International Journal of Scientific Research and Management (IJSRM)
||Volume||11||Issue||11||Pages||953-959||2023||
Website: https://ijsrm.net ISSN (e): 2321-3418
DOI: 10.18535/ijsrm/v11i11.ec01

Design of Rural Revitalization Model Classification System based on Machine Learning

Pengcheng Yang[1], Zhan Wen*[1,2], Cheng Zhang[1], Xiaoming Zhang[1], Dehao Ren[1]

[1] College of Communication Engineering, Chengdu University of Information Technology, Chengdu, 610225, China;

[2] Meteorological information and Signal Processing Key Laboratory of Sichuan Higher Education Institutes of Chengdu University of Information Technology, Chengdu, 610225, China;

Abstract:

There are several models of rural revitalization such as planting, breezing, tourism and so on. In this paper, we propose a design scheme to achieve the most suitable rural revitalization model based on machine learning. This scheme includes three main modules of data collection, data processing, and machine learning model. First, collect the text of successful rural revitalization from Internet. Then process the collected text to extract feature words. Finally, use machine learning algorithm like Logistic Regression, Random Forest, and KNN (K Nearest Neighbors) to classify the best rural revitalization model. The experimental results show that the proposed scheme is effective and KNN can gain the best accuracy. This scheme can bring convenience and benefits to rural revitalization.

Keywords: Rural Revitalization, Machine Learning, Text Data Analysis, Crawler Technology

Introduction

This study focuses on rural revitalization in different regions, collected data on different successful cases of rural revitalization, and built machine learning models based on these data. During the study, the collected raw data were first analyzed and processed to facilitate the training of the machine learning model. Subsequently, machine learning algorithms were utilized to train the data into the model, and then predicted results were compared with actual results to select the machine learning model that best matches the actual results. Textual data came from websites such as the Rural Revitalization Bureau. The primary aim of this study is to assist grassroots staff in initially determining revitalization strategies and providing certain locally feasible revitalization programs to promote subsequent work. The remainder of this paper is organized as follows: The second part introduces research on model structure, detailing research methods and theoretical knowledge. The third part presents the experimental procedure. The fourth part analyzes and summarizes experimental results.

Model Structure Design of the Rural Revitalization Program

The project's scheme design identifies relevant research subjects through social research, then gathers successful rural revitalization cases through methods such as crawler programs. These cases are classified and labeled. Following data processing, the Jieba word segmentation package is utilized to eliminate stop words, followed by PCA to extract keywords and establish a feature word library. K Nearest Neighbors, logistic regression, and regression forest models are employed for training and testing, ultimately selecting an optimal model design system.

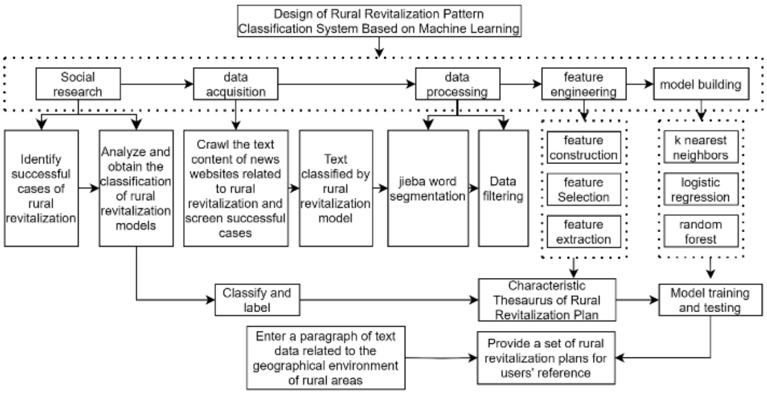

Design of Rural Revitalization Pattern Classification System Based on Machine Learning
Social research
data acquisition
data processing
feature engineering
model building
Identify successful cases of rural revitalization
Analyze and obtain the classification of rural revitalization models
Crawl the text content of news websites related to rural revitalization and screen successful cases
Text classified by rural revitalization model
jieba word segmentation
Data filtering
feature construction
feature Selection
feature extraction
k nearest neighbors
logistic regression
random forest
Classify and label
Characteristic Thesaurus of Rural Revitalization Plan
Model training and testing
Enter a paragraph of text data related to the geographical environment of rural areas
Provide a set of rural revitalization plans for users' reference

Figure 1: Model structure diagram

2.1 Data processing

(1) Chinese word segmentation

Text segmentation is the process of recombining continuous word sequences into word sequences according to certain specifications. We know that in English writing, spaces are used as natural separators between words, while Chinese is just words, sentences and paragraphs, which can be simply separated by obvious separators, except for words without formal separators. Although English also has the problem of phrase division, at the word level, Chinese is much more complex than English. [1].

(2) Keyword extraction

The Keyword Extraction API provides an interface for extracting the keywords. With this API, the core content of the text can be extracted from a large amount of information. It can be an entity with a specific meaning, such as the name, location, film, etc., or some basic key words in the text [2]. With this API, the extracted keywords can be sorted from high to low according to the weights in the text. The higher the ranking, the higher the weight, and the more accurate the core text content that can be extracted.

(3) Delete stop words

In information retrieval, in order to save storage space and improve the search efficiency, some words or some words are automatically filtered out before and after processing natural language data (or text), which is called stop words. These stop words are entered manually and not generated automatically. The generated stop words will form a list of stop words.

(4) Text vectorization

Computers cannot understand human language, but computers can understand numbers. NLP is often unstructured and messy text data, while machine learning does not directly process raw text data. Text data must be converted into numbers, such as vectors

(5) Principal component analysis (PCA)

Principal component analysis (PCA) is a feature dimension reduction method. It replaces the original features with a series of concise new features [3]. These fresh characteristics are linear combinations of the original attributes. It maximizes sample variance, striving to make new traits uncorrelated. In this manner, diverse characteristics' impact on the model can readily be investigated and examined, successfully decreasing model complexity and enhancing model training speed.

2.2 Machine learning model

In machine learning, there are three categories: supervised learning, semi-supervised learning, and unsupervised learning. We opt for supervised learning, which establishes a predictive model by studying known data sets' feature characteristics and result metrics to forecast and measure unknown data's characteristics and results [4]. In this paper, we use KNN, logistic regression, and regression forest models to do experiments.

Experiment Procedure and the Results

3.1 Experimentation procedure

Based on the general process of machine learning training data, the experimental procedure of this paper is as follows.

- Text data on different rural revitalization programs were collected through the crawler program.
- The collected text data is classified and stored, and the word segmentation step is started.
- The 11 categories of the open Chinese thesaurus used

Tsinghua University, divided the keywords of each article into 11 categories through the pouch model and generates an excel table

- Treatment vector without dimension (or PCA vector)

Reduction was used to train the logistic regression models.

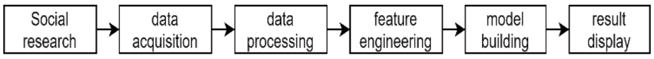


Figure 2: Experimental procedure

data collection: The data used in this experiment is captured through a Python crawler program on successful cases of rural revitalization, and influencing factors are analyzed from a large number of existing studies. Rural revitalization is divided into four categories: e-commerce, tourism, modern agriculture, and rural service industry [5].

data preprocessing: After crawling the text, it is classified and stored first. Since the raw data is text, it is necessary to split the text into separate words and remove the stop words in the text. It is still a similar, meaningless word. Text segmentation was performed using the Python Jieba word segmentation package[6].

Word segmentation: The top n keyword is extracted from each document (n can be customized in the program). These keywords can be selected based on the TF IDF value and can be called through the python package Jieba[7]. Keywords can represent each document, that is, the scheme used in rural revitalization.

Text vectorization: After extracting the keywords, the txt file (keyword txt) representing all the schemes is converted into the digital form by using the word bag model. The Tsinghua University Thesaurus divides the vocabulary into many types. Write a program to compare each keyword TXT document with the Tsinghua University Thesaurus, see which category the keyword data belongs to, and calculate the vocabulary under which category (word frequency statistics). Revitalization strategies and implementation situations are different, and the key words are also different. By calculating the frequency of keywords in each type of document, we can get a document vectorization representation of the rural revitalization plan (each document).

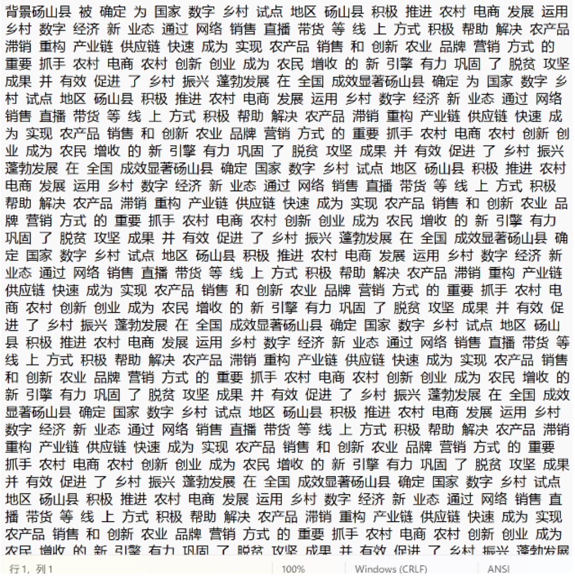
principal component analysis: After obtaining the vectorized data (data in digital form), each row in the table is a document corresponding to a regional implementation scheme. If it is an e-commerce company, the label after the data is 1, and the label of the tourism data is 0. Principal component analysis (PCA) to compress the data and prevent overfitting.

Model training: In this experiment, we chose to use three algorithms: K Nearest Neighbors, logistic regression, and random forest for machine learning. Train three models using a training set, test the model using a testing set after the model is established, and select the best model by comparing the accuracy of the three models.

result display: The final display of this experiment is to provide a user interface via the web page. Users can input the text of a region into the corresponding region to obtain the corresponding scheme results of the model.

3.2 Experimental results

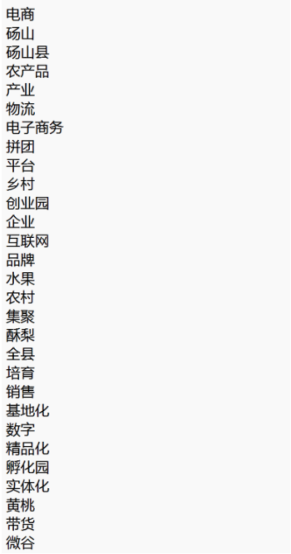
After running the program, the program will split the input text and give the following results。



```
背景砀山县 被 确定 为 国家 数字 乡村 试点 地区 砀山县 积极 推进 农村 电商 发展 运用
乡村 数字 经济 新 业态 通过 网络 销售 直播 带货 等 线 上 方式 积极 帮助 解决 农产品
滞销 重构 产业链 供应链 快速 成为 实现 农产品 销售 和 创新 农业 品牌 营销 方式 的
重要 抓手 农村 电商 农村 创新 创业 成为 农民 增收 的 新 引擎 有力 巩固 了 脱贫 攻坚
成果 并 有效 促进 了 乡村 振兴 蓬勃发展 在 全国 成效显著砀山县 确定 为 国家 数字 乡
村 试点 地区 砀山县 积极 推进 农村 电商 发展 运用 乡村 数字 经济 新 业态 通过 网络
销售 直播 带货 等 线 上 方式 积极 帮助 解决 农产品 滞销 重构 产业链 供应链 快速 成
为 实现 农产品 销售 和 创新 农业 品牌 营销 方式 的 重要 抓手 农村 电商 农村 创新 创
业 成为 农民 增收 的 新 引擎 有力 巩固 了 脱贫 攻坚 成果 并 有效 促进 了 乡村 振兴
蓬勃发展 在 全国 成效显著砀山县 确定 国家 数字 乡村 试点 地区 砀山县 积极 推进 农村
电商 发展 运用 乡村 数字 经济 新 业态 通过 网络 销售 直播 带货 等 线 上 方式 积极
帮助 解决 农产品 滞销 重构 产业链 供应链 快速 成为 实现 农产品 销售 和 创新 农业 品
牌 营销 方式 的 重要 抓手 农村 电商 农村 创新 创业 成为 农民 增收 的 新 引擎 有力
巩固 了 脱贫 攻坚 成果 并 有效 促进 了 乡村 振兴 蓬勃发展 在 全国 成效显著砀山县 确
定 国家 数字 乡村 试点 地区 砀山县 积极 推进 农村 电商 发展 运用 乡村 数字 经济 新
业态 通过 网络 销售 直播 带货 等 线 上 方式 积极 帮助 解决 农产品 滞销 重构 产业链
供应链 快速 成为 实现 农产品 销售 和 创新 农业 品牌 营销 方式 的 重要 抓手 农村 电
商 农村 创新 创业 成为 农民 增收 的 新 引擎 有力 巩固 了 脱贫 攻坚 成果 并 有效 促
进 了 乡村 振兴 蓬勃发展 在 全国 成效显著砀山县 确定 国家 数字 乡村 试点 地区 砀山
县 积极 推进 农村 电商 发展 运用 乡村 数字 经济 新 业态 通过 网络 销售 直播 带货 等
线 上 方式 积极 帮助 解决 农产品 滞销 重构 产业链 供应链 快速 成为 实现 农产品 销售
和 创新 农业 品牌 营销 方式 的 重要 抓手 农村 电商 农村 创新 创业 成为 农民 增收 的
新 引擎 有力 巩固 了 脱贫 攻坚 成果 并 有效 促进 了 乡村 振兴 蓬勃发展 在 全国 成效
显著砀山县 确定 国家 数字 乡村 试点 地区 砀山县 积极 推进 农村 电商 发展 运用 乡村
数字 经济 新 业态 通过 网络 销售 直播 带货 等 线 上 方式 积极 帮助 解决 农产品 滞销
重构 产业链 供应链 快速 成为 实现 农产品 销售 和 创新 农业 品牌 营销 方式 的 重要
抓手 农村 电商 农村 创新 创业 成为 农民 增收 的 新 引擎 有力 巩固 了 脱贫 攻坚 成果
并 有效 促进 了 乡村 振兴 蓬勃发展 在 全国 成效显著砀山县 确定 国家 数字 乡村 试点
地区 砀山县 积极 推进 农村 电商 发展 运用 乡村 数字 经济 新 业态 通过 网络 销售 直
播 带货 等 线 上 方式 积极 帮助 解决 农产品 滞销 重构 产业链 供应链 快速 成为 实现
农产品 销售 和 创新 农业 品牌 营销 方式 的 重要 抓手 农村 电商 农村 创新 创业 成为
农民 增收 的 新 引擎 有力 巩固 了 脱贫 攻坚 成果 并 有效 促进 了 乡村 振兴 蓬勃发展
```

Figure 3: Text word segmentation results

The program then extracts the key words from the participle results and gives the results as follows.



电商
砀山
砀山县
农产品
产业
物流
电子商务
拼团
平台
乡村
创业园
企业
互联网
品牌
水果
农村
集聚
酥梨
全县
培育
销售
基地化
数字
精品化
孵化园
实体化
黄桃
带货
微谷

Figure 4: Keyword extraction results

Establish a feature vocabulary after vectorization to facilitate subsequent model selection and training.

We conducted a simulation experiment on the rural revitalization model using the KNN model, changing the parameters of n-neighbors. Finally, after statistical analysis of the experimental results, we visualized them to obtain the results in Figures 4-10. From the figure, it can be seen that with the increase of n-neighbors, the accuracy of the KNN model reached its highest point when n-neighbors approached 10. Subsequently, with the increase of n-neighbors, the matching accuracy of the KNN model decreased, and when n-neighbors approached 10, The accuracy of model matching can reach around 25%.

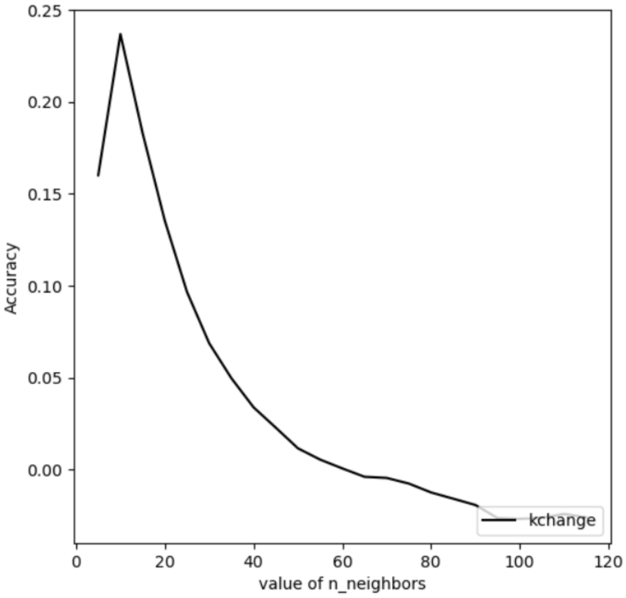


Figure 5: k nearest neighbors accuracy results

During the experiment, we continuously test and adjust the model parameters, fitting the logistic regression model under different parameter configurations (changes in C values), and continuously testing the accuracy of the obtained training set. From the results, it can be seen that as the regularization intensity parameter (C) changes, the model matching accuracy also continuously changes. After visualizing the results, we can conclude that in this model, the matching accuracy reaches its highest when the regularization intensity parameter (C) approaches 1.

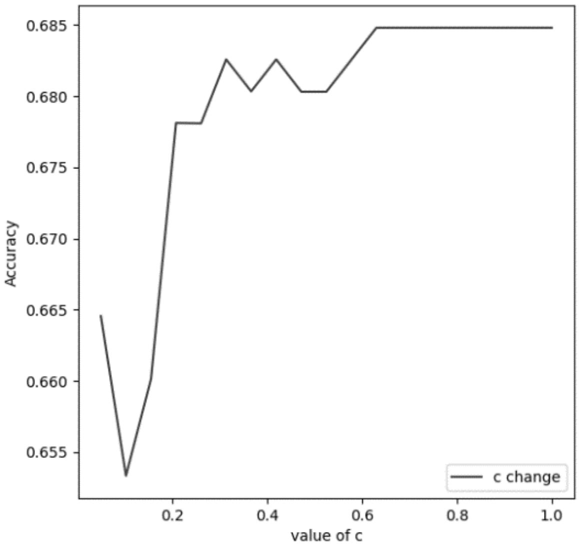


Figure 6: Logistic regression model testing

In the process of random forest experiments, we selected random cases of four rural revitalization models, namely "e-commerce", "tourism", "modern agriculture", and "rural service industry", as features (X), and matched them with the successful case library of rural revitalization we established, with their "matching results" listed as targets (Y). Finally, through statistical analysis, the model matching result was obtained

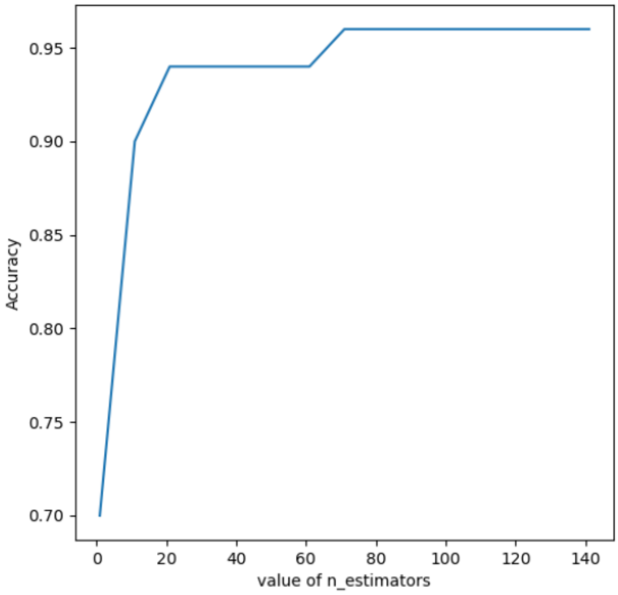


Figure 7: Random Forest Matching Test

After obtaining the vectorized data for training the model through data processing, we established three models: logistic regression, KNN, and random forest. We conducted statistical analysis on the accuracy and construction time of the three models, and the final results are shown in the table below.

Table 1 Experiment Results

model	logistic regression	k nearest neighbors	random forest
accuracy	68.4%	25.1%	89.1%
average run time	0.04s	0.01s	0.09s

From the chart results, it can be seen that random forest has the highest accuracy, so random forest is chosen as the machine learning model

Conclusions

This experiment is based on the collection of relevant data of rural revitalization, mainly focusing on e-commerce, tourism, modern agriculture, and tourism as cases for the design and implementation of machine learning. The system is a precision poverty alleviation auxiliary evaluation system designed using machine learning algorithms. It utilizes web crawlers to gather online text data regarding economic levels, conducting text processing through methods such as Chinese word segmentation, keyword extraction, and text vectorization. These methods are used to build a feature dictionary and develop a machine learning model. Finally, a mobile application is created to provide an input interface. The application of machine learning in rural revitalization can offer effective data analysis and prediction capabilities, contributing to the automation and precision of agricultural production while improving agricultural production efficiency and economic benefits. Additionally, machine learning can play a role in fields such as rural environmental monitoring, ecological protection, rural planning, and other areas providing technical support for rural revitalization efforts. Through experiments, it becomes evident that machine learning classification requires a large sample size for effective learning. When the sample size is small, it contains numerous irrelevant features that may lead to low classification accuracy or overfitting issues. This also highlights that no single machine learning algorithm can solve all problems; instead, they must be selected and optimized according to actual situations [8].

Acknowledgements

This work was supported by the Sichuan Science and Technology Program, Soft Science Project

(No.2022JDR0076). We also would like to thank the sponsors of Industry-school Cooperative Education Project of Ministry of Education (No. 202002230010, 202101014067, 202101291013, 220500643270506)

References

1. Xu Lin Hong. Research on Automatic Recognition of Citation Emotions Based on Machine Learning Algorithms: Taking the Field of Natural Language Processing as an Example. modern information, 2020,40(01)
2. Wang Ding. Analysis and research of natural language processing technology. Introduction to Scientific and Technological Innovation, 2020,17 (07): 141-142
3. Feng Xiaodi. Natural language processing techniques based on reinforcement learning. Digital World, 2020 (03): 9-10
4. Havlicek V ,Córcoles, Antonio D, Temme K ,et al.Supervised learning with quantum enhanced feature spaces[J]. 2018.DOI:10.1038/s41586-019-0980-2.
5. Dai Chengyingzi.Identification of successful entrepreneurs in rural areas based on machine learning[D], Master's Dissertation of Chongqing University of Technology,2022
6. Liu Zhiming, Liu Lu, Empirical Study on Chinese Weibo Emotion Classification Based on Machine Learning [J] Computer Engineering and Applications. 1002-8331 (2012) 01-0001-04
7. Use machine learning techniques to classify and predict the body density of states and chemical properties. AP. Mathematical compression.412: 126587, 2022.
8. Yang Jian Feng,Qiao Pei Rui,Li Yong Mei,Wang Ning. A Review of Machine Learning Classification Problems and Algorithm Research. Statistics & Decision. 1002-6487（2019）06-0036-05