Postprint on the Application of Random Forest in the Integration of Fragmented Information in Universities

Authors: Zhang Wende, Cheng Han, Liu Tian, Zeng Jinjing

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] In response to the trend of fragmentation in university information presentation, this paper proposes a process for integrating fragmented university information and constructs a feature selection model for such integration using the random forest algorithm. [Method/Process] Based on an analysis of the current development status and existing problems in university information integration, the principles and advantages of the random forest algorithm are examined and applied to the feature selection model within the process of integrating fragmented university information. The model is validated through a case study on the identification of impoverished university students. [Results/Conclusions] The random forest algorithm exhibits high accuracy and effectiveness in feature selection for university information integration, thereby offering a novel approach for the integration of fragmented university information.

Full Text

Preamble

ChinaXiv Partner Journal, Vol. 62, No. 7, April 2018

Application of Random Forest in the Fragmented Integration of University Information*

Zhang Wende[1], Cheng Han[1], Liu Tian[2], Zeng Jinjin[2]

[1]Institute of Information Management, Fuzhou University, Fuzhou 350108
[2]Library of Fujian Agriculture and Forestry University, Fuzhou 350002

Abstract

[**Purpose/Significance**] Facing the trend of information fragmentation in universities, this paper proposes a process for fragmented university information integration and applies the random forest algorithm to construct a feature selection model for this integration. [**Method/Process**] Based on the current development status and existing problems of university information integration, we analyze the principles and advantages of the random forest algorithm and apply it to the feature selection model in the fragmented integration process, validating the model through the example of identifying financially needy students. [**Result/Conclusion**] The random forest algorithm demonstrates high accuracy and effectiveness in feature selection for university information integration, providing a new approach for fragmented university information integration.

Keywords: random forest; fragmentation; information integration; feature selection
Classification Number: G203
DOI: 10.13266/j.issn.0252-3116.2018.07.014

Introduction

Since the early 1990s, university informatization in China has developed rapidly under the high priority of government departments. Throughout the development of university informatization, the integration process can be divided into three stages: data-based integration, information-based integration, and knowledge-based integration [1-3]. In the early stages of university informatization construction, independent information systems were established to meet the needs of individual departments. As "information silos" formed, the integration of heterogeneous information systems in universities gradually attracted attention. The primary goal of the data-based integration stage was to use middleware and data warehouse technologies to eliminate system distribution and heterogeneity, achieving unified storage of heterogeneous information system data. Information-based integration developed on the basis of data integration, focusing on business requirements to integrate information resources that satisfy specific business processes through enterprise architecture and other methods, and applying data mining and machine learning algorithms. Chang Tongshan [4] believes that data mining technology can help university administrators better analyze data to obtain hidden, useful information and knowledge, ultimately improving decision-making efficiency. Liao Fenglu and Zhou Qing [5] used a Naive Bayes model to predict students' employability, providing assistance for student employment services. Shi ?, Qian Yuan, and Sun Ling [6] applied data mining techniques such as association rule algorithms and clustering algorithms to online learning supervision, providing references for understanding online learning effectiveness and improving learning processes. Shu Zhongmei and Xu Xiaodong [7] used data mining methods including stepwise regression and decision tree analysis to analyze university student satisfaction, exploring factors

influencing student satisfaction and providing references for talent cultivation in universities. He Shiming and Shen Jun [8] utilized BP neural networks and clustering analysis techniques to mine useful patterns and knowledge hidden in data, proposing a learning evaluation method suitable for online teaching to provide decision support for teaching assessment. Liu Meiling, Li Xi, and Li Yongsheng [9] proposed a grade clustering analysis method based on the K-Means algorithm to illustrate the application of data mining technology in educational systems.

Looking at all current research in the field of university data mining, whether decision trees, neural networks, or association and clustering algorithms, they all share the following problems: (1) Data mining algorithms are only applied to partial educational issues, such as teaching evaluation and learning efficiency supervision, meaning the algorithm models can only solve one or a class of problems and lack strong universality and generalizability; (2) The accuracy, efficiency, and implementation complexity of existing educational data mining algorithms need improvement, and they have high data requirements, facing certain obstacles in practical applications.

This study applies the Random Forest (RF) algorithm, leveraging its advantages of good generalization and robustness, insensitivity to noise, and high precision and accuracy to construct a Random Forest-based integration feature selection model. We conduct experimental analysis on university financial need identification data to verify the effectiveness and accuracy of the Random Forest-based feature selection model for university information integration, aiming to better perform university information integration and provide personalized decision support.

2. Process for Fragmented University Information Integration

2.1 Fragmented University Information Integration

With the gradual implementation of university informatization construction, many universities have conducted beneficial practical explorations in information integration. For example, Nanjing Agricultural University utilized enterprise architecture theory to build a campus informatization application architecture platform. However, with the arrival of the big data era, university information has gradually become fragmented, and existing information integration systems suffer from poor dynamic scalability and difficulty in providing personalized decision support. Simultaneously, facing the impact of information fragmentation, universities have deficiencies in many aspects such as discipline construction, research management, and student management, leading to a series of problems including insufficient departmental collaboration, divergent objectives, and low management efficiency. Therefore, this study defines a "knowledge fragment" as the smallest granular knowledge segment obtained in school information services through the "fragmented" integration of structured,

semi-structured, and unstructured data [10]. When users propose requirements, we only need to decompose the key features of the requirement, extract and integrate "knowledge fragments" based on these features, and form visualized query results submitted to users in report form. This fragmented information integration approach can effectively compensate for the shortcomings of existing university integration systems, such as poor scalability, weak autonomy, and low information utilization.

2.2 Process for Fragmented University Information Integration

[Figure 1: see original paper] Process for Fragmented University Information Integration

Based on user requirements, the system queries the historical requirement feature database to determine whether a feature set for the same requirement exists. If it exists, the system extracts corresponding "knowledge fragments" with those features from the knowledge fragment sharing pool according to the historical feature requirement set. If it does not exist, the system needs to use the random forest-based feature selection model to extract a feature set that satisfies the requirement, then extract corresponding "knowledge fragments" based on this feature set, and finally integrate the "knowledge fragments" that match the requirement features and feed them back to users in visualized form.

Therefore, for fragmented university information integration systems, the core lies in the feature selection process for university information integration. The higher the accuracy and effectiveness of the selected features, the more credible and persuasive the integration results will be. As a novel ensemble classifier, random forest has advantages such as requiring fewer training samples, less manual intervention, and high classification accuracy [11], and can handle high-dimensional data and quickly obtain classification results, meeting the needs of fragmented university information integration. To address the fragmented information needs of users, this study proposes a fragmented university information integration process. The integration system consists of two main parts: the information integration process and the user access process.

2.2.1 Information Integration Process The information integration process primarily targets numerous heterogeneous databases in universities, such as personnel management systems, academic management systems, student management systems, research management systems, financial management systems, asset management systems, and campus card systems. It obtains large amounts of structured, semi-structured, and unstructured data from these systems and transforms them into "knowledge fragments" through unified fragmentation processing, storing them in a knowledge fragment sharing pool to achieve fragmented integration of university resources.

2.2.2 User Access Process The user access process primarily queries the historical requirement feature database based on user requirements to deter-

mine whether a feature set for the same requirement exists. If it exists, the system extracts corresponding "knowledge fragments" with those features from the knowledge fragment sharing pool according to the historical feature requirement set. If it does not exist, the system needs to use the random forest-based feature selection model to extract a feature set that satisfies the requirement, then extract corresponding "knowledge fragments" based on this feature set, and finally integrate the "knowledge fragments" that match the requirement features and feed them back to users in visualized form.

3. Construction of Random Forest-Based Feature Selection Model for Fragmented University Information Integration

3.1 Random Forest Algorithm

Random Forest is a machine learning algorithm with good classification performance proposed by American scholar L. Breiman in 2001. Its basic idea is to use the Bagging method to randomly draw different training sample sets with replacement, and construct corresponding decision trees for each sampled set, thereby forming a Random Forest model [12]. Random Forest consists of a set of decision tree classifiers. Random training sample subsets, i.e., using the bagging method to randomly draw n training sample sets of the same size as the original sample set with replacement from the samples, can ensure that approximately 63% of the samples from the initial training set appear in the sample sets. Random feature subspaces mean that when splitting each node of the decision tree, an attribute subset is randomly drawn with equal probability from all attributes, typically taking $\sqrt{M}$ features, where M is the total number of features, and each time an optimal attribute is selected from this subset to split the remaining samples at the current node [14].

Since the process of generating decision trees is independent, the Random Forest algorithm can be processed in parallel. Meanwhile, it has high classification accuracy and fast training speed, can effectively overcome overfitting problems, and is able to evaluate feature importance [15-16]. Therefore, the Random Forest algorithm is suitable for feature selection in the process of fragmented university information integration.

3.2 Feature Importance Calculation

Since Random Forest algorithm selects features completely randomly, i.e., each feature has an equal probability of being selected, it assumes each feature has the same importance for the target requirement. However, in the process of fragmented university information integration, we find that a large number of features increases model complexity without significantly affecting integration results. That is, in reality, each feature has different importance for different integration requirements and different impacts on node splitting. Therefore, it is necessary to ensure integration result accuracy while screening out features with higher importance through feature importance calculation.

Random Forest feature importance scoring statistics can be calculated using two methods: Gini index and Out-of-Bag (OOB) error rate [17-18]. This study calculates feature importance based on the Gini index. Given a set of random variables x_1, x_2, ..., x , the score statistic for variable x is denoted as VIM(Gini), representing the average change in node splitting impurity for the j-th variable across all decision trees in the Random Forest. The calculation process for VIM(Gini) is as follows:

The Gini index for a node is:
GI = Σ P (1-P) Formula (1)
where K is the number of classes in the sample set, and P is the probability estimate that samples at node m belong to class K.

The importance of variable x at node m is:
VIM(Gini) = GI - GI = GI Formula (2)
where GI and GI represent the Gini indices of the two new nodes split from node m.

If variable x appears M times in the i-th tree, the importance of variable x in the i-th tree is:
VIM(Gini) = Σ VIM(Gini) Formula (4)
where n is the number of decision trees in the Random Forest.

3.3 Evaluation Metrics for Feature Selection Performance

For classification prediction problems, commonly used evaluation metrics include recall rate, precision rate, and classification accuracy. For the integration feature selection model in this study, we similarly define a set of evaluation metrics [19]: algorithm recall rate (Rec), algorithm precision rate (Pre), algorithm classification accuracy (Acc), and AUC (Area Under the ROC Curve).

Assuming TP represents the number of actual positive cases correctly identified as positive, FP represents the number of actual negative cases incorrectly identified as positive, FN represents the number of actual positive cases incorrectly identified as negative, and TN represents the number of actual negative cases correctly identified as negative, then:

Algorithm recall rate is: TP / (TP + FN) Formula (5), representing the proportion of positive samples correctly classified among all positive samples.

Algorithm precision rate is: TP / (TP + FP) Formula (6), representing the proportion of positive samples correctly classified among those classified as positive.

Algorithm classification accuracy is: (TP + TN) / (TP + TN + FP + FN) Formula (7), representing the proportion of all samples correctly classified. This metric measures overall classification accuracy, where higher Acc values indicate better classification performance.

Additionally, since sample data used in university information integration typically has imbalanced positive-negative ratios, AUC must also be used as one of the model evaluation indicators [20-21]. Its calculation formula is:
AUC = (Σ rank - M(M+1)) / (MN) Formula (8)
where M is the number of positive samples and N is the number of negative samples.

This study uses K-fold cross-validation [22] to estimate classification accuracy. The complete dataset is divided into K roughly equal subsets. Each time, a different (K-1) subsets are used to train the model, and the remaining one subset is used to test the model. This process is repeated K times, and finally, the mean of the evaluation metrics obtained is used as the indicator estimate for the selected features.

3.4 Design of Integration Feature Selection Model Based on Random Forest

The Random Forest-based feature selection model for fragmented university information integration consists of three main modules: feature extraction module, training module, and testing module. As shown in Figure 2 [Figure 2: see original paper]:

3.4.1 Feature Extraction Module This module primarily extracts features from the sample set to form a complete feature collection, then calculates the importance of all features based on the Gini index and ranks all features in descending order. Since a large number of feature vectors not only fails to affect integration results but also increases model complexity, it is necessary to set a threshold λ, compare the ranking results with the selected threshold, and select the top λ feature vectors to form an optimized feature set for training.

3.4.2 Training Module This module primarily inputs the optimized feature set generated by the feature extraction module, samples a portion of data as training samples, and uses the rest as test samples. It creates a Random Forest classification model on the training samples, ultimately forming a collection of decision trees for the Random Forest.

3.4.3 Testing Module This module inputs test samples into the trained decision tree set, obtains classification results for the feature set, and evaluates the accuracy of classification results by calculating evaluation metrics to determine the quality of feature selection. Finally, based on the optimal evaluation metric results, it determines the number of integration features and forms the optimal feature set.

4. Experiments and Results Discussion

4.1 Experimental Background and Data Description

The process of identifying financially needy students in Chinese universities faces many difficulties, such as determining which indicators among numerous student poverty metrics can reflect the degree of poverty, how to control the influence of subjective human factors, and how to balance the transparency of the identification process with student privacy confidentiality [23]. This experiment uses the identification of financially needy students as an example to verify the accuracy and effectiveness of Random Forest in the feature selection process of university information integration.

Experimental data were collected from 430 students in a particular grade at a university, including student basic information tables, student family situation tables, student consumption tables, student loan tables, student work-study tables, and student financial need application forms. In the dataset, TP denotes the number of samples correctly identified as financially needy students, TN denotes the number of samples correctly identified as non-needy students, FP denotes the number of samples actually needy but incorrectly identified as non-needy, and FN denotes the number of samples actually non-needy but incorrectly identified as needy.

During the experiment, ten-fold cross-validation was adopted, dividing the dataset into training and test sets, with the training set accounting for 90% of the total data used to design and construct the Random Forest algorithm, and the remaining 10% used as the test set to evaluate algorithm performance.

4.2 Feature Importance Ranking for Financial Need Identification

Using the Random Forest model to construct 200 decision trees for feature selection in financial need identification, relevant features were selected and their importance calculated based on the Gini index. With threshold $\lambda = 27$, the top 27 features by importance ranking were selected to form a feature set. The feature descriptions and importance values are shown in Table 1 .

4.3 Experimental Results

Numerous studies have shown that more features are not necessarily better. Too many features not only fails to affect integration results but also increases model complexity. The purpose of the integration feature selection model is to find the optimal set of integration features that satisfy user requirements through sample data learning, thereby making integration results more reliable. Therefore, we need to compare model accuracy under different feature quantities and find the optimal integration feature set. The experiment selected 3, 6, 9, 12, 15, 18, 21, 24, and 27 feature vectors for training, and the evaluation metric results are shown in Figure 3 [Figure 3: see original paper].

Based on Figure 3, the following conclusions can be drawn: (1) When the number of feature vectors is less than 9, accuracy (Acc), recall rate (Rec), precision rate (Pre), and AUC all change significantly. When the number of feature vectors exceeds 9, each metric begins to stabilize. When the number of feature vectors exceeds 15, AUC begins to decrease.

By examining the performance of evaluation metrics in the comparison chart when introducing different numbers of features, it is evident that introducing more feature vectors does not necessarily lead to better model fitting and prediction results. Therefore, feature vector selection is necessary. (2) AUC achieves optimal classification results with 9-15 feature vectors, reaching up to 75%, with Acc and Pre both above 80% and Rec as high as 95%. This experiment extracted the 12 most important feature variables, obtaining satisfactory evaluation results. The evaluation metrics for the optimal feature set are shown in Table 2 .

Table 2 : Evaluation Metrics for Optimal Feature Set

Feature	Importance Score
family_{income}	1.0051E?01
isDisabled	8.5436E?02
hasSeriousIllness	6.0030E?02
isLowIncome	5.2865E?02
monthly_{consumption}	5.0087E?02
isFromPoorArea	4.9823E?02
hasLuxuryConsumption	4.8374E?02
hasMajorDisaster	4.7693E?02
family_{size}	4.6351E?02
numberInSchool	4.5714E?02
isApplyPoorStudent	4.2302E?02
isWorkStudy	4.0310E?02
monthlyWorkIncome	3.9964E?02
class_{job}	2.8399E?02
birthplace	2.8120E?02
isMartyrChild	2.7384E?02
course_{credit}	2.7097E?02
major_{ranking}	2.6133E?02
isSingleParent	2.4649E?02
loan_{timeLimit}	2.3208E?02
health_{condition}	2.1444E?02
healthcondition_{parents}	2.0467E?02
working_{class}	1.7513E?02
isFromCountry	1.7214E?02
isActive	1.0617E?02
(additional features)	1.0609E?02

Feature	Importance Score
(additional features)	7.6888E?03

The combination shows excellent performance in accuracy, recall rate, precision rate, and AUC, indicating that this feature set can provide a good reference basis for identifying financially needy students in universities.

This study proposes the concept of fragmented university information integration, i.e., selecting feature sets that satisfy requirements based on specific business needs, and integrating knowledge fragments according to the optimal feature set, thereby constructing a process for fragmented university information integration. The core of this process lies in how to select optimal integration features. Random Forest's good generalization and robustness, insensitivity to noise, and ability to handle continuous attributes make it highly suitable for constructing university information integration feature selection models. Therefore, this study combines the advantages of Random Forest to construct a Random Forest-based feature selection model for fragmented university information integration and analyzes the main modules of the model.

This study verifies through the experiment of identifying financially needy students that the model has high accuracy and precision in selecting university integration features, providing a feasible approach for university information resource integration. Although Random Forest has good discriminative power, this method causes features with large weights to always be selected, thereby reducing the diversity of feature subspaces and making the correlation between individual decision trees too high, which instead increases generalization error. Therefore, subsequent improvements to the algorithm are needed to enhance feature relevance while reducing generalization error.

References

[1] Ma Wenfeng, Du Xiaoyong. Data-based resource integration [J]. Information and Documentation Services, 2007(1): 41-45.
[2] Ma Wenfeng, Du Xiaoyong, Hu Ning. Information-based resource integration [J]. Information and Documentation Services, 2007(1): 46-50, 70.
[3] Ma Wenfeng, Du Xiaoyong, Lu Xiaohui. Knowledge-based resource integration [J]. Information and Documentation Services, 2007(1): 51-56.
[4] Chang Tongshan. Application of data mining technology in American institutional research [J]. Fudan Education Forum, 2009(2): 72-79.
[5] Liao Fenglu, Zhou Qing. Application of EDM in predicting graduate employability [J]. Education Teaching Forum, 2017(33): 65-66.
[6] Shi ?, Qian Yuan, Sun Ling. Research on online learning process supervision based on educational data mining [J]. Modern Educational Technology, 2016, 26(6): 87-93.
[7] Shu Zhongmei, Xu Xiaodong. Educational data mining of university student

satisfaction from learning analytics perspective [J]. e-Education Research, 2014(5): 39-44.
[8] He Shiming, Shen Jun. Online learning evaluation method based on BP neural network [J]. Microcomputer Development, 2004, 14(12): 26-29.
[9] Liu Meiling, Li Xi, Li Yongsheng. Application of data mining technology in university teaching and management [J]. China Education Informatization, 2016(19): 11-13.
[10] Li Hengbei, Cha Guiting, Mao Liju, et al. Fragmented service-based university informatization architecture and practice [J]. China Education Informatization, 2016(19): 11-13.
[11] Fang Kuangnan, Wu Jianbin, Zhu Jianping, et al. Review of Random Forest methods [J]. Statistics and Information Forum, 2011, 26(3): 32-38.
[12] BREIMAN L. Random forests [J]. Machine learning, 2001, 45(1): 5-32.
[13] ZHOU Zhihua. Machine Learning [M]. Beijing: Tsinghua University Press, 2016: 178-180.
[14] LIU Y, CHEN M. Random forest method and application in stream big data systems [J]. Journal of Northwestern Polytechnical University, 2015, 33(6): 1055-1061.
[15] Wu Chenwen, Wang Wei, Li Changsheng, et al. A feature selection method combining Random Forest and neighborhood rough set [J]. Small Microcomputer System, 2017, 38(6): 1358-1362.
[16] YAO D, YANG J, ZHANG X. Feature selection algorithm based on random forest [J]. Journal of Jilin University (Engineering and Technology Edition), 2014, 44(1): 137-141.
[17] Yang Kai, Hou Yan, Li Kang. Random forest variable importance scoring and its research progress [EB/OL]. [2017-08-25]. http://www.paper.edu.cn/html/releasepaper/2015/07/212/.
[18] ARCHER KJ, KIMES RV. Empirical characterization of random forest variable importance measures [J]. Computational statistics & data analysis, 2008, 52(4): 2249-2260.
[19] Wang Xiaojie, Sun Rencheng, Shao Fengjing. Random forest-based prediction of user abandonment of online courses [J]. Journal of Qingdao University (Engineering & Technology Edition), 2016, 31(4): 17-21.
[20] Zhang Xiaofeng, Hou Yan, Li Kang. Research on random forest variable importance scoring based on AUC statistic [J]. Chinese Journal of Health Statistics, 2016, 33(3): 537-540, 542.
[21] JANUTZA S, STROBL C, BOULESTEIX AL. An AUC-based permutation variable importance measure for random forests [J]. BMC bioinformatics, 2013, 14(3): 433-440.
[22] Wang Yuyan, Wang Dujuan, Wang Yanzhang, et al. Improved random forest ensemble classification method for predicting colorectal cancer survivability [J]. Management Science, 2017, 30(1): 95-106.
[23] Liu Haiyuan. Research on financial aid identification assistance system based on data mining [J]. Computer Knowledge and Technology, 2015, 11(24): 5-7.

Author Contributions

Zhang Wende: Proposed research ideas, revised final version;
Cheng Han: Designed research plan, drafted paper;
Liu Tian: Designed experimental plan and collected data;
Zeng Jinjin: Conducted experimental analysis.

Application of Random Forest in the Fragmented Integration of University Information

Zhang Wende[1], Cheng Han[1], Liu Tian[2], Zeng Jinjin[2]

[1]Institute of Information Management, Fuzhou University, Fuzhou 350108
[2]Library of Fujian Agriculture and Forestry University, Fuzhou 350002

Abstract: [Purpose/significance] Facing the trend of fragmentation of university information, this paper puts forward the integration process of fragmented university information, and applies the random forest algorithm to construct the feature selection model of information fragmented integration in universities. [Method/process] This paper presents the development, research status and existing problems of university information integration. Furthermore, in this paper, we elaborate the principles and advantages of the random forest algorithm, and use it to the feature selection model of information fragmented integration process in universities. Finally, we validate the model by using the example of identifying the students in the need of financial help. [Result/conclusion] Random forest algorithm shows higher accuracy and validity in the selection of features for integrating university information and therefore provides a new way for the integration of fragmented university information.

Keywords: random forest; fragmentation; information integration; feature selection

Note: Figure translations are in progress. See original paper for figures.