MACHINE LEARNING ALGORITHMS FOR HETEROGENEOUS DATA: A COMPARATIVE STUDY

Dr. Poornima Nataraja

Research Supervisor, Department of MCA - VTU, DSCE, Bangalore

Bharathi Ramesh

Assistant Professor, Department of MCA, Surana College, Bangalore

ABSTRACT

In the present digital era massive amount of data is being continuously generated at exceptional and increasing scales. This data has become an important and indispensable part of every economy, industry, organization, business and individual. Further handling of these large datasets due to the heterogeneity in their formats is one of the major challenge. There is a need for efficient data processing techniques to handle the heterogeneous data and also to meet the computational requirements to process this huge volume of data. The objective of this paper is to review, describe and reflect on heterogeneous data with its complexity in processing, and also the use of machine learning algorithms which plays a major role in data analytics.

Keywords: Big data, homogeneous data, heterogeneous data, MapReduce, Pandas, SVM, ANN, BN.

Cite this Article: Dr. Poornima Nataraja and Bharathi Ramesh, Machine Learning Algorithms for Heterogeneous Data: A Comparative Study, *International Journal of Computer Engineering and Technology*, 10(3), 2019, pp. 9-19.

<http://iaeme.com/Home/issue/IJCET?Volume=10&Issue=3>

1. INTRODUCTION

Data are collected and also analyzed to convert into information which is suitable for making decisions. Hence this data provide a rich resource for knowledge discovery and decision support [1]. A database is an organized collection of data which can easily be accessed, stored, managed, and updated. Also this data has interesting knowledge such as patterns, anomalies and associations and considerable structures in large databases, data warehouses or other repositories. The large amount of data in recent years has been increasing drastically with the rapid development of internet. This type of data not only has high-speed characteristics, but also has the diversity and variability [2]. A recent study estimated the data statistics that every minute, Google receives over 4 million queries, e-mail users send more than 200 million messages, YouTube users upload nearly 100 hours of video, face book users

share over 2 million pieces of content and Twitter users generate more than 277,000 tweets [3]. Structured data, constitutes less than 5% of all existing data, which refers to the data in tabular form present in relational databases or spreadsheets.

1.1. Homogeneous Data and Heterogeneous Data

Homogeneity and heterogeneity are the two main concepts with respect to the uniformity in a dataset frequently used in all sciences and statistics. A homogeneous dataset is made up of things that are similar to each other, that mean it is uniform in character or composition (i.e. color, size, weight, height, shape, distribution, texture, architectural design, etc.); the other one which is heterogeneous is manifestly no uniform in any one of these qualities.

2. OBJECTIVE OF THE PROPOSED RESEARCH

- The main aim of this study is to explore a comparative study on the performance issues of few existing Supervised Machine Learning Algorithms with the efficiency factors
- Proposed hybrid algorithm mainly focuses on transforming massive data into valuable knowledge by identifying the hidden patterns of the heterogeneous data

3. LITERATURE SURVEY

The heterogeneous data is becoming more prevalent, which constitutes 95% of big data [4]. A variety of traditional data models and different query languages are not sufficient to process these, since data is often irregular; some data is missing; heterogeneity in datasets [5].

Data challenges related to the heterogeneous data (i.e. diverse and dissimilar forms) is a major challenge. Many researchers have concluded that the huge volume of data which is not consistent and it does not follow a specific template or format – it is collected in different forms and diverse sources. [6]

The following diagram refers to the different forms of data which is heterogeneous and it is a major challenge to process and manage such data. [7].

Figure 1 Data variety/heterogeneous data

3.1. Challenges in Heterogeneous data analysis

As most of the heterogeneous data collection is customized for a particular format, which is a major limitation associated with heterogeneous data analysis [8].

The following are few important challenges in heterogeneous data analysis:

Figure 2 Challenges in heterogeneous data analysis

3.2. Examples of Heterogeneous Data

3.2.1. Unstructured data in Social media channels

Social media is a wide area which combines a variety of online platforms and allows users to create and exchange heterogeneous content. Social media has different categories such as social networks, micro blogs, social news, media sharing, social bookmarking, wikis, review sites and etc [9]. Examples for unstructured data are the data posted by users on social media platforms like customer feedback, product reviews, images, and videos, text etc.

The social media landscape is enormous and changing every second as it deals with heterogeneous data. There is a need of different data processing technologies to address the data processing challenges. [10]

3.2.2. Heterogeneous data in Healthcare

The healthcare industry has huge volume of data nowadays. Healthcare related databases stores information like patient's records [11]. Data is available from diverse sources for example, medical reports in different form. Further, raw medical data is dissimilar in nature and it may be collected from different sources like, images, videos, interviews with the patient, X-ray, computed tomography scans (CT), magnetic resonance images (MRI), ultrasound, laboratory data, physician's observations and evaluations, etc [12]. In today's digital world, it is mandatory that these data should be digitized and analyzed for further decision making purpose [13].

4. METHODOLOGY

Figure 3 Methodology of the proposed research

The main aim of this proposed research is to use an hybrid model which uses different formats of datasets for training and testing. The overall methodology is divided into two tasks;

i) First one is to collect the datasets in heterogeneous form and represent it using relevant preprocessing tool. In order to deal with large quantity of heterogeneous data, input need to be collected, indexed, stored, retrieved, analyzed and also further to be mined to allow a simple and regular access to these data and information for further analysis [14]. Various approaches and techniques are used. Two renowned approaches which can be used to represent heterogeneous data are:

- **Pandas (Panel data):** Pandas is an open-source Library in Python which provides relevant way to manipulate and analyze data using powerful data structures. It is used as a tool for loading data into memory from different file formats
- **MapReduce:** It is a fault-tolerant framework for data processing that enables its users to process the massive amount of unstructured or semi-structured data [15]. By using this tool, data can be collected, indexed, stored, retrieved, and analyzed to mine the relevant data for a simple and continuous access [16].

ii) Second task is to apply hybrid machine learning algorithm in order to identify and classify the heterogeneous data for further use.

5. MACHINE LEARNING ALGORITHMS FOR HETEROGENEOUS DATA

Machine Learning is an interdisciplinary field which has concealed more or less every scientific domain [17]. Artificial intelligence, optimal control, cognitive science, statistics, information theory, optimization theory, and many other domains of mathematics, engineering, and science are few fields in which machine learning has an extensive variety of

applications [18, 19]. Machine learning algorithms play a major role in heterogeneous data analysis. Few are described below:

Figure 4: Methodology of the proposed research

Supervised learning techniques are basically categorized as regression and classification problems. In a regression problem, input variables are mapped to continuous output function whereas in a classification problem, input variables are mapped to discrete categories. Few supervised learning techniques:

5.1. Support Vector Machines

A support vector machine (SVM) is a popular algorithm for classification and regression problems in the large datasets [20]. Especially it is suitable for processing heterogeneous data. This has been applied to tasks such as handwritten digit recognition, object recognition, and text classification, image analysis, etc.

SVM basically used to identify two different classes in a heterogeneous or multidimensional environment to extract the condensed data set [21]. This learning technique uses a huge number of features without much computation. SVM splits the dataset into two vector sets under 'n' dimensional space vector. Also defines some separating hyper planes, (also known as the decision boundaries) that partition the labeled training data into a pre-defined number of classes. SVM is a best method to handle data dimensionality in large datasets [22, 23].

Figure 5 Classification of datasets in SVM

Many researchers have proved the application of SVM in pattern recognition problems. Image classification is one of the major problem of concern in image processing used to predict the categories of the input using its features [24]. Especially in advanced healthcare the SVM is well used algorithm in image data processing like cancer detection [25], different disease diagnosis, extracting hidden features in image data etc.

5.2. Naïve Bayes Network

Bayesian network is a type of network which is used to represent knowledge about uncertain domain. It belongs to the domain of probabilistic graphical models (GMs). In this approach nodes represent the variables and edges represent probabilistic dependencies among those variables [26, 27]. There are several works using Bayesian Network. The majority of them involve feature extraction, image classification, text recognition, classification and retrieval [28]. A Bayesian network model for image processing is the model of extracting the features in order to improve the precision in content-based image retrieval systems. This approach allows retrieving images according to their features [29].

5.3. Artificial Neural Networks

ANNs handle a variety of classification or pattern recognition problems. They are trained to generate an output as a combination between the input variables. Multiple hidden layers that represent the neural connections mathematically are used in this process. Even though ANN used as a standard algorithm in several classification tasks [30], they too suffer from few drawbacks. Its layered structure proves to be very time-consuming lead to very poor performance. Additionally, this specific technique is characterized as a “black-box” technology.

These neural networks use hidden layers with the help of which it solves the classification problem for non linear sets [31]. Following table depicts the brief survey of existing application domains of machine learning algorithms:

Table 1: Application areas of ML algorithms

Algorithm	Application domain	Representative References
Support Vector Machine	Pattern recognition, Image classification, Cancer disease detection, text processing, video analysis, social media data processing, different diagnosis by processing other medical records.	[32], [33], [34], [35], [36], [37], [38]
Artificial Neural Networks	Image Processing, character recognition, Forecasting, text classification, speech recognition, medical data analysis, E-mail Classification, social media data processing etc.	[39], [40], [41], [42], [43], [44], [45]
Bayesian networks	Pattern recognition, Image classification, text processing, video analysis, social media data processing, etc.	[46], [47] [48], [49], [50], [51]

6. OBSERVATIONS & DISCUSSIONS

Big data has grown nine times in volume in just 5 years and its amount in the world will reach 35 trillion gigabytes by 2020 [52]. It is evident that big data is voluminous and heterogeneous. It may be collected in different forms like images, numbers, text, symbols, signals, audio, video, etc. It is worth to mention here that the heterogeneous data collected, requires lot of storage space and needs efficient tools to analyse the data. In fact, un-stored and un-organized data are considered less pragmatic in any domain like social media, medical, business, etc [53].

This paper presents a review on various machine learning algorithms necessary for handling heterogeneous datasets in big data.

According to the survey in 2011[54], machine learning based big data Processing has gained popularity and new developments are on the rise for efficient data processing. Machine learning is used to solve variety of problems. If taken an example of large volume and heterogeneous nature of medical databases; it is not possible that any tool or algorithm can succeed with raw and unorganized data [55]. The data may be collected in various forms like images, physician's observations, interviews with patients, laboratory data, etc. All these help in diagnosis and prognosis of diseases and to maintain the patient's records. [56]. similarly the data from different domain may be in multiple format and many machine learning techniques are used to analyze and organize data. However, this paper mainly focused on these techniques namely; neural networks, support vector machines and Bayesian networks. Each technique has its own advantages and associated disadvantages.

It is hard to analyze that which method is best. Indeed, different techniques in different scenario perform best while the same technique may performs worst in other application with the same set of datasets. The performance of techniques varies from one dataset to other dataset. A comparative study is shown in the below table:

Table 2: Comparative study on SVM, ANN & BNs

ML Algorithm	Advantage	Disadvantages
Support Vector Machine	<p>It is effective in high dimensional spaces and large heterogeneous datasets. Provides better accuracy in comparison to other classifiers.</p> <p>It is more effective in situations where the number of dimensions is greater than the total number of samples.</p> <p>It easily handles complex non linear data points and over fitting is not a problem like in other cases.</p> <p>It is memory efficient because it uses a subset of training sets in support vectors.</p> <p>It is versatile because different kernel functions can be specified for the decision functions.</p>	<p>It gives poor performance when number of samples are more</p> <p>It is computationally expensive and even the training process is time consuming compared to other methods.</p> <p>Selection of right kernel function is difficult because for every dataset different kernel function produce different results.</p> <p>SVM used to solve the problems of binary class. Thus, it solves problem of multi class by breaking it into pair of two classes to which data needs to be properly represented.</p> <p>It does not provide probability estimates directly. These are calculated using a cross validation method.</p>
Neural Networks	<p>It can handle noisy data properly for training.</p> <p>It is capable of obtaining complex relationships between input and output. Without any external help, it can analyse and organize data based</p>	<p>It is not suitable for data with more number of input features with different forms and for complex problems.</p> <p>It is difficult to understand the model and requires high processing time when dealing</p>

	on its own features. Various neural networks can be used for clustering and classifying data which is heterogeneous.	with heterogeneous data.
Bayesian networks	It is fast and accurate for medium number of datasets It requires short computational time for training and very easy to construct. No need of any complicated iterative parameter estimation schemes so it can be applied to large data set. Easy way to interpret knowledge representation It performs well and robust.	It does not give accurate results in the case of dependency among variables in the heterogeneous datasets. When data is scarce, BN learning is inaccurate Theoretically, naïve bayes classifier has minimum error rate comparing to other classifier.

Considerations when choosing an algorithm

- Accuracy
- Training time
- Linearity
- Number of parameters
- Number of features

6.1. Summary

The above table shows the comparative study of some commonly used classification techniques from the existing evidence and theoretical studies [1, 32].

ANN & SVM are easy to design and deploy for the specific classification problem. The precision is high, but the performance of processing time needs to improve, especially in complex problems like image classification such as facial image, medical related image. [39] The training time of ANN & SVM is also a problem in the large dataset.

Many researchers proved that the accuracy of ANN and SVM model decreases when the number of classes increases. BNs-based model usually falls into local minimums, which may further generate inaccurate results for the heterogeneous data processing. Thus, we can use a model-based feature extractor combined with a features of discriminative classifier like SVM to overcome the above issue, which is theoretically proved to have better performance.

Bayesian networks have proven as a best model to handle uncertainty in heterogeneous data and supporting decision making in practice. However, in many applications it is hard to obtain sufficient results using BNs. For example in medical data, in a small hospital there may be insufficient data to learn an effective medical diagnosis network. However, directly applying a network learned in another domain may be inaccurate or impossible because the underlying tasks may have quantitative or qualitative differences.

Among all machine learning algorithms SVM method is one of the best classification algorithm due to its ability of minimizing the empirical classification error and calculating the margin classification space. If it can combine with additional features of other methods can offer effective and efficient solutions to learn heterogeneous data.

7. CONCLUSION

Heterogeneous data/mixture analysis technology is expected to play a significant role in almost all the domains. It is an advanced technology used in big data analysis. Machine learning algorithms play a major role in this regard. The above comparative study describe that, each algorithm has its own set of advantages and disadvantages, as well as its own area of implementation. None of the algorithm can be used to satisfy all the criteria. Integration of two or more algorithms by combining their strength would be more useful for processing heterogeneous data analysis.

REFERENCES

- [1] Thair N. Phyu, "Survey of Classification techniques in Data Mining", in International Multi conference of Engineers and Computer Scientists, Hong Kong, 2009.
- [2] Yun Liu, Qi Wang and Hai-Qiang Chen "Research on IT Architecture of Heterogeneous Big Data" in Journal of Applied Science and Engineering, Vol. 18, No. 2, pp. 135_142 (2015)
- [3] Dietrich, D., Heller, B., Yang, B.: Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. Wiley, Hoboken (2015)
- [4] Junfei Qiu, Qihui Wu, Guoru Ding*, Yuhua Xu and Shuo Feng "A survey of machine learning for big data processing" EURASIP Journal on Advances in Signal Processing (2016) 2016:67
- [6] Abiteboul, Serge, et al. "The Lorel query language for semi structured data." International journal on digital libra
- [7] ries1.1 (1997): 68-88.
- [8] S.D.Gheware, A.S.Kejkar, S.M.Tondare, "Data Mining: Task, Tools, Techniques and Applications", International Journal of Advanced Research Computer and Communication Engineering Vol.3, Issue 10, October 2014.
- [9] Yongjian Fu "Data Mining : Tasks, Techniques and Applications" <http://academic.csuohio.edu/fuy/Pub/pot97.pdf>
- [10] Fujimaki Ryohei, Morinaga Satoshi "The Most Advanced Data Mining of the Big Data Era" in NEC Technical Journal Vol. 7 No.2/2012
- [11] C.C. Aggarwal "An introduction to social network data analytics" C.C. Aggarwal (Ed.), Social network data analytics, Springer, United States (2011), pp. 1-15
- [12] Weiguo Fan, Michael D. Gordon, "The Power of Social Media Analytics", in Communica
- [13] tions of the ACM 57(6):74-81 · June 2014
- [14] Rallapalli, Sreekanth.,& Gondkar, "Map reduce programming for electronic medical
- [15] records data analysis on cloud using apache hadoop, hive and sqoop". International Journal of Latest Technology in Engineering, Management & Applied Science, R. R. (2015) 4(8), 73-76.
- [16] Krzysztof J. Cios, G.William Moore, Uniqueness of medical data mining, Artificial Intelli
- [17] gence in Medicine 26, 1–24, 2002.
- [18] S. Mitra, S.K.Pal&Mitra , P., "Data mining in soft computing framework: A survey, IEEE
- [19] transactions on neural networks", 13(1), 3-14,2002.
- [20] Mohamed, H., Marchand-Maillet, "MRO-MPI: MapRe-duce overlapping using MPI and an
- [21] optimized data exchange policy", Parallel Comput. 39, 851–866 (2013)
- [22] P. Zadrozny and R. Kodali, "Big Data Analytics using Splunk, Berkeley, CA, USA: Apress,

- [23] 2013.
- [24] F. Li, B. C. Ooi, M. T. Özsu and S. Wu, "Distributed data management using MapReduce,"
- [25] ACM Computing Surveys, 46(3), pp. 1-42, 2014.
- [26] C. Rudin and K. L. Wagstaff, "Machine learning for science and society," Mach. Learn., vol. 95, no. 1, pp. 1-9, 2014.
- [27] Russell, S., Norvig, P., & Intelligence, A. "A modern approach. Artificial Intelligence", Prentice-Hall, Englewood Cliffs, 25, 27 (1995).
- [28] Mitchell, T. M. (2006). "The discipline of machine learning" (Vol. 3). Carnegie Mellon University, School of Computer Science, Machine Learning Department.
- [29] V. N. Vapnik, "The Nature of Statistical Learning Theory". Springer Verlag, 1995.
- [30] N. Chistianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines, and other kernel- based learning methods", Cambridge University Press, 2000.
- [31] Boser, B. E., I. Guyon, and V. Vapnik (1992), "A training algorithm for optimal margin classifiers" in Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages. 144 -152. ACM Press 1992.
- [32] Hao Jiang, Wai-Ki Ching, Zeyu Zheng, "Kernel Techniques in Support Vector Machines for Classification of Biological Data", IJITCS, 2011, Vol.3, No.2, pp.1-8.
- [33] Nasser H. Sweilam a,* , A.A. Tharwat b, N.K. Abdel Moniem "Support vector machine for diagnosis cancer disease: A comparative study" Egyptian Informatics Journal (2010) 11, 81– 92
- [34] Friedman N., Geiger, D. Goldszmidt M, "Bayesian network classifiers". Machine Learning 29: pp. 131- 163, 1997.
- [35] Finn V. Jensen, An Introduction to Bayesian Networks, Springer, New York, 1996.
- [36] Jensen, —An Introduction to Bayesian Networks, Springer, 1996.
- [37] P.S. Rodrigues and A.A. Araujo, "A bayesian network model combining color, shape and texture information to improve content based image retrieval systems", 2004
- [38] Q. Zhang and I. Ebroul, "A bayesian network approach to multi feature based image retrieval", First International Conference on Semantic and Digital Media Technologies, GRECE, 2006.
- [39] Divya Tomar and Sonali Agarwal , A survey on data mining approaches for healthcare, International Journal of Bio-Science and Bio-Technology Vol.No.5, pp. 241-266, 2013.
- [40] Haykin. S, "Neural Networks: A Comprehensive Foundation", Prentice Hall, 1999.
- [41] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Elsevier, 2011.
- [42] Seyed-Ali Bahrainian ,Andreas Dengel, "Sentiment Analysis using Sentiment Features" ,IEEE Computer Society ,Volume, Issue No. : 2902-3/13, pp-26-29, 2013
- [43] Xiaohui Yu, Yang Liu, Aijun An" An Adaptive Model for Probabilistic Sentiment Analysis", IEEE Computer Society ,Volume, Issue No. : 4191-4/10, pp-661-667, November 2010.
- [44] Noriaki Kawamae "Hierarchical Approach to Sentiment Analysis", IEEE Computer Society, Volume, Issue No. : 4859-3/12, pp- 138-145, 2012.
- [45] D. Lu, Q. WENG, A survey of image classification methods and techniques for improving classification performance, International Journal of Remote Sensing, 2007, Vol. 28, No. 5, pp.823-870.
- [46] Qing Chen, Real-time Vision-based Hand Gesture Recognition Using Haar- like Features, Instrumentation and Measurement Technology Conference Proceedings, 2007. IMTC 2007. IEEE, 2007, pp.1-6
- [47] Noriaki Kawamae "Hierarchical Approach to Sentiment Analysis", IEEE Computer Society, Volume, Issue No. : 4859-3/12, pp- 138-145, 2012.

- [48] Lashari, S. A., & Ibrahim, R. (2013), "A Framework for Medical Images Classification Using Soft Set", *Procedia Technology*, 11, (2013). 548-556
- [49] Chatap, N. J., & Shrivastava, A. K. (2014). "A Survey on Various Classification Techniques for Medical Image Data" *International Journal of Computer Applications*, 97(15).
- [50] Guoqiang Zhang, B. Eddy Patuwo, Michael Y. Hu, "Forecasting with artificial neural networks: The state of the art" *International Journal of Forecasting* 14 (1998) 35–62
- [51] Abirimi S., Neelamegam P., Kala H., Analysis of Rice Granules using Image Processing and Neural Network Pattern Recognition Tool, *International Journal of Computer Applications*, 2014.
- [52] Brause R., Hamker F., Paetz J.: *Septic Shock Diagnosis By Neural Networks And Rule Based Systems*; in: L.C. Jain: *Computational Intelligence Techniques In Medical Diagnosis And Prognosis*, Springer Verlag 2001, in press
- [53] Brause R., Hanisch E. (Eds.) (2000), *Medical Data Analysis ISMDA 2000*. Springer Lecture Notes in Comp.Sc., LNCS 1933, Springer Verlag, Heidelberg
- [54] Ahan M R, Honnesh Rohmetra , Ayush Mungad, "Social Network Analysis using Data Segmentation and Neural Networks", *International Research Journal of Engineering and Technology (IRJET)* Volume: 05 Issue: 06 | June-2018.
- [55] Khlifia Jayech and Mohamed Ali Mahjoub, "Clustering and Bayesian network for image of faces classification", (IJACSA) *International Journal of Advanced Computer Science and Applications*
- [56] Zhang Q., Ebroul I. "A bayesian network approach to multi based image retrieval". First *International Conference on Semantic and Digital Media Technologies*. GRECE, 2006
- [57] Jayech K , Mahjoub M.A "New approach using Bayesian Network to improve content based image classification systems", *International Journal of Computer Science Issues*, Vo7, Issue 6, November 2010.
- [58] A.V. Nefian, L. Liang, X. Liu, and K. Murphy, *Dynamic Bayesian Networks for Audio-Visual Speech Recognition*," *EURASIP Journal on Applied Signal Processing*, Nov 2002.
- [59] S. Gole, B. Tidke, A survey of Big Data in social media using data mining techniques, in: 2015 *Int. Conf. Adv. Comput. Commun. Syst. (ICACCS -2015)*, 2015: pp. 1–5. doi:10.1109/ICACCS.2015.7324059.
- [60] Schoen, H., Gayo-Avello, D., Metaxas, P.T., Mustafaraj, E., Strohmaier, M. and Gloor, P. (2013), "The power of prediction with social media", *Internet Research*, Vol. 23 No. 5, pp. 528-543
- [61] CM Bishop, *Pattern recognition and machine learning* (Springer, New York, 2006)
- [62] D Che, M Safran, Z Peng, From big data to big data mining: challenges, issues, and opportunities, in *Proceedings of the 18th International Conference on DASFAA (Wuhan, 2013)*, pp. 1–15
- [63] A Sandryhaila, JMF Moura, Big data analysis with signal processing on graphs: representation and processing of massive data sets with irregular structure. *IEEE Signal Proc Mag* 31(5), 80–90 (2014).
- [64] Zupan B., Halter J.A and Bohanec M., "Qualitative model approach to computer assisted reasoning in physiology", In *Proceedings of Intelligent Data Analysis in Medicine and Pharmacology-IDAMAP98*, Brighton, UK, 1998.
- [65] Saul J M, "Legal policy and security issues in the handling of medical data", In: Cios KJ, editor. *Medical data mining and knowledge discovery*. Heidelberg: Springer, pp. 17–31 [chapter 2], 2000.