RESEARCH ARTICLE OPEN ACCESS

A Novel Correlation-based Feature Selection (CFS) with SVM (PolyKernel) Algorithm (CFS- SVM) for Heart Disease Classification

Dr. Isabella Jones

Academic Development Manager, Goldsmiths, University of London

ABSTRACT

To fulfill this aim, 18 machine learning methods divided into 6 different categories, and 3 different feature selection was used in this study. After the Correlation-based Feature Selection (CFS) feature selection, the most successful algorithm was Naive Bayes and Fuzzy RoughSet with a ratio of 84.818%. However, after using Chi-Square feature selection, the most successful algorithm was found to be the RBF Network algorithm with 81.188% ratio. Consequently, it is recommended that specialist doctors who want to classify heart disease should use the SVM (PolyKernel) algorithm if they are not going to use feature selection whereas they should use should the Naive Bayes algorithm if they are going to use CFS as a feature selection. Additionally, if they are to use Fuzzy Rough Set and Chi-Square as the feature selection, it is recommended that they use the RBF Network algorithm.

***Keywords*:** Machine Learning, SVM (PolyKernel), Correlation-based Feature Selection (CFS), Heart Disease, CFS-SVM, Classification.

I. INTRODUCTION

As with all diseases, early diagnosis is very important in people with heart disease. Because when it is too late for the treatment of the disease, both large amounts of money are spent and it causes problems in recovery [1-5]. The main problems on this issue are that in the first stage of the disease, the diagnosis of the disease is not made with the help of computer-aided systems and the process is very slow since the final decision is made by the doctors [6-13]. In addition, results are directly affected due to conditions such as the education taken, working conditions, number of patients per physician [45-56]. In addition, the low rate of making the right decision at the first stage can be seen as another problem. In order to make a full diagnosis, people who apply to health institutions are repeatedly tested. This state means both time and financial loss. The way to minimize the damage in these situations is to benefit from computer aided smart systems. The main reason why such smart systems have not been used until now is that artificial intelligence and computer systems were not so common. Besides, the fact that the technology was not so advanced and accessible made the transition to these smart systems difficult [58-60]. This does not mean that the human factor will be completely eliminated in decision-making processes. On the contrary, the human factor will become even more effective. Because ultimately, it is the people who produce these smart systems. The only goal here is to speed up the process and make decisions with high accuracy by minimizing human errors [23-28].

II. METHODS

In 2013 Thangamani, M., T et al. [34], has been proposed to identify the disease name with the symptoms specified and extract the sentence from the article and get the Relation that exists between DiseaseTreatment and classify the information into cure, prevent, side effect to the user.This electronic documentis a “live” template.

In 2007, Abraham, R., et al. [35], a new feature selection algorithm to improve the classification accuracy of Naïve Bayes with respect to medical datasets. Our experimental results with 17 medical datasets suggest that on an average the new CHI-WSS algorithm gave best results. The proposed algorithm utilizes discretization and simplifies the’ wrapper’ approach based feature selection by reducing the feature dimensionality through the elimination of irrelevant and least relevant features using chi-square statistics. For our experiments we utilize two established measures to compare the performance of statistical classifiers namely; classification accuracy (or error rate) and the area under ROC to demonstrate that the proposed algorithm using generative Naïve Bayesian classifier on the average is more efficient than using discriminative models namely Logistic Regression and Support Vector Machine.

In 2017 Kikuchi, H., et al. [36], has been proposed the real medical dataset related to stroke and attempt to apply multiple regression with six predictors of age, sex, the medical scales, e.g., Japan Coma Scale, and the modified Rankin Scale. Our contributions of this paper include (1) to propose a practical privacy-preserving protocols for linear multiple regression with vertically partitioned datasets, and (2) to show the feasibility of the proposed

system using the real medical dataset distributed into two parties, the hospital who knows the technical details of diseases during the patients are in the hospital, and the local government who knows the residence even after the patients left hospital. (3) to show the accuracy and the performance of the PPDM system which allows us to estimate the expected processing time with arbitrary number of predictors.

The number of parameters used in diagnosis has an effect on the accuracy rate of the algorithm.

Feature selection is the process of evaluating which parameters are effective and how effective they are on the result. There are many feature selections in the literature. The important thing here is not to use feature selection in studies. It is necessary to find out whether feature selection has a positive effect on the result of the study. Depending on that:

The results of the algorithms used in the studies are evaluated only on the correct percentages of success in most of them. However, these evaluations mislead readers. Success rates are necessarily important. Yet, it is also necessary to decide which of the algorithms are better in Type I and which in Type II error types.

The concept of classification can be defined as distributing data between classes that are defined under certain rules on a data set. There are many classification methods in the literature. The important point here is to determine the correct classification algorithm appropriate to the data set and the success rate of the algorithm used is high.

Regression is the description of the relationship between a response variable and one or more explanatory variables.

Naive Bayes; Bayesian network consists of a structural model and a set of conditional probabilities. The structural model is a directed graph in which nodes represent attributes and arcs represent attribute dependencies.

The decision tree is a method which is easy to understand and interpret classification (6). The decision tree method is one of the most popular algorithms in classification algorithms.

Naive Bayes Tree (NBTree) is a hybrid algorithm, which deploys a naive Bayes classifier on each leaf node of the built decision tree and has demonstrated remarkable classification performance (8).

Fuzzy-Rough Nearest Neighbour (FRNN) is the extension of the K-nearest neighbor algorithm by using the fuzzy-rough uncertainty. The fuzzy uncertainty concept is used to measure the distance between the test pattern and the neighbor. It also helps to represent the neighbour to be in many classes. Due to the lack of features some of the neighbors and the test patterns may be indistinguishable hence the concept of rough uncertainty is used. The neighborhood structure is artificial, so the roughness emerges (9).

The Fuzzy Nearest Neighbor (FuzzyNN) classifier is well known for its effectiveness in supervised learning problems. K-NN classifies by comparing new incoming examples with a similarity function using the samples of the training set. The fuzzy version of the kNN accounts for the underlying uncertainty in the class labels, and it is composed of two different stages. The first one is responsible for calculating the fuzzy membership degree for each sample of the problem in order to obtain smoother boundaries between classes. The second stage classifies similarly to the standard kNN algorithm but uses the previously calculated class membership degree (10).

Innovative Feature selection

Feature selection is an important set of algorithms used to achieve more consistent results by improving the correct classification rates or performances of the methods used in machine learning systems. In this study, CFS, Fuzzy Rough Set and Chi-Square algorithms are used as feature selection algorithm.

Correlation-based Feature Selection (CFS) is a simple correlation-based filtering algorithm. One point to note here is that features with low correlation should be ignored. It should be ensured that the remaining features are highly correlated with each other. CFS's feature subset evaluation function equation below is repeated here for ease of reference:

At this point, it is aimed to employ a method to calculate reducts for fuzzy rough sets, and only the minimal elements positioned in the discernibility matrix are taken into consideration. Initially, the definition of the relative discernibility relations of the conditional attribute is carried out, then to qualify the minimal elements in the discernibility matrix relative discernibility relations are employed. Followingly, to calculate the minimal elements an algorithm is created.

Chi-Square; is one of the most popular feature selection algorithms known in the literature. What lies on the basis of the algorithm is the calculation of the chi-square value between each feature and the target feature. With this calculation, the best chi-square score is determined and the desired number of properties is selected.

Conclusion

The aim of this study is to determine the machine learning algorithm with the highest accuracy rate of machine learning algorithms, which is one of the computer-aided smart systems in the diagnosis of heart disease. And also, it aims to determine how feature selection affects the performance of these machine learning algorithms and to

make determinations about the diagnosis of disease by using TP, FP, Kappa Statistics and ROC analysis as performance criteria.

REFERENCE

[1] Feng, Y., Yi, J. H., & Wang, G. G. (2019). Enhanced Moth Search Algorithm for the Set-Union Knapsack Problems. IEEE Access, 7, 173774-173785.

[2] Sivaram, M., Batri, K., Amin Salih, M., &Porkodi, V. (2019). Exploiting the Local Optima in Genetic Algorithm using Tabu Search. Indian Journal of Science and Technology, 12(1), 1-13.

[3] Venkatraman, S., &Surendiran, B. (2020). Adaptive hybrid intrusion detection system for crowd sourced multimedia internet of things systems. Multimedia Tools and Applications, 79(5), 3993-4010.

[4] Sujitha, B., Parvathy, V. S., Lydia, E. L., Rani, P., Polkowski, Z., & Shankar, K. (2020). Optimal deep learning based image compression technique for data transmission on industrial Internet of things applications. Transactions on Emerging Telecommunications Technologies, e3976.

[5] Ezhilarasu, P., Krishnaraj, N., &Dhiyanesh, B. (2015). Arithmetic Coding for Lossless Data Compression–A Review. International Journal of Computer Science Trends and Technology, 3(3).

[6] Porkodi, V., Singh, A. R., Sait, A. R. W., Shankar, K., Yang, E., Seo, C., & Joshi, G. P. (2020). Resource Provisioning for Cyber–Physical–Social System in Cloud-Fog-Edge Computing Using Optimal Flower Pollination Algorithm. IEEE Access, 8, 105311-105319.

[7] Gao, D., Wang, G. G., &Pedrycz, W. (2020). Solving fuzzy job-shop scheduling problem using DE algorithm improved by a selection mechanism. IEEE Transactions on Fuzzy Systems.

[8] Sivaram, M., Mohammed, A. S., Yuvaraj, D., Porkodi, V., Manikandan, V., &Yuvaraj, N. (2019, February). Advanced expert system using particle swarm optimization based adaptive network based fuzzy inference system to diagnose the physical constitution of human body. In International Conference on Emerging Technologies in Computer Engineering (pp. 349-362). Springer, Singapore.

[9] Jiménez, A. C., García-Díaz, V., González-Crespo, R., &Bolaños, S. (2018). Decentralized Online Simultaneous Localization and Mapping for Multi-Agent Systems. Sensors, 18(8), 2612.

[10] Venkatraman, S., Surendiran, B., & Kumar, P. A. R. (2020). Spam e-mail classification for the Internet of Things environment using semantic similarity approach. The Journal of Supercomputing, 76(2), 756-776.

[11] Lydia, E. L., Raj, J. S., PandiSelvam, R., Elhoseny, M., & Shankar, K. (2019). Application of discrete transforms with selective coefficients for blind image watermarking. Transactions on Emerging Telecommunications Technologies, e3771.

[12] Ezhilarasu, P., Prakash, J., Krishnaraj, N., Kumar, D. S., Babu, K. S., &Parthasarathy, C. (2015). A Novel Approach to Design the Finite Automata to Accept the Palindrome with the Three Input Characters. Indian Journal of Science and Technology, 8(28).

[13] Devaraj, A. F. S., Elhoseny, M., Dhanasekaran, S., Lydia, E. L., & Shankar, K. (2020). Hybridization of firefly and Improved Multi-Objective Particle Swarm Optimization algorithm for energy efficient load balancing in Cloud Computing environments. Journal of Parallel and Distributed Computing.

[14] Zou, D., Wang, G. G., Sangaiah, A. K., & Kong, X. (2017). A memory-based simulated annealing algorithm and a new auxiliary function for the fixed-outline floorplanning with soft blocks. Journal of Ambient Intelligence and Humanized Computing, 1-12.

[15] Kumar, A., Ahuja, H., Singh, N. K., Gupta, D., Khanna, A., & Rodrigues, J. J. (2018). Supported matrix factorization using distributed representations for personalised recommendations on twitter. Computers & Electrical Engineering, 71, 569-577.

[16] Sivaram, M., Porkodi, V., Mohammed, A. S., Manikandan, V., &Yuvaraj, N. (2019). Retransmission DBTMA protocol with fast retransmission strategy to improve the performance of MANETs. IEEE Access, 7, 85098-85109.

[17] Venkatraman, S., & Kumar, P. A. R. (2019). Improving Adhoc wireless sensor networks security using distributed automaton. Cluster Computing, 22(6), 14551-14557.

[18] Lydia, E. L., Govindaswamy, P., Lakshmanaprabu, S., &Ramya, D. (2018). Document clustering based on text mining K-means algorithm using euclidean distance similarity. J. Adv. Res. Dyn. Control Syst.(JARDCS), 10(2), 208-214.

[19] Ortin, F., Mendez, S., García-Díaz, V., & Garcia, M. (2014). On the suitability of dynamic languages for hot-reprogramming a robotics framework: a Python case study. Software: Practice and Experience, 44(1), 77-104.

[20] Krishnaraj, N., Ezhilarasu, P., & Gao, X. Z. Hybrid Soft Computing Approach for Prediction of Cancer in Colon Using Microarray Gene Data. Current Signal Transduction Therapy, 11(2).

[21] Le Nguyen, B., Lydia, E. L., Elhoseny, M., Pustokhina, I., Pustokhin, D. A., Selim, M. M., ... & Shankar, K. (2020). Privacy Preserving Blockchain

Technique to Achieve Secure and Reliable Sharing of IoT Data. CMC-COMPUTERS MATERIALS & CONTINUA, 65(1), 87-107.
[22] Chavhan, S., Gupta, D., Chandana, B. N., Khanna, A., & Rodrigues, J. J. (2019). IoT-based Context-Aware Intelligent Public Transport System in a metropolitan area. IEEE Internet of Things Journal.
[23] Gu, Z. M., & Wang, G. G. (2020). Improving NSGA-III algorithms with information feedback models for large-scale many-objective optimization. Future Generation Computer Systems, 107, 49-69.
[24] Porkodi, V., Khan, J., Mohammed, A. S., Bhuvana, J., & Sivaram, M. OPTIMIZED COOPERATIVE QOS ENHANCED DISTRIBUTED MULTIPATH ROUTING PROTOCOL.
[25] Geerthik, S., Venkatraman, S., & Gandhi, R. (2016). AnswerRank: Identifying Right Answers in QA system. International Journal of Electrical and Computer Engineering, 6(4), 1889.
[26] Samad, A., Salima, R., Lydia, E. L., & Shankar, K. (2020). Definition and Features of Rural Marketing Strategies for Encourage Development in Rural Areas. TEST Engineering & Management, 82, 4983-4988.
[27] Palani, E., Nagappan, K., &Alhadidi, B. (2016). Segmentation and Texture Analysis for Efficient Classification of Breast Tumors from Sonograms. Current Signal Transduction Therapy, 11(2), 84-90.
[28] Rajagopal, A., Ramachandran, A., Shankar, K., Khari, M., Jha, S., Lee, Y., & Joshi, G. P. (2020). Fine-tuned residual network-based features with latent variable support vector machine-based optimal scene classification model for unmanned aerial vehicles. IEEE Access, 8, 118396-118404.
[29] Mondragon, V. M., García-Díaz, V., Porcel, C., & Crespo, R. G. (2018). Adaptive contents for interactive TV guided by machine learning based on predictive sentiment analysis of data. Soft Computing, 22(8), 2731-2752.
[30] Feng, Y., Yu, X., & Wang, G. G. (2019). A Novel Monarch Butterfly Optimization with Global Position Updating Operator for Large-Scale 0-1 Knapsack Problems. Mathematics, 7(11), 1056.
[31] Mohammed, A. S., & Sivaram, P. (2018). Securing the Sensor Networks Along With Secured Routing Protocols for Data Transfer in Wireless Sensor Networks.
[32] Geerthik, S., Venkatraman, S., & Gandhi, K. R. (2016, February). Reward rank: A novel approach for positioning user answers in community question answering system. In 2016 International Conference on Information Communication and Embedded Systems (ICICES) (pp. 1-6). IEEE.
[33] Sivaram, M., Lydia, E. L., Pustokhina, I. V., Pustokhin, D. A., Elhoseny, M., Joshi, G. P., & Shankar, K. (2020). An optimal least square support vector machine based earnings prediction of blockchain financial products. IEEE Access, 8, 120321-120330.
[34] Ghantasala, G. P., &KrishnaRaj, N. Support Vector Machine Based Automatic Mammogram Classification Using Hybrid Optimization Algorithm.
[35] Sikkandar, M. Y., Alrasheadi, B. A., Prakash, N. B., Hemalakshmi, G. R., Mohanarathinam, A., & Shankar, K. (2020). Deep learning based an automated skin lesion segmentation and intelligent classification model. Journal of Ambient Intelligence and Humanized Computing, 1-11.
[36] Zhang, Z., Wang, G. G., Zou, K., & Zhang, J. (2014). A solution quality assessment method for swarm intelligence optimization algorithms. The Scientific World Journal, 2014.
[37] Sivaram, Murugan et al. 'Data Fusion Using Tabu Crossover Genetic Algorithm in Information Retrieval'. 1 Jan. 2020 : 1 – 10.
[38] Khamparia, A., Pandey, B., Tiwari, S., Gupta, D., Khanna, A., & Rodrigues, J. J. (2020). An integrated hybrid CNN–RNN model for visual description and generation of captions. Circuits, Systems, and Signal Processing, 39(2), 776-788.
[39] Geerthik, S., Gandhi, K. R., &Venkatraman, S. (2016, December). Domain expert ranking for finding domain authoritative users on community question answering sites. In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-5). IEEE.
[40] Muruganantham, A., Nguyen, P. T., Lydia, E. L., Shankar, K., Hashim, W., &Maseleno, A. (2019). Big data analytics and intelligence: A perspective for health care.
[41] Ramkumar, V., &Krishnaraj, N. Weight Based LSA to Retrieve Information from Web Pages Based On Document Score.
[42] Balakiruthiga, B., Deepalakshmi, P., Mohanty, S. N., Gupta, D., Kumar, P. P., & Shankar, K. (2020). Segment routing based energy aware routing for software defined data center. Cognitive Systems Research.
[43] Chu, H. C., Wang, G. G., & Deng, D. J. (2016). The social networking investigation of metadata of forensic artifacts of a typical WeChat session under Windows. Security and Communication Networks, 9(18), 5698-5709.
[44] Sivaram, M., Yuvaraj, D., Mohammed, A. S., Manikandan, V., Porkodi, V., &Yuvaraj, N. (2019). Improved Enhanced Dbtma with Contention-Aware Admission Control to Improve the Network Performance

in Manets. CMC-COMPUTERS MATERIALS & CONTINUA, 60(2), 435-454.
[45] Gochhayat, S. P., Lal, C., Sharma, L., Sharma, D. P., Gupta, D., Saucedo, J. A. M., &Kose, U. (2019). Reliable and secure data transfer in IoT networks. Wireless Networks, 1-14.
[46] Subbarayalu, V., Surendiran, B., &Arun Raj Kumar, P. (2019). Hybrid Network Intrusion Detection System for Smart Environments Based on Internet of Things. The Computer Journal, 62(12), 1822-1839.
[47] Rosa, A. T. R., Pustokhina, I. V., Lydia, E. L., Shankar, K., & Huda, M. (2019). Concept of electronic document management system (EDMS) as an efficient tool for storing document. Journal of Critical Reviews, 6(5), 85-90.
[48] Espada, J. P., Diaz, V. G., Crespo, R. G., Bustelo, B. C. P. G., &Lovelle, J. M. C. (2015). An intelligent Mobile Web Browser to adapt the mobile web as a function of the physical environment. IEEE Latin America Transactions, 13(2), 503-509.
[49] Kumar, R. S., Krishnaraj, N., &Keerthana, G. (2017). Assessment of Quality of Service in Communication Network and Evaluating Connectivity Among IP Networks. Asian Journal of Applied Science and Technology (AJAST), 1(3), 319-322.
[50] Elhoseny, M., Rajan, R. S., Hammoudeh, M., Shankar, K., &Aldabbas, O. (2020). Swarm intelligence–based energy efficient clustering with multihop routing protocol for sustainable wireless sensor networks. International Journal of Distributed Sensor Networks, 16(9), 1550147720949133.
[51] Chu, H. C., Wang, G. G., & Park, J. H. (2015). The digital fingerprinting analysis concerning google calendar under ubiquitous mobile computing era. Symmetry, 7(2), 383-394.
[52] Manikandan, V., Sivaram, M., Mohammed, A. S., &Porkodi, V. (2020). Nature Inspired Improved Firefly Algorithm for Node Clustering in WSNs. CMC-COMPUTERS MATERIALS & CONTINUA, 64(2), 753-776.
[53] Kuppusamy, P., Venkatraman, S., Rishikeshan, C. A., & Reddy, Y. P. (2020). Deep learning based energy efficient optimal timetable rescheduling model for intelligent metro transportation systems. Physical Communication, 101131.
[54] Asih, E. S., Nguyen, P. T., Lydia, E. L., Shankar, K., Hashim, W., &Maseleno, A. (2019). Mobile E-commerce website for technology-based buying selling services. International Journal of Engineering and Advanced Technology, 8(6), 884-888.
[55] Lydia, E. L., &Swarup, M. B. (2015). Big data analysis using hadoop components like flume, mapreduce, pig and hive. International Journal of Science, Engineering and Computer Technology, 5(11), 390.
[56] Sengar, S. S., Hariharan, U., &Rajkumar, K. (2020, March). Multimodal Biometric Authentication System using Deep Learning Method. In 2020 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 309-312). IEEE.
[57] Maseleno, A., Hashim, W., Perumal, E., Ilayaraja, M., & Shankar, K. (2020). Access control and classifier-based blockchain technology in e-healthcare applications. In Intelligent Data Security Solutions for e-Health Applications (pp. 151-167). Academic Press.
[58] Li, J., Lei, H., Alavi, A. H., & Wang, G. G. (2020). Elephant Herding Optimization: Variants, Hybrids, and Applications. Mathematics, 8(9), 1415.
[59] Mohammed, A. S., Kareem, S. W., Al Azzawi, A. K., & Sivaram, M. (2018). Time series prediction using SRE-NAR and SRE-ADALINE. Journal of Advanced Research in Dynamical and Control Systems, Pages, 1716-1726.
[60] Shankar, K., Elhoseny, M., Chelvi, E. D., Lakshmanaprabu, S. K., & Wu, W. (2018). An efficient optimal key based chaos function for medical image security. IEEE Access, 6, 77145-77154.
[61] Geerthik, S., Gandhi, R., &Venkatraman, S. (2006). CATEGORY BASED EXPERT RANKING: A NOVEL APPROACH FOR EXPERT IDENTIFICATION IN COMMUNITY QUESTION ANSWERING.
[62] Laxmi, C. V., &Somasundaram, K. (2014). Application Level Scheduling (AppLeS) in Grid with Quality of Service (QoS). International Journal of Grid Computing & Applications, 5(2), 1.
[63] Kumar, R. S., Krishnaraj, N., &Keerthana, G. Highly Energy Efficient and Scalable Distributed Clustering Procedure for Dense Wireless Sensor Networks.
[64] Krishnaraj, N., Kumar, K. A., & Kumar, P. K. (2018). DESIGN OF ADAPTIVE SCHEDULER TO IMPROVE PERFORMANCE OF COMPUTATIONAL GRIDS. International Journal of Pure and Applied Mathematics, 119(18), 1741-1751.
[65] Shankar, K., & Eswaran, P. (2016, January). A new k out of n secret image sharing scheme in visual cryptography. In 2016 10th International Conference on Intelligent Systems and Control (ISCO) (pp. 1-6). IEEE.
[66] Wei, C. L., & Wang, G. G. (2020). Hybrid Annealing Krill Herd and Quantum-Behaved Particle Swarm Optimization. Mathematics, 8(9), 1403.
[67] Sivaram, M., Yuvaraj, D., Mohammed, A. S., &Porkodi, V. Estimating the Secret Message in the Digital Image. International Journal of Computer Applications, 975, 8887.

[68] Nieto, Y., Gacía-Díaz, V., Montenegro, C., González, C. C., & Crespo, R. G. (2019). Usage of machine learning for strategic decision making at higher educational institutions. IEEE Access, 7, 75007-75017.