2) (2018) 373-389.
[6] Gay, J. The new mathematics and an old culture. Mind, Culture, and Activity, 27(2) (2020) 113-116.
[7] Bautista, A., Yau, X., & Wong, J. High-quality music teacher professional development: A review of the literature. Music Education Research, 19(4) (2017) 455-469.