Study and Analysis of Recent Data Mining based Machine Learning Methods

Shubham Sand
Department of Information Technology
Sinhgad collage of Engineering

Abstract— The Data Mining using Machine Learning Methods it is important topic for a research and many methods are established for the improvement in accuracy and results. The data Mining information and knowledge from large databases are described in many research papers as a main research concept in machine learning and the multiple numbers of industries has main area with an opportunity to make different researches. The Researchers in different fields show the great interest in data mining using machine learning. Several applications are available in information providing services and some concepts are data warehousing, online services over the Internet and it also called as various data mining techniques to understand the user behaviour to improve the services and business opportunity. This article provides a survey for a database researcher's point of view on the data mining techniques developed recently. A classification of the available data mining using machine learning techniques are provided and a comparative study of such techniques are presented.

Key words: Data Mining, Machine Learning

I. INTRODUCTION

Now days, the opinion mining is used world widely in real time applications. The Machine Learning is a mature and well-recognized research and mainly focused with the analysis of approach, design and other precision in data. The data mining is a technique that are used to computing procedures of finding patterns in huge data sets that includes function at the intersection of machine learning, database model and statics. The Data mining is an important technique where intelligent function is provides to extract data patterns. It is an integrative sub branch of computer science. The main objective of the data mining technique to eliminates information for the datasets and also converts it into an intelligible scheme for future use [1]. In this phase consists of some numbers of steps, the first step is raw analysis, it includes database and data management aspects, data pre-processing, interestingness metrics, post-processing of discovered scheme, inference and model, visualization, complexity considerations and online updating. In the KDD (knowledge discovery) analysis step is known as data mining. The main objective of these methods is that entries points placed in low-dimensional manifold, and the graph is utilized for an estimation of the denoted manifold. The adjacent point pair joins huge weight edges move to have the similar labels and inversely. In these directions, the labels related with data can be communicated via the data mining.

The ML (machine learning) / DM (Data mining) methods are explained and some project of these methods to cyber IDS challenges. The computation complexity of ML (machine learning) / DM (Data mining) is explained, in this research paper supplies a datasets of differential precedents for ML/DM methods and a set of commendation on the good methods to utilized depending on the attributes of the cyber issue to find out. The traditional machine learning algorithm take as input a feature vector, which shown an object in concepts of the categorical or numeric attributes. The important machine learning work is to mapping from this feature vector to an outcomes guess of some form. These could be class labels, a regression score, or a latent vector or unsupervised cluster identity. In the statistical relation learning, presented of an object that include its association to other objects. Thus the data is in the form of a graph, consisting of nodes and edge labelled. The important objectives of data mining using machine learning algorithm involves prediction of properties of nodes, prediction of missing edge and clustering nodes based on their connected designs. The works obtained in multiple setting thus examines of biological pathways and also social networks.
Machine learning technique is similar to the computational statistics, which also concentrated on guess making via the utilized of computers. It has powerful link to mathematical optimization, these are transfer method, concepts and project domain to the branch. In the field of data analytics, the machine learning is a function utilized to formulated complex approach and algorithms that provides themselves to guess; these are utilized in commercial is called as predictive analytics. These analytical approaches enable researchers, engineers, information scientists and analysis to "generate reliable, repeatable decision and output" and uncover "hidden insights" via learning from ancient association and trends in the data.

The section II, explain the various technique. In section III, the comparative study and analysis is discussed in tabular format. In IV, conclusion and future suggestions discussed.

II. RELATED WORKS

In this section, the various technique are explained from 2012 to 2017 reported in domain of Data Mining based on Machine Learning by considering the different sub domains such as analysis, mining and classification.

A. Wes Copeland et.al (2012)

In [1], in this paper the author describe the importance pregnancy of women for both her and her doctor to be appreciative if there are some kinds are issue with the growing formation. The present direction to finding the issues by using the invasive and non-invasive technique. The UAMS (University of Arkansas for Medical Sciences) has currently implements a non-invasive technique is known as the SARA (Squid Array for Reproductive Assessment) that are utilized to collect formation heartbeat data. These raw data, although, examines by a people to calculated if there is issue with a given formation. In this innovation paper, the author investigates a method to allow a computer

to calculate if a formation is in unhealthy or healthy state by the implementation of an approach that enable for frequent analysis by using data mining.

B. Zhiquan Qi et.al (2015)

In [2], the authors described Semi supervised learning (SSL) problem, these builds to utilized of huge amount of cheap unlabelled information and also some unlabelled information for training, in the past year, the data mining and machine learning most popular and lots of people work on these two technology. Utilized the MR (manifold regularization) Belkin et al. innovates a recent supervised classification machine learning algorithm: Lap SVMs (Laplacian support vector machines) and represents the state-of-the-arts efficiency in SSL filed. The modify the Lap SVMs, we innovated the Flap SVM (fast Laplacian SVM) determines for classification. Differentiates with the standard Lap SVM, ours function has some improved benefits as below: 1) the Flap SVM does not require to allocate with the extra matrix and load the execution associated to the variable switching, which create it large comfortable for huge scale problem; 2) Flap SVM's double issue has the similar classic formulation as that of standard SVMs. That means the kernel trick can be provided explicitly in to the optimization system; 3) The Flap SVM can be completely determine by successive across relaxation technology, which concentrates linearly to a solution and can procedure huge data that required not reside in memory.

C. Jayaram Raghuram et.al (2014)

In [3], the author proposed recent technique for semi supervised learning form pair same data is represented. These are main restriction of existing system; this solution cannot obtain better generation of the sample data to unconstrained data. The author solve the restriction by constraining the solution to conform to continuous class partition of the feature space, these are need details constraint generation and formation to unconstrained samples. These can obtain through a parameterized mean-field nearness to the posterior distribution across parts allotments, with the parameterization select to test the presentation power of the select mixture density family. Dissimilar multiple existing system flexibly models classes by using a variable numbers of parts, which enables it to acquire information Complex class boundaries.

Also, unlike most of the methods, evaluates the numbers of latent classes represent in the information. Practical on generated information and datasets from the UC Irvine machine learning repository display that technique obtained beneficial growth in classification performance differentiates with the existing system.

D. Anna L. Buczak et.al (2015)

In [4], this research paper describes a concentrate literature study of DM (data mining) and ML (machine learning) technique for cyber analytics in help of IDS. The short concept descriptions of ML/DM technique are supplied. Depends upon the multiple numbers of quotations or the pertinence of a new technique, in this research papers shown the each technique were identified, review and read. Because data are so significant in ML/DM systems, few some well-known cyber data sets utilized in ML/DM are explained. The complexity of ML/DM algorithms is addressed, explanation of problem for using ML/DM for cyber security is given, and several commendations on when to utilize a given technique are supplies.

E. LiuMingxia Liu et.al (2015)

In [5], the authors defined the ECOC (Error-correcting output coding) is one of the mostly utilized schemes for handle with multi-class issue by break down the original multi-class issue in to series of binary sun-problems. In traditional ECOC-based technique, the binary classifiers according to these sub-issue are normally prepared differentially without examines the association between these classifiers. Although, as these classifiers are confirmed on the similar training data, there may be few intrinsic association between them. The utilizing thus association thus association are potentially increases the formation efficiency of separate classifiers, and, thus, boost ECOC learning algorithm. In this research paper, the author traverse to mining and uses thus association through a joints classifier learning technique, by combined training of binary classifiers and the learning of the association between them. The evaluation of the innovation system, they working a series of practical on 11 datasets from the UCI machine repository and 2 datasets from actual-world image detection works. The practical outputs shown the efficiency of the innovation technique, differentiates with state-of-the-art technique for ECOC-based multi-class classification.

F. Shiliang Sun, et.al (2015)

In [6], authors proposed the Semi supervised learning has been an effective innovation concepts in ML (machine learning) and DM (data mining). The important reason is that labelling examples is costly and time-consuming, while there are huge numbers of unlabelled examples present in multiple numbers of experiments problems. So far, the Laplacian regularization has been mostly utilized in semi supervised learning. In this research paper, they innovates a recent regularization technique is known as tangent space intrinsic manifold regularization. The basic components includes in the generation of the regularization are local tangent space representations, which are evaluation by local objectives parts analysis, and the attached that associates neighbours tangent spaces. Concurrently, they utilizes its application to semi supervised classification and innovates two recent learning algorithm is known as tangent space intrinsic manifold regularized SVM and tangent space intrinsic manifold regularized twin SVMs. The author efficiently combined the tangent space intrinsic manifold regularization consideration. The practical's outputs of semi supervised classification problems display the efficient and effective innovation algorithm.

G. David M. Johnson et.al (2016)

In [7], the author describe the Metric learning is a important issue for multiple numbers of machine learning and data mining projects, and has been govern by the Mahalanobis technique. The recent improved in non-linear metric learning have shown the potential power of non-Mahalanobis distraction function, especially tree-based

functions. The author innovate a decent non-linear metric learning technique that utilized an repetitive, hierarchical variants of semi-supervised max-margin clustering to build a forest of cluster hierarchies, where each separate hierarchy can be interpreted as a weak metric over the data. By introducing randomness during hierarchy training and combining the output of many of the resulting semi-random weak hierarchy metrics, they obtain a powerful and robust nonlinear metric model. The system has main two proposes methods: first, it is semi-supervised, integrating data from unconstrained and constrained points. Second, they take a relaxed mechanism to constraint fulfilment; enable the technique to fulfil various subsets of the constraints at various levels of hierarchy rather than trying to concurrently content all of them. These organize to a robust learning algorithm. They differentiates our proposed methods to a number of state-of-art benchmarks on k-nearest neighbour classification, large-scale image retrieval and semi supervised clustering issue, and discover that our proposed improve outcomes differentiates or higher to the state-of-the-art.

H. Maximilian Nickel et.al (2016)

In [8], authors proposed Relational machine learning studies system for the graph-structured or statistical analysis of relational and data. In this innovation paper, they supply an analysis of how match statistical method are trained on huge information graph, and then utilizes to assume recent facts about the world. In specific, the author describes two basic types of statistical relational approach, both of which is scale to large datasets. The first approach based on feature approach thus as tensor factorization and multi-way neural networks.

The second is depends on mining analysis design in the graph. The author also demonstrates how to integrate these latent and analysis model to obtain increase modelling power at reduce execution cost. Lastly, they describe how thus statistical models of graph are integrated with text-based data extraction technique for automatically building information graph from web. Also describe Google's information vault application as an example of thus integration.

I. Adyan Marendra Ramadhani et.al (2017)

In [9], the author discuss about the social media immense and popularity among all the services today. Data from SNS (Social Network Service) are used for a lot of objectives such as prediction or sentiment analysis. Twitter is a SNS that has a huge data with user posting, with this significant amount of data, it has the potential of research related to text mining and could be subjected to sentiment analysis. But handling such a huge amount of unstructured data is a difficult task; machine learning is needed for handling such huge of data. Deep learning is of the machine learning method that uses the deep feed forward neural network with many hidden layers in the term of neural network with the result of the experiment about 75%.

J. Gauri D. Kalyankar et.al (2017)

In [10], authors identify and discuss the health care industries large volume of data is generating. It is necessary to collect, store and process this Data to discover knowledge from it and utilize it to take significant decisions. Diabetic Mellitus (DM) is from the Non Communicable Diseases (NCD), and lots of people are suffering from it. Now days, for developing countries such as India, DM has become a big health issue. The DM is one of the critical diseases which has long term complications associated with it and also follows with various health problems. With the help of technology, it is necessary to build a system that store and analyse the diabetic data and predict possible risks accordingly. Predictive analysis is a method that integrates various data mining techniques, machine learning algorithms and statistics that use current and past data sets to gain insight and predict future risks. In this work machine learning algorithm in Hadoop Map Reduce environment are implemented for Pima Indian diabetes data set to find out missing values in it and to discover patterns from it. This work will be able to predict types of diabetes are widespread, related future risks and according to the risk level of patient the type of treatment can be provided.

K. Jing-Jing Li, et.al (2017)

In [11], authors are focused on the Multi-label learning, it play an important work in these domain of data mining, text and multimedia machine learning. Yet, multiple numbers of multimedia models have been innovates, several of them examined to de-emphasize the affect of noisy feature in the learning procedure. To address this problem, the research paper scheme a innovation method named representative multi-label learning algorithm. Alternatively of examines all features, the innovation algorithm concentrates only on the representative ones, through integrated an kernel formulation, affinity propagation algorithm, multi-label SVM into the learning approach. Particularly, it first accepts a propagation algorithm to choose a set of representative features and capture the association between features. Then, the algorithm builds the representative kernel function to determine the comparable among data instance. Lastly, a multi-label SVM is provided to solve the learning issue. Depends on the representative multi-label learning algorithm, the author scheme a representative multi-label learning combined approach to growth the correctness, strong and stableness. The practical outputs display that the innovation algorithm works well on most of the datasets and outperforms the differentiates multi-label learning approaches are highly growth the correctness of the final system.

L. Sergio Ramírez-Gallego et.al (2017)

In [12], author presented mining massive and high-speed data streams between the important recent problems in machine learning. This calls for technique shown a greatly execution efficiency, with capability to regularly upgrade their scheme and control ever-arriving big number of instance. In this innovation paper, they represent a recent distributed and incremental based on the nearest neighbour algorithm, accepted to thus a required structure. This system, developed in Apache Spark, involves a distributed metric-space manner to perform faster searches. In contributed, they innovates an efficient and effective incremental instance selection method for huge information

stream that regularly update and deleting out-dated example from the case-base. The practical study shown that a set of real-life massive data streams explain the benefits of the innovation outcomes and display that capability to supplies the first efficient nearest neighbour solution for high-speed big and streaming data.

III. COMPARATIVE ANALYSIS

This section presents the comparative analysis in tabular form for the most recent techniques (Reported after 2007) in the form of methodology adopted, datasets used, classifiers etc.

Ref. No	Name	Publication Year	Methods
[1]	A Method For Fetal Assessment Using Data Mining and Machine Learning	2012	Arkansas for Medical Sciences (UAMS), Squid Array for Reproductive Assessment (SARA)
[2]	Successive Over relaxation for Laplacian Support Vector Machine	2015	Semi supervised learning (SSL), manifold regularization (MR), Classification, machine learning, support vector machines (SVMs).
[3]	Instance-Level Constraint-Based Semi supervised Learning With Imposed Space-Partitioning	2014	Constraint propagation, instance-level constraints, pairwise sample constraints, semi supervised learning, Space-partitioning.
[4]	A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection	2015	machine learning (ML) and data mining (DM), Cyber Analytics
[5]	Joint Binary Classifier Learning for ECOC-based Multi-class classification.	2015	Multi-class Classification, joint binary classifier and error-correcting output coding.
[6]	Semi supervised Support Vector Machines With Tangent Space Intrinsic Manifold Regularization	2015	Manifold learning, semi supervised classification, support vector machine (SVM), tangent space intrinsic manifold Regularization,twin SVM (TSVM).
[7]	Semi-Supervised Nonlinear Distance Metric Learning via Forests of Max-Margin Cluster Hierarchies	2016	Clustering, classification, and association rules, data mining, image/video retrieval, machine learning, similarity measures.
[8]	A Review of Relational Machine Learning for Knowledge Graphs	2016	Graph-based models; knowledge extraction; knowledge graphs; latent feature models; statistical relational learning
[9]	Twitter Sentiment Analysis using Deep Learning Methods	2017	Twitter, Sentiment, Deep Learning, Neural Network, Tweets
[10]	Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop	2017	Healthcare industry, Hadoop, MapReduce, Machine Learning, Predictive Analysis
[11]	A Multi-Label Learning Method Using Affinity Propagation and Support Vector Machine	2017	multi label learning, Affinity propagation, SVM, classifier ensemble.
[12]	Nearest Neighbor Classification for High-Speed Big Data Streams Using Spark	2017	Apache Spark, machine learning, big data, nearest Neighbor, data streams, distributed Computing, instance reduction.

IV. CONCLUSION AND FUTURE WORKS

In this research paper represented the significance of data mining based on machine learning and its challenges using the large scale datasets. In this research briefly described several aspects of machine learning and data mining, aiming to supplies the situation and main understanding of the concepts represented in this innovation paper. With the evaluates to data mining innovation, each and every year the innovation community addresses new open problems and new problem areas. In the future, the data mining based on machine learning to envisage intensive development and increased usage of data mining in particular area section,

thus as text and web data analysis, bioinformatics and multimedia analysis. On another side, the data mining are used for building surveillance systems recent research also concentrates on developing algorithms for mining databases without compromising sensitive information. A shift towards automated use of data mining in practical systems is also expected to become very common.

REFERENCES

[1] Wes Copeland, Chia-Chu Chiang, "A Method For Fetal Assessment Using Data Mining and Machine Learning" 978-1-4673-2588-2/12/$31.00 ©2012 IEEE.
[2] Zhiquan Qi, Yingjie Tian, and Yong Shi, "Successive Over relaxation for Laplacian Support Vector Machine", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 26, NO. 4, APRIL 2015.
[3] Jayaram Raghuram, David J. Miller, and George Kesidis, "Instance-Level Constraint-Based Semisupervised Learning with Imposed Space-Partitioning", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 25, NO. 8, AUGUST 2014.
[4] Anna L. Buczak*, Member IEEE, Erhan Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", DOI 10.1109/COMST.2015.2494502, IEEE Communications Surveys & Tutorials.
[5] Mingxia Liu, Daoqiang Zhang*, Songcan Chen,andHui Xue, "Joint BinaryClassifier Learning for ECOC-based Multi-class Classification", DOI 10.1109/TPAMI.2015.2430325, IEEE Transactions on Pattern Analysis and Machine Intelligence.
[6] Shiliang Sun and Xijiong Xie," Semisupervised Support Vector Machines With Tangent Space Intrinsic Manifold Regularization", 2162-237X © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
[7] David M. Johnson, Caiming Xiong, and Jason J. Corso, "Semi-Supervised Nonlinear Distance Metric Learning via Forests of Max-Margin Cluster Hierarchies", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 28, NO. 4, APRIL 2016.
[8] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich, "A Review of Relational Machine Learning for Knowledge Graphs", Vol. 104, No. 1, January 2016 | Proceedings of the IEEE.
[9] Adyan Marendra Ramadhani, Hong Soon Goo, "Twitter Sentiment Analysis uses Deep Learning Methods", 2017 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia.
[10] Gauri D. Kalyankar, Shivananda R. Poojara, Nagaraj V. Dharwadkar, "Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop" International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017).
[11] Jing-Jing Li, Farrikh Alzami, Yue-Jiao Gong, and Zhiwen Yu, "A Multi-Label Learning Method Using Affinity Propagation and Support Vector Machine" DOI 10.1109/ACCESS.2017.2676761, IEEE Access.
[12] Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michał Wo´zniak, José Manuel Benítez, and Francisco Herrera, "Nearest Neighbor Classification for High-Speed Big Data Streams Using Spark", 2168-2216 _c 2017 IEEE.

