Identification and Modeling of College Students' Loneliness from the Perspective of Big Data a Machine Learning-Based Approach

Yuexiang Liu[1], Huajie Sui[1(✉)], and Luming Feng[2]

[1] Jiangxi University of Traditional Chinese Medicine, Nanchang, China
510563704@qq.com

[2] Department of Applied Psychology, Jiangxi University of Traditional Chinese Medicine, Nanchang, China

Abstract. With the advent of the era of big data, the information society inevitably intersects and integrates with everyone's life. Compared with the traditional self-statement scale for psychological measurement, big data network information has advantages such as excellent ecological validity. In this paper, we use the consumption data and access control data of college students in a university as the data source and the employment situation as the target variable to identify and predict the loneliness of college students. In the context of the current COVID-19 epidemic, which is generally in quarantine, this paper provides a realistic basis for this study. By extracting the features from the data, we address the limitations of the machine learning modeling approach for the autonomous identification and prediction of loneliness symptoms and propose further development prospects. This paper provides a practical basis for this research. By extracting features from the data information, we propose the limitations of the machine learning modeling approach for the autonomous identification and prediction of symptoms of loneliness, and propose the future development of this approach.

Keywords: Machine Learning · Loneliness · Algorithms

1 Introduction

With the rapid development of higher education in China, a large group of college students has received more and more attention from the state and society. Especially in recent years, the COVID-19 epidemic is still serious and most universities are closed, which inevitably leads to an increase in the number of isolation symptoms among college students. In the past, the identification of psychological problems mostly relied on psychological questionnaires or other self-assessment scales, but people often tended to show misleading answers due to the social approval effect, and subjects were more inclined to show misleading answers due to social expectation intervention, so they did not show their true thoughts well.

In recent years, thanks to the rapid development of computer technology and statistics courses and other disciplines, artificial intelligence, machine learning, deep learning

Z. Zhan et al. (Eds.): ICBDIE 2022, AHCS 5, pp. 418–425, 2023.
https://doi.org/10.2991/978-94-6463-034-3_43

and big data technologies are also progressing rapidly. Based on machine learning, it indicates a good ability to solve the problem. Machine learning, an important branch of Artificial Intelligence, optimizes the algorithm through the computer simulation of human autonomous learning process and new data, so as to improve the accuracy of the model prediction. It automatically analyses the original data, grasps the law, and then uses the law to predict the unknown data methods.

With the popularization and development of Internet technology, users' behaviors can be stored electronically in cyberspace in real time, resulting in rich user behavior data in natural contexts. The results of many studies have shown that users' behavioral data on social media have a lot of psychological implications. For example, browsing time on social media is positively correlated with users' willingness to socialize, and the number of friends on social networking sites is negatively correlated with users' shyness [4]. In addition to online behavioral characteristics, text messages posted by users on forums or other social platforms [5], for example, have been shown to be significantly correlated with psychological characteristics, with effect levels above moderate [1], suggesting the feasibility of using big data information to build computational models for identifying psychological indicators.

In this paper, we analyze and verify the feasibility and effectiveness of the machine algorithm based on the data sources of campus card consumption and access control information of college students in a university [6], and use the employment situation as a valid standard, and look forward to its future application areas and development trends.

2 Loneliness

Loneliness is a common unpleasant experience that arises from a lack of important social relationships or dissatisfaction with current social relationships. Surveys show that loneliness has increased significantly over the past 20 years, with the rate of increase being more pronounced in the college population. [3] found that among 991 respondents, 146 (14.7%) felt lonely frequently, 686 (69.2%) felt lonely occasionally, and the percentage of students who felt lonely frequently or occasionally was as high as 83.9%. [2] used the same questionnaire to survey 705 college students in five universities in southwest China, and found that the percentage of students with medium to high loneliness was 83%.

It is evident that the isolation of college students has become more serious with the progress of the times. Especially in recent years, the COVID-19 epidemic has swept through the world, and especially the university students, who are the target of this study, are mostly managed in closed schools or isolated at home, which has aggravated the problem of loneliness among university students.

3 Machine Learning

Machine Learning is a multi-disciplinary discipline, which is the core of artificial intelligence. It aims to enable computers to learn by themselves, to learn from existing data information, to obtain potential patterns, and to apply these patterns to the analysis and prediction of unknown data.

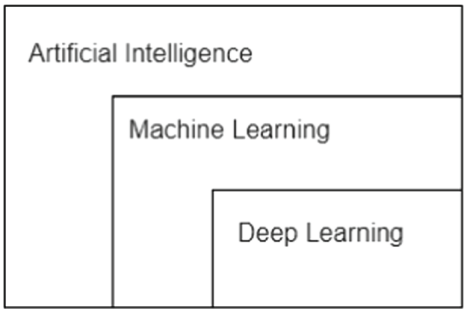

Fig. 1. Relationship between deep learning and machine learning (Draw by the author)

According to the depth of the model structure, there are traditional machine learning and deep learning. The relationship between deep learning and machine learning is shown in Fig. 1.

Commonly used machine learning algorithms include decision trees, support vector machines, K-nearest neighbors, logistic regression, multilayer perceptron, random forests, and K-means. For example, the number of friends on social networking sites is negatively correlated with the shyness of users [4]. Text messages posted by users on social media [5] were found to be significantly associated with psychological traits with above moderate effect sizes [1].

4 Modelling Process

The process of machine learning modeling includes Data Collection, Feature Extraction, Feature Selection, Modeling, Validation, and Output.

4.1 Data Sources

This study takes a university as an example [6] and collects 25 types of data information, including basic student information, graduate job search information, campus card consumption information, and library access information, from a total of 17,828 undergraduate students in the classes of 2011 to 2014. The research idea is to use the inverse method to identify the lonely group by ranking the non-lonely group. The data features are used to indicate that the same cafeteria swipe time is close and more often, and the library access swipe time is close and more often. Close swipe time was defined as within 5 min.

4.2 Data Processing

First, three months of consumption records were selected as sample data. The structure of the consumption relationship details table is shown in Fig. 2 [6], M represents the month, X1 represents the student, T1 represents the consumption swipe time of X1, X2 represents all students within 5 min of the consumption time with X1, and T2 represents

M	X1	X2	T1	T2
201404	201 0018	201 002	2014/4/24 11:04	2014/4/24 11:04
201404	201 018	201 06	2014/4/24 11:04	2014/4/24 11:04
201404	201 018	201 48	2014/4/24 11:04	2014/4/24 11:05
201404	201 018	201 07	2014/4/24 11:04	2014/4/24 11:05
201404	201 018	201 22	2014/4/24 11:04	2014/4/24 11:05
201404	201 018	201 12	2014/4/24 11:04	2014/4/24 11:05
201404	201 018	201 16	2014/4/24 11:04	2014/4/24 11:05
201404	201 018	201 17	2014/4/24 11:04	2014/4/24 11:05
201404	201 018	201 09	2014/4/24 11:04	2014/4/24 11:05
201404	201 018	201 15	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	200 12	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	201 04	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	201 17	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	201 15	2014/4/24 11:04	2014/4/24 11:05
201404	201 018	201 08	2014/4/24 11:04	2014/4/24 11:05
201404	201 018	201 17	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	201 53	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	200 27	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	201 20	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	201 29	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	201 44	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	201 67	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	20 024	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	20 019	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	201 026	2014/4/24 11:04	2014/4/24 11:05
201404	20 018	20 055	2014/4/24 11:04	2014/4/24 11:05
201404	201 018	20 026	2014/4/24 11:04	2014/4/24 11:05

Fig. 2. Screenshot of the consumption relationship breakdown

X1	X2	rl
201 0018	201 0002	213
20 0005	20 017	190
201 0014	20 015	181
201 032	20 068	166
201 0015	20 071	151
201 0001	20 043	158
201 0042	20 025	152
201 0014	20 036	150
201 0016	20 058	150
201 0029	20 046	149
201 0034	20 028	149
20 0002	20 004	143
20 0069	20 031	143
20 0082	20 009	143
20 0036	20 077	143
20 0042	20 026	141
20 0025	20 015	137
20 0043	20 006	136
20 0039	20 067	122
20 0057	20 055	121
20 0038	20 005	113

Fig. 3. Screenshot of canteen consumption friend circle relationship

the consumption swipe time of another student and within 5 min of T2. Based on this table, we count the number of encounters between two students and the total number of consumptions in the cafeteria, and then take the encounter relationship details of each cafeteria with the number of encounters greater than 10 to get the total consumptions table, and then sum up by X1 and X2 groups to form the friend circle relationship of cafeteria consumptions, as shown in Fig. 3 [6]. Based on this method, we can also get the friend circle relationship of the library.

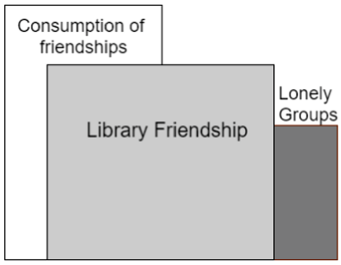

Fig. 4. Screenshot of the method for finding a lonely student

4.3 Data Analysis

By randomly selecting several sets of data, all were verified to be true friendships.

Returning to this project, using 17828 all undergraduate students as the sample data, there are 20585 people who have a friend relationship form for cafeteria consumption and 43,840 people who have a friend relationship form for the library (both figures here are larger than the sample data because, according to the data processing principle mentioned earlier, all undergraduate students as X1, X2 can be the whole university students and faculty within the range of the swipe time). There are 15,312 people who get more friends by taking the intersection between the cafeteria consumption friends relationship and the library's friends circle relationship, and there are 1932 people who are neither in the cafeteria consumption friends relationship table nor in the library's friends circle relationship table. The details are shown in Fig. 4 [6].

By analyzing the distribution profile of the 1932 students suspected of being lonely, the non-employment rate was 5.95%. Among them, the data of 1194 students suspected of loneliness in 2011 class had system errors and needed to be deleted, leaving 738 students, 76 of whom were not employed, with an unemployment rate of 10.30%. In contrast, only 608 of the 15,312 students with more friendships were not successfully employed, with an unemployment rate of 3.97%. It can be concluded that the unemployment rate of students with fewer friendships is higher than that of students with more friendships.

4.4 Data Validation

The above conclusions can be drawn through canteen consumption friendships and library friendships. Through the sample data i.e. 17,828 students from 2011 to 2014 class, 1127 were not employed. The success of employment was considered as the target variable and 25 indicators (gender, ethnicity, faculty, per capita family income, number of scholarships, number of campus card consumption, number of library access, etc.) were used as independent variables to identify the factors affecting employment using a decision tree model.

Decision trees are a common non-parametric supervised learning method used in machine learning for classification and regression. The goal is to create a model that derives simple decision rules from the data to predict the value of the target variable. Decision trees are easy to illustrate and understand. Decision Tree Classifier from the python algorithm library is used. Before calling the algorithm, the train test split division

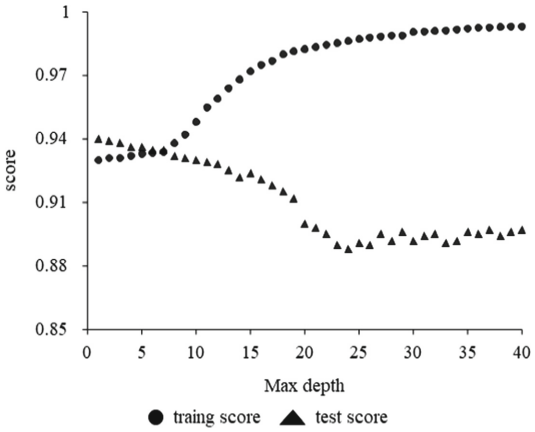

Fig. 5. Accuracy of training set and accuracy of test set at different depths

function is used to randomly divide the data into a training set and a test set. The training set is used to construct the decision tree and the test set is used to calculate the error rate and analyze the effect of the trained decision tree model. The decision tree model is required to find the appropriate max_depth. The prediction of each value is calculated and plotted, and the obtained results are shown in Fig. 5. Note: This figure is drawn with reference to the data of Fig. 5 in the literature [6].

4.5 Research Discussion

The purpose of this research project is to identify the list of lonely people through big data and help the university to provide psychological and employment support to students on as large a scale as possible for college leaders to make decisions and make positive interventions.

5 Strengths and Losses

5.1 Strengths

Psychological modeling based on big data information has unique advantages over paper-and-pencil measurement methods. Self-reported responses may be subject to social approval effects, and subjects may give misleading responses. In contrast, modeling with big data analysis allows for non-intrusive measurement of subjects and is therefore more ecologically valid.

Psychological modeling is performed by computer for uniform feature extraction, and the computational process is highly consistent. The large time span of traceability allows for cross-sectional or follow-up studies.

5.2 Losses

Although psychological recognition modeling based on big data information is feasible, it still has shortcomings. The scenario of the new method is dominated by school students, which may bring about group bias; again, there is also the problem of limited accuracy at present. Most of the recognition models still use the self-assessment scale scores as the validity standard (Kosinski, 2015). The accuracy of the self-assessment results affects the goodness of the model. Finally, the recognition accuracy of mental models still needs to be improved.

6 Conclusion and Outlook

Machine learning can predict the tendency of college students to feel loneliness based on the information collected from big data. The performance of the data platform in major universities, where college students are the main group, is expected. The two-way combination of machine learning to filter and extract the big data information and then to make a diagnosis with the help of clinicians is more suitable for practical application scenarios. In future research, the optimization of loneliness prediction models should be focused. User input information collected by smartphones and physiological data collected by wearable devices may be important additions to the prediction model; exploring EEG data from the perspective of brain science is beneficial to further reveal the brain neural mechanism behind their behaviors. It provides effective reference for further clinical interventions and contributes methodological solutions to smart medical care in China.

Acknowledgements. Funding for this study was obtained from the Key Laboratory of Psychology of TCM and Brain Science, Jiangxi Administration of traditional Chinese Medicine, Jiangxi University of Chinese Medicine, 1688 Meiling Avenue, Nanchang China.

References

1. Carvalho, L. D. F., & Pianowski, G. (2017). Pathological personality traits assessment using Facebook: Systematic review and meta-analyses. Computers in Human Behavior, 71, 307–317.
2. Li, The relationship between online interactions and loneliness among college students: the mediating role of coping styles [J]. Chinese Journal of Health Psychology, 2016, 24(2): 239–243
3. Luo, An investigation and analysis of the psychology of loneliness among college students [J]. Journal of Zhejiang University (Science Edition), 1999, 26(3): 112–115
4. Orr, E. S., Sisic, M., Ross, C., Simmering, M. G., Arseneault, J. M., & Orr, R. R. (2009). The influence of shyness on the use of Facebook in an undergraduate sample. CyberPsychology & Behavior, 12(3), 337–340.
5. Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. Journal of Research in Personality, 46(6), 710–718.
6. Yu Lin, Xu Ting, Li Chao, Liao Lili, Permission, Xie Panke. An early warning model of student loneliness in the context of big data–a case study of Huazhong Normal University [J]. Modern Information Technology, 2019, 3(23): 1–4.