December 2025: Top 10 Cited Articles in Database Management Systems Research Articles

International Journal of DatabaseManagement Systems (IJDMS)

*** WJCI Indexed***`

ISSN : 0975-5705 (Online); 0975-5985 (Print)

https://airccse.org/journal/ijdms/index.html

Citations, h-index, i10-index

Citations 3739 h-index 29 i10-index 81

A CRITICAL STUDY OF SELECTED CLASSIFICATION ALGORITHMS FOR LIVER DISEASE DIAGNOSIS

Bendi Venkata Ramana[1] , Prof. M.Surendra Prasad Babu[2] , Prof. N. B. Venkateswarlu[3]

[1]Associate Professor, Dept.of IT, AITAM, Tekkali, A.P. India

[2]Dept. of CS&SE, Andhra University, Visakhapatnam-530 003, A.P, India

[3] Professor, Dept. of CSE, AITAM, Tekkali, A.P., India.

ABSTRACT:

Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. Automatic classification tools may reduce burden on doctors. This paper evaluates the selected classification algorithms for the classification of some liver patient datasets. The classification algorithms considered here are Naïve Bayes classifier, C4.5, Back propagation Neural Network algorithm, and Support Vector Machines. These algorithms are evaluated based on four criteria: Accuracy, Precision, Sensitivity and Specificity.

KEYWORDS:

Classification Algorithms, Data Mining, Liver diagnosis

For More Details : http://airccse.org/journal/ijdms/papers/3211ijdms07.pdf

Volume Link : http://airccse.org/journal/ijdms/current2011.html

Citation Count – 305

REFERENCES

[1] Rong-Ho Lin. An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine 2009;47:53—62.

[2] BUPA Liver Disorders Dataset. UCI repository of machine learning databases. Available from ftp://ftp.ics.uci.edu/pub/machinelearningdatabases/ liverdisorders/bupa.data, last accessed: 07 October 2010.

[3] Prof.M.S.Prasad Babu, Bendi Venkata Ramana, Boddu Raja Sarath Kumar, New Automatic Diagnosis of Liver Status Using Bayesian Classification

[4] Paul R. Harper, A review and comparison of classification algorithms for medical decision making

[5] Schiff's Diseases of the Liver, 10th Edition Copyright ©2007 Lippincott Williams & Wilkins by Schiff, Eugene R.; Sorrell, Michael F.; Maddrey, Willis C.

[6] P. Domingos, M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning 29 (2–3) (1997) 103–130.

[7] Weka-3-4-10jre : data mining with open source machine learning software © 2002-2005 David Scuse and University of Waikato

[8] 16th Edition HARRISON'S PRINCIPLES of Internal Medicine

[9] Wendy Webber Chapman,* Marcelo Fizman,† Brian E. Chapman,‡ and Peter J. Haug†, A Comparison of Classification Algorithms to Automatically Identify Chest X-Ray Reports That Support Pneumonia.

[10] Kemal Polat, Seral Sahan, Halife Kodaz and Salih Gunes, Breast Cancer and Liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism.

[11] Michael J. Sorich,† John O. Miners,*,‡ Ross A. McKinnon,† David A. Winkler,§ Frank R. Burden,| and Paul A. Smith‡ Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP- Glucuronosyltransferase Isoforms

[12] Paul R. Harper, A review and comparison of classification algorithms for decision making

[13] Mitchell TM. Machine learning. Boston, MA: McGraw-Hill, 1997.

[14] Lung-Cheng Huang, Sen- Yen Hsu and Eugene Lin, A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data (2009).

CUSTOMER DATA CLUSTERING USING DATA MINING TECHNIQUE

Dr. Sankar Rajagopal

Enterprise DW/BI Consultant Tata Consultancy Services, Newark, DE, USA

ABSTRACT:

Classification and patterns extraction from customer data is very important for business support and decision making. Timely identification of newly emerging trends is very important in business process. Large companies are having huge volume of data but starving for knowledge. To overcome the organization current issue, the new breed of technique is required that has intelligence and capability to solve the knowledge scarcity and the technique is called Data mining. The objectives of this paper are to identify the high-profit, high-value and low-risk customers by one of the data mining technique - customer clustering. In the first phase, cleansing the data and developed the patterns via demographic clustering algorithm using IBM I-Miner. In the second phase, profiling the data, develop the clusters and identify the high-value low-risk customers. This cluster typically represents the 10-20 percent of customers which yields 80% of the revenue.

KEYWORDS:

Data mining, customer clustering and I-Miner

For More Details : http://airccse.org/journal/ijdms/papers/3411ijdms01.pdf

Volume Link : http://airccse.org/journal/ijdms/current2011.html

REFERENCES

[1] Lefait, G. and Kechadi, T, (2010) "Customer Segmentation Architecture Based on Clustering Techniques" Digital Society, ICDS'10, Fourth International Conference, 10 -02-2010.

[2] Fraley, Andrew, and Thearting, Kurt (1999). Increasing customer value by integrating data mining and campaign management software. DataManagement, 49–53.

[3] P. Bhargavi and S.Jyothi, (2009) "Applying Naïve Bayes Data Mining Technique for Classification of Agricultural land Soils" IJCSNS International Journal of computer Science and Network Security, VOL. 9 No.8, August 117-122.

[4] I.Krishna Murthy, "Data Mining- Statistics Applications: A Key to Managerial Decision Making", Article/Report indiastat.com, April-May 2010.

[5] Association Analysis of Customer Services from the Enterprise Customer Management SystemICDM-2006.

[6] Terry Harris, (2008)"Optimization creates lean green supply chains", Data Mining Book

[7] Matt Hartely (2005)"Using Data Mining to predict inventory levels" Data Mining Book

[8] Hu, Tung-Lai, & Sheub, Jiuh-Biing (2003). "A fuzzy-based customer classification method for demand-responsive logistical distribution operations", Fuzzy Sets and Systems, 139, 431– 450.

[9] Hwang, Hyunseok, Jung, Taesoo, & Suh, Euiho (2004). "An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry", Expert Systems with Applications, 26, 181–188.

[10] Jiao, Jianxin, & Zhang, Yiyang (2005). "Product portfolio identification based on association rule mining" Computer-Aided Design, 37, 149–172.

[11] Jonker, Jedid-Jah, Piersma, Nanda, & Poel, Dirk Van den (2004). "Joint optimization of customer segmentation and marketing policy to maximize long-term profitability". Expert Systems with Applications, 27, 159–168.

[12] Kim, Su-Yeon, Jung, Tae-Soo, Suh, Eui-Ho, & Hwang, Hyun-Seok (2006). "Customer segmentation and strategy development based on customer life time value": A case study. Expert Systems with Applications, 31, 101–107.

[13] Kim, Yong Seog, & Street, W. Nick (2004). "An intelligent system for customer targeting: A data mining approach". Decision Support Systems, 37, 215–228.

[14] Kim, Yong Seog, Street, W. Nick, Russell, Gary J., & Menczer, Filippo (2005). "Customer Targeting: A Neural Network Approach Guided by Genetic Algorithms." Management Science, 51(2), 264– 276.

[15] Kuo, R. J., An, Y. L., Wang, H. S., & Chung, W. J. (2006). "Integration of self-organizing feature maps neural net work and genetic K-means algorithm for market segmentation". Expert Systems with Applications, 30, 313–324.

[16] Pham, D.T. and Afify, A.A. (2006) "Clustering techniques and their applications in engineering". Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science,

[17] A.K. Jain, M.N. Murty, and P. J. Flynn,(1999) "Data clustering: a review", ACM Computing Surveys (CSUR), Vol.31, Issue 3,, 1999.

[18] Grabmeier, J. and Rudolph, A. (2002) "Techniques of cluster algorithms in data mining" Data Mining and Knowledge Discovery, 6, 303-360.

[19] Han, J. and Kamber, M. (2001) "Data Mining: Concepts and Techniques", (Academic Press, San Diego, California, USA).

[20] Jiyuan An , Jeffrey Xu Yu , Chotirat Ann Ratanamahatana , Yi-Ping Phoebe Chen,(2007) "A dimensionality reduction algorithm and its application for interactive visualization", Journal of Visual Languages and Computing, v.18 n.1, , February p.48-70.

[21] Nargess Memarsadeghi , Dianne P. O'Leary,(2003) "Classified Information: The Data Clustering Problem", Computing in Scienceand Engineering, v.5 n.5, September p.54-60,

[22] Yifan Li , Jiawei Han , Jiong Yang, (2004) "Clustering moving objects", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, August 22-25, Seattle, WA, USA

[23] Sherin M. Youssef , Mohamed Rizk , Mohamed El-Sherif, (2008) "Enhanced swarm-like agents for dynamically adaptive data clustering", Proceedings of the 2nd WSEAS International Conference on Computer Engineering and Applications, p.213-219, January 25-27, Acapulco, Mexico

[24] Marcel Brun , Chao Sima , Jianping Hua , James Lowey , Brent Carroll , Edward Suh , Edward R. Dougherty,(2007) Model-based evaluation of clustering validation measures, Pattern Recognition, v.40 n.3, March p.807-824.

AUTHOR:

Sankar Rajagopal received M.Sc., degree in Electronics, M.E degree in Materials Science and Ph.D in Metallurgical Engineering from Bharathidasan University in 1990, 1992 and Indian Institute of Technology-Madras, Chennai 1999 respectively. After completion of his doctoral degree, he joined in TATA Consultancy Services as a Software Consultant. Then he has elevated to position Enterprise DW/BI Architect. His areas of research interests include Metal Matrix Composites, Mechanical alloying, Material Informatics, Mechanical Behaviors of Materials, Nanotechnology and Data Mining and Knowledge Discovery. He has published about 15 contributed peer reviewed papers at National / International Journals and Conferences. He received best oral presentation awards in conferences.

HIGH CAPACITY DATA HIDING USING LSB STEGANOGRAPHY AND ENCRYPTION

Shamim Ahmed Laskar[1] and Kattamanchi Hemachandran[2]

Department of Computer Science Assam University, Silchar, Assam, India

ABSTRACT

The network provides a method of communication to distribute information to the masses. With the growth of data communication over computer network, the security of information has become a major issue. Steganography and cryptography are two different data hiding techniques. Steganography hides messages inside some other digital media. Cryptography, on the other hand obscures the content of the message. We propose a high capacity data embedding approach by the combination of Steganography and cryptography. In the process a message is first encrypted using transposition cipher method and then the encrypted message is embedded inside an image using LSB insertion method. The combination of these two methods will enhance the security of the data embedded. This combinational methodology will satisfy the requirements such as capacity, security and robustness for secure data transmission over an open channel. A comparative analysis is made to demonstrate the effectiveness of the proposed method by computing Mean square error (MSE) and Peak Signal to Noise Ratio (PSNR). We analyzed the data hiding technique using the image performance parameters like Entropy, Mean and Standard Deviation. The stego images are tested by transmitting them and the embedded data are successfully extracted by the receiver. The main objective in this paper is to provide resistance against visual and statistical attacks as well as high capacity.

KEYWORDS

Steganography, Cryptography, plain text, encryption, decryption, transposition cipher, Least Significant Bit, Human Visual System, Mean square error and Peak Signal to Noise Ratio.

REFERENCES

[1] Anderson , R. J. and Petitcolas, F. A.P. (1998) "On The Limits of Steganography", IEEE Journal of Selected Areas in Communications, Vol.16 No.4, pp.474-481, ISSN 0733-8716.

[2] Petitcolas, F.A.P., Anderson, R. J. and Kuhn, M.G. (1999) "Information Hiding -A Survey", Proceedings of the IEEE, Special issue on Protection of Multimedia Content, vol. 87, no. 7, pp.1062-1078.

[3] Johnson, N.F. and Jajodia, S. (1998) "Exploring Steganography: Seeing the Unseen", IEEE, Computer, vol. 31, no. 2, pp. 26-34.

[4] Raphael, A. J. and Sundaram, V. "Cryptography and Steganography – A Survey", Int. J. Comp. Tech. Appl., Vol 2 (3), pp. 626-630 , ISSN:2229-6093.

[5] Gutte, R. S. and Chincholkar, Y. D. (2012) "Comparison of Steganography at One LSB and Two LSB Positions", International Journal of Computer Applications, Vol.49,no.11, pp.1 -7.

[6] Laskar, S.A. and Hemachandran, K. (2012), "An Analysis of Steganography and Steganalysis Techniques", Assam University Journal of Sscience and Technology, Vol.9, No.II, pp.83 -103, ISSN: 0975-2773.

[7] Younes, M.A.B. and Jantan, A. (2008), "Image Encryption Using Block-Based Transformation Algorithm," International Journal of Computer Science, Vol. 35, Issue.1, pp.15- 23.

[8] Walia, E., Jain, P. and Navdeep. (2010), " An Analysis of LSB & DCT based Steganography", Global Journal of Computer Science and Technology, Vol. 10 Issue 1 , pp 4 -8.

[9] Khare, P., Singh, J. and Tiwari, M. (2011), "Digital Image Steganography", Journal of Engineering Research and Studies, Vol. II, Issue III, pp. 101-104, ISSN:0976-7916.

[10] Sokouti, M., Sokouti, B. and Pashazadeh, S. (2009), "An approach in improving transposition cipher system", Indian Journal of Science and Technology, Vol.2 No. 8, pp. 9-15, ISSN: 0974- 6846.

[11] Kharrazi, M., Sencar, H. T. and Memon, N. (2006), "Performance study of common image steganography and steganalysis techniques", Journal of Electronic Imaging, SPIE Proceedings Vol. 5681.15(4), 041104 pp.1-16.

[12] R., Chandramouli, and Nasir Memon.(2001), "Analysis of LSB based image steganography techniques." In Image Processing, 2001. Proceedings. 2001 International Conference on, IEEE, vol. 3, pp. 1019-1022.

[13] Giddy, J.P. and Safavi- Naini, R. (1994), " Automated Cryptanalysis of Transposition Ciphers", The Computer Journal, Vol.37, No.5, pp. 429-436.

[14] Johnson, N. F. and Katzenbeisser, S. (2000), "A survey of steganographic techniques", In Information Hiding, Artech House, Norwood, MA, pp. 43-78.

[15] Chandramouli, R. and Menon, N. (2001), "Analysis of LSB based image steganography techniques", IEEE Proceedings on Image Processing, Vol.3, pp.1019-1022.

[16] Carvajal-Gamez , B.E., Gallegos-Funes, F. J. and Lopez-Bonilla, J. L. (2009), " Scaling

Factor for RGB Images to Steganography Applications", Journal of Vectorial Relativity, Vol. 4, no. 3, pp.55-65.

[17] Ulutas, G., Ulutas, M. and Nabiyev,V. (2011), "Distortion free geometry based secret image sharing", Elsevier Inc, Procedia Computer Science 3, pp.721–726.

[18] Tiwari, N. and Shandilya, M. (2010), "Evaluation of Various LSB based Methods of Image Steganography on GIF File Format", International Journal of Computer Applications (0975 – 8887) Vol. 6, no.2, pp.1-4.

[19] Rabah, K. (2004), "Steganography – The Art of Hiding Data", Information Technology Journal, Vol.3, no.3, pp. 245-269.

[20] Deshpande, N., Kamalapur, S. and Daisy, J. (2006), "Implementation of LSB steganography and Its Evaluation for Various Bits", 1st International Conference on Digital Information Management, pp.173-178.

[21] Karen, Bailey, and Kevin Curran.(2006) "An evaluation of image based steganography methods" Multimedia Tools and Applications, Springer Vol.30, no. 1, pp. 55-88.

[22] Celik, M. U., Sharma, G., Tekalp, A.M. and Saber, E. (2005), "Lossless Generalized-LSB Data Embedding", IEEE Transaction on Image Processing, Vol. 14, No. 2, pp. 253-266.

[23] Huang, Y. S., Huang, Y. P., Huang, K.N. and Young, M. S. (2005), "The Assessment System of Human Visual Spectral Sensitivity Curve by Frequency Modulated Light", Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pp. 263-265.

[24] Chan, Chi-Kwong, and L. M. Cheng. (2004), "Hiding data in images by simple LSB substitution." Pattern Recognition Vol. 37, no. 3, pp. 469-474.

[25] Brisbane, G., Safavi-Naini, R. and Ogunbona, P. 2005. "High-capacity steganography using a shared colour palette", IEEE Proceedings on Vision, Image and Signal Processing, Vol.152, No.6, pp.787- 792.

[26] Curran, K. and Bailey, K. (2003), "An Evaluation of Image Based Steganography Methods", International Journal of Digital Evidence Fall 2003, Volume 2, Issue 2, www.ijde.org.

[27] Dickman, S.D. (2007), "An Overview of Steganography", JMU-INFOSEC-TR-2007-002, http://citeseerx.ist.psu.edu/viewdoc/summary? doi=10.1.1.137.5129.

[28] Dunbar, B. (2002). "A detailed look at Steganographic Techniques and their use in an Open-Systems Environment", SANS Institute 2002, pp.1-9, http://www.sans.org.

[29] Lee, Y-K. ; Bell, G., Huang, S-Y., Wang, R-Z. and Shyu, S-J. (2009), "An Advanced LeastSignificant-Bit Embedding Scheme for Steganographic Encoding",PSIVT 2009, LNCS 5414, Springer, pp. 349–360.

[30] Smith, C. (2001), “Basic Cryptanalysis Techniques”, SANS Institute 2001, GSEC Version 1.2f, http://www.sans.org.

[31] Kaur, R., Singh, B. and Singh, I. (2012), “A Comparative Study of Combination of Different Bit Positions In Image Steganography”, International Journal of Modern Engineering Research, Vol.2, Issue.5, pp-3835-3840.

[32] Kruus, P., Caroline, S., Michael, H. and Mathew, M. (2002), “A Survey of Steganographic Techniques for Image Files”, Advanced Security Research Journal, Network Associates Laboratories, pp.41-51.

[33] Kharrazi, M., Sencar, H. T. and Memon, N. (2004), “Image Steganography: Concepts and Practice”, WSPC/Lecture Notes Series: 9in x 6in, pp.1-31.

[34] B, Li., J, He. and J, Huang. (2011), “A Survey on Image Steganography and Steganalysis”, Journal of Information Hiding and Multimedia Signal Processing, Vol. 2, No. 2, pp. 142-172.

[35] Friedman, W.F. (1967), "Cryptology", Encyclopedia Britannica, Vol. 6, pp. 844-851, 1967.

[36] Kahate, A. (2008), “Cryptography and Network Security”, 2nd Edition, Tata McGraw-Hill.

[37] Gonzalez, R. C. and Woods, R. E. (2002), “Digital Image Processing”, 2nd edition, Prentice Hall, Inc.

A SURVEY ON EDUCATIONAL DATA MINING AND RESEARCH TRENDS

Rajni Jindal and Malaya Dutta Borah

Department of Computer Engineering, Delhi Technological University, N. Delhi, India

ABSTRACT

Educational Data Mining (EDM) is an emerging field exploring data in educational context by applying different Data Mining (DM) techniques/tools. It provides intrinsic knowledge of teaching and learning process for effective education planning. In this survey work focuses on components, research trends (1998 to 2012) of EDM highlighting its related Tools, Techniques and educational Outcomes. It also highlights the Challenges EDM.

KEYWORDS

Educational Data Mining (EDM), EDM Components, DM Methods, Education Planning

For More Details : http://airccse.org/journal/ijdms/papers/5313ijdms04.pdf

Volume Link : http://airccse.org/journal/ijdms/current2013.html

REFERENCES

[1] Amershi, S., and Conati, C., (2009) "Combining unsupervised and supervised classification to build user models for exploratory learning environments" Journal of Educational Data Mining.Vol.1, No.1, pp. 18-71.

[2] Mercheron, A., and Yacef, K. (2005), "Educational Data Mining: a case study" in Proc. Conf. on Artificial Intelligence in Education Supporting Learning through Intelligent and Socially Informed Technology. IOS Press, Amsterdam, The Netherlands, pp. 467-474.

[3] Amershi, S., Conati, C. and Maclaren, H., (2006) "Using Feature Selection and Unsupervised Clustering to Identify Affective Expressions in Educational Games", in Proc .Of The Intelligent Tutoring Systems Workshop on Motivational and Affective Issues. pp. 21 -28.

[4] Al-shargabi, A.A. and Nusari, A. N (2010), "Discovering Vital Patterns From UST Students Data by Applying Data Mining Techniques", in Proc. Int. Conf. On Computer and Automation Engineering, China: IEEE, 2010, 2,547-551.DOI:10.1109/ICCAE.2010.5451653.

[5] Badri, M., and El Mourad, T (2011) "Measuring job satisfaction among teachers in Abu Dhabi: design and testing differences" in Proc.Int. Conf. on NIE, 4th Redesigning Pedagogy.Singapore.

[6] Baker, R.S, Corbett, A.T., Koedinger, K.R (2004) , "Detecting Student Misuse of Intelligent Tutoring Systems" in Proc. Lecture Notes in Computer ScienceVol.3220,531-540.

[7] aker, R.S.J.D.,and Yacef, K.(2009), "The state of Educational Data Mining in 2009:A review and future vision" Journal of Educational Data Mining, Vol.1,No. 1,pp.3-17.

[8] Nelson,B., Nugent, R., Rupp, A. A.(2012), "On Instructional Utility, Statistical Methodology, and the Added Value of ECD: Lessons Learned from the Special Issue", Journal of Educational Data Mining.Vol.4, No.1,pp.224-230.

[9] Baradwaj,B.K., and Pal,S.(2011), "Mining Student Data to Analyze Students' Performance" .International Journal of advanced Computer Science and applications.2,6. International Journal of Database Management Systems (IJDMS) Vol.5, No.3, June 2013 69

[10] Aggrawal,C.C, and Yu, P.S.(2009), "A Survey of Uncertain Data Algorithm and Applications" IEEE Transactions on Knowledge and Data Engineering, Vol.21,No. 5,pp.609 - 623.

[11] Lee, C.H., Lee, G., Leu, Y.(2009), "Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning" .Expert Syst. Appl. J.Vol.36, pp.1675-1684.

[12] Chrysostomu K. el al.(2009), "Investigation of users' preference in interactive multimedia learning systems: a data mining approach",Taylor and Francis online journal Interactive learning environments. Vol. 17,No. 2.

[13] Conati,C., Muldner,K., and Carenini G.,(2006)., ".From Example Studying to Problem Solving via Tailored Computer-Based Meta-Cognitive Scaffolding: Hypotheses and Design", Technology, Instruction, Cognition, and Learning - Special Issue on Problem Solving Support in Intelligent Tutoring System.Vol.4, No.1-54.

[14] Cocea,M., and Weibelzahl,S (2009), "Log file analysis for disengagement detection in e- learning

environments", Springer Journal User Modeling and User Adapted Interaction. Vol.19, No.4, pp.341-385. DOI: 10.1007/s 11257-009-9065-5.

[15] Romero, C., and Ventura, S. (2007), " Educational Data Mining : A survey from 1995 to 005" Expert Systems with Applications. Vol. 33, pp.135-146.

[16] Romero,C. et al.,(2008), "Data Mining in course management systems: Moodle case study and tutorial", Computer and Education, Elsevier publication. Vol. 51, No. 1,pp.368-384.

[17] Romero, C., and Ventura, S. (2010), " Educational Data Mining: A review of the state of the Art", IEEE Trans.on on Sys. Man and Cyber.-Part C: Appl. and rev., Vol.40, No.6, pp. 601-618.

[18] Romero, C., and Ventura S.(2013), " Data Mining in Education". WIREs Data Mining and Know.Dis., Vol.3,pp.12-27.

[19] Kothari, C.R,(2004)Research Methodology Methods and Techniques, 2nd ed., New Age International Publishers, New Delhi,pp.99-111.

A XGBOOST RISK MODEL VIA EATURE SELECTION AND BAYESIAN HYPER-PARAMETER OPTIMIZATION

Yan Wang[1], Xuelei Sherry Ni[2]

[1]Graduate College, Kennesaw State University, Kennesaw, USA
[2]Department of Statistics and Analytical Sciences, Kennesaw State University, Kennesaw, USA

ABSTRACT

This paper aims to explore models based on the extreme gradient boosting (XGBoost) approach for business risk classification. Feature selection (FS) algorithms and hyper-parameter optimizations are simultaneously considered during model training. The five most commonly used FS methods including weight by Gini, weight by Chi-square, hierarchical variable clustering, weight by correlation, and weight by information are applied to alleviate the effect of redundant features. Two hyper-parameter optimization approaches, random search (RS) and Bayesian tree-structured Parzen Estimator (TPE), are applied in XGBoost. The effect of different FS and hyper-parameter optimization methods on the model performance are investigated by the Wilcoxon Signed Rank Test. The performance of XGBoost is compared to the traditionally utilized logistic regression (LR) model in terms of classification accuracy, area under the curve (AUC), recall, and F1 score obtained from the 10-fold cross validation. Results show that hierarchical clustering is the optimal FS method for LR while weight by Chi-square achieves the best performance in XG-Boost. Both TPE and RS optimization in XGBoost outperform LR significantly. TPE optimization shows a superiority over RS since it results in a significantly higher accuracy and a marginally higher AUC, recall and F1 score. Furthermore, XGBoost with TPE tuning shows a lower variability than the RS method. Finally, the ranking of feature importance based on XGBoost enhances the model interpretation. Therefore, XGBoost with Bayesian TPE hyper-parameter optimization serves as an operative while powerful approach for business risk modeling.

KEYWORDS

Extreme gradient boosting; XGBoost; feature selection; Bayesian tree-structured Parzen estimator; risk modeling

For More Details : https://aircconline.com/ijdms/V11N1/11119ijdms01.pdf

Volume Link : https://airccse.org/journal/ijdms/current2019.html

REFERENCES

[1] E. I. Altman and A. Saunders, "Credit risk measurement: Developments over the last 20 years, " *Journal of banking & finance*, vol. 21, no. 11-12, pp. 1721-1742, 1997.

[2] R. A. Walkling, "Predicting tender offer success: A logistic analysis," *Journal of financial and Quantitative Analysis*, vol. 20, no. 4, pp. 461-478, 1985.

[3] S. Finlay, "Multiple classifier architectures and their application to credit risk assessment," *European Journal of Operational Research*, vol. 210, no. 2, pp. 368-378, 2011.

[4] Y. Wang and J. L. Priestley, "Binary classification on past due of service accounts using logistic regression and decision tree," 2017.

[5] Y. Wang, X. S. Ni, and B. Stone, "A two-stage hybrid model by using artificial neural networks as feature construction algorithms," *arXiv preprint arXiv:1812.02546*, 2018.

[6] Y. Zhou, M. Han, L. Liu, J.S. He, and Y. Wang, "Deep learning approach for cyberattack detection," *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE*, pp. 262-267, 2012.

[7] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3446-3453, 2012.

[8] G. Paleologo, A. Elisseeff, and G. Antonini, "Subagging for credit scoring models," *European journal of operational research*, vol. 201, no. 2, pp. 490-499, 2010.

[9] G. Wang, J. Ma, L. Huang, and K. Xu, "Two credit scoring models based on dual strategy ensemble trees," *Knowledge-Based Systems*, vol. 26, pp. 61-68, 2012.

[10] T. Chen, T. He, M. Benesty et al., "Xgboost: extreme gradient boosting," *R pack-age version 0.4- 2*, pp. 1{4, 2015.

[11] M. Zieba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Systems with Applications*, vol. 58, pp. 93-101, 2016.

[12] Y. Xia, C. Liu, Y. Li, and N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," *Expert Systems with Applications*, vol. 78, pp. 225-241, 2017.

[13] S. Piramuthu, "Evaluating feature selection methods for learning in data mining applications," *European journal of operational research*, vol. 156, no. 2, pp. 483-494, 2004.

[14] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281-305, 2012.

[15] J. Bergstra, D. Yamins, and D. D. Cox, "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms," in *Proceedings of the 12th Python in Science Conference*. Citeseer, 2013, pp. 13-20.

[16] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *Journal of Retailing and Consumer Services*, vol. 27, pp. 11-23, 2015.

[17] F. Akthar and C. Hahne, "RapidMiner 5 operator reference," *Rapid-I GmbH*, vol. 50, p. 65, 2012.

[18] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, Automatic subspace clustering of high dimensional data for data mining applications. *ACM*, 1998, vol. 27, no. 2.

[19] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124-136, 2015.

[20] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785-794.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.

[22] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kegl, "Algorithms for hyper-parameter optimization," in *Advances in neural information processing systems*, 2011, pp. 2546-2554.

[23] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, "Hyperopt: a python library for model selection and hyperparameter optimization," *Computational Science & Discovery*, vol. 8, no. 1, p. 014008, 2015.

[24] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148-175, 2016.

[25] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyper-parameter optimization in hundreds of dimensions for vision architectures," 2013.

[26] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-weka: Combined selection and hyperparameter optimization of classification algorithms," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 847-855.

[27] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *International Conference on Learning and Intelligent Optimization*. Springer, 2011, pp. 507-523.

[28] L. Breiman, Classification and regression trees. Routledge, 2017.

[29] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802-813, 2008.

[30] G. G. Moisen, E. A. Freeman, J. A. Blackard, T. S. Frescino, N. E. Zimmermann, and T. C. Edwards Jr, "Predicting tree species presence and basal area in utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods," *Ecological modelling*, vol. 199, no. 2, pp. 176-187, 2006.

[31] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote sensing of Environment*, vol. 62, no. 1, pp. 77-89, 1997.

[32] E. A. Freeman and G. G. Moisen, "A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa," *Ecological Modelling*, vol. 217, no. 1-2, pp. 48-58, 2008.

[33] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness

and correlation," 2011.

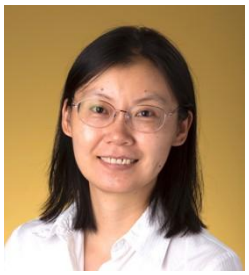
[34] S. Beguer a, "Validation and evaluation of predictive models in hazard assessment and risk management," *Natural Hazards*, vol. 37, no. 3, pp. 315-329, 2006.

[35] Y. Sasaki et al., "The truth of the f-measure," *Teach Tutor mater*, vol. 1, no. 5, pp. 1-5, 2007.

[36] E. W. Weisstein, "Bonferroni correction," 2004.

AUTHORS

Yan Wang is a Ph.D. candidate in Analytics and Data Science at Kennesaw State University. Her research interest contains algorithms and applications of data mining and machine learning techniques in financial areas. She has been a summer Data Scientist intern at Ernst & Young and focuses on the fraud detections using machine learning techniques. Her current research is about exploring new algorithms/models that integrates new machine learning tools into traditional statistical methods, which aims at helping financial institutions make better strategies. Yan received her M.S. in Statistics from university of Georgia.



Dr. Xuelei Sherry Ni is currently a Professor of Statistics and Interim Chair of Department of Statistics and Analytical Sciences at Kennesaw State University, where she has been teaching since 2006. She served as the program director for the Master of Science in Applied Statistics program from 2014 to 2018, when she focused on providing students an applied leaning experience using real-world problems. Her articles have appeared in the Annals of Statistics, the Journal of Statistical Planning and Inference and Statistica Sinica, among others. She is the also the author of several book chapters on modeling and forecasting. Dr. Ni received her M.S. and Ph.D. in Applied Statistics from Georgia Institute of Technology.

EXPERIMENTAL EVALUATION OF NOSQL DATABASES

Veronika Abramova[1] , Jorge Bernardino[1,2] and Pedro Furtado[2] [1]Polytechnic Institute of Coimbra - ISEC / CISUC, Coimbra, Portugal [2]University of Coimbra – DEI / CISUC, Coimbra, Portugal

ABSTRACT

Relational databases are a technology used universally that enables storage, management and retrieval of varied data schemas. However, execution of requests can become a lengthy and inefficient process for some large databases. Moreover, storing large amounts of data requires servers with larger capacities and scalability capabilities. Relational databases have limitations to deal with scalability for large volumes of data. On the other hand, non-relational database technologies, also known as NoSQL, were developed to better meet the needs of key-value storage of large amounts of records. But there is a large amount of NoSQL candidates, and most have not been compared thoroughly yet. The purpose of this paper is to compare different NoSQL databases, to evaluate their performance according to the typical use for storing and retrieving data. We tested 10 NoSQL databases with Yahoo! Cloud Serving Benchmark using a mix of operations to better understand the capability of non-relational databases for handling different requests, and to understand how performance is affected by each database type and their internal mechanisms.

KEYWORDS

NoSQL databases, SQL databases, performance evaluation, database models, YCSB

REFERENCES

[1] Fayech, I. and Ounalli, H.: Towards a Flexible Database Interrogation. International Journal of Database Management Systems (IJDMS) Vol.4, No.3, June 2012 .

[2] http://nosql-database.org/.

[3] Vimala, S., Khanna Nehemiah, H., Bhuvaneswaran, R. S., and Saranya, G.: Design Methodology for Relational Databases: Issues Related to Ternary Relationships in Entity-relationship Model and Higher Normal Forms. International Journal of Database Management Systems (IJDMS) Vol.5, No.3, June 2013.

[4] Stonebraker, M.: SQL databases vs. NoSQL databases. Communications of the ACM, Vol. 53 No. 4, Pages 10-11.

[5] Gajendran, S.: A Survey on NoSQL Databases, 2012, http://ping.sg/story/A-Survey-on-NoSQLDatabases---Department-of-Computer-Science.

[6] Cooper, B., Silberstein, A., Tam, E., Ramakrishnan, R., and Sears, R.: Benchmarking cloud serving systems with YCSB. In Proceedings of the 1st ACM Symposium on Cloud Computing (SoCC '10). ACM, New York, NY, USA, 143-154.

[7] Strozzi, C.: NoSQL – A relational database management system, 2013, http://www.strozzi.it.

[8] Chang, F., Jeffrey, D., Ghemawat, S., Hsieh, W., Wallach, D., Burrows, M., Chandra, T., Fikes, A. and Gruber, R.: Bigtable: A Distributed Storage System for Structured Data. ACM Transactions on Computer Systems, 26(2), Article 4.

[9] Decandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., and Vogels, W.: Dynamo: amazon's highly available key- value store. In Proceedings of twenty-first ACM SIGOPS Symposium on Operating Systems principles (SOSP '07). ACM, New York, NY, USA, 205-220.

[10] Notes from the NoSQL Meetup, 2013, http://developer.yahoo.com/blogs/ydn/posts/2009/06/nosql_meetup.

[11] Hecht, R. and JABLINSKI, S.: NoSQL Evaluation A Use Case Oriented Survey. Proceedings International Conference on Cloud and Service Computing, pp. 12-14.

[12] Han, J.: Survey on NOSQL Databases. Proceedings 6th International Conference on Pervasive Computing and Applications, pp. 363-366.

[13] Leavitt, N.: Will NoSQL Databases Live up to Their Promise?. Computer Magazine, Vol. 43 No. 2, pp. 12-14.

[14] Floratou, A., Teletia, N., Dewitt, D., Patel, J. and Zhang, D.: Can the elephants handle the NoSQL onslaught?. Proc. VLDB Endow. 5,1712-1723.

[15] Tudorica, B.G. and Bucur, C.: A comparison between several NoSQL databases with comments and notes. Roedunet International Conference (RoEduNet), pp.1 -5. [16] Pritchett, D.:

BASE: An Acid Alternative. ACM Queue 6(3), 48-55.

[17] Cook, J. D.: ACID versus BASE for database transactions, 2009, http://www.johndcook.com/blog/2009/07/06/brewer-cap-theorem-base.

[18] Browne, J.: Brewer's CAP Theorem, 2009, http://www.julianbrowne.com/article/viewer/brewers-captheorem.

[19] Indrawan-Santiago, M.: Database Research: Are We at a Crossroad? Reflection on NoSQL. NetworkBased Information Systems (NBiS), 15th International Conference on Network-Based Information Systems, pp.45-51.

[20] Zhang, H. and Tompa, F.W.: Querying XML documents by dynamic shredding. InProceedings of the 2004 ACM symposium on Document engineering (DocEng '04). ACM,New York, NY, USA, 21-30.

[21] Crockford, D.: JavaScript: The Good Parts. Sebastopol, CA: O'Reilly Media.

[22] Lamb, C.: Oracle NoSQL Database in 5 minutes, 2013, https://blogs.oracle.com/charlesLamb/entry/oracle_nosql_database_in_5.

[23] Armstrong, T., Ponnekanti, V., Dhruba, B., and Callaghan, M.: LinkBench: a database benchmark based on the Facebook social graph. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (SIGMOD '13). ACM, New York, NY, USA, 1185-1196.

[24] Dayarathna, M. and Suzumura, T.: XGDBench: A benchmarking platform for graph stores in exascale clouds. Cloud Computing T

Citation Count – 93

APPROXIMATE K-NEAREST NEIGHBOUR BASED SPATIAL CLUSTERING USING K-D TREE

Dr. Mohammed Otair

Department of Computer Information Systems, Amman Arab University, Amman, Jordan

ABSTRACT

Different spatial objects that vary in their characteristics, such as molecular biology and geography, are presented in spatial areas. Methods to organize, manage, and maintain those objects in a structured manner are required. Data mining raised different techniques to overcome these requirements. There are many major tasks of data mining, but the mostly used task is clustering. Data set within the same cluster share common features that give each cluster its characteristics. In this paper, an implementation of Approximate kNN-based spatial clustering algorithm using the K-d tree is proposed. The major contribution achieved by this research is the use of the k-d tree data structure for spatial clustering, and comparing its performance to the brute-force approach. The results of the work performed in this paper revealed better performance using the k-d tree, compared to the traditional brute-force approach.

KEYWORDS

Spatial data, Spatial Clustering, Approximate kNN, K-d tree, brute-force.

For More Details : http://airccse.org/journal/ijdms/papers/5113ijdms08.pdf

Volume Link : https://airccse.org/journal/ijdms/current2013.html

REFERENCES

[1] Anoop Jain , Parag Sarda , & Jayant R. Haritsa, (2003) "Providing Diversity in K-Nearest Neighbour Query", Tech. Report TR-2003-04.

[2] Antonin Guttman, (1984), "R-Trees: A Dynamic Index Structure for Spatial Searching". SIGMOD Conference, 47-57.

[3] Arya & Mount, (1993) "Algorithms for fast vector quantization", Proc. of DCC '93: Data Compression Conference, eds. J. A. Storer and M. Cohn, IEEE Press, 381-390.

[4] Beckmann N., Kriegel H.-P., Schneider R., Seeger B., (1990) "The R*-tree: An Efficient and Robust Access Method for Points and Rectangles" , Proc. ACM SIGMOD Int. Conf. on Management of Data, Atlantic City, NJ, pp. 322-331.

[5] Christian Böhm, (2002) "Powerful Database Primitives to Support High Performance Data Mining", Tutorial, IEEE Int. Conf. on Data Mining.

[6] David M. Mount , (2010) "ANN Programming Manual", University of Maryland.

[7] Dunham M., (2002) "Data Mining: Introductory and Advanced Topics", New Jersey, Prentice Hall.

[8] Erica Kolatch, (2001) "Clustering Algorithms for Spatial Databases: A Survey", citeseerx.ist.psu.edu.

[9] Ester M., Kriegel H.-P., Xu X, (1998) "Incremental Clustering for Mining in a Data Warehousing Environment", Proceedings of the 24th VLDB Conference.

[10] Garcia, V., Debreuve, E., and Barlaud, M., (2008) "Fast k nearest neighbor search using GPU", IEEE Computer Society Conference, 1-6.

[11] Graham Nolan, (2009) "Improving the k-Nearest Neighbour Algorithm with CUDA", Honours Programme, The University of Western Australia.

[12] J. Sander, M. Ester, H. Kriegel, and X. Xu, (1998) "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications". Journal of Data Mining and Knowledge Discovery, Vol. (2), Issue (2), 169-194.

[13] J.L. Bentley, Friedman, J.H., Finkel, R.A., (1977) "An algorithm for finding best matches in logarithmic expected time", ACM Transactions on Mathematical Software 3(3), 209–226.

[14] J.L. Bentley, (1975) "Multidimensional binary search trees used for associative searching", Comm. ACM, 18(9):509 517.

[15] Jaim Ahmed, (2009) "Efficient K-Nearest Neighbor Queries Using Clustering With Caching", Master Thesis, The University of Georgia.

[16] Marius Muja and David G. Lowe, (2009) "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration", in International Conference on Computer Vision Theory and Applications (VISAPP'09).

[17] N. Beckmann, H. P. Kriegel, R. Schneider, and B. Seeger. (1990) "The R*-tree: an efficient and robust access method for points and rectangles". ACM SIGMOD, pages 322-331.

[18] Nitin Bhatia, Vandana, (2010) "Survey of Nearest Neighbor Techniques", International Journal of Computer Science and Information Security, Vol. 8, No. 2.

[19] R. F Sproull, (1991) “Refinements to Nearest Neighbor Searching”, Technical Report, International Computer Science, ACM (18) 9, pp 507-517.

[20] Rina Panigrahy, (2006) "Nearest Neighbor Search using Kd-trees", citeseerx.ist.psu.edu.

[21] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman and A. Wu, (1998) "An optimal algorithm for approximate nearest neighbor searching", Journal of the ACM, 45(6):891-923.

[22] S. Dhanabal, S. Chandramathi, (2011) "A Review of various k-Nearest Neighbor Query Processing Techniques", International Journal of Computer Applications, Volume 31– No.7.

[23] S. N. Omohundro, (1989) “Five Ball Tree Construction Algorithms”, Technical Report.

[24] Shekhar S, Zhang P, Huang Y, Vatsavai R., (2003) "Trends in Spatial Data Mining". Department of Computer Science and Engineering, University of Minnesota, Minneapolis.

[25] Shekhar S, Zhang P., (2004) "Spatial Data Mining: Accomplishments and Research Needs". University of Minnesota. GIS Science.

[26] Shekhar S., Chawla S., (2003) "Spatial Databases: A tour". Pearson education Inc, Upper Saddle River, New Jersey.

[27] Steven S. Skiena, (2010) "The Algorithm Design Manual" , 2nd Edition, Stony Brook, NY 11794-4400.

[28] T. Bailey and A. K. Jain, (1978) "A note on Distance weighted k-nearest neighbor rules", IEEE Trans. Systems, Man Cybernatics, Vol.8, pp 311-313.

[29] T. Liu, A. W. Moore, A. Gray, (2006) "New Algorithms for Efficient High Dimensional Non-Parametric Classification", Journal of Machine Learning Research, pp 1135-1158.

[30] T. M. Cover and P. E. Hart, (1967) "Nearest Neighbor Pattern Classification", IEEE Trans. Inform. Theory, Vol. IT-13, pp 21-27.

[31] Toussaint GT (2005). "Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining". International Journal of Computational Geometry and Applications 15 (2): 101–150.

[32] William R. Mark, Gordon Stoll, (2006) "Fast kd-tree Construction with an Adaptive Error-Bounded Heuristic", Warren Hunt, IEEE Symposium on Interactive Ray Tracing.

[33] Yu-Chen Fu, Zhi-Yong Hu, We1 Guo, Dong-Ru Zhou, (2003) "QR-tree: a hybrid spatial index structure", Proceedings of the Second International Conference on Machine Learning and Cybernetics.

[34] Zhou, K., Hou, Q., Wang, R., and Guo, B. (2008) "Real-time kd-tree construction on graphics hardware". ACM Trans. Graph. 27, 5, 1-11.

AUTHOR

Mohammed A. Otair is an Associate Professor in Computer Information Sy: University-Jordan. He received his B.Sc. in Computer Science from IU-Jordan an 2000, 2004, respectively, from the Department of Computer Information Syste major interests are Mobile Computing, Data Mining and Databases Neural Netw Web-computing, E-Learning. He has more than 29 publications.

A DATA QUALITY METHODOLOGY FOR HETEROGENEOUS DATA

Batini Carlo[1], Barone Daniele[1], Cabitza Federico[1] and Grega Simone[2]

[1] Università degli Studi di Milano Bicocca, Milan, Italy

[2] Nextt Lab S.r.l.Via Benedetto Croce 19, 00142 Roma, Italy

ABSTRACT

We present a Heterogenous Data Quality Methodology (HDQM) for Data Quality (DQ) assessment and improvement that considers all types of data managed in an organization, namely structured data represented in databases, semistructured data usually represented in XML, and unstructured data represented in documents. We also define a meta-model in order to describe the relevant knowledge managed in the methodology. The different types of data are translated in a common conceptual representation. We consider two dimensions widely analyzed in the specialist literature and used in practice: Accuracy and Currency. The methodology provides stakeholders involved in DQ management with a complete set of phases for data quality assessment and improvement. A non trivial case study from the business domain is used to illustrate and validate the methodology.

KEYWORDS

Data quality, Methodology, structured data, semistructured data, unstructured data

For More Details : http://airccse.org/journal/ijdms/papers/3111ijdms05.pdf

Volume Link : https://airccse.org/journal/ijdms/current2011.html

REFERENCES

[1] Abiteboul, S. (1997). "Querying semi-structured data". In *Proceedings of the 6th International Conference on Database Theory*, Delphi, Greece.

[2] Abiteboul, S., Buneman, P. and Suciu, D. (2000). *Data on the Web: from relations to semistructured data and XML*. Morgan Kaufmann Publishers Inc.

[3] Anderson, J. C., Rungtusanatham, M. and Schroeder, R. G. (1994). "A theory of quality management underlying the deming management method". *The Academy of Management Review,* Vol. 19, No. 3, pp472-509.

[4] Avenali, A., Bertolazzi, P., Batini, C. and Missier, P. (2008). "Brokering infrastructure for minimum cost data procurement based on quality -quantity models". *Decision Support Systems,* Vol. 45, No.1, pp95-109.

[5] Ballou, D., Wang, R., Pazer, H. and Tayi, G. K. (1998). "Modeling information manufacturing systems to determine information product quality". *Management Science,* Vol. 44, No. 4, pp462- 484.

[6] Barone, D., Batini, C. and De Amicis, F. (2006). "An analytical framework to analyze dependencies among data quality dimensions". In *Proceedings of the 11th International Conference on Information Quality*.

[7] Batini, C., Cabitza, F., Cappiello, C. and Francalanci, C. (2008). "A comprehensive data quality methodology for web and structured data". *International Journal of Innovative Computing and Applications* Vol. 1, No.3, pp205-218.

[8] Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. (2009). "Methodologies for data quality assessment and improvement". *ACM Computing Survey*, Vol. 41, No. 3, pp1-52.

[9] Batini, C. and Scannapieco, M. (2006). *Data Quality: Concepts, Methods, and Techniques.* Springer Verlag.

[10] Cappiello, C., Francalanci, C., Pernici, B. and Plebani, P. (2003). "Data quality assurance in cooperative information systems: a multi-dimension certificate". In *Proceedings of the International Workshop on Data Quality in Cooperative Information Systems*, Siena, Italy.

[11] Elmagarmid, A. K., Ipeirotis, P. G. and Verykios, V. S. (2007). "Duplicate record detection: A survey". *IEEE Transactions on Knowledge and Data Engineering,* Vol.19, No. 1, pp1-16.

[12] Elmasri, R. and Navathe, S. B. (1994). *Fundamentals of Database Systems, 2nd Ed*. Benjamin- Cummings.

[13] English, L. P. (1999). *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons Inc.

[14] Falorsi, P. D., Pallara, S., Pavone, A., Alessandroni, A., Massella, E. and Scannapieco, M. (2003). "Improving the quality of toponymic data in the italian public administration". In *Prooceedings of the International Workshop on Data Quality in Cooperative Information Systems*.

[15] Hegewald, J., Naumann, F. and Weis, M. (2006). "XStruct: Efficient schema extraction from multiple and large xml documents". In *Prooceedings of 22th International Conference on Data Engineering Workshops*.

[16] Hengst, M. D. and Vreede, G. D. (2004). "Collaborative business engineering: A decade of lessons from the field". *Journal of Management Information Systems*, Vol.20, No.4, pp85-114.

[17] Krawczyk, H. and Wiszniewski, B. (2003). "Visual gqm approach to quality-driven development of electronic documents". In *Proceedings of the 2nd International Workshop on Web Document Analysis*.

[18] Lee, Y. W., Strong, D. M., Kahn, B. K. and Wang, R. Y. (2002). "Aimq: A methodology for information quality assessment". *Information & Management*, Vol.40, No.2, pp133-146.

[19] Long, J. and Seko, C. (2002). "A new method for database data quality evaluation at the canadian institute for health information (CIHI)". In *Proceedings of 7th International Conference on Information Quality*.

[20] Loshin, D. (2008). *Master Data Management*. Morgan Kaufmann.

[21] Maurino, A., Batini, C., Barone, D., Mastrella, M. and Ruffini, C. (2007). "A framework and a methodology for data quality assessment and monitoring". In *Proceedings of the 12th International Conference on Information Quality*. Boston, MA, USA.

[22] McCallum, A. (2005). "Information extraction: distilling structured data from unstructured text". *ACM Queue*, Vol.3, No.9, pp48-57.

[23] Naumann, F., Freytag, J. C. and Leser, U. (2004). "Completeness of integrated information sources". *Information Systems*, Vol.29, No.7, pp583-615.

[24] Pareto, L. and Boquist, U. (2006). "A quality model for design documentation in model-centric projects". In *Proceedings of the 3rd international workshop on Software quality assurance*. New York, NY, USA.

[25] Penna, G. D., Marco, A. D., Intrigila, B., Melatti, I. and Pierantonio, A. (2006). "Interoperability mapping from XML schemas to ER diagrams". *Data Knowledge Engineering*, Vol.59, No.1, pp166-188.

[26] Pipino, L., Lee, Y. W. and Wang, R. Y. (2002). "Data quality assessment". *Communications of ACM*, Vol.45, No.4, pp211-218.

[27] Redman, T.C. (1997). *Data Quality for the Information Age*. Artech House, Inc.

[28] Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M. and Baldoni, R. (2004). "The daquincis architecture: a platform for exchanging and improving data quality in cooperative information systems". *Information Systems*, Vol.29, No.7, pp551-582.

[29] Shankaranarayan, G., Wang, R. Y. and Ziad, M. (2000). "Modeling the manufacture of an information product with IP-MAP". In *Proceedings of the 5th International Conference on Information Quality*. Massachusetts Institute of Technology, USA.

[30] Shankaranarayanan, G. and Cai, Y. (2006). "Supporting data quality management in decision- making". *Decision Support Systems*, Vol.42, No.1, pp302-317.

[31] Shvaiko, P. and Euzenat, J. (2005). "A survey of schema-based matching approaches". *Journal on Data Semantics IV*, pp146-171.

[32] Stoica, M., Chawat, N. and Shin, N. (2004). "An investigation of the methodologies of business process reengineering". *Information Systems Education Journal*, Vol.2, pp1-11.

[33] Wang, R. Y. (1998). "A product perspective on total data quality management". *Communications of ACM*, Vol.41, No.2, pp58-65.

[34] Wang, R. Y., Lee, Y. W., Pipino, L. and Strong, D. M. (1998). "Manage your information as a product". *Sloan Management Review*, Vol.39, No.4, pp95-105.

[35] Wang, R. Y. and Strong, D. M. (1996). "Beyond accuracy: what data quality means to data consumers". *Journal of Management Information Systems*, Vol.12, No.4, pp5-34.

[36] Yeh, D. and Li, Y. (2005). "Extracting entity relationship diagram from a table-based legacy database". In *Proceedings of the 9th European Conference on Software Maintenance and Reengineering*, Manchester, UK.

Citation Count – 74

CLOUD DATABASE DATABASE AS A SERVICE

Waleed Al Shehri

**Department of Computing, Macquarie University Sydney,
NSW 2109, Australia**

ABSTRACT

Cloud computing has been the most adoptable technology in the recent times, and the database has also moved to cloud computing now, so we will look into the details of database as a service and its functioning. This paper includes all the basic information about the database as a service. The working of database as a service and the challenges it is facing are discussed with an appropriate. The structure of database in cloud computing and its working in collaboration with nodes is observed under database as a service. This paper also will highlight the important things to note down before adopting a database as a service provides that is best amongst the other. The advantages and disadvantages of database as a service will let you to decide either to use database as a service or not. Database as a service has already been adopted by many e-commerce companies and those companies are getting benefits from this service.

KEYWORDS

Database, cloud computing, Virtualization, Database as a Service (DBaaS).

For More Details : http://airccse.org/journal/ijdms/1011s1.pdf

Volume Link : http://airccse.org/journal/ijdms/Currentissue.html

REFERENCES

[1] Bloor, R. 2011. WHAT IS A CLOUD DATABASE? Retrieved 25th November 2012 from http://www.algebraixdata.com/wordpress/wp-content/uploads/2010/01/AlgebraixWP2011v06.pdf
[2] Curino, C., Madden, S. and et.al. Relational Cloud: A DatabaseasaService for the Cloud. Retrieved 24th November 2012 from http://www.cidrdb.org/cidr2011/Papers/CIDR11_Paper33.pdf
[3] Finley, K. 2011. 7 Cloud-Based Database Services. Retrieved 23rd November 2012 from http://readwrite.com/2011/01/12/7-cloud-based-database-service
[4] Hacigumus, H., Iyer, B. and Mehrotra, S. 2004. Ensuring the Integrity of Encrypted Databases in the Database-as-a-Service Model. Retrieved 24th November 2012 from http://link.springer.com/chapter/10.1007%2F1-4020-8070-0_5?LI=true
[5] Hacıgumus, H., Iyer, B. and Mehrotra, S. Providing Database as a Service. Retrieved 25th November 2012 from http://archive.systems.ethz.ch/www.systems.ethz.ch/education/past-courses/fs09/HotDMS/pdf/daas.pdf
[6] Harris, D. 2012. Cloud Databases 101: Who builds 'em and what they do. Retrieved 25th November 2012 from http://gigaom.com/cloud/cloud-databases-101-who-builds-em-and-what-they-do/
[7] Hogan, M. 2008. Cloud Computing & Databases:How databases can meet the demands of cloud computing. Retrieved 23rd November 2012 from http://www.scaledb.com/pdfs/CloudComputingDaaS.pdf
[8] Mykletun, E. and Tsudik, G. 2006. Aggregation Queries in the Database-As-a-Service Model. Retrieved 24th November 2012 from http://link.springer.com/chapter/10.1007%2F11805588_7?LI=true
[9] Oracle. 2011. Retrieved 23rd November 2012 from http://www.oracle.com/technetwork/topics/entarch/oes-refarch-dbaas-508111.pdf
[10] Pizzete, L. and Cabot, T.2012. Database as a Service: A Marketplace Assessment. Retrieved 23rd November 2012 from http://www.mitre.org/work/tech_papers/2012/11_4727/cloud_database_service_dbaas.pdf
[11] Postgres Plus. 2012. Cloud Database: Getting started Guide. Retrieved 23rd November 2012 from http://get.enterprisedb.com/docs/Postgres_Plus_Cloud_Database_Getting_Started_Guide.pdf
[12] Rouse, M. 2012. Cloud Database. Retrieved 25th November 2012 from http://searchcloudapplications.techtarget.com/definition/cloud-database-database-as-a-service
[13] Saini, G.P. 2011. Cloud Computing: Database as a Service. Retrieved 24th November 2012 from http://cloudcomputing.sys-con.com/node/1985543
[14] VMware. 2012. Getting Started with Database-as-a-Service. Retrieved 23rd Novermber 2012 from http://www.vmware.com/pdf/vfabric-data-director-20-database-as-a-service-guide.pdf
[15] Zhang, J. 2011. Database in the Cloud Retrieved 25th November 2012 from http://www.ibm.com/developerworks/data/library/dmmag/DMMag_2011_Issue2/cloudDBaaS/ Image Source
[16] Bloor, R. (Author). 2011. WHAT IS A CLOUD DATABASE ? Retrieved 25th November 2012 from http://www.algebraixdata.com/wordpress/wp-content/uploads/2010/01/AlgebraixWP2011v06.pdf
[17] Pizzete, L. and Cabot, T. (Authors). 2012. Database as a Service: A Marketplace Assessment. Retrieved 23rd November 2012 from http://www.mitre.org/work/tech_papers/2012/11_4727/cloud_database_service_dbaas.pdf

AUTHOR

Waleed Al Shehri received his bachelor degree in computer science from King Abdulaziz University, Jeddah, Saudi Arabia (2005), MSc degree in information technology form Macquarie university, Sydney, Australia (2011). His current research interests in databases and cloud computing specially Database as a Service (DBaaS) . Currently working in the Department of IT in Royal Saudi Air Force (RSAF).

DATA MINING TECHNIQUES: A SOURCE FOR CONSUMER BEHAVIOR ANALYSIS

Abhijit Raorane[1] & R.V.Kulkarni[2]

[1]Department of computer science, Vivekanand College, Tarabai park Kolhapur
[2]Head of the Department, Chh. Shahu Institute of business Education and Research Centre Kolhapur.

ABSTRACT

Various studies on consumer purchasing behaviors have been presented and used in real problems. Data mining techniques are expected to be a more effective tool for analyzing consumer behaviors. However, the data mining method has disadvantages as well as advantages. Therefore, it is important to select appropriate techniques to mine databases. The objective of this paper is to know consumer behavior, his psychological condition at the time of purchase and how suitable data mining method apply to improve conventional method. Moreover, in an experiment, association rule is employed to mine rules for trusted customers using sales data in a super market industry

KEYWORDS

Consumer behavior, Data mining, Association Rule, Super market

For More Details : https://airccse.org/journal/ijdms/papers/3311ijdms04.pdf

Volume Link : https://airccse.org/journal/ijdms/current2011.html

REFERENCES

1. Consumer behavior The psychology of marketing Lars parner

2. Consumer behavior – Himanshu S. M. Dec 24 2008

3. Market basket Analysis in Multiple store environment- Yen- liang Chen, Kwei Tung, Ren-Jie Shen, Ya- Han Hu. Decision support system (2005) Elsevier

4. A Data Mining Approach to consumer behavior- Junzo Watada, Kozo Yamashiro. Procedings of the first International Conference on Innovative computing Information (2006)

5. Mining of users access behavior for frequent sequential pattern from web logs. S. Vijaylakshmi, V. Mohan, S. Suresh Raja. International Journal of Database Management System (IJDM) Vol 2, August 2010.

6. Mining utility-oriented association rules: An efficient approach based on profit and quantity Parvinder S. Sandhu, Dalvinder S. Dhaliwal and S. N. Panda International Journal of the Physical Sciences Vol. 6(2), pp. 301-307, 18 January, 2011

7. B. Yıldız and B. Ergenç (Turkey) in "Comparison of Two Association Rule Mining Algorithms without Candidate Generation" International Journal of Computing and ICT Research 2010. International Journal of Computing and ICT Research, ISSN

8. "Extraction Of Interesting Association Rules Using Genetic Algorithms" Peter P. Wakabi-Waiswa* Venansius Baryamureeba, International Journal of Computing and ICT Research, Vol. 2 No. 1, June 2008

9. User centric approach to itemset utility mining in Market Basket Analysis Jyothi Pillai / International Journal on Computer Science and Engineering (IJCSE 1 Jan 2011

10. Efficient Association Rule Mining for Market Basket Analysis Shrivastava A., Sahu R Global Journal of e-Business & Knowledge Management Year:2007,Volume:3,Issue:1

11. A data mining approach to consumer behavior – Janzo Watada, Kozo Yamashiro 2006

12. Enhancing consumer behavior analysis by data mining techniques – Nan-chan Hsich, Kuo-Chang cha 2009

13. Classification- Mike Chapple

14. Data mining techniques- CDIG- Business intelligence Data mining techniques

15. Association Rules mining- S. Kotsiantis

16. Mining association with the collective strength approach- Aggarwal C.C., Yu, P.S. 2001

17. Fast Algorithms for Mining Association Rules- Agrawal R. and Srikat l. Sept 1994

18. R.Agrawal, T. Imilienski and A. Swami. Mining Associations Rules between Sets of Items in large databases. Proc. Of the ALM SIGNOD. Int'l conf. on management of Data, pages 207-216,May1993

19. Fast Algorithms for Mining Association Rules R. Agarwal and R. Srikant algorithm for mining association rules. In proceedings of the 20th VLDB conference Santiago ,Chile, 1994

20. Sergey Brin , Rajeev Motwani and Craig silvertrin Beyound Market Baskets : Generalizing association rules for correlation.SIGMOD Record (ACM Special Interest Group on Management of Data).26 (2):265,1997

21. Sergey Brin ,Rajeev Motwani, Jetlrey D.Ullman, and Shulem Tsur. Dynamic interest counting and implication rules for market basket data..SIGMOD Record (ACM Special Interest Group on Management of data). 26(2):255,1997

22. Giadici Paulo. Applied Data mining :Statistical Methods for business and industry-ISBN 9812-53- 178-5

23. Data mining Concepts and Techniques Jiauei Han , Michele Kamber,Simon Fraser University ISBN 1-55860-489-8-2001