April 2026:Top10 Cited Articles in Database Management Systems Research Articles

International Journal of Database Management 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

BendiVenkataRamana[1],Prof.M.SurendraPrasadBabu[2],Prof.N.B.Venkateswarlu[3]

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

[2]Dept.ofCS&SE,AndhraUniversity,Visakhapatnam-530003,A.P,India

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

ABSTRACT:

PatientswithLiverdiseasehavebeencontinuouslyincreasingbecauseofexcessiveconsumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. Automatic classificationtoolsmayreduceburdenondoctors.Thispaperevaluatestheselected 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:

ClassificationAlgorithms,DataMining,Liverdiagnosis

ForMoreDetails:http://airccse.org/journal/ijdms/papers/3211ijdms07.pdf

VolumeLink:http://airccse.org/journal/ijdms/current2011.html

CitationCount–305

REFERENCES

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

[2] BUPALiverDisordersDataset.UCIrepositoryofmachinelearningdatabases.Available fromftp://ftp.ics.uci.edu/pub/machinelearningdatabases/liverdisorders/bupa.data,lastaccessed:07 October 2010.

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

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

[5] Schiff'sDiseasesoftheLiver,10thEditionCopyright©2007LippincottWilliams& Wilkins by Schiff,EugeneR.; Sorrell, Michael F.; Maddrey,WillisC.

[6] P.Domingos,M.Pazzani,Ontheoptimalityofthesimple Bayesianclassifierunderzero-oneloss,MachineLearning29(2–3)(1997) 103–130.

[7] Weka-3-4-10jre:dataminingwithopensourcemachinelearningsoftware©2002-2005 David Scuse and University of Waikato

[8] 16thEditionHARRISON'SPRINCIPLESofInternalMedicine

[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 predictionof drugandchemicalmetabolismbyhumanUDP-GlucuronosyltransferaseIsoforms

[12] PaulR.Harper,Areviewandcomparisonofclassificationalgorithmsfordecisionmaking

[13] MitchellTM.Machinelearning.Boston,MA:McGraw-Hill,1997.

[14] Lung-ChengHuang, Sen- Yen Hsuand EugeneLin, A comparison of classification methods for predictingChronicFatigueSyndromebasedon genetic data(2009).

CitationCount–168

CUSTOMERDATACLUSTERINGUSINGDATAMININGTECHNIQUE

Dr.SankarRajagopal

EnterpriseDW/BIConsultantTataConsultancyServices,Newark,DE,USA

ABSTRACT:

Classificationandpatternsextractionfromcustomerdataisveryimportantforbusinesssupport anddecisionmaking.Timelyidentificationofnewlyemergingtrendsisveryimportantin businessprocess.Largecompaniesarehavinghugevolumeofdatabutstarvingforknowledge. To overcome the organization current issue, the new breed of technique is required that has intelligenceandcapabilitytosolvetheknowledgescarcityandthetechniqueiscalledData mining. The objectivesof this paper are to identify the high-profit, high-value and low-riskcustomersby oneof thedataminingtechnique- customerclustering.In thefirst phase,cleansing the dataanddevelopedthepatternsviademographicclusteringalgorithmusingIBMI-Miner.In the second phase, profiling the data, develop the clusters and identify the high-value low-risk customers.Thisclustertypicallyrepresentsthe10-20percentofcustomerswhichyields80%of the revenue.

KEYWORDS:

Datamining,customerclusteringand I-Miner

ForMoreDetails:http://airccse.org/journal/ijdms/papers/3411ijdms01.pdf

VolumeLink:http://airccse.org/journal/ijdms/current2011.html

REFERENCES

[1] Lefait,G.andKechadi,T,(2010)"CustomerSegmentationArchitectureBasedonClustering Techniques"DigitalSociety,ICDS'10,FourthInternationalConference,10-02-2010.

[2] Fraley, Andrew, and Thearting, Kurt (1999). Increasing customer value by integrating data miningand campaignmanagementsoftware.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, August117-122.

[4] I.KrishnaMurthy,"DataMining-StatisticsApplications:AKeytoManagerialDecision Making",Article/Reportindiastat.com,April-May2010.

[5] AssociationAnalysisofCustomerServicesfromtheEnterpriseCustomerManagement SystemICDM-2006.

[6] TerryHarris,(2008)"Optimizationcreatesleangreensupplychains",DataMiningBook

[7] MattHartely(2005)"UsingDataMiningtopredictinventorylevels"DataMiningBook

[8] Hu,Tung-Lai,&Sheub,Jiuh-Biing(2003)."Afuzzy-basedcustomerclassificationmethod fordemand-responsivelogisticaldistributionoperations",FuzzySetsandSystems,139,431– 450.

[9] Hwang, Hyunseok, Jung, Taesoo, & Suh, Euiho (2004). "An LTV model and customer segmentationbasedoncustomervalue:Acasestudyonthewirelesstelecommunication industry", Expert Systemswith Applications,26,181–188.

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

[11] Jonker, Jedid-Jah, Piersma, Nanda, & Poel, Dirk Van den (2004). "Joint optimization of customersegmentationandmarketingpolicytomaximizelong-termprofitability".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,YongSeog,&Street,W.Nick(2004)."Anintelligentsystemforcustomertargeting: 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,andP.J.Flynn,(1999)"Dataclustering:areview",ACM Computing Surveys (CSUR), Vol.31, Issue 3,,1999.

[18] Grabmeier,J.andRudolph,A.(2002)"Techniquesofclusteralgorithmsindatamining" DataMiningandKnowledgeDiscovery,6,303-360.

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

[20] JiyuanAn,JeffreyXuYu,ChotiratAnnRatanamahatana,Yi-PingPhoebeChen,(2007) "Adimensionalityreductionalgorithmanditsapplicationforinteractivevisualization",Journal of Visual Languagesand Computing,v.18 n.1, , February p.48-70.

[21] Nargess Memarsadeghi , Dianne P. O'Leary,(2003) "Classified Information: The Data ClusteringProblem",ComputinginScienceandEngineering,v.5n.5,Septemberp.54-60,

[22] Yifan Li , Jiawei Han , Jiong Yang, (2004) "Clustering moving objects", Proceedings of the tenthACMSIGKDDinternationalconferenceon Knowledgediscovery and data mining,August 22-25, Seattle, WA, USA

[23] Sherin M. Youssef , Mohamed Rizk , Mohamed El-Sherif, (2008) "Enhanced swarm-likeagents for dynamically adaptive data clustering", Proceedings of the 2nd WSEAS International Conferenceon ComputerEngineeringand Applications,p.213-219,January25-27,Acapulco, Mexico

[24] MarcelBrun, ChaoSima,JianpingHua, JamesLowey,BrentCarroll,EdwardSuh, Edward R. Dougherty,(2007) Model-based evaluation of clustering validation measures, Pattern Recognition, v.40 n.3, Marchp.807-824.

AUTHOR:

SankarRajagopalreceivedM.Sc.,degreeinElectronics,M.EdegreeinMaterialsScienceand Ph.Din MetallurgicalEngineeringfromBharathidasanUniversityin 1990,1992andIndian Instituteof Technology-Madras,Chennai1999respectively.After completionof hisdoctoral degree,hejoinedinTATAConsultancyServicesasaSoftwareConsultant.Thenhehaselevated to position Enterprise DW/BI Architect. His areas of research interests include Metal Matrix Composites, Mechanical alloying, Material Informatics, Mechanical Behaviors of Materials, Nanotechnology andData Miningand KnowledgeDiscovery.Hehas publishedabout15 contributedpeerreviewedpapersatNational/InternationalJournalsandConferences.He received best oral presentation awards inconferences.

CitationCount -146

HIGHCAPACITYDATAHIDINGUSINGLSB STEGANOGRAPHYANDENCRYPTION

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

DepartmentofComputerScienceAssamUniversity,Silchar,Assam,Indi

a

ABSTRACT

Thenetworkprovidesamethodofcommunicationtodistributeinformationtothemasses.With thegrowthofdatacommunicationovercomputernetwork,thesecurityofinformationhas 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 obscuresthecontentofthemessage.Weproposeahighcapacitydataembeddingapproachby the combination of Steganography and cryptography. In the process a message is first encryptedusing transposition cipher method and then the encrypted message is embeddedinside an image usingLSBinsertionmethod.Thecombinationof thesetwomethodswillenhancethesecurityof thedataembedded.Thiscombinationalmethodologywillsatisfytherequirementssuchas capacity,securityandrobustnessforsecuredatatransmissionoveranopenchannel.A comparativeanalysisismadetodemonstratetheeffectivenessof theproposedmethodby computingMeansquareerror(MSE)andPeakSignaltoNoiseRatio(PSNR).Weanalyzedthe data hiding technique using the image performance parameters like Entropy, Mean and Standard Deviation.Thestegoimagesaretestedbytransmittingthemandtheembeddeddataare successfully extracted by the receiver. The main objective in this paper is to provide resistanceagainst visual and statistical attacks as well ashighcapacity.

KEYWORDS

Steganography, Cryptography, plain text, encryption, decryption, transposition cipher, Least SignificantBit, HumanVisualSystem,MeansquareerrorandPeakSignaltoNoiseRatio.

ForMoreDetails:http://airccse.org/journal/ijdms/papers/4612ijdms05.pdf

VolumeLink:http://airccse.org/journal/ijdms/current2012.html

REFERENCES

[1] Anderson , R. J. and Petitcolas,F. A.P. (1998)“On The Limitsof Steganography”,IEEE Journalof SelectedAreasin Communications,Vol.16No.4,pp.474-481,ISSN0733-8716.

[2] Petitcolas, F.A.P., Anderson, R. J. and Kuhn, M.G. (1999) “Information Hiding -A Survey”, Proceedingsof theIEEE,Specialissueon Protectionof MultimediaContent,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.andSundaram,V.“CryptographyandSteganography–ASurvey”,Int.J. Comp.Tech. Appl., Vol 2 (3),pp. 626-630,ISSN:2229-6093.

[5] Gutte,R. S.and Chincholkar,Y.D. (2012)“Comparisonof Steganographyat OneLSBand TwoLSBPositions”,InternationalJournalof ComputerApplications,Vol.49,no.11,pp.1 -7.

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

[7] Younes,M.A.B.andJantan,A.(2008),"ImageEncryptionUsingBlock-Based TransformationAlgorithm,"InternationalJournalofComputerScience,Vol.35,Issue.1,pp.15- 23.

[8] Walia,E.,Jain,P.andNavdeep.(2010),“AnAnalysisofLSB&DCTbased Steganography”,GlobalJournalofComputerScienceandTechnology,Vol.10Issue1,pp4-8.

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

[10] Sokouti,M.,Sokouti,B.andPashazadeh,S.(2009),“Anapproachinimproving 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 steganographyandsteganalysistechniques”,JournalofElectronicImaging,SPIEProceedings Vol. 5681.15(4), 041104pp.1-16.

[12] R., Chandramouli, and Nasir Memon.(2001), "Analysis of LSB based image steganography techniques."InImageProcessing,2001.Proceedings.2001InternationalConferenceon,IEEE, vol. 3, pp.1019-1022.

[13] Giddy,J.P.andSafavi-Naini,R.(1994),“AutomatedCryptanalysisofTransposition Ciphers”,TheComputerJournal,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.andMenon,N.(2001),“AnalysisofLSBbasedimagesteganography techniques”,IEEEProceedingsonImageProcessing,Vol.3,pp.1019-1022.

[16] Carvajal-Gamez,B.E.,Gallegos-Funes,F.J.andLopez-Bonilla,J.L.(2009),“ Scaling FactorforRGBImagestoSteganographyApplications”,Journalof VectorialRelativity,Vol.4, no. 3, pp.55-65.

[17] Ulutas,G.,Ulutas,M.andNabiyev,V.(2011),“Distortionfreegeometrybasedsecretimage sharing”,ElsevierInc,ProcediaComputerScience3,pp.721–726.

[18] Tiwari, N. and Shandilya, M. (2010), “Evaluation of Various LSB based Methods of Image SteganographyonGIFFileFormat”,InternationalJournalofComputerApplications(0975– 8887) Vol. 6, no.2,pp.1-4.

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

[20] Deshpande, N., Kamalapur, S. and Daisy, J. (2006), “Implementationof 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" MultimediaToolsandApplications, SpringerVol.30, no. 1, pp.55-88.

[22] Celik,M.U.,Sharma,G.,Tekalp,A.M.andSaber,E. (2005),“LosslessGeneralized-LSB DataEmbedding”,IEEETransactionon ImageProcessing,Vol.14, No. 2,pp. 253-266.

[23] Huang,Y.S.,Huang,Y.P.,Huang,K.N.andYoung,M.S.(2005),“TheAssessment SystemofHumanVisualSpectralSensitivityCurvebyFrequencyModulatedLight”, Proceedingsofthe2005IEEEEngineeringinMedicineandBiology27thAnnualConference, 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.andOgunbona,P.2005.“High-capacitysteganographyusing asharedcolourpalette”,IEEEProceedingsonVision,ImageandSignalProcessing,Vol.152, No.6, pp.787-792.

[26] Curran,K.andBailey,K.(2003),“AnEvaluationofImageBasedSteganography Methods”,InternationalJournalofDigitalEvidenceFall2003,Volume2,Issue2,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).“AdetailedlookatSteganographicTechniquesandtheiruseinanOpen-SystemsEnvironment”,SANSInstitute2002,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-BitEmbeddingSchemeforSteganographicEncoding”,PSIVT2009,LNCS5414, Springer, pp. 349–360.

[30] Smith, C. (2001), “Basic Cryptanalysis Techniques”, SANS Institute 2001, GSEC Version1.2f, http://www.sans.org.

[31] Kaur,R.,Singh,B.andSingh,I.(2012),“AComparativeStudyofCombinationof 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 AssociatesLaboratories, pp.41-51.

[33] Kharrazi,M.,Sencar,H.T.andMemon,N.(2004),“ImageSteganography:Conceptsand Practice”,WSPC/LectureNotesSeries:9inx6in,pp.1-31.

[34] B,Li.,J,He.andJ,Huang.(2011),“ASurveyonImageSteganographyandSteganalysis”, JournalofInformationHidingandMultimediaSignalProcessing,Vol.2,No.2,pp.142-172.

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

[36] Kahate,A.(2008),“CryptographyandNetworkSecurity”,2ndEdition,TataMcGraw-Hill.

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

CitationCount–136

ASURVEY ONEDUCATIONALDATAMINING ANDRESEARCHTRENDS

RajniJindalandMalayaDuttaBorah

DepartmentofComputerEngineering,DelhiTechnologicalUniversity,N.Delhi, India

ABSTRACT

Educational Data Mining (EDM) is an emerging field exploring data in educational context by applyingdifferentDataMining(DM)techniques/tools.Itprovidesintrinsicknowledgeof teaching and learning process for effective education planning. In this survey work focuses on components,researchtrends(1998to2012)ofEDMhighlightingitsrelatedTools,Techniques and educational Outcomes.It also highlights the ChallengesEDM.

KEYWORDS

EducationalDataMining(EDM),EDMComponents,DMMethods,EducationPlanning

ForMoreDetails:http://airccse.org/journal/ijdms/papers/5313ijdms04.pdf

VolumeLink:http://airccse.org/journal/ijdms/current2013.html

REFERENCES

[1] Amershi,S.,andConati,C.,(2009)"Combiningunsupervisedandsupervisedclassification to build user models for exploratory learning environments" Journal of Educational Data Mining.Vol.1, No.1, pp. 18-71.

[2] Mercheron,A., andYacef,K.(2005),"EducationalDataMining:a casestudy"in Proc.Conf. on Artificial Intelligence in Education Supporting Learning through Intelligent and SociallyInformedTechnology.IOSPress,Amsterdam,TheNetherlands,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 TutoringSystemsWorkshopon MotivationalandAffectiveIssues.pp. 21-28.

[4] Al-shargabi,A.A. and Nusari,A. N (2010), "DiscoveringVital 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.,andElMourad,T(2011)"MeasuringjobsatisfactionamongteachersinAbu 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 TutoringSystems"in Proc.LectureNotesin ComputerScienceVol.3220,531-540.

[7] aker,R.S.J.D.,andYacef,K.(2009),"ThestateofEducationalDataMiningin2009:A reviewandfuturevision"Journalof Educational DataMining,Vol.1,No.1,pp.3-17.

[8] Nelson,B., Nugent, R., Rupp, A. A.(2012), "On Instructional Utility, Statistical Methodology,and the AddedValueof ECD: LessonsLearnedfromthe SpecialIssue", Journalof Educational Data Mining.Vol.4, No.1,pp.224-230.

[9] Baradwaj,B.K.,andPal,S.(2011),"MiningStudentDatatoAnalyzeStudents'Performance" .InternationalJournalofadvancedComputerScienceandapplications.2,6.InternationalJournalof Database Management Systems (IJDMS) Vol.5, No.3, June 2013 69

[10] Aggrawal,C.C,and Yu, P.S.(2009), "A Surveyof UncertainDataAlgorithmand Applications"IEEETransactionsonKnowledgeandDataEngineering,Vol.21,No.5,pp.609- 623.

[11] Lee,C.H.,Lee,G.,Leu,Y.(2009),"Applicationofautomaticallyconstructedconceptmap of learningtoconceptualdiagnosisof e-learning".ExpertSyst.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.,andCareniniG.,(2006).,".FromExampleStudyingtoProblem 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/s11257-009-9065-5.

[15] Romero,C.,andVentura,S.(2007),"EducationalDataMining:Asurveyfrom1995to 005"ExpertSystemswithApplications.Vol.33,pp.135-146.

[16] Romero,C.etal.,(2008),"DataMiningincoursemanagementsystems:Moodlecasestudy andtutorial",Computerand Education,Elsevier publication.Vol.51,No. 1,pp.368-384.

[17] Romero,C.,andVentura,S.(2010),"EducationalDataMining:Areviewofthestateofthe Art",IEEETrans.ononSys.ManandCyber.-PartC:Appl.andrev.,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.

CitationCount–119

AXGBOOSTRISKMODELVIAEATURESELECTION AND BAYESIANHYPER-PARAMETEROPTIMIZATION

YanWang[1],XueleiSherryNi[2]

[1]GraduateCollege,KennesawStateUniversity,Kennesaw,USA
[2]DepartmentofStatisticsandAnalyticalSciences,KennesawStateUniversity, 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 byinformationareappliedtoalleviatetheeffectofredundantfeatures.Twohyper-parameteroptimization 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 optimizationshowsasuperiorityoverRSsinceitresultsinasignificantlyhigheraccuracyandamarginally 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

ForMoreDetails:https://aircconline.com/ijdms/V11N1/11119ijdms01.pdf

VolumeLink:https://airccse.org/journal/ijdms/current2019.html

REFERENCES

[1] E.I.AltmanandA.Saunders,"Creditriskmeasurement:Developmentsoverthelast20 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,"Multipleclassifierarchitecturesandtheirapplicationtocreditriskassessment," *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," *Europeanjournal of operational research*, vol. 201, no. 2, pp. 490-499, 2010.

[9] G.Wang,J.Ma,L.Huang,andK.Xu,"Twocreditscoringmodelsbasedondualstrategy ensemble trees," *Knowledge-Based Systems*, vol. 26, pp. 61-68, 2012.

[10] T.Chen,T.He,M.Benestyetal.,"Xgboost:extremegradientboosting,"*Rpack-ageversion0.4- 2*, pp. 1{4, 2015.

[11] M. Zieba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generationinapplicationtobankruptcyprediction,"*ExpertSystemswithApplications*,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,"Evaluatingfeatureselectionmethodsforlearningindataminingapplications," *Europeanjournalofoperationalresearch*,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,andM.Khanbabaei,"Ahybriddataminingmodeloffeatureselection algorithms and ensemble learning classifiers for credit scoring," *Journal of Retailing and Consumer Services*, vol. 27, pp. 11-23, 2015.

[17] F.AktharandC.Hahne,"RapidMiner5operatorreference,"*Rapid-IGmbH*,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.ChenandC.Guestrin,"Xgboost:A scalabletreeboostingsystem,"in *Proceedingsofthe22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785-794.

[21] L.Breiman,"Randomforests,"*Machinelearning*,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,andN.DeFreitas,"Takingthehumanoutofthe loop:AreviewofBayesianoptimization,"*ProceedingsoftheIEEE*,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,Classificationandregressiontrees.Routledge,2017.

[29] J.Elith,J.R.Leathwick,andT.Hastie,"Aworkingguidetoboostedregressiontrees,"*Journalof 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,"Selectingandinterpretingmeasuresofthematicclassificationaccuracy,"*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

binaryclassificationintermsofpredictedprevalenceandkappa,"*EcologicalModelling*,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.

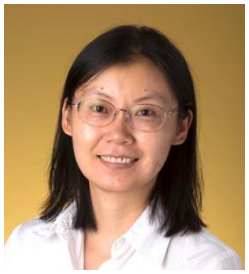
[34] S. Beguer a, "Validation and evaluation of predictive models in hazard assessment and risk management," *Natural Hazards*, vol. 37, no. 3, pp. 315-329, 2006.

[35] Y.Sasakietal.,"Thetruthofthef-measure,"*TeachTutormater*,vol.1,no.5,pp.1-5,2007.

[36] E.W.Weisstein,"Bonferronicorrection,"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 miningandmachinelearningtechniquesinfinancialareas.Shehasbeenasummer 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, whereshehasbeenteachingsince2006.Sheservedastheprogramdirectorforthe 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.

CitationCount–101

EXPERIMENTALEVALUATIONOFNOSQLDATABASES

Veronika Abramova[1] , Jorge Bernardino[1,2] and Pedro Furtado[21]PolytechnicInstituteofCoimbra-ISEC/CISUC,Coimbra,Portugal [2]University of Coimbra – DEI / CISUC, Coimbra, Portugal

ABSTRACT

Relational databases are a technology used universally that enables storage, management andretrieval of varied data schemas. However, execution of requests can become a lengthy andinefficient processfor some large databases. Moreover,storinglarge amountsof data requires servers with larger capacities and scalability capabilities. Relational databaseshave limitations to deal with scalability for large volumes of data. On the other hand, non-relational database technologies,alsoknownasNoSQL,weredevelopedtobettermeettheneedsofkey-value storageof large amountsof records.But thereis a large amountof NoSQLcandidates,andmost havenotbeencomparedthoroughlyyet.Thepurposeofthispaperistocomparedifferent NoSQL databases, to evaluate their performance according to the typical use for storing and retrievingdata.Wetested10NoSQLdatabaseswithYahoo!CloudServingBenchmarkusinga mix of operations to better understand the capability of non-relational databases for handlingdifferentrequests,andtounderstandhowperformanceisaffectedbyeachdatabasetypeand their internal mechanisms.

KEYWORDS

NoSQLdatabases,SQLdatabases,performanceevaluation,databasemodels, YCSB

ForMoreDetails:http://airccse.org/journal/ijdms/papers/6314ijdms01.pdf

VolumeLink:http://airccse.org/journal/ijdms/current2014.html

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

CitationCount–93

APPROXIMATEK-NEARESTNEIGHBOURBASED SPATIALCLUSTERINGUSINGK-DTREE

Dr.MohammedOtair

DepartmentofComputerInformationSystems,AmmanArabUniversity,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

Spatialdata,SpatialClustering,ApproximatekNN,K-dtree,brute-force.

ForMoreDetails:http://airccse.org/journal/ijdms/papers/5113ijdms08.pdf

VolumeLink:https://airccse.org/journal/ijdms/current2013.html

REFERENCES

[1] Anoop Jain , ParagSarda , & Jayant R. Haritsa, (2003) "Providing Diversityin 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 AccessMethodforPointsandRectangles",Proc.ACMSIGMODInt.Conf.onManagementofData, 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] DavidM.Mount,(2010)"ANNProgrammingManual",UniversityofMaryland.

[7] DunhamM.,(2002)"DataMining:IntroductoryandAdvancedTopics",NewJersey,PrenticeHall.

[8] EricaKolatch,(2001)"ClusteringAlgorithmsforSpatialDatabases:ASurvey",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., andBarlaud, M.,(2008)"Fast knearestneighbor searchusingGPU",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: TheAlgorithmGDBSCANandits Applications". Journalof DataMiningandKnowledgeDiscovery, 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] JaimAhmed,(2009)"EfficientK-NearestNeighborQueriesUsingClusteringWithCaching",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 Theoryand 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] RinaPanigrahy,(2006)"NearestNeighborSearchusingKd-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)“FiveBallTreeConstructionAlgorithms”,TechnicalReport.

[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.BaileyandA.K.Jain,(1978)“AnoteonDistanceweightedk-nearestneighborrules”,IEEETrans. 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 andCybernetics.

[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.

CitationCount–89

A DATA QUALITY METHODOLOGY FOR HETEROGENEOUS DATA

BatiniCarlo[1],BaroneDaniele[1],CabitzaFederico[1]andGregaSimone[2]

[1]UniversitàdegliStudidiMilanoBicocca,Milan,Italy

[2]NexttLabS.r.l.ViaBenedettoCroce19,00142Roma,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 withacompleteset ofphases for dataqualityassessmentandimprovement.A nontrivialcasestudyfrom the business domain is used to illustrate and validate the methodology.

KEYWORDS

Dataquality,Methodology,structureddata,semistructureddata,unstructureddata

ForMoreDetails:http://airccse.org/journal/ijdms/papers/3111ijdms05.pdf

VolumeLink:https://airccse.org/journal/ijdms/current2011.html

REFERENCES

[1] Abiteboul,S.(1997).“Queryingsemi-structureddata”.In*Proceedingsofthe6thInternational Conference on Database Theory*, Delphi, Greece.

[2] Abiteboul,S.,Buneman,P.andSuciu,D.(2000).*DataontheWeb:fromrelationsto semistructured data and XML*. Morgan Kaufmann Publishers Inc.

[3] Anderson, J. C., Rungtusanatham, M. and Schroeder, R. G. (1994). “A theory of quality managementunderlyingthedemingmanagementmethod”.*TheAcademyofManagement Review,* Vol. 19, No. 3, pp472-509.

[4] Avenali, A., Bertolazzi, P., Batini, C. and Missier, P. (2008). “Brokering infrastructure for minimumcostdataprocurementbasedonquality-quantitymodels”.*DecisionSupportSystems,* Vol. 45, No.1, pp95-109.

[5] Ballou, D., Wang, R., Pazer, H. and Tayi, G. K. (1998). “Modeling information manufacturing systemstodetermineinformationproductquality”.*ManagementScience,*Vol.44,No.4,pp462- 484.

[6] Barone, D., Batini, C. and De Amicis, F. (2006). “An analytical framework to analyze dependenciesamongdataqualitydimensions”.In*Proceedingsofthe11thInternational Conference on Information Quality*.

[7] Batini,C.,Cabitza,F.,Cappiello,C.andFrancalanci,C.(2008).“Acomprehensivedataquality methodologyfor web andstructured data”. *InternationalJournalofInnovative Computingand Applications* Vol. 1, No.3, pp205-218.

[8] Batini,C.,Cappiello,C.,Francalanci,C.andMaurino,A.(2009).“Methodologiesfordata quality assessment and improvement”. *ACM Computing Survey*, Vol. 41, No. 3, pp1-52.

[9] Batini,C.andScannapieco,M.(2006).*DataQuality:Concepts,Methods,andTechniques.* SpringerVerlag.

[10] Cappiello,C.,Francalanci,C.,Pernici,B.andPlebani,P.(2003).“Dataqualityassurance in cooperative information systems: a multi-dimension certificate”. In *Proceedings of the International Workshop onData Quality inCooperative Information Systems*, Siena, Italy.

[11] Elmagarmid,A.K.,Ipeirotis,P.G.andVerykios,V.S.(2007).“Duplicaterecorddetection: A survey”. *IEEE Transactions on Knowledge and Data Engineering,* Vol.19, No. 1, pp1-16.

[12] Elmasri,R.andNavathe,S.B.(1994).*FundamentalsofDatabaseSystems,2ndEd*.Benjamin- Cummings.

[13] English,L.P.(1999).*Improvingdatawarehouseandbusinessinformationquality:methods for reducing costs and increasing profits*. John Wiley & Sons Inc.

[14] Falorsi,P.D.,Pallara,S.,Pavone,A.,Alessandroni,A.,Massella,E.andScannapieco, M. (2003).“Improvingthe qualityof toponymic datainthe italianpublicadministration”. In *ProoceedingsoftheInternationalWorkshoponDataQualityinCooperativeInformation Systems*.

[15] Hegewald, J., Naumann, F. and Weis, M. (2006). "XStruct: Efficient schema extraction from multipleandlargexmldocuments".In*Prooceedingsof22thInternationalConferenceonData Engineering Workshops*.

[16] Hengst, M. D. and Vreede, G. D. (2004). "Collaborative business engineering: A decade of lessonsfromthefield".*JournalofManagementInformationSystems*,Vol.20,No.4,pp85-114.

[17] Krawczyk,H.andWiszniewski,B.(2003).“Visualgqmapproachtoquality-drivendevelopment of electronic documents”. In *Proceedingsof the 2ndInternational Workshop on Web Document Analysis*.

[18] Lee,Y.W.,Strong,D.M.,Kahn,B.K.andWang,R.Y.(2002).“Aimq:Amethodologyfor information quality assessment”. *Information & Management*, Vol.40, No.2, pp133-146.

[19] Long,J.andSeko,C.(2002).“Anewmethodfordatabasedataqualityevaluationatthe canadian institute for health information (CIHI)”. In *Proceedings of 7th International Conference on Information Quality*.

[20] Loshin,D.(2008).*MasterDataManagement*.MorganKaufmann.

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

[22] McCallum,A.(2005).“Informationextraction:distillingstructureddatafromunstructuredtext”. *ACMQueue*,Vol.3,No.9,pp48-57.

[23] Naumann,F.,Freytag,J.C.andLeser,U.(2004).“Completenessofintegratedinformation sources”. *Information Systems*, Vol.29, No.7, pp583-615.

[24] Pareto, L. and Boquist, U. (2006). “A qualitymodel for design documentation in model-centric projects”.In*Proceedingsofthe3rdinternationalworkshoponSoftwarequalityassurance*.New York, NY, USA.

[25] Penna,G.D.,Marco,A.D.,Intrigila,B.,Melatti,I.andPierantonio,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.andWang,R.Y.(2002).“Dataqualityassessment”.*Communicationsof ACM*, Vol.45, No.4, pp211-218.

[27] Redman,T.C.(1997).*DataQualityfortheInformationAge*.ArtechHouse,Inc.

[28] Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M. and Baldoni, R. (2004). “The daquincisarchitecture:aplatformforexchangingandimprovingdataqualityincooperative information systems”. *Information Systems*, Vol.29, No.7, pp551-582.

[29] Shankaranarayan, G., Wang, R.Y.and Ziad, M.(2000).“Modelingthe manufacture of an informationproductwithIP-MAP”.In*Proceedingsofthe5thInternationalConferenceon Information Quality*. Massachusetts Institute of Technology, USA.

[30] Shankaranarayanan,G.andCai,Y.(2006).“Supportingdataqualitymanagementindecision - making”. *Decision Support Systems*, Vol.42, No.1, pp302-317.

[31] Shvaiko,P.andEuzenat,J.(2005).“Asurveyofschema-basedmatchingapproaches”.*Journal on Data Semantics IV*, pp146-171.

[32] Stoica,M.,Chawat,N.andShin,N.(2004).“Aninvestigationofthemethodologiesofbusiness process reengineering”. *Information Systems Education Journal*, Vol.2, pp1-11.

[33] Wang,R.Y.(1998).“Aproductperspectiveontotaldataqualitymanagement”. *CommunicationsofACM*,Vol.41,No.2,pp58-65.

[34] Wang,R.Y.,Lee,Y.W.,Pipino,L.andStrong,D.M.(1998).“Manage yourinformationasa product”. *Sloan Management Review*, Vol.39, No.4, pp95-105.

[35] Wang,R.Y.andStrong,D.M.(1996).“Beyondaccuracy:whatdataqualitymeanstodata consumers”. *Journal of Management Information Systems*, Vol.12, No.4, pp5-34.

[36] Yeh,D.andLi,Y.(2005).“Extractingentityrelationshipdiagramfromatable-basedlegacy database”. In *Proceedings of the 9th European Conference on Software Maintenance and Reengineering*, Manchester, UK.

CitationCount–74

CLOUD DATABASE DATABASE AS A SERVICE

WaleedAlShehri

DepartmentofComputing,MacquarieUniversity Sydney,
NSW 2109, Australia

ABSTRACT

Cloud computing has been the most adoptable technology in the recent times, and the database has also moved to cloudcomputingnow, so we willlook into thedetailsofdatabase asa service and itsfunctioning. Thispaperincludesallthebasicinformationaboutthedatabaseasaservice.The workingofdatabaseasa service and the challenges it is facing are discussed with an appropriate. The structure of database incloud computing and its working in collaboration with nodes is observed underdatabase 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 decideeitherto usedatabaseasa service ornot. Databaseasa service hasalreadybeen adopted by many e-commerce companies and those companies are getting benefits from this service.

KEYWORDS

Database,cloudcomputing,Virtualization,DatabaseasaService(DBaaS).

ForMoreDetails:http://airccse.org/journal/ijdms/1011s1.pdf

VolumeLink:http://airccse.org/journal/ijdms/Currentissue.html

REFERENCES

[1] Bloor,R.2011.WHATISACLOUDDATABASE?Retrieved25thNovember2012from 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. Retrieved24th November 2012 from http://www.cidrdb.org/cidr2011/Papers/CIDR11_Paper33.pdf

[3] Finley,K.2011.7Cloud-BasedDatabaseServices.Retrieved23rdNovember2012from http://readwrite.com/2011/01/12/7-cloud-based-database-service

[4] Hacigumus,H.,Iyer,B.andMehrotra,S.2004.EnsuringtheIntegrityofEncryptedDatabasesinthe 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.andMehrotra,S.ProvidingDatabaseasaService.Retrieved25thNovember 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.CloudComputing&Databases:Howdatabasescanmeetthedemandsofcloud computing. Retrieved 23rd November 2012 from http://www.scaledb.com/pdfs/CloudComputingDaaS.pdf

[8] Mykletun,E.andTsudik,G.2006.AggregationQueriesintheDatabase-As-a-ServiceModel. 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.andCabot,T.2012.DatabaseasaService:AMarketplaceAssessment.Retrieved23rd November 2012 from http://www.mitre.org/work/tech_papers/2012/11_4727/cloud_database_service_dbaas.pdf

[11] PostgresPlus.2012.CloudDatabase:GettingstartedGuide.Retrieved23rdNovember2012from 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.CloudComputing:DatabaseasaService.Retrieved24thNovember2012from http://cloudcomputing.sys-con.com/node/1985543

[14] VMware.2012.GettingStartedwithDatabase-as-a-Service.Retrieved23rdNovermber2012from 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/ ImageSource

[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.andCabot,T.(Authors).2012.DatabaseasaService:AMarketplaceAssessment. 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 workinginthe Department ofIT in Royal Saudi Air Force (RSAF).

CitationCount–71

DATA MINING TECHNIQUES: A SOURCE FOR CONSUMER BEHAVIOR ANALYSIS

AbhijitRaorane[1]&R.V.Kulkarni[2]

[1]Departmentofcomputerscience,VivekanandCollege,TarabaiparkKolhapur
[2]HeadoftheDepartment,Chh.ShahuInstituteofbusinessEducationandResearch Centre Kolhapur.

ABSTRACT

Various studies on consumer purchasing behaviors have been presented and used in real problems. Data mining techniquesare expected to bea more effective toolforanalyzing consumerbehaviors. 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

Consumerbehavior,Datamining,AssociationRule,Supermarket

ForMoreDetails:https://airccse.org/journal/ijdms/papers/3311ijdms04.pdf

VolumeLink:https://airccse.org/journal/ijdms/current2011.html

REFERENCES

1. ConsumerbehaviorThepsychologyofmarketingLarsparner
2. Consumerbehavior–HimanshuS.M.Dec242008
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. Procedingsof the first International Conference on Innovative computingInformation (2006)
5. Mining of users access behavior for frequent sequential pattern from web logs. S. Vijaylakshmi, V. Mohan, S. SureshRaja. InternationalJournalofDatabase ManagementSystem(IJDM) Vol2, August 2010.
6. Mining utility-oriented association rules: An efficient approach based on profitand quantity Parvinder S. Sandhu, Dalvinder S. Dhaliwal and S. N. PandaInternational 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 AssociationRuleMining forMarketBasket AnalysisShrivastava A.,SahuRGlobalJournal of e-Business & Knowledge Management Year:2007,Volume:3,Issue:1
11. Adataminingapproachtoconsumerbehavior–JanzoWatada,KozoYamashiro2006
12. Enhancing consumer behavior analysis bydata mining techniques – Nan-chan Hsich, Kuo-Chang cha 2009
13. Classification-MikeChapple
14. Dataminingtechniques-CDIG-BusinessintelligenceDataminingtechniques
15. AssociationRulesmining-S.Kotsiantis
16. Miningassociationwiththecollectivestrengthapproach-AggarwalC.C.,Yu,P.S.2001
17. FastAlgorithmsforMiningAssociationRules-AgrawalR.andSrikatl.Sept1994
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 MiningAssociation 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. Dynamicinterestcounting 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 UniversityISBN 1-55860-489-8-2001