Survey on Machine Learning Algorithm for Reducing Data Storage System

Miss. Amruta

Student, Department of Computer Engineering, Bharati Vidyapeeth Deemed University College of Engineering, Pune, 411043, Maharashtra, India.

Abstract — *Nowadays few computers have shown to be particular helpful for arithmetic and logic implementations of data, their accuracy and efficiency for applications such as e.g. face, object and speech recognition, are not that impressive, especially when compared to what the human brain can do. In this paper, Machine learning algorithms have been useful, especially for these types of applications, so they operate in a similar way to the human brain, by learning the data provided and storing it for future recognition of analysis of data. Now, there has been a strong focus on increasing the process of data storage and retrieval of data, only neglecting the value of the provided information and the amount of data required to store for management of data storage system. Hence, this paper examines the chance to reduce data storage through the use of separation and combine it with a presented similarity detection algorithm in machine learning. The Principal Component Analysis (PCA) /Fisher Linear Discrimination analysis is used to reduce the dimensionality of the feature vector.*

Key words: **Machine Learning, Data Storage Efficiency, Principle Component Analysis (PCA), Pattern Recognition etc.**

I. INTRODUCTION

In this system, Human beings are a very imaginative thinker which helps to overcome obstacles; though their operational capacities are controlled in time. For example, humans are only able to work for a limited number of hours per day, even as machines can be programmed to operate continuously, and this is where machines have confirmed mostly useful to us, humans and machines to control the program **[1].** In this method we investigates the task to design a suitable classification system is difficult as it is not even known if a desired accuracy performance can be obtained with the given sensor data. Initial, it is not clear which texture or shape features make a distinction possible. Secondly, a suitable classifier concept with adequate parameters has to be chosen to achieve good recognition rates and generalization abilities. Additionally, dimension reduction techniques such as Principal Component Analysis may increase the performance of the system **[1] [2].**

The various machine learning methods are divided into supervised and unsupervised machine learning. Of course, there is also individual QOS quality of service features for classification. This paper has compared with the dissimilar identification method based on supervised and unsupervised machine learning. Further it has analyzed the element which has effect on the supervised machine learning **[3].**Normally, these types of algorithms can be classified as either supervised or unsupervised learning methods. In supervised learning **[3]**, the user stores the information in the machine to calculate the output values based on earlier experiences.

This paper presents a proposal of ECG analysis, which determines a spontaneous termination of a trial fibrillation prediction. Supervised neural networks are trained to develop this task, where a comparison is carried out between multilayer perception (MLP) and supervised self organized maps (SOM). Principal component analysis (PCA) is implemented to reduce the input dimensionality.PCA is a standard tool in modern data analysis because it is a simple, non-parametric method for extracting relevant information from confusing data sets **[4].** The machine does not store any information that is not used and hence requires less storage space to be able to predict the output value.

The selected algorithm of machine learning also influences the storage principles of data used during training as well as the actual recognition phase of features. These algorithms are obviously challenged through the fact that they need to recognize e.g. the same face under various different lighting conditions, make-up, facial expressions, etc. In order to deal with

these challenges, one either has to restrict oneself towards specific features, and add as much contextual information on top of that to deal with these various environmental factors. The large amount of information being stored, does not only affect the actual data storage requirements, but also the processing, as during the recognition, one desires to work through all of this information to identify possible similarity before a final decision can be made. Though, presently, there is a limited amount of work with regards to reducing the amount of data being stored, which is the mainly focus on this paper.

In reducing the required amount of data storage, one could clearly recover the machine learning algorithm, although it would be necessary to make a superior judgment on which information is helpful, and which is not. The results of this system for different stages of the PCA, as well as the grouping with the machine learning algorithm will then be explored, after that which the paper will achieve the implementation of this storage requirement and current a few future work. In this survey, Section II gives the Literature survey for data storage requirement and also listed the different methods

used in this survey of machine learning system of data management.

II. LITERATURE SURVEY

This author **[1] Saritha Kinkiri & Wim J.C. Melis** have proposed the best methodology of Machine learning algorithms has been valuable, particularly for these type of applications, as they work in a similar way to the human brain, by learning the data provided and storing it for pattern recognition in future. Therefore, this paper examines the better chance to reduce data storage during the use of separation and merge it with an implementing similarity of detection algorithm. The separation is achieved through the use of, Principal Component Analysis (PCA), which not only reduces the data storage requirements. Usually, these algorithms can be divided as either supervised or unsupervised. The machine does not store any data that is not used and hence requires less storage space to be able to predict the output value. In order to reduce the storage requirements, and due to the importance of data being context dependent, it is important to be able to derive context as if one can identify context then it becomes possible to store the data with regards to the context in which you operate. This would then result in a reduction in data storage requirements, because the context is stored as generic information, while for each item only the deviation from this "central" context needs to be stored. The final outcome of this system shows that the data storage requirement and detection accuracy improves which depends on data management system.

In this paper, the author proposes **[2] Fabian B¨urger1, Christoph Buck2, Josef Pauli1 and Wolfram Luther2** methodology of the optimization process of an object classification task for an image-based steel quality measurement system. The goal is to distinguish hollow from solid defects inside of steel samples by using texture and shape features of reconstructed 3D objects. In order to optimize the classification results we propose a holistic machine learning framework .The framework consists of three layers, namely feature subset selection, feature transform and classifier which subsequently reduce the data dimensionality. In real world machine learning tasks it is usually the case that any combination of the aforementioned problems can occur. It is time-consuming manual work to evaluate all possible combinations of solutions to achieve the optimal classifier performance. These observations motivate a holistic view on machine learning with an automatic optimization process of the components. In this survey, there are two categories of approaches in the context of classifier framework optimization, namely search-based and meta learning algorithms. In search algorithms, a classifier system is optimized by trying and evaluating a set of the system's hyper-parameters. Usually, the final classifier accuracy based on the training data is used as the objective function. Different search strategies and framework components have incorporated within this optimization process. Furthermore, an additional feature selection component has been incorporated using genetic algorithms, particle swarms and simulated annealing. A larger framework for biomedical spectra classification is proposed that incorporates data visualization, preprocessing, feature extraction and selection, classifier development and aggregation. The authors use several strategies to optimize the hyper-parameters and configurations. However, their approach is focused especially on the development of spectral features and interpretability in the diagnostics field**.[2]**

This author **[3] DONG Shi, ZHOU DingDing, DING Wei** introduced a method of Network traffic identification. Network traffic identification is an important application research direction for network management and measure, the current network traffic identification methods roughly can be classified into four categories.(1)port based method;(2)DPI(Deep packets inspection);(3)host behavior method;(4)flow-based method based on machine learning. It has become a hot research between domestic and foreign experts who take the traffic identification as research direction, which proceed distinguish, QOS, intrusion detection, traffic monitoring, billing and management. From the beginning of the study port-based method, this method used well-known port numbers to identify Internet traffic. This technique has been shown to be ineffective for some applications such as the current generation Of P2P applications. So Payload-based Analysis technology was proposed to overcome the shortcoming of port-based, which adopts method based on deep packet detection methods, but this method has still drawbacks that it can't cope with some encrypted traffic and can't obtain the new service type. Recently traffic identification and classification have new method with a number of new applications and service increasing, Machine learning methods have been applied to the traffic identification. This paper has studied and analyzed the machine learning algorithm for network traffic identification and mainly studied unsupervised and supervised machine learning.

The Author **[4] Germ´an E. Melo A, Ricardo A. Osorio M, Alvaro D. Orjuela C.** presents a proposal of ECG analysis, which determines a spontaneous termination of atrial fibrillation Prediction. Supervised neural networks are trained to develop this task, where a comparison is carried out between multilayer perceptron (MLP) and supervised self organized maps (SOM). Principal component analysis (PCA) is implemented to reduce the input dimensionality. Results show maximum classification rates of 100% for MLP in the cases without and with PCA. For SOM the maximum classification rates are in 65% and 75% for case without and with PCA, respectively. That is why effective therapeutic strategies against AF will increase as the elderly population grows .Medical evidence suggests that an episode of atrial fibrillation with spontaneous termination can produce a chronic AF in the future. Because of, the identification of the moment in which the episode will happen is an issue of interest in health's community. Considerations for design and parametric analysis results of classifiers based on supervised neural networks, which predict spontaneous termination of atrial

fibrillation, were established. The results indicate the MLP has an unstable statistical behavior than the supervised SOM, since the MLP obtains optimal hit rates (100%) in certain experiments. The supervised SOM statistical behavior gets more stable but less accurate, since the best hit rate does not exceed 75%. This is seen in the mean and standard deviation of the experiments, showing that the standard deviation for the Supervised SOM is much lower than that found for the MLP.

The Author **[5] Mrs.N.G.Chitaliya and Prof.A.I.Trivedi** suggested the method simply with the increasing demands of visual surveillance systems, vehicle & people identification at a distance has gained more attention for the researchers recently. Extraction of Information from images and image sequences are vary important for the analysis according to the application. This research proposes feature extraction and classification method using Wavelet. The DWT is used to generate the feature images from individual wavelet sub bands. The feature images constructed from Wavelet Coefficients are used as a feature vector for the further process. The Principal Component Analysis (PCA) /Fisher Linear Discrimination analysis is used to reduce the dimensionality of the feature vector. Reduced feature vector are used for further classification using Euclidian distance classifier and neural network Classifier. Vehicle detection and classification is always a complicated and uncertain area within mobile robotics as well as any traffic surveillance system due to the interference of illumination and blurriness. Moreover, the performance requirements are no longer left in a prototype works at research lab but exposed to the real world problems. This demand makes the task greatly challenging for requirements of speed and accuracy. In our system, deriving best feature vectors of object is particularly of interest for Mobile Robotics application as well as visual surveillance system. There are three core objectives that make our approach innovative and effective; they are accuracy in terms of classification, speed for performing real time application and producing desired results related to the position of vehicle.

This author **[6] MING-JING YANG, HUI-RU ZHENG, HAI-YING WANG, SALLY MCCLEAN, NIGEL HARRIS** have proposed the best methodology of Feature reduction has been generally recognized as an effective approach to improve the performance of classification problems. There are several reasons for feature reduction. High dimensionality of data will lead to high complexity of the classifier. The mutual correlation between the features causes the prediction power of combination of features lower than expected. Moreover, for a limited number of training samples, keeping the number of features small is one of the key points in designing a classifier with good generalization performance. In this paper, a novel feature reduction method was proposed. A hybrid approach of feature ranking and Feature generation combined with multilayer perceptron (MLP) neural-networks was applied in two gait datasets (i.e. footswitch gait and accelerometer gait datasets). The result of the proposed method was compared with the classification results of feature ranking and PCA employed separately. The hybrid approach achieved the best performance. Feature reduction is important in gait analysis using machine learning methods. The hybrid approach of feature ranking and feature generation has the best performance. It can use the small feature subset to achieve best classification performance when it was carried out in footswitch dataset.

In this paper, **[7] W.hua, M.cuiqin and Z. Lijun** the author presented the effective method of Machine learning is the core problem of artificial intelligence data storage research, this paper suggests the definition of machine learning and its basic structure, and describes an different type of machine learning methods, including rote learning, inductive learning, analogy learning, explained learning, learning based on neural network. Machine learning compared with human learning, machine learning learns faster, the accumulation of knowledge is more facilitate the results of learning spread easier. Learning is the process that processes the outside information to knowledge; first it obtains the information of outside environment and then processes the information to knowledge.
There are two ways to determine the value of neural network: one is determined through the design calculations; another is determined by the study of network through certain rules. This paper recommends the concept of machine learning, the basic model and its application in many fields.

III. PROPOSED SYSTEM

This system proposes an efficient method of PCA that defines a simple system of machine learning in data storage requirements is described in below: In this proposed system to reduce the data storage requirements, and the importance of information being context dependent, it is significant to be able to derive context as if one can recognize context and pattern recognition after that it becomes probable to store the data with the context in which you activate. To classify these contexts and improve on data storage efficiency, this paper obtains the use of Principle Component Analysis (PCA) to reduce the data storage requirements for a machine-learning algorithm that is used for face recognition. PCA encapsulates huge data sets by creating new vectors, called principle components that are a linear combination of the original information, which results in a reduction of the data's capacity. As a result, PCA helps to reduce redundancy of data, filters noise in the data and compresses the data.

In this way, to achieve this compression of data, PCA takes information that is linked, and recognizes what one could call a "lowest common denominator". All data is then stored as a difference from this "lowest common denominator" which considerably reduces the large amount of data that wants to be stored in that system of management. Based on this type of application of PCA, one should be appreciate data that there are important savings with the data storage requirements.

That's why, in this paper they presented the good result achieved in which compares different PCA-type algorithms and needs the original information to be stored. Hence this proposed system in data storage requirement is effectively implemented with identifies the context and face recognition method of object to preprocessing that data and improve the efficiency of data.

IV.CONCLUSION AND FUTURE SCOPE

In this paper, the machine learning algorithms are implementing clearly on the mechanism of similarity of detecting information. Here, this paper has shared PCA method with an actual machine learning algorithm of management, to recognize the impact of using framework and to achieve the impact of this combination with better detection accuracy and storage requirements of data. The final results show that the data storage requirement and detection accuracy improves which depends on data storage.

That's why; the future work will mostly focus on how to classify the context or pattern recognition automatically in data management system as well as how to combine the major principle of PCA directly into the machine learning algorithm of data storage.

REFERENCES

1. Akash U. Suryawanshi, P. D. N. K. (2018). Review on Methods of Privacy-Preserving auditing for storing data security in cloud. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), ISSN, 7(4), 247–251.
2. Archana, R. C., Naveenkumar, J., & Patil, S. H. (2011). Iris Image Pre-Processing And Minutiae Points Extraction. International Journal of Computer Science and Information Security, 9(6), 171–176.
3. Ayush Khare, D. N. J. (2017). Perspective Analysis Recommendation System in Machine Learning. International Journal of Emerging Trends & Technology in Computer Science, 6(2), 184–187.
4. AyushKhare Nitish Bhatt, DrNaveen Kumar, J. G. (2017). Raspberry Pi Home Automation System Using Mobile App to Control Devices. International Journal of Innovative Research in Science, Engineering and Technology, 6(5), 7997–8003.
5. AyushKhare, J. G., Bhatt, N., & Kumar, N. (2017). Raspberry Pi Home Automation System Using Mobile App to Control Devices. International Journal of Innovative Research in Science, Engineering and Technology, 6(5), 7997–8003.
6. Bhore, P. R., Joshi, S. D., & Jayakumar, N. (2016). A Survey on the Anomalies in System Design: A Novel Approach. International Journal of Control Theory and Applications, 9(44), 443–455.
7. Bhore, P. R., Joshi, S. D., & Jayakumar, N. (2017a). A Stochastic Software Development Process Improvement Model To Identify And Resolve The Anomalies In System Design. Institute of Integrative Omics and Applied Biotechnology Journal, 8(2), 154–161.
8. Bhore, P. R., Joshi, S. D., & Jayakumar, N. (2017b). Handling Anomalies in the System Design: A Unique Methodology and Solution. International Journal of Computer Science Trends and Technology, 5(2), 409–413.
9. Desai, P., & Jayakumar, N. (n.d.). AN EXTENSIBLE FRAMEWORK USING MOBILITYRPC FOR POSSIBLE DEPLOYMENT OF ACTIVE STORAGE ON TRADITIONAL STORAGE ARCHITECTURE.
10. Desai, P. R., & Jayakumar, N. K. (2017). A Survey on Mobile Agents. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 5(XI), 2915–2918.
11. Divyansh Shrivastava Amol K. Kadam, Aarushi Chhibber, Naveenkumar Jayakumar, S. K. (2017). Online Student Feedback Analysis System with Sentiment Analysis. International Journal of Innovative Research in Science, Engineering and Technology, 6(5), 8445–8451.
12. Gawade, M. S. S., & Kumar, N. (2016). Three Effective Frameworks for semi-supervised feature selection. International Journal of Research in Management & Technology, 6(2), 107–110.
13. GAWADE, S., & JAYKUMAR, N. (2017). ILLUSTRATION OF SEMI-SUPERVISED FEATURE SELECTION USING EFFECTIVE FRAMEWORKS. Journal of Theoretical & Applied Information Technology, 95(20).
14. Jaiswal, U., Pandey, R., Rana, R., Thakore, D. M., & JayaKumar, N. (2017). Direct Assessment Automator for Outcome Based System. International Journal of Computer Science Trends and Technology (IJCS T), 5(2), 337–340.
15. Jayakumar, D. T., & Naveenkumar, R. (2012). SDjoshi,". International Journal of Advanced Research in Computer Science and Software Engineering," Int. J, 2(9), 62–70.
16. Jayakumar, M. N., Zaeimfar, M. F., Joshi, M. M., & Joshi, S. D. (2014). INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET). Journal Impact Factor, 5(1), 46–51.
17. Jayakumar, N. (2014). Reducts and Discretization Concepts, tools for Predicting Student's Performance. Int. J. Eng. Sci. Innov. Technol, 3(2), 7–15.
18. Jayakumar, N. (2015). Active storage framework leveraging processing capabilities of embedded storage array.
19. Jayakumar, N., Bhardwaj, T., Pant, K., Joshi, S. D., & Patil, S. H. (2015). A Holistic Approach for Performance Analysis of Embedded Storage Array. Int. J. Sci. Technol. Eng, 1(12), 247–250.
20. Jayakumar, N., Iyer, M. S., Joshi, S. D., & Patil, S. H. (2016). A Mathematical Model in Support of Efficient offloading for Active Storage Architectures. In International Conference on Electronics,

Electrical Engineering, Computer Science (EEECS) : Innovation and Convergence (Vol. 2, p. 103).
21. Jayakumar, N., & Kulkarni, A. M. (2017). A Simple Measuring Model for Evaluating the Performance of Small Block Size Accesses in Lustre File System. Engineering, Technology & Applied Science Research, 7(6), 2313–2318.
22. Jayakumar, N., Singh, S., Patil, S. H., & Joshi, S. D. (2015). Evaluation Parameters of Infrastructure Resources Required for Integrating Parallel Computing Algorithm and Distributed File System. IJSTE-Int. J. Sci. Technol. Eng, 1(12), 251–254.
23. KAKAMANSHADI, M. G., J, N., & PATIL, S. H. (2011). A METHOD TO FIND SHORTEST RELIABLE PATH BY HARDWARE TESTING AND SOFTWARE IMPLEMENTATION. International Journal of Engineering Science and Technology, 3(7), 5765–5768.
24. Khare, A., & Jayakumar, N. (2017). Perspective Analysis Recommendation System in Machine Learning. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 6(2), 184–187.
25. Komalavalli, R., Kumari, P., Navya, S., & Naveenkumar, J. (2017). Reliability Modeling and Analysis of Service-Oriented Architectures. International Journal of Engineering Science, 5591.
26. Kumar, N., Angral, S., & Sharma, R. (2014). Integrating Intrusion Detection System with Network Monitoring. International Journal of Scientific and Research Publications, 4, 1–4.
27. Kumar, N., Kumar, J., Salunkhe, R. B., & Kadam, A. D. (2016). A Scalable Record Retrieval Methodology Using Relational Keyword Search System. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (p. 32). ACM.
28. Namdeo, J., & Jayakumar, N. (2014). Predicting Students Performance Using Data Mining Technique with Rough Set Theory Concepts. International Journal, 2(2).
29. Naveenkumar, J. (2011). Keyword Extraction through Applying Rules of Association and Threshold Values. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), ISSN, 1021–2278.
30. Naveenkumar, J. (2015). SDJ, 2015. Evaluation of Active Storage System Realized Through Hadoop. International Journal of Computer Science and Mobile Computing, 4(12), 67–73.
31. Naveenkumar, J., & Joshi, S. D. (2015). Evaluation of Active Storage System Realized Through Hadoop. Int. J. Comput. Sci. Mob. Comput, 4(12), 67–73.
32. Naveenkumar, J., Makwana, R., Joshi, S. D., & Thakore, D. M. (2015a). OFFLOADING COMPRESSION AND DECOMPRESSION LOGIC CLOSER TO VIDEO FILES USING REMOTE PROCEDURE CALL. International Journal of Computer Engineering and Technology, 6(3), 37–45.
33. Naveenkumar, J., Makwana, R., Joshi, S. D., & Thakore, D. M. (2015b). Performance Impact Analysis of Application Implemented on Active Storage Framework. International Journal of Advanced Research in Computer Science and Software Engineering, 5(2), 550–554.
34. Naveenkumar, J., & Raval, K. S. (2011). Clouds Explained Using Use-Case Scenarios. INDIACom-2011 Computing for Nation Development, 3.
35. Naveenkumar J, P. D. S. D. J. (2015). Evaluation of Active Storage System Realized through MobilityRPC. International Journal of Innovative Research in Computer and Communication Engineering, 3(11), 11329–11335.
36. NAVEENKUMAR, M. J., Bhor, M. P., & JOSHI, D. R. S. D. (2011). A Self Process Improvement For Achieving High Software Quality. International Journal of Engineering Science, 3(5), 3850–3853.
37. Osho Tripathi Dr. Naveen Kumar Jayakumar, P. G. (2017). GARDUINO- The Garden Arduino. InternatIonal Journal of Computer SCIenCe and TeChnology, 8(2), 145–147.
38. Prashant Desai, N. J. (2018). AN EXTENSIBLE FRAMEWORK USING MOBILITYRPC FOR POSSIBLE DEPLOYMENT OF ACTIVE STORAGE ON TRADITIONAL STORAGE ARCHITECTURE. IIOAB Journal, 9(3), 25–30.
39. R. Salunkhe N. Jayakumar, and S. Joshi, A. D. K. (2015). "Luster A Scalable Architecture File System: A Research Implementation on Active Storage Array Framework with Luster file System. In ICEEOT.
40. RAVAL, K. S., SURYAWANSHI, R. S., NAVEENKUMAR, J., & THAKORE, D. M. (2011). The Anatomy of a Small-Scale Document Search Engine Tool: Incorporating a new Ranking Algorithm. International Journal of Engineering Science and Technology, 3(7), 5802–5808.
41. Rishikesh Salunkhe, N. J. (2016). Query Bound Application Offloading: Approach Towards Increase Performance of Big Data Computing. Journal of Emerging Technologies and Innovative Research, 3(6), 188–191.
42. Salunkhe, R., Kadam, A. D., Jayakumar, N., & Thakore, D. (2016). In search of a scalable file system state-of-the-art file systems review and map view of new Scalable File system. In Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on (pp. 364–371). IEEE.
43. Sawant, Y., Jayakumar, N., & Pawar, S. S. (2016). Scalable Telemonitoring Model in Cloud for Health Care Analysis. In International Conference on Advanced Material Technologies (ICAMT) (Vol. 2016).
44. Singh, A. K., Pati, S. H., & Jayakumar, N. (2017). A Treatment for I/O Latency in I/O Stack. International Journal of Computer Science Trends and Technology (IJCS T), 5(2), 424–427.
45. Yogesh Sawant, P. D. N. kumar. (2016). Crisp Literature Review One andScalableFramework:Active Model to Create Synthetic Electrocardiogram Signals. International Journal of Application or Innovation in Engineering & Management, 5(11), 73–80.

46. Zaeimfar, S. (2014). Workload Characteristics Impacts on file System Benchmarking. Int. J. Adv, 39–44.

AUTHOR

Miss. Amruta
Bharati Vidyapeeth Deemed University College of Engineering, Pune, 411043, Maharashtra, India.
Student : Department of Computer Engineering,