International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue-17
December 2017

Analysis on different Data Mining methods for the Internet of Things

B.Arathi

Assistant professor,CSE Dept Kamala Institute of Technology and science,Huzurabad

Abstract:*The big information generated by way of the Internet of Things (IoT) is taken into consideration of excessive commercial enterprise price, and data mining algorithms may be carried out to IoT to extract hidden data from records. Starting from home, workplace, industry automation to healthcare and smart city internet of factors has revolutionized the arena by way of interconnecting them. As a result, it generates massive volumes of information. For many, this statistics has huge enterprise cost and data. This is wherein data mining comes into play which makes such kind of systems smarter sufficient for better performance and greater opportunities and services. This paper introduces to the Internet of Things technology and states the need for data mining in an international wherein the whole thing is delivered over the internet and explains the process and suitable algorithms required for the Internet of things.*

Keywords- Data mining, Internet of things, Knowledge Data Discovery.

I. INTRODUCTION

The Internet of Things (IOT) and its relatedtechnologies can seamlessly integrate classicalnetworks with network instruments and devices. Thedata in the Internet of Things can be categorized intoseveral types: RFID data stream, address identifiers,descriptive data, positional data, environment dataand sensor network data etc. [1]. Today, IOT bringsthe great challenges for managing, analysing andmining data. In IOT systems, data qualitymanagement is a critical technology to provide highquality and trusted data to business-level analysis,optimization and decision making. In order toimprove quality of data, anomaly detectiontechniques are widely used to remove noises andinaccurate data. For anomaly detection, having moredata means it's easier to detect an unusual eventagainst the background of normal events [3].Data Clustering refers to grouping of data basedon specific features and its value. In IOT, Dataclustering is an intermediate step for identifyingpatterns from the collected data. It's most commonprocess in unsupervised machine learning. Clusteringmethods are divided into 4 major categories such as:partitioning methods, hierarchical methods, densitybased methods and grid based methods. Otherclustering techniques also exist such as: fuzzyclustering, artificial neural network and genericalgorithms.The problem of Data classification is stated as:given a set of training data points along withassociated label for an unlabelled test instances.Classification algorithm contain 2 phases:

Trainingphase and
Testing phase.

On the basis of training dataset, segmentation is done which encodes knowledgeabout the structure of the groups in form of targetvariable. Thus classification problem is referred to assupervised learning.The feature selection is the process used torecognized pattern and allows us to identifyattributes that affect quality index the most. Aftersome initial level of experiment feature selection ispreferable, identify what are attributes that affects aspecific problem most and then perform dataclassification, time series prediction or anomalydetection more easily as it reduce the dimensionalityin mining the problem. Features selection is to find asatisfactory feature subset from the candidate featureset, so that to reach an optimal classification accuracyand computing complexity control.A time series is collection of temporal dataobjects, which includes characteristics such as: largedata size, high dimensionality, and updatingcontinuously. Representation, similarity measuresand indexing are 3 components of time series taskrelies on. Time series representation reduces thedimension and it divides into 3 categories: modelbased representation, non-adaptive datarepresentation and adaptive data representation. Thesimilarity measure is carried out in proper mannersuch as: research directions include subsequencematching and full subsequence matching. Theindexing of time series is linked with representationand similar measure tools [2].

II. RELATED WORK

Te goal of classifcation is to accurately predictthe target class for each case in the data [15]. For example, aclassifcation model could be used to identify loan applicantsas low, medium, or high credit risks [16].There are many methods to classify the data, including decision tree induction, frame-based or rule-basedexpert systems, hierarchical classifcation, neural networks,Bayesian network, and support vector machines (see Figure2).

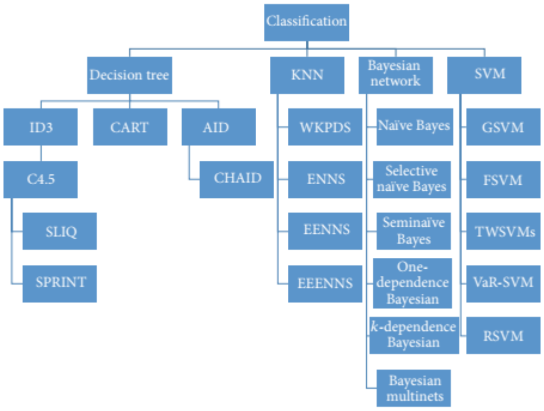

Figure 2: Te research structure of classifcation.

(i) A decision tree is a flow-chart-like tree structure,where each internal node is denoted by rectangles andleaf nodes are denoted by ovals. All internal nodeshave two or more child nodes. All internal nodescontain splits, which test the value of an expressionof the attributes. Arcs from an internal node to itschildren are labeled with distinct outcomes of the test.Each leaf node has a class label associated with it.Iterative Dichotomiser 3 or ID3 is a simple decisiontree learning algorithm [17]. C4.5 algorithm is animproved version of ID3; it uses gain ratio as splittingcriteria [18]. Te difference between ID3 and C4.5algorithm is that ID3 uses binary splits, whereas C4.5algorithm uses multiway splits. SLIQ (SupervisedLearning In Quest) is capable of handling large datasets with ease and lesser time complexity [19, 20],SPRINT (Scalable Parallelizable Induction of Decision Tree algorithm) is also fast and highly scalable,and there is no storage constraint on larger data setsin SPRINT. Other improvement researches arefnished. Classifcation and Regression Trees(CART) is a nonparametric decision tree algorithm.

It produces either classifcation or regression trees,based on whether the response variable is categorical or continuous. CHAID (chi-squared automaticinteraction detector) and the improvement researcher focus on dividing a data set into exclusive andexhaustive segments that differ with respect to theresponse variable.(ii) Te KNN (K-Nearest Neighbor) algorithm is introduced by the Nearest Neighbor algorithm which isdesigned to fnd the nearest point of the observedobject. Te main idea of the KNN algorithm is to fndthe K-nearest points. There are a lot of differentimprovements for the traditional KNN algorithm,such as the Wavelet Based K-Nearest Neighbor PartialDistance Search (WKPDS) algorithm, EqualAverage Nearest Neighbor Search (ENNS) algorithm, Equal-Average Equal-Norm Nearest Neighborcode word Search (EENNS) algorithm, theEqual-Average Equal-Variance Equal-Norm NearestNeighbor Search (EEENNS) algorithm andother improvements.

(iii) Bayesian networks are directed acyclic graphs whosenodes represent random variables in the Bayesiansense. Edges represent conditional dependencies;nodes which are not connected represent variables which are conditionally independent of eachother. Based on Bayesian networks, these classifershave many strengths, like model interpretability andaccommodation to complex data and classifcationproblem settings. Te research includes naïve Bayes, selective na¨ıve Bayes , semina¨ıveBayes, one-dependence Bayesian classifers, K-dependence Bayesian classifers, Bayesiannetwork-augmented na¨ıve Bayes, unrestrictedBayesian classifers, and Bayesian multinets.

(iv) Support Vector Machines algorithm is supervised learning model with associated learning algorithmsthat analyze data and recognize patterns, which isbased on statistical learning theory. SVM producesa binary classifer, the so-called optimal separatinghyperplanes, through an extremely nonlinear mapping of the input vectors into the high-dimensionalfeature space. SVM is widely used in textclassifcation , marketing, pattern recognition, and medical diagnosis. A lot of furtherresearch is done, GSVM (granular support vectormachines), FSVM (fuzzy support vectormachines), TWSVMs (twin support vectormachines), VaR-SVM (value-at-risk support

III. PROPOSED WORK

There are two ways in which the processes of thedata mining is explained one is the KDD processeswith seven stages where as the other process model is
the Cross Industry Standard Process for Data Mining (CRISP – DM) which has six stages inclusive ofBusiness Understanding as the name suggests thisprocess model deals with the Industry standards so a
basic business understanding is inevitable astraditionally companies are mined to see future trends
and better opportunities in the company.For solving our present scenario which is tomanage the huge data from IoT and apply suitabledata mining technique, we will first look up the sevenstages in the KDD processwhich are as follows:

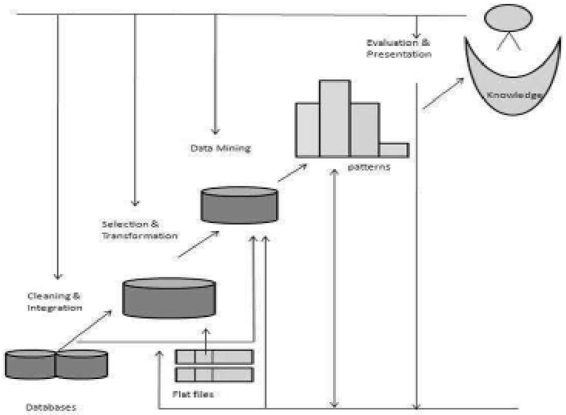

Fig. 2. The figure depicts the basic process model of KnowledgeData Discovery which comprises of cleaning, integration, selection,transformation of data followed by pattern evaluation andpresentation.

Cleaning: The erratic data which has norole in providing valuable information isto be removed.

• **Integration**: This process is to associatevarious types of data.

• **Selection**: In this step the pertinent data isto be restored from the database to achieveproper knowledge by analyzingappropriate data.

• **Transformation of data**: The termtransformation itself states that there is achange in the state of data, i.e., the data'sformat is changed from the source systemto the destination system by performingvarious operations on it such as mapping
or summation.

• **Data Mining**: As mentioned above, thisstep is to extract information from thedatabase on the basis of the requiredpatterns using suitable algorithms.

• **Evaluation**: Through which pattern thedata is being extracted and information isgenerated is evaluated to ensure thecorrectness of the information.

• **Presentation**: Finally, the informationrequired is plotted in the form of graphs orother statistical methods for betterunderstanding.

The above mentioned seven stage KDD processesare the typical process stages under which data miningis performed. Further discussion is upon how thismodel is suitable for IoT.

B.Suitable Data Mining Processes for IoT: We live in a world where the speed with which thebusiness

needs to move is much faster than the time ittakes to conceive and launch new solutions in theareas of big data, data mining, cloud, and IoT [3]. Tofind relatively small chunks of data in peta byte sizeddatabases generated from an IoT system is likelooking for a black cat in a coal cellar. To get in thegame, variety of data mining algorithms should bebuilt with various capabilities to get insights andreduce the risk of project failures. Till today there are
many studies which have been trying to solve theproblem of acquiring of big data on IoT systems.Most of the mining techniques are developed toexecute on a single system, so these KDD systemscannot be applied directly to process big data of theIoT system, whereas for a small system undoubtedlythese KDD processes can be applied directly.

To develop a high geared data mining structure ofKDD for an IoT system the following three points [5]
are to be considered to elect the suitable miningtechnology, and they are –

- First and the foremost it is essential tounderstand the definition of the problem,their limitations and required informationand so forth.
- Secondly, the major concern would be tounderstand what kind of data is to berequired like the representation, size ofdata, processing of different data etc.,
- Thirdly on the basis of the abovementioned points, a suitable data miningalgorithm is to be chosen to bring outsensible and required information fromthe raw data.

C. Data Mining Algorithms

- **Classification:** It is a function of datamining that delegates items intocategorical labels. It helps us to predict thecategory of a particular item in a dataset.Let’s consider a scenario where amarketing manager of an automobilecompany wants to analyze the probabilityof a customer buying a type of car on the
basis of his/her profile. A classificationmodel can be utilized to predict the typeof car; family, sports, truck or van, that acustomer is likely to buy on the basis ofhis/her age and family background.There are various classification modelssuch as decision tree, neural networks, IF-THEN rules depending upon their use.

- **Clustering**: Unlike classification,clustering is typically defined ascategorizing the data into some sensible,meaningful groups or classes. This helpsto achieve an easy perceptive for the usersby grouping naturally. The best examplefor this could be a search engine which isbased on clustering, that can categorizeendless web pages into news, images,videos, reviews etc.,There are various clustering models suchas kMeans clustering, k-Medoidsclustering, Densitybased clustering andHierarchical clustering that can be useddepending upon their use.

- **Association Analysis**: Market basket is thebest relatable module to association.Market basket analysis is observedroutinely in supermarket chains where theitems which are likely to be boughttogether with another set of items arealways placed together such as toothbrushand toothpaste are always in the samesection. This helps in decision making. Atfirst the data is processed incessantly, forfirst catalog of association analysis.To discover inter transactional associationapriori algorithm has been used followedup with association discovery. Otheralgorithms used are pattern growth, eventoriented, event-based, partition based, FPGrowth, Fuzzy set and incrementalmining.
- **Time Series Analysis**: When data pointsare present in consecutive time interval,time series analysis is applied to extractmeaningful related to specific patterns orstatistics. Stock market index value isanalyzed in a time series manner. Timeseries analysis is also used in forecasting,to analyze dependent events; that is topredict future values based on past events.
- **Outlier Detection**: Intermittently thereexists a data which is not complaisant withgeneral behavior or model of the data.This kind of data is different fromremaining set of data which is called asoutlier. This type of data contains usefulinformation regarding aberrant behaviorof the system comprised of outliers.Outlier analysis can be used to

extrapolateoutliers, to calculate distance amongobjects, distribution of input space.The above mentioned data mining functionalities withthe listed algorithms are the most commonly usedalgorithms in any field to mine the data and extractthe required information/

IV. CONCLUSION

In this paper, we've mentioned the brand newrising generation that is the Internet of Things (IoT),later moving on to how statistics mining is an essentiala part of IoT which makes those systems smarter bydiscussing the overall strategies of information mining. Alsowe've got seen key factors to maintain in mind whiledeciding on the appropriate algorithm for an IoT gadget.Further discussion becomes about the broadly used informationmining functionalities with their particular algorithmsand various IoT applications concerning it to the properdata mining capability applied to beautify thesystem for better offerings.

REFERENCES

1] S. Haller, S. Karnouskos, and C. Schroth,"The Internet of Things in an enterprise context,"Future Internet Systems (FIS), LCNS, vol. 5468.Springer, 2008, pp. 14-8.

[2] PlamenNedeltchev,"It is inevitable. It ishere. Are we ready?" The Internet of Everything is thenew Economy [Online].Available :http://www.cisco.com/c/en/us/solutions/collateral/en te rprise/ciscooncisco/Cisco_IT_Trends_IoE_Is_the_New_Economy.html

[4] Shen Bin, Liu Yuan, Wang Xiaoyi,"Introduction to Research on Data Mining", Researchon Data Mining Models for the Internet of Things[Online].Available:https://www.ceid.upatras.gr/webpages/faculty/vasilis/Courses/SpatialTemporalDM/Papers/InternetOfThings05476146.pdf

[5] Chun-Wei Tsai, Chin-Feng Lai, Ming-ChaoChiang, and Laurence T. Yang, "Basic Idea of UsingData Mining for IoT,"Data Mining for Internet ofThings: A Survey, IEEE Communications Surveys &Tutorials, vol.16.

[6] L. Duan, W. N. Street, and E.Xu, "HealthCare Information Systems: Data Mining Methods inthe Creation of a Clinical Recommender System,"Enterprise Information Systems, vol.5, no.2, pp.169-181, 2011.

[7] B. K. Schuerenberg, "An informationexcavation. Las Vegas payer uses data miningsoftware to improve HEDIS reporting and providerprofiling," Health Data Management, vol. 11, no. 6,pp. 80–82, 2003.

[8]P. Deng, J. W. Zhang, X. H. Rong, F. Chen, "A model oflarge-scale Device Collaboration system based on PI-Calculusfor green communication", Telecommunication Systems, 52(2),pp.1313-1326, 2013.

[9]P. Deng, J. W. Zhang, X. H. Rong, F. Chen, "Modeling thelarge-scale device control system based on PI-calculus",Advanced Science Letters, 4(6-7), pp.2374-2379, 2011.

[10]J. Zhang, P. Deng, J. Wan, B. Yan, X. Rong, F. Chen, "A novelmultimedia device ability matching technique for ubiquitouscomputing environments", EURASIP Journal on WirelessCommunications and Networking, 2013(1), pp.1-12, 2013.

[11]G. Kesavaraj, S. Sukumaran, "A study on classificationtechniques in data mining", In Proceedings of Computing,Communications and Networking Technologies (ICCCNT),2013Fourth International Conference on, pp. 1-7, 2013.

[12]S. Song, "Analysis and acceleration of data mining algorithmson high performance reconfigurable computing platforms",Ph.D, Iowa State University, 2011.

[13]J. R. Quinlan, "Induction of decision trees", Machine Learning,1(1), pp.81-106, 1986.

[14]J. R. Quinlan, C4. 5: programs for machine learning, vol. 1, pp.,Morgan kaufmann, 1993.

[15]M. Mehta, R. Agrawal, J. Rissanen, SLIQ: A fast scalableclassifier for data mining, Springer (18-32), 1996.

[16]B. Chandra, P. P. Varghese, "Fuzzy SLIQ decision treealgorithm", Systems, Man, and Cybernetics, Part B: Cybernetics,IEEE Transactions on, 38(5), pp.1294-1301, 2008.

[17]J. Shafer, R. Agrawal, M. Mehta, "SPRINT: A scalable parallelclassier for data mining", In Proceedings of Twenty-secondInternational Conference on Very Large Data Bases, pp. 544-555,1996.

[18]K. Polat, S. Güneş, "A novel hybrid intelligent method basedon C4. 5 decision tree classifier and one-against-all approachfor multi-class classification problems", Expert Systems withApplications, 36(2), pp.1587-1592, 2009.

[19]S. Ranka, V. Singh, "CLOUDS: A decision tree classifier forlarge datasets", Knowledge discovery and data mining, pp.2-8,1998.

[20]M. van Diepen, P. H. Franses, "Evaluating chi-squaredautomatic interaction detection", Information Systems, 31(8),pp.814-831, 2006.