FEATURE SELECTION HYBRID METHOD FOR OPTIMIZATION ALGORITHMS IN THE FIELD OF MEDICAL RECORDS

Yuda Syahidin, Ade Irma Suryani

Manajemen Informasi Kesehatan, Rekam Medis dan Informasi Kesehatan, Piksi Ganesha Politechnic, Bandung, Indonesia

Abstract

Medical Records or Electronic Health Records refers to the collection of patient health information in a digital format. The problem of classifying medical record data involves high-dimensional features. This raises a problem in determining which features have a correlation with the predicted results. Embedded technique uses learning model construction and feature selection simultaneously. The Wrapper technique performs feature evaluation by utilizing machine learning algorithms. The experimental results produce accuracy values for each embedded and wrapper technique. In this research, it is proposed to develop a hybrid technique that aims to find feature significance by applying machine learning techniques to increase the accuracy of predictions for disease classification. The proposed hybrid model combines the results of feature weighting and compares feature performance with several known classification techniques. The test results resulted in an increase in the accuracy value according to the disease classification dataset through the hybrid feature weight evaluation (HFWE) model.

Keywords: Electronic Health Records, Embeded, Wrapper, Hybrid

Introduction

The health of patients is a priority issue and medical experts are constantly trying to implement new technologies and achieve important results. The use of medical data data that enables the analysis of large volumes of medical data that can be used in research areas such as clinical decision support phenotypic information extraction disease inference and personalized healthcare(Shickel dkk, 2017). Predictive analytics is one of the important areas of medical science to provide better services to patients (Panesar, n.d.). In recent years most of the methods used to analyze data from electronic health records which contain a lot of information about patient health have used predictive analytics to extract information using machine learning and statistical techniques. such as the results of clinical trials (Jensen dkk, 2012),(Xiao dkk, 2018),(Dinov, 2018). The performance of prediction algorithms depends on data representation and feature selection (Bengio dkk, 2013),(Polyzotis dkk, 2018). The challenge in health data is to find patterns that generate predictive models to support clinical decisions and much health data remains untapped. Electronic health reporting still lacks the scope and efficiency of medical record data analysis and because there is still unstructured health data there is still a lack of systems for health decision making(Goldstein dkk, 2017), (Weiskopf dkk, 2013), (Latif dkk, 2020).

There are additional approaches to feature selection using filter and banding methods to implement a stepwise search strategy to discriminate feature quality and feature evaluation. All the features in the data sheet are important for establishing a model hypothesis for prediction. The filtering method is a selection method that has an independent machine learning method and is a selection method based on the relationship of variables and a machine learning algorithm. A feature selection study used convolution techniques to reduce the features selected by Yang et al.(Yan & Zhang, 2015), Feature Selection Packing Techniques in Diabetes estimate, Le, et al (Le dkk, 2021), cluster ranking technique in feature selection by Anwar et al (Haq dkk, 2019), and Genetic algorithms used in feature selection to help predict mortality patterns by Ghorbani et al (Ghorbani dkk, 2020).

However, these different feature selection methods due to inappropriate data reduction methods and limitations for feature selection still have challenges with high-dimensional data problems and require more in-depth research and continuous research to combine features. Based on this there is a need for methods to develop feature selection algorithms to facilitate feature selection for medical record data with both structured and unstructured features.

Literature Review

1. ELECTRONIC HEALTH RECORDS

Electronic Health Record (EHR) is a person's health information that is stored digitally and is instantly and securely available to authorized users. EHRs contain patient diagnoses, medications, vital signs, treatment plans, progress notes, radiological images, and test results (Jensen dkk, 2012). Classification of records exists as unstructured and structured EHR data. Unstructured EHR data is written based on the clinical context that describes the patient's condition and is most useful for clinical documentation.

EHR data consists of various types of data, from structured information such as prescription drug data consisting of dates and doses to unstructured data such as clinical narratives that describe the medical reasons behind medical record documents. The relationship between structured data and unstructured data can be seen in Figure 1 below.

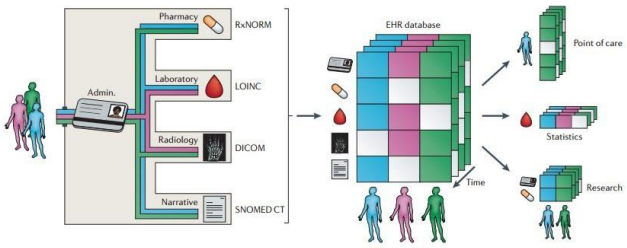

Figure 1. Electronic health record content (Jensen dkk, 2012)

A patient's EHR can be viewed as a storehouse of information regarding his health status in a computer- readable form. When connected to the health care system it produces various types of patient-related data. Patient data is stored in a database and can be viewed in a format that suits the needs and authority of certain user groups. The term Electronic Health Record (EHR), or Electronic Health Record, refers to the collection of patient health information in a digital format. EHRs can be categorized in terms of functionality: (i) basic EHR without clinical records, (ii) basic EHR with clinical records and (iii) comprehensive systems. EHRs, even in their simplest form, provide researchers with a rich collection of data. Data can be shared across networks and can include, as previously described, a wide variety of information. EHR is primarily designed for internal hospital administration tasks and many different schemes exist in different structures (Poongodi dkk, 2021).

2. FEATURE SELECTION

Feature selection is a critical process in machine learning, designed to remove irrelevant, redundant, and noisy features and retain a small fraction of features from the main feature space. Thus, effective feature selection can help reduce computational complexity, improve model accuracy, and improve model interpretability. In machine learning and data science more generally, feature selection (also known as variable/feature selection, attribute selection or subset selection) is the process by which data is automatically or manually selected for a relevant subset of features for use in building a machine learning model. In fact, it is one of the core concepts in machine learning that has a huge impact on model performance and is key to creating reliable machine learning models (Zheng & Casari, n.d.).

Feature selection can be described in two steps as follows:

a) Combination of search techniques to generate new feature subsets.
b) Take measurements of features to evaluate or judge how well different subsets of features are.

Feature selection methods can be divided into three categories (Max Kuhn, 2019):

a) Filter Methods

Rely on the characteristics of the features without using any machine learning algorithms. The filter method selects features from the dataset independently for the machine learning algorithm. This method only depends on the characteristics of the variables, so that features are filtered from the data before learning begins.

b) Wrapper Methods

Based on consideration of selecting a feature set as a search problem, then using predictive machine learning algorithms to select the best feature subset. This method trains a new model on each feature subset with the aim of generating the best performing feature subset for a given machine learning algorithm.

c) Embedded Methods

The Embedded method considers the interaction of features and models. This method performs feature selection as part of the model construction process.

State-of-the-Art

1) Feature Selection – Genetic Algorithm (Ghorbani dkk, 2020)

The hybrid model uses a Genetic Algorithm as a feature selection method and an ensemble model based on a combination of Stacking and Boosting. This process selects an optimal subset of the relevant features for use in predictive model development.

Figure 2. Model Framework New Hybrid (Ghorbani dkk, 2020).

2) Feature Selection - Wrapper (Le dkk, 2021)

This research uses the wrapper method to select features to remove features that have nothing to do with the diabetes dataset and this method helps optimize the number of attributes for the Multilayer Perceptron algorithm. Below is picture 3 of the Wrapper Method and picture 4 of the framework for early diabetes prediction.

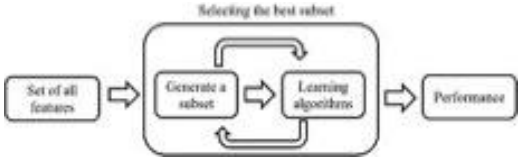

Figure 3. Wrapper Method (Le dkk, 2021)

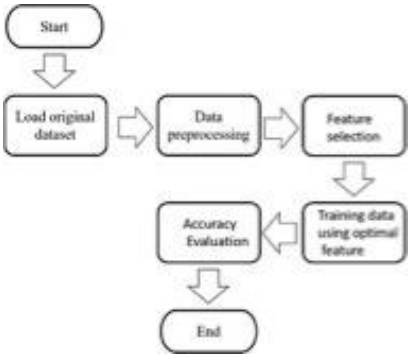

Figure 4. Framework untuk early diabetes Prediction (Le dkk, 2021).

3) Feature Selection – Imbalanced Data Classification (Dudkk, 2020)

This study uses filter techniques in selecting features for unbalanced data. The results of this feature selection will be used for the prediction model for hospital readmission. Below is figure 5. Framework Imbalanced data classification

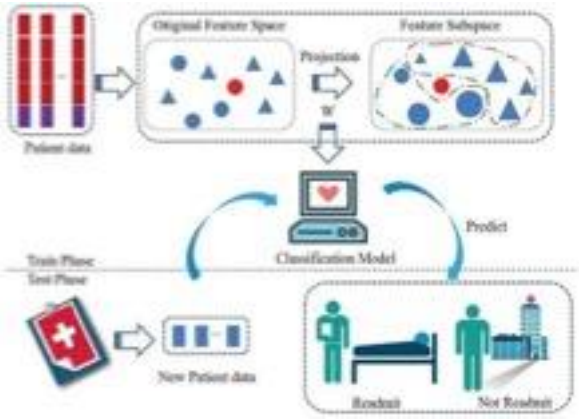

Figure 5. Framework Imbalanced data classification (Du dkk, 2020)

4) Feature Selection - Classification models for heart disease prediction(Gárate-Escamila dkk, 2020)

This study uses the dimension reduction method and finds features that have a correlation with heart disease by applying feature selection techniques. Combining PCA (principal component analysis) techniques can help with data dimension problems. Below is a picture of 6 approaches to feature selection.

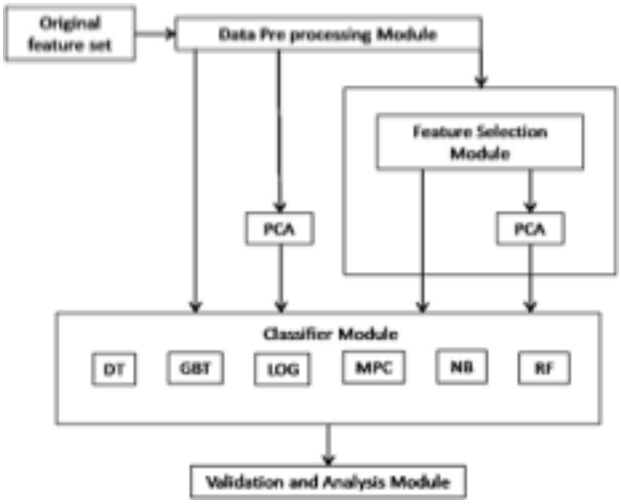

Figure 6. Classification models for heart disease prediction (Gárate- Escamila dkk, 2020)

Research Method

In selecting features for the dataset that will be modeled against machine learning algorithms based on the feature selection technique used, namely filter, wrapper and embedded techniques. The following is a picture of the approach method in feature selection (Galli, 2020).

Figure 7. Feature Selection Approach Method (Galli, 2020).

The Hybrid method in feature selection depends on the combination of the selected approaches. Techniques to speed up the selection process and combine embedded techniques with selected learning algorithms through wrapper techniques. The embedded engineering approach to feature selection makes it possible to apply learning in the selection process as shown in Figure 7 below.

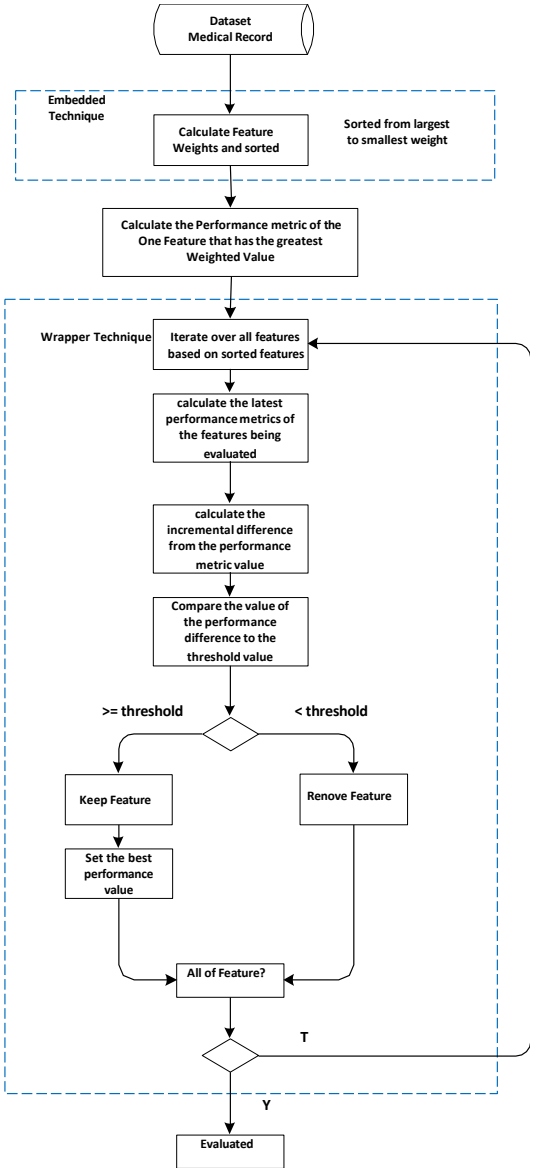

Figure 7. The proposed Hybrid Approach method.

Results and Discussion

As part of this study, initial trials were conducted on a collection of datasets to test the proposed Hybrid Feature Evaluation (HBFE) model. The required dataset is a dataset that fits the criteria in the Health sector. In this study, the Hybrid model uses heart disease, breast cancer and Hepatocellular carcinoma. The dataset is taken from the UCI Machine Learning Repository. The following is Table 1 Summary Dataset UCI Repository.

Table 1 Summary Dataset UCI Repository

Dataset	Number of instances	Number of features	Number of classes
Heart Disease	207	13	2
Breast Cancer	569	30	2
Hepatocellular carcinoma	165	49	2

Preliminary trials weighed the characteristics of each data set. In the next step the weighted features are sorted from the highest weight value to the lowest. After conducting the experiments according to the steps of the designed research the results are obtained in the form of comparison between hybrids and non-hybrids. From the experimental results it can be seen that the comparison of the accuracy values of the tested datasets is based on the weighting of the features associated with the target (label). Following are the comparison tables of the embedded, wrapper and the Hybrid Feature Evaluation technique. The following is Table 2 Comparison of technique, figure 8 comparison of accuracy model and figure 9 number of picure.

Table 2 Comparison of Embedded, Wrapper and Hybrid Feature Evaluation models against UCI ML Disease Datasets

Dataset	Alg.	Technique			No of Feature		
		EBD ACC (%)	WRP ACC (%)	HBFE ACC (%)	EBD	WRP	HBFE
Heart Disease	RF	86.04	82.41	**91.66**	8	12	**7**
	LR	80.23	79.12	**88.6**	8	12	**6**
	SVM	81.39	80.21	**88.37**	8	12	**8**
	GBT	83.72	81.31	**89.15**	8	12	**8**
Breast Cancer	RF	96.49	97.07	**97.66**	9	25	**3**
	LR	97.07	95.7	**97.49**	9	25	**5**
	SVM	97.08	95.9	**97.59**	9	25	**5**
	GBT	95.9	98.24	**98.39**	9	25	**8**
Hepatocellular carcinoma	RF	82	72	**90.32**	21	40	**14**
	LR	78	74	**83.87**	21	40	**21**
	SVM	78	72	**83.97**	21	40	**11**
	GBT	76	74	**89.2**	21	40	**12**

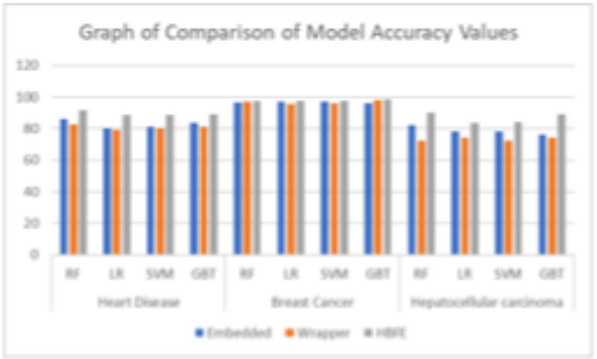

Figure 8. Comparison of Model Accuracy Values

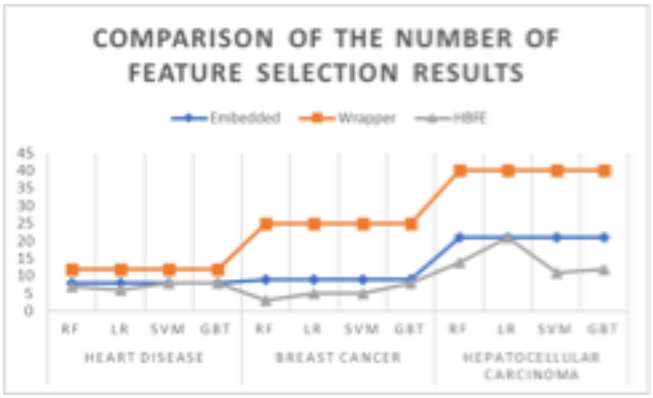

Figure 9. Comparison of the number of feature selection results

The embedded technique and wrapper technique have better accuracy results compared to the embedded technique for each medical record dataset. When compared with the embedded technique, it can be seen with the wrapper technique that there is a decrease in the accuracy value. This is because the wrapper technique is influenced by the parameters of the n-feature to be selected. The proposed hybrid technique is an embedded technique in determining feature weights with NCAFW then combined with a wrapper technique to evaluate features through learning algorithms, producing good accuracy values for each medical record dataset.

Conclusion

Research in determining the attributes in the dataset with the feature selection method at this time has already been done by making a very good contribution and from the results of the research there are still opportunities for further research. The approach through previous research can be seen as an opportunity for improvement or improvement of methods that can contribute to the feature selection technique (feature section). This study aims to contribute to the hybrid method for feature selection algorithms in the field of medical record data or medical records obtained due to problems and challenges to structured and unstructured medical record data so that it can help improve its accuracy and performance in each prediction model. Future research can be developed to develop hybrid techniques to increase the time in selecting features and handling unbalanced datasets.

References

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Dinov, I. D. (2018). Data science and predictive analytics: Biomedical and health applications using R. In Data Science and Predictive Analytics: Biomedical and Health Applications using R. https://doi.org/10.1007/978-3-319-72347-1

Du, G., Zhang, J., Luo, Z., Ma, F., Ma, L., & Li, S. (2020). Joint imbalanced classification and feature selection for hospital readmissions. Knowledge-Based Systems, 200, 106020. https://doi.org/10.1016/j.knosys.2020.106020

Galli, S. (2020). Python Feature Engineering Cookbook. Packt.

Gárate-Escamila, A. K., Hajjam El Hassani, A., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19. https://doi.org/10.1016/j.imu.2020.100330

Ghorbani, R., Ghousi, R., Makui, A., & Atashi, A. (2020). A New Hybrid Predictive Model to Predict the Early Mortality Risk in Intensive Care Units on a Highly Imbalanced Dataset. IEEE Access, 8, 141066–141079. https://doi.org/10.1109/ACCESS.2020.3013320

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. Journal of the American Medical Informatics Association, 24(1), 198–208. https://doi.org/10.1093/jamia/ocw042

Haq, A. U., Zhang, D., Peng, H., & Rahman, S. U. (2019). Combining Multiple Feature-Ranking Techniques and Clustering of Variables for Feature Selection. IEEE Access, 7, 151482–151492. https://doi.org/10.1109/ACCESS.2019.2947701

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395–405. https://doi.org/10.1038/nrg3208

Latif, J., Xiao, C., Tu, S., Rehman, S. U., Imran, A., & Bilal, A. (2020). Implementation and Use of Disease Diagnosis Systems for Electronic Medical Records Based on Machine Learning: A Complete Review. IEEE Access, 8, 150489–150513. https://doi.org/10.1109/ACCESS.2020.3016782

Le, T. M., Vo, T. M., Pham, T. N., & Dao, S. V. T. (2021). A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic. IEEE Access, 9, 7869–7884. https://doi.org/10.1109/ACCESS.2020.3047942

Max Kuhn, K. J. (2019). Feature Engineering and Selection: A Practical Approach for Predictive Models. (Chapman & Hall/CRC Data Science Series) 1st Edition.Panesar, A. (n.d.). Machine Learning and AI for Healthcare.

Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2018). Data lifecycle challenges in production machine learning: A survey. SIGMOD Record, 47(2), 17–28. https://doi.org/10.1145/3299887.3299891

Poongodi, T., Sumathi, D., Suresh, P., & Balusamy, B. (2021). Deep learning techniques for electronic health record (EHR) analysis. Studies in Computational Intelligence, 903, 73–103. https://doi.org/10.1007/978-981-15-5495-7_5

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. arXiv, 22(5), 1589–1604.

Weiskopf, N. G., Hripcsak, G., Swaminathan, S., & Weng, C. (2013). Defining and measuring completeness of electronic health records for secondary use. Journal of Biomedical Informatics, 46(5), 830– 836. https://doi.org/10.1016/j.jbi.2013.06.010

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. In Journal of the American Medical Informatics Association (Vol. 25, Nomor 10, hal. 1419–1428). https://doi.org/10.1093/jamia/ocy068

Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors and Actuators, B: Chemical, 212, 353–363. https://doi.org/10.1016/j.snb.2015.02.025

Zheng, A., & Casari, A. (n.d.). Feature Engineering for Machine Learning.