Lecture Notes on Data Engineering and Communications Technologies 156

Bernard J. Jansen
Qingyuan Zhou
Jun Ye *Editors*

Proceedings of the 2nd International Conference on Cognitive Based Information Processing and Applications (CIPA 2022)

Volume 2

Springer

Lecture Notes on Data Engineering and Communications Technologies

Volume 156

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It will publish latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series will have a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

Indexed by SCOPUS, INSPEC, EI Compendex.

All books published in the series are submitted for consideration in Web of Science.

Bernard J. Jansen · Qingyuan Zhou · Jun Ye
Editors

Proceedings of the 2nd International Conference on Cognitive Based Information Processing and Applications (CIPA 2022)

Volume 2

Springer

Editors
Bernard J. Jansen
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar

Jun Ye
School of Computer Science
and Cyberspace Security
Hainan University
Haikou, Hainan, China

Qingyuan Zhou
School of Economics and Management
Changzhou Institute of Mechatronic
Technology
Changzhou, China

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-981-19-9375-6 ISBN 978-981-19-9376-3 (eBook)
https://doi.org/10.1007/978-981-19-9376-3

Conference Committee

General Chair

Lin Shen, Changzhou Vocational Institute of Mechatronic Technology, China

Local Organizing Chairs

Jin Lou, Changzhou Vocational Institute of Mechatronic Technology, China
Weijie Gu, Changzhou Vocational Institute of Mechatronic Technology, China

Program Chairs

Bernard J. Jansen, Qatar Computing Research Institute, Qatar
Qingyuan Zhou, Changzhou Vocational Institute of Mechatronic Technology, China
Jun Ye, Hainan University, China

Publicity Chairs

Haibo Liang, Southwest Petroleum University, China
Sulin Pang, Jinan University, China

Program Committee

Ameer Al-Nemrat, University of East London, UK
Robert Ching-Hsien Hsu, Chung Hua University, China
Neil Yen, University of Aizu, Japan
Meng Yu, The University of Texas at San Antonio, USA
Shunxiang Zhang, Anhui University of Science and Technology, China
William Liu, Auckland University of Technology, New Zealand
Mustafa Mat Deris, Universiti Tun Hussein Onn Malaysia, Malaysia
Zaher Al Aghbari, Sharjah University, UAE
Guangli Zhu, Anhui University of Science and Technology, China
Raja Al Jaljouli, College of Computer Science and Engineering, Kingdom of Saudi Arabia
Abdul Basit Darem, University of Mysore, India
Vijay Kumar, VIT, India
Xiangfeng Luo, Shanghai University, China
Jemal Abawajy, Deakin University
Ahmed Mohamed Khedr, University of Sharjah, UAE
Xiao Wei, Shanghai University, China
Sabu M. Thampi, Indian Institute of Information Technology and Management, India
Huan Du, Shanghai University, China
Shamsul Huda, Deakin University, Australia
Zhiguo Yan, Fudan University, China
Rick Church, UC Santa Barbara, USA
Tom Cova, University of Utah, USA
Susan Cutter, University of South Carolina, USA
Yi Liu, Tsinghua University, China
Kuien Liu, Pivotal Inc., USA
Wei Xu, Renmin University of China, China
V. Vijayakumar, Professor and Associate Dean, SCSE, VIT, Chennai, India
Abdullah Azfar, KPMG Sydney, Australia
Florin Pop, University Politehnica of Bucharest, Romania
Kim-Kwang Raymond Choo, The University of Texas at San Antonio, USA
Mohammed Atiquzzaman, University of Oklahoma, USA
Rafiqul Islam, Charles Sturt University, Australia
Morshed Chowdhury, Deakin University, Australia

Preface

Cognition is emerging as a new and promising methodology with the development of cognitive-inspired computing, cognitive-inspired interaction, and systems, which have the potential to enable a large class of applications. These applications have emerged with great potential to change our lives. However, recent advances in artificial intelligence (AI), fog computing, big data, and cognitive computational theory show that multidisciplinary cognitive-inspired computing still struggles with fundamental, long-standing problems, such as computational models and decision-making mechanisms based on the neurobiological processes of the brain, cognitive sciences, and psychology. How to enhance human cognitive performance with machine learning, common sense, natural language processing, etc., is worth exploring.

The 2nd International Conference on Cognitive-based Information Processing and Applications (CIPA 2022) is held in Changzhou, China, from September 22 to 23, 2022. The conference communicated the theory, technology, and application of artificial intelligence, including precision mining, intelligent computing, deep learning, and all other theories, models, and technologies related to artificial intelligence.

The purpose of CIPA 2022 is to provide a forum for the presentation and discussion of innovative ideas, cutting-edge research results, and novel techniques, methods, and applications on all aspects of technology and intelligence in intelligent computing.

At least two independent experts reviewed each paper. The conference would not have been a reality without the contributions of the authors. We sincerely thank all the authors for their valuable contributions. We want to express our appreciation to all members of the program committee for their valuable efforts in the review process that helped us to guarantee the highest quality of the selected papers for the conference.

We want to express our thanks for the strong support of the publication chairs, the organizing chairs, the program committee members, and all the volunteers.

Our special thanks are also due to the editor of Springer, Ramesh Nath Premnath, for his assistance throughout the publication process.

Doha, Qatar — Bernard J. Jansen
Changzhou, China — Qingyuan Zhou
Haikou, China — Jun Ye

Keynotes

Bernard J. Jansen is a principal scientist in the social computing group of the Qatar Computing Research Institute. He is a graduate of West Point and has a Ph.D. in computer science from Texas A&M University. Professor Jansen is the editor-in-chief of the journal, Information Processing and Management (Elsevier), a member of the editorial boards of seven international journals, and the former editor-in-chief of the journal, *Internet Research* (Emerald). He has received several awards and honors, including an ACM Research Award, six application development awards, and a university-level teaching award, along with other writing, publishing, research, teaching, and leadership honors. Dr. Jansen has authored or co-authored 300 or so research publications, with articles appearing in a multi-disciplinary range of journals and conferences. He is the author of the book, Understanding Sponsored Search: A Coverage of the Core Elements of Keyword Advertising (Cambridge University Press).

Jemal Abawajy is a faculty member at Deakin University and has published more than 100 articles in refereed journals and conferences as well as a number of technical reports. He is on the editorial board of several international journals and edited several international journals and conference proceedings. He has also been a member of the organizing committee for over 60 international conferences and workshops, serving in various capacities, including the best paper award chair, the general co-chair, the publication chair, the vice-chair, and program committee. He is actively involved in funded research in building secure, efficient, and reliable infrastructures for large-scale distributed systems. Toward this vision, he is working in several areas, including pervasive and networked systems (mobile, wireless network, sensor networks, grid, cluster, and P2P), e-science and e-business technologies and applications, and performance analysis and evaluation.

Contents

About the Editors

Bernard J. Jansen is a Principal Scientist in the social computing group of the Qatar Computing Research Institute. He is a graduate of West Point and has a Ph.D. in computer science from Texas A&M University. Professor Jansen is editor-in-chief of the journal, *Information Processing and Management* (Elsevier), a member of the editorial boards of seven international journals, and former editor-in-chief of the journal, *Internet Research* (Emerald). He has received several awards and honors, including an ACM Research Award, six application development awards, and a university-level teaching award, along with other writing, publishing, research, teaching, and leadership honors. Dr. Jansen has authored or co-authored 300 or so research publications, with articles appearing in a multi-disciplinary range of journals and conferences. He is author of the book, *Understanding Sponsored Search: A Coverage of the Core Elements of Keyword Advertising* (Cambridge University Press).

Qingyuan Zhou received Ph.D. degrees from Sichuan University (SCU), Chengdu in 2014, respectively. Currently, he is a professor in Changzhou Institute of Mechatronic Technology, China. He has joined and accomplished 5 national and 6 provincial research programs. He has also authored/coauthored over 60 papers in international/national journals and conferences including *Computers in Human Behavior*, *Electronic Commerce Research*, *Intelligent Automation and Soft Computing* etc. His current research interests include electronic commerce, artificial intelligence, and computational economics.

Jun Ye received his B.S. degree in Applied Mathematics at Chongqing University. M.S. degree in Cryptography at Guilin University of Electronic Technology. Ph.D. in Xidian University, He is a high level talent of Hainan Province, and he is working at school of Computer Science and Cyberspace Security of Hainan University. His current research interests include computer science and information security. He

has authored or co-authored more than 20 high level publications, and he is also a reviewer of many well-known journals. He is one of the high level talents of Hainan Province, and got the "First prize" of science and technology progress prize of Hainan Province in 2019.

Cognitive Environment, Sensing and Data

Data-Driven Fuzzy Clustering Approach in Logistic Regression Model

Xiaofei Li, Yunying Guo, and Jinrui Wei

Abstract Logistic regression, as one of the special cases of generalized linear model, has important role in multi-disciplinary fields for its powerful interpretability. Although there are many similar methods such as linear discriminant analysis, decision tree, boosting and SVM, we always face a trade-off between more powerful interpretability and accurate predictability inevitably. Because the inference and prediction are the two main purposes of a model and the latter has been paid so much attention, this paper studies the logistic regression and its imperfection mending. It is found that there happens to be the non-existence of maximum likelihood estimates when there is relatively clearer partition in logistic regression model. Therefore, a data-driven fuzzy clustering method has been proposed. Then data sets from UCI machine learning repository are employed to evaluate the performance of proposed approach. The experiment results indicate that fuzzy clustering logistic regression model improves prediction accuracy in comparison with decision tree and linear discriminant analysis. At the same time, interpretability of logistic regression has been reserved in this model.

Keywords Data driven · Fuzzy clustering · Logistic regression model

1 Introduction

When the response variable is categorical, which is known as classification, many techniques are available such as linear discriminant analysis, decision tree, boosting and SVM. Among them, logistic regression model has relatively more powerful

X. Li · Y. Guo
School of Big Data Management and Application, Dalian Neusoft University of Information, Dalian 116025, Liaoning, China

J. Wei (✉)
School of Statistics, Dongbei University of Finance and Economics, Dalian 116025, Liaoning, China
e-mail: a13942868163@126.com

B. J. Jansen et al. (eds.), *Proceedings of the 2nd International Conference on Cognitive Based Information Processing and Applications (CIPA 2022)*, Lecture Notes on Data Engineering and Communications Technologies 156,
https://doi.org/10.1007/978-981-19-9376-3_1

interpretability than others, while its predictive ability may be in the shadow of these methods. Because there is always a trade-off between interpretability and predictability in modeling, most of these models put more focus on accurate prediction and treat the function as a black box. In fact, the algorithms in data mining and statistical learning aim at that to a great extent. However, in many situations, inference is the object of modeling. We expect to see the relationship between response variable and predictor variables, the meaning of the model parameters and attributes' relative importance.

Maximum likelihood has been used to estimate the model, but it obtains numerical solution rather than the closed form generally. And there is a challenge that numerical algorithms could fail to converge. For instance, pattern recognition [1], Logit regression [2] and discrete data analysis [3] have encountered the warring message. References [4–6] study this problem extensively on the existence of maximum likelihood estimates. They give the existence conditions, necessary or sufficient, but did not tell us what we should do if the conditions cannot be satisfied. Inspired by these studies, clustering method has been employed. And in order to obtain corresponding probability, which has been transformed to odds, fuzzy cluster has been proposed. Interestingly, we acquire clearer interpretability by the fuzzy method.

The rest of the paper is organized as follows. In Sect. 2, on account of non-existence of maximum likelihood estimates for numerical algorithms failing to converge in some cases, data-driven fuzzy clustering approach is proposed for the solution of non-existence of maximum likelihood estimates. Then a case of non-existence of solution has been explored by data-driven fuzzy clustering approach, and some comparison with decision tree and linear discriminate analysis has been made in Sect. 3. Finally, the conclusion and remarks are drawn in Sect. 4.

2 Method Introduction

The issues of existence of maximum likelihood estimates in logistic regression models have received considerable attention in the literature [7, 8]. Concerning multinomial logistic regression models, reference [9] has proved existence theorems under consideration of the possible configurations of data points, which separated into three detailed and mutually exclusive categories: complete separation, quasi-complete separation and overlap. Unfortunately, there does not exist the solution of maximum likelihood estimates all the time toward multinomial logistic regression, such as the data structure of completely and quasi-completely separated. Unfortunately, there does not exist the solution of maximum likelihood estimates all the time toward multinomial logistic regression, such as the data structure of completely and quasi-completely separated.

But what should we do if there is non-existence of maximum likelihood estimates in logistic regression model? In these cases, how could we make an interpretive estimation the same as this model?

In the examples below, you could see the warning messages about "fitted probabilities numerically 0 or 1 occurred" or "algorithm did not converge" when we use function glm() to run in R. With the help document of this function, where the site is http://127.0.0.1:17145/library/stats/html/glm.html, it offers a reference [9]. However, there is only with the description of this problem. References lead the way in which we investigate the existence conditions and their complex proof of maximum likelihood estimates. The reason is obvious, while we find nothing about next step. Just giving up for conditions is not satisfied. We will show an alternative using fuzzy clustering method.

Fuzzy clustering is a generalization of cluster [10]. Instead of hard clustering that "object x belongs to cluster U", fuzzy clustering method allows for some ambiguity that "object x belongs to cluster U with accuracy 92%". For, where cluster $U = \{x_1, x_2, \ldots, x_n\}$ is a finite data set, membership coefficients $m_{xU} = m(x)$ give the grade of membership of x in U. $m_{xU} = 0$ means that x is not included in cluster U, and hence, $m_{xU} = 1$ means that x is fully included (hard partition). The fuzziest point lies in $m_{xU} = 1/k$, where k stands for the number of clusters. So we could describe the grade of fuzzy solution from a hard clustering by the normalized Dunn's partition coefficient.

$$F_k = \sum_{j=1}^{n} \sum_{v=1}^{k} \frac{m_{jv}^2}{n} \quad \mathrm{NF}_k = \frac{kF_k - 1}{k - 1} \tag{1}$$

If $m_{jv} = 1/k$, the case of complete fuzzy clustering, F_k takes on its minimal value $1/k$, which yields $\mathrm{NF}_k = 1$. For the hard partition, where $m_{jv} = 1$ or 0, $\mathrm{NF}_k = 0$.

There are not any representative objects in fuzzy clustering method. Instead, it attempts to minimize the function.

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} m_{iv}^r m_{jv}^r d(i,j)}{2\sum_{j=1}^{n} m_{jv}^r}, \text{ st. } m_{jv} \geq 0, \ \sum_{v=1}^{k} m_{jv} = 1 \tag{2}$$

To find the dissimilarities $d(i, j)$ and the membership coefficients m_{jv}, $k = 2$ in that case and r is the membership exponent. Fuzzy L_1 method ($r = 1$) has some advantages over fuzzy L_2 ($r = 2$), for example, more robust to outliers or better recognizable of non-spherical clusters. Note that it increases crisper clustering and leads to slow convergence when $r \to 1$, while $r \to \infty$ leads to complete fuzzy (sometimes even the default $r = 2$). In that case, a warning message is advised to turn down the r.

To avoid the interference of measurement unit, we take normalization by $\left(x_j - \mathrm{mean}_j\left(x_j\right)\right)/\mathrm{sd}_j\left(x_j\right)$ before running a fuzzy clustering, when variables are continuous.

$$\mathrm{sd}\left(x_j\right) = \sqrt{\frac{1}{n-1}\sum_{j}\left(x_j - \mathrm{mean}_j\left(x_j\right)\right)^2} \tag{3}$$

The latter is less sensitive to the outlier. Then Minkowski distances (A generalization of the Euclidean and the Manhattan metric) between the observations are computed to construct the dissimilarity matrix. When variables are categorical, simple matching coefficient (also known as M-coefficient or affinity index) is employed in computing dissimilarity.

Now we could transform the membership coefficient m_{jv} to odd m_{j1}/m_{j0}, where $m_{j0} = 1 - m_{j1}$, then regress the logarithm of odds $\ln\left(m_{j1}/m_{j0}\right)$ on input variables

$$\ln\left(m_{j1}/m_{j0}\right) = x_j^T \beta, \quad j = 1, 2, \ldots, n \tag{4}$$

$$\hat{m}_{j1} = \frac{e^{x_j^T \beta}}{1 + e^{x_j^T \beta}}, \quad j = 1, 2, \ldots, n \tag{5}$$

Then it simply assigns an observation j to cluster-1 if $\hat{m}_{j1} > 0.5$, and to cluster-0 otherwise.

$$\hat{\beta}_r = \frac{\partial \ln(\text{odds})}{\partial x_r} = \frac{\partial(\text{odds})/\text{odds}}{\partial x_r} \quad r = 1, 2, \ldots, p \tag{6}$$

It gives odds percentage changes in the response to a unit change in x_r, ceteris paribus.

3 Experiments

The banknote data is about genuine and forged banknote images. Features of images were extracted by wavelet transform. And variance, skewness and Kurtosis of wavelet transformed image were given, also entropy of image. These input variables are all continuous. The output variable class is binary, as a function of these quantitative variables with 1372 observations. A logistic model obtained warning message like "fitted probabilities numerically 0 or 1 occurred".

The summary of FCLR is given in Table 1. Adjusted *R*-squared is 86.02%. *F* statistic is three star significant. Overall error in sample is 2.33%. That is fantastic fitting! Decision tree and linear discriminant analysis have been compared to the approach we come up with here. These two methods are chosen based on their ability of interpretation. That is why boosting is absent. Among these models, DT is a little bit worse than LDA and FCLR.

According to the classical theory of statistics, if a model (or a statistic) is "good", it (empirical distribution or model parameters) will converge to the real values at different intensity as the number of sample goes infinity. That is also a basic assessment criterion for a model or statistic, saying consistency. The realistic situation, however, seldom can reach the level limit of sample size. In fact, it is not necessary to reach the infinite perfectly. The point which is not too far away from home is the

Table 1 Ability of learning and generalization in banknote data

	Var. and error	DT	LDA	FCLR
Learning ability	Intercept	–	–	-0.8436*** (0.0222)
	Variance	–	− 0.8386	0.4124*** (0.0062)
	Skewness	–	− 0.4606	0.2156*** (0.0046)
	Kurtosis	–	− 0.5977	0.2797*** (0.0056)
	Entropy	–	− 0.0047	-0.0127 (0.0093)
	Error-1	0.0066	0.0000	0.0000
	Error-0	0.0577	0.0420	0.0420
	Error-T	0.0350	0.0233	0.0233
Generalization	Error-5CV-1	0.0213	0	0
	Error-5CV-0	0.0367	0.0431	0.0421
	Error-5CV-T	0.0298	0.0233	0.0233

one we have almost met the view of the end of the road. In these cases, the fitting and prediction are consistent. However, in some limited sample size, fitting and prediction are not consistent. Because, when the sample size n is limited and fixed, as the model becomes more and more complicated, the degree of fitting (ability within in the sample, learning ability) can always be improved constantly (training error is a decreasing function of model complexity), but the predictability (ability out of the sample, generalization ability) is not necessary (overfitting) and may even worse. That is why an excellent fitness model may be "wrong", while a not much good one may be "right". Data we used is virtually finite, which inevitably leads to sample error. Then it brings us to the George Box's famous statement "all models are wrong but some are useful". Therefore, in case of finite number of samples, the object of modeling turns to minimize the test error rather than the training error. But we could not compute the test error directly. Because there is not test data most of time, we estimate it by two primitive methods. The one is indirect adjustment on the basis of the training error, such as adjusted R^2, AIC, BIC and Mallow's Cp. All of them introduce punishment for high dimension on the sum of squared residuals. The other is direct estimation by cross-validation. But first of all, the data set has been split into training set and validation set. Here we use fivefold cross-validation to evaluate the generalization.

Table 1 shows whether to prediction or fitting, FCLR is similar to LDA but slightly better than DT as a whole. To check out if FCLR is stable, or not, we plot every fold of validation error in Fig. 2. You could see that DT is unstable in Fig. 1. That is why DT is called weak learner, while relatively, FCLR and LDA are stable.

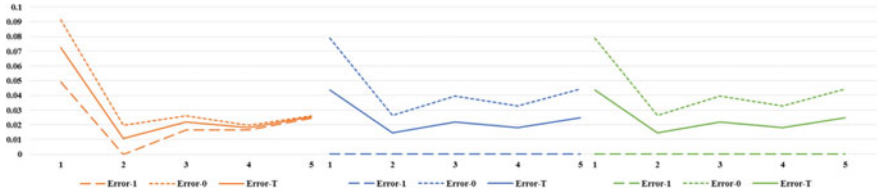

Fig. 1 Line chart of fivefold CV error for banknote data. Cluster-1 recognized error (long dash), cluster-0 recognized error (dotted line) and total recognized error (solid line). Left: decision tree (orange). Center: linear discriminant analysis (blue). Right: fuzzy clustering approach for logistic regression

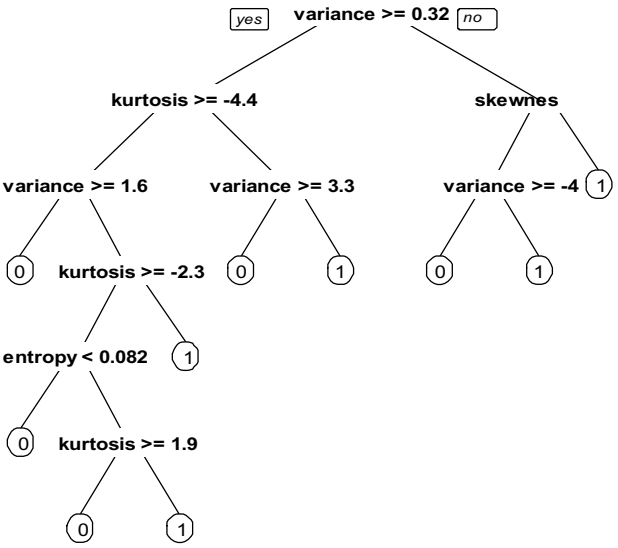

Fig. 2 Decision tree for banknote data

For accurate approximation, we try to find more explanation. Even sometimes, in order to obtain interpretability, we give up some forecast accuracy. It is interesting to note that fuzzy brought a further explanation. In FCLR, we can see not only the relative importance of explanatory variables, but also the effect of predictors on the response variable. Compared with Figs. 1 and 2, we could see that the importance of variable obtained by FCLR is consistent with DT: variance, then the skewness and Kurtosis and finally the entropy. Now we turn to the interpretation of the model parameters. Taking coefficient of variance as example, it means that the odds will increase by 41.24% in response to an additional unit of variance holding all other variables constant. Or through exponentiation, we get odds multiply 1.51 when variance increases one unit. In the same way, if entropy is significant, it means that odds will decrease 1.27% in response to an additional unit of entropy holding all other

variables constant. Or through exponentiation, we get odds 98.74% discount when entropy increase one unit.

In aspect of model prediction accuracy, compared with DT, the learning ability of the FCLR and LDA model in the training set and the generalization ability of the test machine are the most prominent. The FCLR model has stronger explanatory power simultaneously. Simulation results show the effectiveness and efficiency of the presented data-driven fuzzy clustering algorithm.

4 Conclusion

There are two objects in modeling, which are interpretation and prediction. Most of the algorithms and their advanced version focus on the latter, while we pay more attention on the interpretability in this paper. Logistic regress model has been extended to the case of non-existence of maximum likelihood estimates based on fuzzy clustering. One reason we use the term "data driven" is that it is flexible to data. The experiment results show that FCLR improves prediction accuracy in comparison with DT and LDA. At the same time, interpretability of logistic regression has been reserved in this model.

Acknowledgements This work was supported by DUFE202126, L20BTJ003 and SF-Y202113.

References

1. Du H et al (2016) Structured discriminant analysis dictionary learning for pattern classification. Knowl-Based Syst 216(5):98–101
2. Schuster NA, Twisk JWR, ter Riet G et al (2016) Noncollapsibility and its role in quantifying confounding bias in logistic regression. BMC Med Res Methodol 136(21):1037
3. Sergeev VA, Korgin NA (2021) Identification of integrated rating mechanisms as an approach to discrete data analysis. IFAC-PapersOnLine 54(13):134–139
4. Day NE, Kerridge DF (1967) A general maximum likelihood discriminant. Biometrics 23(2):313–323
5. Santner TJ, Duffy DE, A note on A. Albert and J. A (1986) Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. Biometrika 73(3):755–758
6. Sur P, Emmanuel J (2019) Candès: a modern maximum-likelihood theory for high-dimensional logistic regression. Proc Natl Acad Sci 116(29):14516–14525
7. Tang W, Ye Y (2020) The existence of maximum likelihood estimate in high-dimensional generalized linear models with binary responses. Electron J Stat 14(1):4028–4053
8. Candès EJ, Sur P (2020) The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regressions. Ann Stat 48(1):27–42
9. Vernables WN, Ripley BD (2003) Modern applied statistics with S. 52(4):704–705
10. Feng Q, Chen L, Chen C et al (2020) Deep fuzzy clustering—a representation learning approach. IEEE Trans Fuzzy Syst 28(7):1420–1433

Digital Storage of Minority Image Based on Hadoop Technology

Xiaodong Wu and Li Fan

Abstract Hadoop technology is a distributed storage system architecture, which has the advantages of high reliability, low cost, and high scalability and can realize efficient and reliable distributed storage of massive data. By means of video inheritance, combined with modern media technology, and video communication through media such as TV and the Internet, it is possible to expand the influence of minority cultures, attract more audiences, and make them better developed. This paper designs a digital image storage system for ethnic minorities based on Hadoop technology. Based on Hadoop architecture, the conceptual structure and logical structure of digital storage are designed. The storage system adopts the Hadoop architecture to meet the storage and computing requirements of massive data information and at the same time realizes the extended database environment through the SQL Server database.

Keywords Hadoop technology · Image digitization · Storage system · SQL server

1 Introduction

In the era of big data, the amount of information is growing rapidly, resulting in an explosive growth in the amount of corporate and personal data. Hadoop is a big data distributed processing framework that uses distributed clusters to process massive data. Hadoop can use multiple ordinary computers to form a logically storable and computational cluster without requiring special infrastructure. Hadoop technology is a distributed storage system architecture, which has the advantages of high reliability, low cost, and high scalability and can realize efficient and reliable distributed storage of massive data. Abroad, Facebook runs Hadoop to store multidimensional data and internal logs in clusters to support its data analysis and machine learning. In addition, the background processing of website attributes is implemented through Hadoop, and Hadoop is used as the engine to promote the transmission of website information In

X. Wu (✉) · L. Fan
College of Journalism and Communications, Bohai University, Jinzhou, Liaoning, China
e-mail: wuxiaodong728@163.com

B. J. Jansen et al. (eds.), *Proceedings of the 2nd International Conference on Cognitive Based Information Processing and Applications (CIPA 2022)*, Lecture Notes on Data Engineering and Communications Technologies 156,
https://doi.org/10.1007/978-981-19-9376-3_2