IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Comparative Study on Data Mining Algorithms for Healthcare Information System

K. Mohan Kumar[1*], S. Jamuna[2]

[1] PG & Research Dept. of Computer Science, Rajah Serfoji Government College, Thanjavur,TN, India.
[2] PG & Research Dept. of Computer Science, Rajah Serfoji Government College, Thanjavur,TN, India.

Available online at: www.ijcseonline.org

Accepted: 17/Aug/2018, Published: 31/Aug/2018

***Abstract*—** Data Mining Is A Process Using High Volume Of Data For Needful Information. Most Popular Data Mining Techniques Are Rule Mining, Clustering, Classification And Sequence Pattern. Number Of Tests Should Be Done For A Patient To Detect A Disease. So, Large Volume Of Information Is Stored By The Health Care Information System For Further Reference. Due To The Complication Of Healthcare Information And The Slow Acquisition Of Technology, This Industry Lags Behind Other Industries In Implementing Effective Data Analysis And Extraction Strategies. Mining Information From The Large Health Databases Gives The Best Healthcare Information, Reduces Time And Saves The Humans From Complicated Diseases Like Cancer. In This Circumstance Proper Data Mining Technique Is Needed For The Best Performance. This Research Work Focuses On The Advantages And Disadvantages Of Various Data Mining Prediction Algorithms.

***Keywords*—**Data Mining, Healthcare System, Prediction, Techniques.

I. INTRODUCTION

In Data Mining, Data Extraction Is An Important Step In Discovering Hidden Prediction Information From Large Data Sets. Data Mining Is The Process Of Analyzing To Find Out The Hidden Factors, And Then The Data Would Become Data Tombs [1]. A Data Mining System May Generate Lots Of Patterns. Data Mining Tools Are Also Called Analytical Tools For Measuring Data. It Is Used By Users To Study Data From Many Different Patterns. In Practical, Data Extraction Is The Process Of Finding Internal Relationship Or Examples Among A Few Of Fields In Substantial Healthcare Databases. The Process Of Data Mining Is Consisting Of Several Steps; They Are Cleaning, Data Integration, Data Selection, Data Transformation, Knowledge Presentation And Pattern Evolution. The Data May Be Collected From Various Applications Including Science And Engineering, Management, Business Houses, Government Administration And So On. Some Data Patterns May Be Mined From Spatial, Time-Related, Text, Biological, Multimedia, Web And Legacy Database [2].

Some Of The Major Concepts Based On Mining Techniques Are Used For Medical Analysis Are Association Rule Mining, Classification, Clustering, Trend Analysis, Deviation Analysis And Similarity Measure. Data Consists Of Large Set Of Facts And Number Of Dimensions. Dimensions Are The Entries On Which An Organization Which Maintains The Record And They Will Be Hierarchical [3].

DATAMINING PROCESS

Data Mining Is Also Known As Knowledge Discovery In Database, Refers To Finding Knowledge From Large Datasets. It Is The Method Used To Operate On Large Set Of Data To Creating Hidden Information's And Relationships In Decision Making. In Data Mining, Discovering New Knowledge Is Done By The Following Seven Sequential Steps [4].

A. Data Cleaning: This Is The First Step Used To Eliminate Noise Data And Irrelevant Data From Collected Raw Data.

B. Data Integration: In This Step, Various Data Sources Are Combined Into Meaningful And Useful Data.

C. Data Selection: In This Step, The Relevant Data For The Analysis Are Retrieved From Various Resources.

D. Data Transformation: In This Step, Data Is Converted Or Clustered Into Required Forms For Mining By Performing Different Operations Such As Smoothing, Normalization Or Aggregation.

E. Data Mining: In This Step, Several Techniques And Smart Tools Are Applied To Extract Patterns Or Data Rules.

F. Pattern Evaluation: In This Step, Representative Knowledge Of The Attraction Information Is Identified Based On The Determined Metric.

G. Knowledge Representation: This Is The Final Stage Of Visualization And Representation Of Knowledge.

II. DATAMINING TECHNIQUES

The Data Mining Techniques Are Used In Mining Tasks. Association, Classification, Clustering, Prediction, Etc. Are Some Of The Data Mining Techniques. The Data Mining Techniques Mainly Classified As Descriptive And Predictive Models [5]. A Descriptive Model Presents The Data Form Which Is Essentially A Summary Of The Data Points, Finds Patterns In The Data And Understands The Relationships Between Attributes Represented By The Data, It Includes Tasks Such As Clustering, Association Rules, Summarizations, And Sequence Discovery. The Predictive Model Works By Predicting The Value Of The Data Using Known Results Found In Different Data Sets. The Predictive Data Mining Model Includes Classification, Prediction, Regression And Analysis Of Time Series [6].

A. Classification

Classification Is The Most Commonly Applied Technique, In Data Mining. It Finds Rules That Partition Data Into Some Groups. The Common Characteristics Of Classification Tasks Are Decision Trees, Neural Networks, Genetic Algorithms, Supervised Learning, Categories Which Depends On Variable And Assigning New Data. The Application Includes Credit Risk Analysis, Fraud Detection, Banking And Modelling Business Etc. [7].

B. Clustering

Clustering Is A Collection Of Similar Data Objects, Dissimilar Objects Is An Another Cluster. It Is Way Of Finding The Similarities Between Data According To The Organizing Data, Categorize Data For Model Construction And Data Compression, Etc. Develop These Algorithms And Classify Them Into Partitioning Methods, Layering Methods, Density, And Grid-Based Methods. The Datasets May Be Numerical Or Categorical K-Means, Hierarchical, Are Some Of The Well-Known Data Clustering Algorithms [8][9].

C. Association Rule Mining

This Is A Most Of Researched Technique For Discovering Hidden Relationship Between Entities In Large Dataset. In This Technique, The Presence Of Another Model, I.E. Item Is Related To Another In Terms Of Cause And Effect. The Main Aim Is To Create All The Rules That Have Greater Than Or Equal To Minimum Support Or Confidence In A Database. Support Means The Percentage Of Total Transactions Of Two Different Items. Confidence Means How Much Particular Item Is Depending On Another. There Is No Significance For The Patterns With Low Confidence And Support [10][11].

D. Regression

Regression Is Another Predictive Data Mining Model Is Also Known As Supervised Learning Technique. This Technique Analyse The Dependency Of Some Attribute Values, Which Is Dependent Upon The Value Of Other Attributes Mainly, Present In Same Item. In This Techniques Target Values Are Known [12].

E. Time Series *Data Analysis*

In Time Series Analysis, Changes And Its Values Are Dependent On Time. The Values Are Typically Measured At Equal Time Interval Based On Hour, Day, And Week. A Sequence Database Which Consists Of Events Is In Ordered Manner, Sometimes Having Frequent Interval Of Time [13].

DATAMINING TOOLS

Data Mining Professionals And Their Organizations Have Access To Many Data Mining Tools That Allow Installing And Using A Variety Of Basic Tools In Different Ways [14]. The Following Are The Most Popular Open Source Tools In Data Mining.

A. MATLAB

This Tool As Of Now Supports Different Usage Of Various Phases Of The Information Mining Process, Including Different Toolboxes Made By Specialists In The Field. An Underlying Finish Of This Investigation Is That MATLAB Is A Powerful And Flexible Bundle For Satisfying The Prerequisites Of The Information Mining Process. It Is Clear, That There Is A Need For The Extension And Synthesis Of The Existing Tools. The Synthesis Of Information Mining Devices Sketched Out And Exhibited In This Implementation Takes Into Thesis A Significantly More All Encompassing Way To Deal With Information Mining In MATLAB Than Has Been Accessible Beforehand. This Work Guarantees That Information Mining Turns Into An Inexorably Clear Undertaking, As The Proper Instruments For A Given Analysis Become Apparent. As A Consistent Augmentation Of The Synthesis Gave, A Concise Dialog Is Given With Respect The Production Of Data Mining Toolbox Stash For MATLAB [15].

B. WEKA

Weka Is The One Of The Popular Tool In Data Mining, It Was Developed In A Non-Java Version For Analysing Agricultural Data. This Tool Is Used For The Different Applications In Data Mining Like Predictive Modelling And

Data Analyzing. This Software Is Under Free Of Cost Which Is The Big Advantage To Compare To Rapidminer. It Is A Graphical User Interface Makes It Better Understanding Tool For Data Mining Process [16].

C. R-Programming

This Is A Programming Language And Free Software For Statistical Computing And Graphics. It Is Supported By The R Foundation For Statistical Computing. R Language Is Used For Writing Lots Of Modules Of The Software Itself. R Programming Software Is Fee, And It Is Developed For Statistical Packages And Analyzing The Data Which Is Highly Extensible. It Provides The Different Statistical Techniques That Include Linear And Non-Liner Modelling Data Mining Process [17].

D. Rapid Miner

Rapidminer Is A Tool Which Is Written In Java Programming Language And It Is The Advanced Level In Analysing The Data Through Its Template. It Also Provides Data Pre-Processing And Visualization, Predictive Analysing And So On. It Is The One Of Best Business Analytics Software [18].

E. ORANGE

This Is An Open Source Tool For Data Mining Which Is Python Based Used For The Purpose Of Knowledge Extraction. It Likewise Has Segments For Machine Learning And Additional Items For Bio-Informatics And Content Mining. Orange Is Upheld On Macos, Windows And Linux And Can Likewise Be Introduced From The Python Package Index Repository [19].

F. KNIME

KNIME Has The Ability To Perform Three Main Tasks In Data Pre-Processing. They Are Extraction, Transformation, And Loading. It Provides A GUI (Graphical User Interface) That Allows Assembly Nodes To Perform Data Processing. It Is A Platform For Analysis Of Information, Integration And Reporting. KNIME Is Written In Java, And Based On Eclipse, KNIME Is Easy To Extend And Add Add-Ons [20].

III. PREDICTION ALGORITHMS IN DATA MINING

The Main Aim Of This Prediction Algorithms Are To Find Hidden Information Based On Current Values. Some Of The Predictive Tools Are Neural Networks, Regression, Support Vector Machines (SVM), And Discriminant Analysis [21][22]. Recently, Data Mining Techniques Such As Neural Networks, Fuzzy Logic Systems, Genetic Algorithms, And Approximate Set Theory Have Been Used For Predictive Control And Detection Tasks. These Algorithms Will Predict The Probability Of A Given Data Situation. If The Probability Is Equal To 1, It Means That The Data (Partial) Is Normal; Otherwise, If The Probability Is Equal To 0, The (Partial) Data Is Considered To Be Unconventional [23].

IV. METHODOLOGY

Data Mining Involves Several Important Techniques Such As Association, Classification, Clustering, Prediction, Sequential Patterns, And Decision Trees. Many Algorithms Developed In Various Periods For Data Mining Process. In This Study The Pros And Cons Of Twenty Popular Prediction Algorithms Are Analysed Under Its Classification And Tabulated.

V. RESULTS AND DISCUSSION

The Following Table-1 Lists The Key Advantages And Disadvantages Of Data Mining Prediction Algorithms In Each Category.

Table-1. Advantages & Disadvantages Of Prediction Algorithms

S.No.	Algorithm	Advantages	Disadvantages
Regression Algorithms			
1	Linear Regression	Easy To Understand And Implement	It Is Limited To Linear Relationships It Is Very Sensitive To Outliers It Assumes That The Data Are Independent
2	Logistic Regression	Easy To Understand And Implement	It Requires Each Data Point To Be Independent Of All Other Data Points Prone To Overfitting
3	Autoregressive Integrated Moving Average (ARIMA)	Adapts Statistical Approach Strictly Increases The Forecasting Accuracy In Minimal Parameters	High Cost Unstable With Respect To Changes In Observation And Changes In Model Specification It Works For Short Run
4	Multivariate Adaptive Regression Splines	It Works Well Even With A Large Number Of Predictors Automatically Detect Interactions Between Variables Efficient And Efficient Powerful Outliers	Hard To Understand Easy To Make More Adjustments The Model Is Prone To Losing Data
Instance Based Algorithms			

S.No.	Algorithm	Advantages	Disadvantages
5	K-Nearest Neighbor (KNN)	Powerful To Noisy Data Effective If The Training Data Is Large	Need To Determine The Value Of Parameter K Must Determine The Type Of Distance High Computational Cost
6	Kernel Regression	It Is Nonparametric	If The Independent Variables Are Unevenly Distributed, It Is Easy To Produce Deviations.
7	Support Vector	The Accuracy Of The Forecast Is Usually Very High Quickly Evaluate The Objective Function Which Are Learning	Long Training Time Hard To Understand Learning Function (Weight)
Decision Tree Algorithms			
8	Classification And Regression Trees (CART)	Easy To Understand And Implement Classification Is Easy To Manage	Oversupply Diagonal Decision Limits May Be A Problem
9	Iterative Dichotomiser 3 (ID3)	Create Comprehensible Prediction Rules From Training Data Build The Fastest And Shortest Tree	Oversupply Prediction Of Continuous Data Can Be Computationally Expensive...
10	C 4.5	It Can Be Used For Classification And Continuous Values It Handles Noise	Minor Changes In Data Can Lead To Different Decision Trees It Does Not Work Well For Small Training Data
Bayesian Algorithms			
11	Naive Bayes	Easy To Implement Need Very Little Data To Train The Model	Loss Of Precision Due To Assumption Of Conditional Independence
12	Bayesian Network (BN)	Have Rigorous Probabilistic Foundation Reasoning Process Is Semi-Transparent	Computationally Expensive Performance Is Poor On High Dimensional Data
Clustering Algorithms			
13	K-Means	Computationally Fast Easy To Implement Works Well With High Dimensions	Difficult To Predict The Number Of Cluster (K-Value) Order Of The Data Has An Impact On Final Result Sensitive To Scale
14	Expectation Maximization (EM)	Easy To Implement	Slow Linear Convergence Initial Estimation Should Be Carefully Chosen
15	Hierarchical Clustering	Easy To Implement Easy To Decide The No. Of Clusters By Looking At The Dendrogram	Not Suitable For Large Dataset Very Sensitive To Outliers
Artificial Neural Network Algorithms			
16	Perceptron	Guaranteed Convergence When Linearly Separable Very Fast On Test Data	Thrashes When Not Linearly Separable
17	Back-Propagation	If The Selected Weight Is Small At The Beginning, The Calculation Time Will Decrease Batch Update Of Weights Provides Smoothing In Weight Correction	Output Can Be Fuzzy Or Non-Numeric Prone To Local Minima, Resulting In Poor Solution
18	Hopfield Network	Massive Parallel Computation	Computational Efficiency Is Not Consistent
Ensemble Algorithms			
19	Adaboost	Easy To Implement Not Prone To Overfitting	Sensitive To Noisy Data And Outliers
20	Random Forest	Reduce Overfitting Less Difference	More Complex Unimaginable

The Above Table-1 Proves That Every Algorithm Has Its Own Advantage And Disadvantages.

VI. CONCLUSION

The Comparative Study Of Various Data Mining Prediction Algorithms Shows That The Powerful And Efficient Algorithm Is Essential To Manipulate Data For Correct Prediction. The Tools Which Are Developed For Prediction Like Health Care System Should Adopt The Advantages And Try To Omit The Disadvantages For Best Prediction. Such Types Of Tools Only Can Give Good Performance In The Fields Such As Healthcare To Identify The Kind Of Disease, Education To Identify The Behaviour Of Teaching And Learning, Etc.

REFERENCES

[1] Larsson EG, Selén Y. "Linear Regression With A Sparse Parameter Vector". IEEE Transactions On Signal Processing. 2007 Feb;55(2):451-60.

[2] Wu J, Huo Q. "A Study Of Minimum Classification Error (MCE) Linear Regression For Supervised Adaptation Of MCE-Trained Continuous-Density Hidden Markov Models". IEEE Transactions On Audio, Speech, And Language Processing. 2007 Feb;15(2):478-88.

[3] Zhang S, Zhang L, Qiu K, Lu Y, Cai B. "Variable Selection In Logistic Regression Model. Chinese Journal Of Electronics". 2015 Oct 1;24(4):813-7.

[4] Zhang J, Jiang J. "Rank-Optimized Logistic Matrix Regression Toward Improved Matrix Data Classification". Neural Computation. 2018 Feb;30(2):505-25.
[5] Li C, Chiang TW. "Complex Neurofuzzy ARIMA Forecasting—A New Approach Using Complex Fuzzy Sets". IEEE Transactions On Fuzzy Systems. 2013 Jun;21(3):567-84.
[6] Gong S, Gao Y, Shi H, Zhao G. "A Practical MGAARIMA Model For Forecasting Real-Time Dynamic Rain Induced Attenuation". Radio Science. 2013 May 1;48(3):208-25.
[7] Fu C. "Business Valuation Based On Intellectual Capital: A Hierarchical Clustering-MARS Approach. In Management And Service Science (MASS)", 2011 International Conference On 2011 Aug 12 (Pp. 1-6). IEEE.
[8] Crino S, Brown DE. "Global Optimization With Multivariate Adaptive Regression Splines". IEEE Transactions On Systems, Man, And Cybernetics, Part B (Cybernetics). 2007 Apr;37(2):333-40.
[9] R.S. Walse, G.D. Kurundkar, P. U. Bhalchandra, "A Review: Design And Development Of Novel Techniques For Clustering And Classification Of Data", International Journal Of Scientific Research In Computer Science And Engineering, Vol.06, Issue.01, Pp.19-22, 2018
[10] Nikita Jain, Vishal Srivastava "Data Mining Techniques: A Survey Paper" IJRET: International Journal Of Research In Engineering And Technology, Volume: 02 Issue: 11 | Nov-2013.
[11] Pawan S. Wasnik, S.D.Khamitkar, Parag Bhalchandra, S. N. Lokhande, Ajit S. Adte, "An Observation Of Different Algorithmic Technique Of Association Rule And Clustering", International Journal Of Scientific Research In Computer Science And Engineering, Vol.06, Issue.01, Pp.28-30, 2018.
[12] Suárez A, Lutsko JF. "Globally Optimal Fuzzy Decision Trees For Classification And Regression". IEEE Transactions On Pattern Analysis And Machine Intelligence. 1999 Dec;21(12):1297-311.
[13] Goin JE. "Classification Bias Of The K-Nearest Neighbor Algorithm". IEEE Transactions On Pattern Analysis And Machine Intelligence. 1984 May(3):379-81.
[14] J. Han And M. Kamber. "Data Mining, Concepts And Techniques", Morgan Kaufmann, 2000.
[15] D. J. Higham And N. J. Higham. MATLAB Guide. Siam, Second Edition Edition, 2005.
[16] R Development Core Team.. R: A Language And Environment For Statistical Computing [Computer Software And Manual]. Vienna, Austria: R Foundation For Statistical Computing. Available From Http://Www .R-Project.Org. 2007
[17] Jindal And Dutta Borah, ―A Survey On Educational Data Mining And Researchtrends, In International Journal Of Database Management Systems,2013, Vol.5, No.3. Http://Dx.Doi. Org/10.5121/Ijdms.2013.5304.
[18] Ha, S., Bae, S., And Park, S , Web Mining For Distance Education, In Proc.Int. Conf. On Management Of Innovation And Technology, IEEE.,2000, Pp.715-719. Http://Dx.Doi. Org/10.1109/Embeddedcom-Scalcom.2009.98.
[19] Zhang, H., Raitoharju, J., Kiranyaz, S., & Gabbouj, M. (2016). Limited Random Walk Algorithm For Big Graph Data Clustering. Journal Of Big Data, 3(1), 26.
[20] Witten, I., H., Frank, E., Hall, M., A.. Data Mining: Practical Machine Learning Tools And Techniques.2011.
[21] Ma X, Ye Q, Yan H. "L2P-Norm Distance Twin Support Vector Machine". IEEE Access. 2017;5:23473-83.
[22] Kampa K, Hasanbelliu E, Cobb JT, Principe JC, Slatton KC. "Deformable Bayesian Network: A Robust Framework For Underwater Sensor Fusion". IEEE Journal Of Oceanic Engineering. 2012 Apr;37(2):166-84.
[23] Aarti Sharma Et Al, "Application Of Data Mining – A Survey Paper", International Journal Of Computer Science And Information Technologies', Vol. 5 (2), 2014.

Authors Profile

Dr. K. Mohan Kumar Received Master Of Computer Science, Ph.D In Computer Science From Bharathidasan University, Tiruchirappalli, India And M.Phil Computer Science From Manonmaniyam Sundaranar University, Thirunelveli, India. He Is Currently Working As Head, PG And Research Department Of Computer Science, Rajah Serfoji Government College, Thanjavur, T.N, India. His Main Research Work Focuses On Iot, Cloud Computing, Network Security, Big Data Analytics And Computational Intelligence Based Education. He Has Published More Than 50 Research Papers In Reputed International Journals. He Has 23 Years Of Teaching Experience And 18 Years Of Research Experience.

Mrs. S.Jamuna Pursed Bachelor Of Computer Science, Master Of Computer Science And M.Phil Computer Science From Bharathidasan University, Thiruchirappaili. Now, She Is Doing Ph.D In PG And Research Department Of Computer Science, Rajah Serfoji Government College, Thanjavur Affiliated To Bharathidasan University, T.N, India. She Is Having 11 Years Teaching Experience. His Main Research Work Focuses On Data Mining, Data Analytics, And Networks.

