IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 36, No. 6, December, 2022, pp. 953-957
Journal homepage: http://iieta.org/journals/ria

A Framework for Blended Sub Feature Engineering for Chronic Disease Prediction Using in-Memory Computing

Gunturi S. Raghavendra[1*], Shanthi Mahesh[2], Manukonda Venkata Poorna Chandrasekhara Rao[3]

[1] Computer Science and Engineering, RVR & JC College of Engineering, Guntur 522019, India
[2] Information Science and Engineering, Atria Institute of Technology, Bangalore 560024, India
[3] Computer Science and Business Systems, RVR & JC College of Engineering, Guntur 522019, India

Corresponding Author Email: raghavendragunturi@rocketmail.com

https://doi.org/10.18280/ria.360617

Received: 19 November 2022
Accepted: 20 December 2022

Keywords:
chronic disease, feature selection, sub feature engineering, in-memory computing, prediction

ABSTRACT

Chronic diseases are among the most frequent major health concerns. Early detection of chronic illnesses can help to avoid or lessen their repercussions, potentially lowering death rates. It's an innovative technique to use machine learning algorithms to identify dangerous variables. The problem with existing feature selection procedures is that each method gives a unique collection of features that influence model validity, and current methods are incapable of performing effectively on large multidimensional datasets. We would want to present a novel model that uses a feature selection strategy to choose ideal features from large multidimensional data sets to deliver credible forecasts of chronic diseases while preserving the uniqueness of the data. To assure the success of our proposed model, we used balanced classes by applying hybrid balanced class sampling methods to the original dataset, as well as methods to provide valid data for the training model, characterization and data conversion are required. Our model was run and assessed on datasets with binary and multi-valued classifications. We utilized a variety of datasets (Parkinson's disease, arrhythmia, breast cancer, kidney disease, and diabetes). To select suitable features, the hybrid feature model is used, which includes six ensemble models and involves voting on attributes. The accuracy of the original dataset before applying the framework is recorded and compared to the accuracy of the reduced set of characteristics. The findings are given individually to allow for comparisons. We can conclude from the results that our proposed model performed best on multi-valued class datasets rather than binary class characteristics.

1. INTRODUCTION

Chronic illnesses are widely recognized as the most severe and deadly diseases in humans. The rising prevalence of chronic illnesses with high death rates poses a considerable danger and burden to healthcare systems globally. Chronic illnesses are more common in males than females, especially in middle- and old-age populations, while youngsters with similar health conditions exist. When educated on appropriate data, machine learning algorithms can be excellent in identifying illnesses [1]. Heart disease datasets are freely available for model comparison. With the advent of machine learning and artificial intelligence, academics may now construct the greatest prediction models by employing the massive databases that are now available. The latest evidence on heart-related issues in adults and children has emphasized the importance of lowering chronic disease mortality. Because the clinical datasets supplied are inconsistent and redundant, proper pre-processing is essential. It is critical to choose the key traits that may be employed as risk factors in prediction models. Various supervised models using feature selection methods such as AdaBoost (AB), Decision Tree (DT), Gradient Boosting (GB), Stochastic Gradient Descent (SGD), Lasso Regression (Lassos), and Random Forest (RF) are used in this work, along with classifiers. The findings are compared to previous research [1].

2. RESEARCH OBJECTIVE AND SCOPE

The goal of this study is to create an efficient multi-level feature selection technique. that eliminates extraneous characteristics without compromising data originality to acquire proper features that help in quicker processing and output.

The following are the actions that must be taken:

- To test the performance of a more dependable feature selection model, five datasets are employed.
- Six selection strategies are incorporated, and a hybrid model is built to extract the most important characteristics from medical references based on rank values.
- The framework's performance is assessed using binary and multivalued class datasets.

3. LITERATURE REVIEW

Feature extraction applications pose new challenges in the selection of streaming features [2]. The feature extraction applications have several characteristics, including a) characteristics are evaluated consecutively with a set number of occurrences; and b) the trained model does not exist in advance. In a text mining assignment for spam filtering, for

example, additional features (e.g., words) are dynamically created and must therefore be exploited to filter out the spam instead of waiting for every characteristic to be collected. Traditional methodologies, which have not been developed for streaming information applications, cannot be employed in this situation since they demand that the whole extracted feature set be known beforehand to evaluate the effective attributes effectively and scientifically [2, 3]. Parkinson's disease is a widespread neurological disorder. One of the early warning indicators is a speech or voice issue. Acoustics and voice communication system technologies can evaluate and measure the effects of Parkinson's disease on the voice. The current research proposes a unique feature extraction framework based on two different stages of the subset of features approach for diagnosing voice loss in Parkinson's patients. The PCA and ECFS techniques are computed individually during the initial level of selection, and the selected features from each method are treated as separate lists, namely the ECFS chosen features sublist and the PCA picked features sublist, in the first set [4, 5]. Traditional feature selection techniques presume that prior to learning, all data instances and attributes are known. However, we are much more likely to encounter streams of data, feature streams, or both in many real-world applications. Feature streams are features that appear sequentially over time while the number of training samples remains constant. Existing streaming feature selection algorithms concentrate on removing unneeded and redundant features and selecting the most relevant features, but they ignore feature interaction. A quality might have a poor connection with the goal idea on its own, but when paired with other characteristics, it may have a high correlation with the goal concept [6, 7]. Feature selection is a crucial part of good learning algorithms, and it also helps to explain machine-driven judgments. Algorithms such as decision trees and the Least Absolute Shrinkage and Selection Operator (LASSO) may be used to choose features during training. These embedded techniques, however, are limited to a subset of machine learning models. Wrapper-based approaches, which are often computationally expensive, can extract features from machine learning models. Many randomized strategies are now being developed to improve their efficiency [8]. Because rolling bearings are one of the most important components of rotating machinery, this study proposes a balanced multidimensional feature fusion-based feature selection and fusion technique. To begin, features are taken from many domains to construct the initial high-dimensional feature collection. Given the huge number of erroneous and redundant features in the initial feature set, a feature selection strategy was developed that incorporated support vector machines (SVM), single feature analysis, correlation coefficients, and principal component analysis. weighted load analysis [8]. Multi-label feature selection has increased in popularity in deep learning, statistic processing, and related domains, and it is currently widely used for a broad range of tasks ranging from music identification to text mining, picture annotation, and so on. Traditional multi-label attribute selection procedures, on the other hand, employ a cumulative accumulation strategy to create solutions, which has the disadvantage of overvaluing the redundancy of some candidate features. As a data preparation strategy, feature selection has been proven to be successful and effective in preparing data (especially high-dimensional data) for various machine learning and data mining concerns. The aims of feature engineering include producing simpler and more transparent models, increasing data mining performance, and giving clean, intelligible data. The growing amount of big data has presented some important challenges and opportunities for feature selection. In this research, we give a complete and systematic assessment of current advances in attribute selection studies. Motivated by current difficulties and opportunities in the big data era, we investigate feature selection research from a data perspective and examine sample feature selection strategies for traditional data, structured data, heterogeneous data, and streams [9]. Because of the vast number of different, highly changeable inputs in today's decision-making system, feature selection is unavoidable. It is vital to pick the suitable feature set to prevent computational strain and educate algorithms to function correctly. If only a few elements are chosen, it is possible that adequate results may not be obtained; if many characteristics are selected, productivity may decrease. To improve accuracy, more salient points can be included [10].

4. RESEARCH METHODOLOGY

Following data pre-processing and exploratory data analysis, the data set is examined for missing values and negative values, which are subsequently handled by the machine learning pipeline. To address this imbalance, rectify these concerns, and minimize extended execution times, three alternative balancing strategies, such as smite, random, and smotek sampling procedures, are applied. This aids in the creation of the finest data set. Classifiers with balanced classes will have better performance. All ensemble feature selection models with classifiers are developed to compare the binary and multi-valued class label datasets. For evaluating the data set, many training models have been provided so that we may select the best characteristics for our dependable data set. Furthermore, the most relevant aspects of a chronic illness patient have been provided in this diagnosis approach.

5. JUSTIFICATION OF THE PROPOSED TECHNIQUE

Recursive feature selection and the six ensemble feature selection methods were used to create this system. There have previously been several studies on various sorts of feature selection algorithms based on classifiers. We picked three of the most often used techniques (DT, RF, and LASSO) as well as three less commonly used approaches (SGD, GB, and ADA). A previous study has discovered that existing feature algorithms have a high level of expected accuracy when compared to other existing approaches. In addition, a limited number of tests have revealed that the ensemble work of wrapper-based feature selection may perform rather well with extremely high accuracy. Except for DT and kNN, none of the research endeavors came close to our proposed techniques as a basic classifier. As a result, all of the prior approaches have been studied further in this work, utilizing ensemble techniques to make the proposed model more efficient. Although the literature review shows that hypotheses put forward yielded good prediction accuracy, it was not high enough in comparison to our investigation.

5.1 The purpose of feature selection

The goal of feature selection in ML is to determine the

optimal collection of characteristics that will allow one to develop meaningful and constructive models of the subject under investigation. Machine learning feature selection approaches are classified as follows:

5.1.1 Supervised

These approaches are utilized for labeled data and to categorize the important characteristics for boosting the effectiveness of supervised models such as classification and regression.

5.1.2 Unsupervised

These approaches are applied to unlabeled data.

6. IMPLEMENTATION

Figure 1 displays two crucial stages in the feature selection process: subset construction and subset assessment. The subset synthesis engine finds feature subset candidates, while the subset evaluation engine evaluates the quality of the subsets. Finally, a halting condition is evaluated at each iteration to terminate the procedure. There are three sorts of methods for selecting features: filter, wrapper, and hybrid. Wrapper approaches rely on a classification algorithm to evaluate feature subsets. In general, because the feature selection process is tuned for the given classification method, the wrapper technique outperforms the filter approach [11]. To assess and choose feature subsets, filter techniques employ independent criteria that are based on the general properties of the data rather than a classification algorithm. Common evaluation functions are often distance, mutual information (MI), dependence, or entropy measurements computed directly from training data [11]. It employs a filter-based strategy to choose highly representative features and a wrapper-based technique to add candidate features and assess candidate subsets in order to identify the best ones. Starting with an empty set, the sequential forward search (SFS) approach adds one feature subset throughout each round until a new extracted feature that maximizes the criteria function value is found, whereas the backward elimination search (SBS) approach commences with a full feature subset and removes a feature on each iteration until a preset criterion is fulfilled [11]. Simple tools like Pandas, Pyplot, and Scikit-learn are used to build the model, which is written in Pyspark (Python) and runs on an Apache Spark cluster.

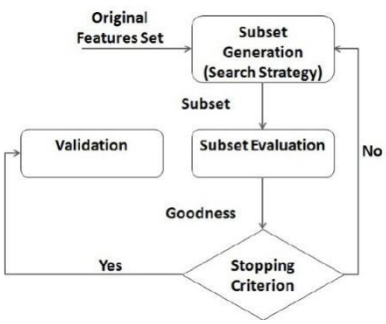

Figure 1. Process of feature selection

6.1 Dataset

Data is the first and most important component in using machine learning algorithms to get reliable results. The dataset used in this study was obtained from the "UCI machine learning repository," which is a well-known data repository. Six distinct datasets are available: Parkinson's illness, arrhythmia, breast cancer, renal disease, diabetes... To obtain more exact findings, we combined all of them into one study. As a text label, more than 1190 instances from their database are collected, together with 14 distinguishing qualities. For example, the 13 attributes of the diabetes dataset are utilized as diagnostic inputs, whereas the "number" property is employed as an output All or most of the following six medically significant variables were included in all or most records: Resting blood pressure (trestbps), fasting blood sugar (fbs), kind of chest pain (cp), and resting electrocardiographic data (restecg). Table 1 describes the different attributes and datasets used.

Table 1. Input Datasets and Count of Attributes

Dataset name	No of original attributes
Parkinson Dataset	755
Arrythmia Dataset	280
Kidney	53
Diabetes	14
Breast Cancer (GSE)	3000

6.2 Reason for selection of above datasets

The above datasets were chosen because they all belong to chronic disease datasets with attributes ranging from tens to thousands. The framework was tested on variable counts of attributes.

7. HYBRID-FRAMEWORK

7.1 Recursive feature elimination (RFE)

It is a popular feature selection method. RFE is popular because it is easy to set up and use, and it is effective at identifying features (columns) in a training dataset that are relevant for predicting the target variable. When using RFE, there are two critical configuration variables to consider: the number of features to select and the technique used to aid in selecting features. Both model parameters can be explored, although their importance in the method's effectiveness cannot be overstated [12, 13].

7.2 RFE with random forest

Random Forest is a well-known machine learning technique for supervised learning. It may be used in machine learning for both classification and regression problems. It is based on the concept of ensemble learning, which is the process of combining several classifiers to solve a complex problem and improve the model's performance [14, 15].

7.3 RFE with decision tree

A decision tree is a non-parametric supervised learning approach that may be used for both classification and regression. It has a tree structure with a root node, branches, internal nodes, and leaf nodes. A decision tree begins with a root node that has no incoming branches. Outgoing branches from the root node feed into internal nodes, also known as decision nodes. Based on the specified features, both node

types perform assessments to yield homogeneity subsets known as leaf nodes or terminal nodes [16,17].

7.4 RFE with Lasso

"LASSO" is an acronym that stands for "Least Absolute Shrinkage and Selection Operator." This model employs shrinkage. Shrinkage simply means that the data points are realigned by imposing a penalty that causes the coefficients to be reduced to zero if they are not significant. It makes use of the L1 regularization penalty method. This type of regression is most suited for models with a high degree of multi-collinearity or when we need to automate model selection aspects, such as parameter removal or feature selection [18, 19].

Lasso loss functions can be represented mathematically.

$$\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \alpha \sum_{j=1}^{m}|\beta_j|$$

7.5 Reasons for selecting above methods

The above methods deal with large datasets and reduce the chances of overfitting the data. Recursive feature elimination with the above methods provides output effectively.

8. RESULTS AND DISCUSSION

Table 2. Reduced set of attributes after applying framework

Dataset name	No of original attributes	Reduced attributes (Using hybrid feature selection)
Parkinson Dataset	755	32
Arrythmia Dataset	280	16
Kidney	53	6
Diabetes	14	5
Breast GSE	3000	289

The framework is assessed using binary and multi-valued class properties. Table 2 Discuss reduced features: The number of characteristics in the Parkinson data set has decreased from 755 to 32 attributes. The multivalued Arrythmia data set's 279 characteristics are reduced to 16 features. The kidney's 53 qualities have been reduced to 6. Diabetes has been lowered from 14 to 6. Breast cancer GSE decreased from 3000 to 158. This framework worked better on multi-valued classes than on binary class characteristics.

The above datasets are applied on proposed model and has attained output of above number of features.

8.1 Comparing framework with individual methods

Table 3 discusses the comparison of the framework with existing methods as follows:

1. The Parkinson dataset is given as input for existing methods like univariate selection (filter-based feature selection), regular RFE, RFE with decision trees, and RFE with random forests, and results are compared with those of the proposed framework.
2. The Arrythmia dataset is given as input for existing methods like univariate selection (filter-based feature selection), regular RFE, RFE with decision trees, and RFE with random forests, and results are compared with those of the proposed framework.
3. The kidney dataset is given as input for existing methods like univariate selection (filter-based feature selection), regular RFE, RFE with decision trees, and RFE with random forests, and results are compared with those of the proposed framework.
4. The Breast GSE dataset is given as input for existing methods like univariate selection (filter-based feature selection), regular RFE, RFE with decision trees, and RFE with random forests, and results are compared with those of the proposed framework.

8.2 Advantages of framework

The proposed framework is reusable, and it requires no input from the user except a mask value (Threshold and voting count).

Table 3. Comparison of existing and proposed methods of feature selection

Feature selection method	Dataset name	No of attributes	Reduced attributes	Framework
Univariate selection	Parkinson Dataset	755	98	32
RFE	Parkinson Dataset	755	128	
RFE With DT	Parkinson Dataset	755	377	
RFE With RF	Parkinson Dataset	755	245	
Univariate selection	Arrythmia Dataset	279	46	16
RFE	Arrythmia Dataset	279	78	
RFE With DT	Arrythmia Dataset	279	92	
RFE With RF	Arrythmia Dataset	279	67	
Univariate selection	Kidney	53	23	6
RFE	Kidney	53	9	
RFE With DT	Kidney	53	14	
RFE With RF	Kidney	53	19	
Univariate selection	Breast GSE	3000	568	
RFE	Breast GSE	3000	768	289
RFE With DT	Breast GSE	3000	456	
RFE With RF	Breast GSE	3000	890	

8.3 Extension of framework

This framework is currently applied to chronic diseases, but it has the potential to be extended to other applications such as stock market and social media analytics, where attributes are more prevalent.

8.4 Limitations of framework

This framework is currently applicable to only structured and semi-structured data and can be extended to images.

8.5 Risk factors of framework

There are no risk factors for implementation of framework it is developed using spark (Apache spark python programming).

9. CONCLUSION

To ensure the success of our proposed model, we employed balanced classes to provide credible data for the training model by employing hybrid balanced class sampling techniques on the original dataset, as well as data preparation and transformation methods. On datasets with binary and multivalued classifications, we ran and tested our model. We used several datasets (Parkinson's, arrythmia, breast cancer, kidney, diabetes). The hybrid feature model is used to select appropriate features, which includes LassoCV, decision trees, random forests, gradient boosting, Ada-boost, stochastic gradient descent, and attribute voting. Before using the framework, the accuracy of the original dataset is recorded and compared to the accuracy of the reduced set of characteristics.

The above framework deals with several cons of using traditional methods of feature selection, like input from the user and performance decreasing with increased datasets.

REFERENCES

[1] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F.J.M., Ignatious, E., Shultana, S., Beeravolu, A.R., De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9: 19304-19326. https://doi.org/10.1109/ACCESS.2021.3053759

[2] Lima, H.C., Otero, F.E., Merschmann, L.H., Souza, M.J. (2021). A novel hybrid feature selection algorithm for hierarchical classification. IEEE Access, 9: 127278-127292. https://doi.org/10.1109/ACCESS.2021.3112396

[3] Almusallam, N., Tari, Z., Chan, J., Fahad, A., Alabdulatif, A., Al-Naeem, M. (2021). Towards an unsupervised feature selection method for effective dynamic features. IEEE Access, 9: 77149-77163. https://doi.org/10.1109/ACCESS.2021.3082755

[4] Ashour, A.S., Nour, M.K.A., Polat, K., Guo, Y., Alsaggaf, W., El-Attar, A. (2020). A novel framework of two successive feature selection levels using weight-based procedure for voice-loss detection in Parkinson's disease. Ieee Access, 8: 76193-76203. https://doi.org/10.1109/ACCESS.2020.2989032

[5] Zhou, P., Li, P., Zhao, S., Wu, X. (2020). Feature interaction for streaming feature selection. IEEE Transactions on Neural Networks and Learning Systems, 32(10): 4691-4702. https://doi.org/10.1109/TNNLS.2020.3025922

[6] Wang, Z., Xiao, X., Rajasekaran, S. (2020). Novel and efficient randomized algorithms for feature selection. Big Data Mining and Analytics, 3(3): 208-224. https://doi.org/10.26599/BDMA.2020.9020005

[7] Li, Y., Dai, W., Zhang, W. (2020). Bearing fault feature selection method based on weighted multidimensional feature fusion. IEEE Access, 8: 19008-19025. https://doi.org/10.1109/ACCESS.2020.2967537

[8] Gao, W., Hu, J., Li, Y., Zhang, P. (2020). Feature redundancy based on interaction information for multi-label feature selection. IEEE Access, 8: 146050-146064. https://doi.org/10.1109/ACCESS.2020.3015755

[9] Li, L., Cheng, K.W., Wang, S.H., Morstatter, F., Trevino, R.P., Tang, J.L., Liu, H. (2018). Feature selection. ACM Computing Surveys, 50(6): 1-45. https://doi.org/10.1145/3136625

[10] Vandana, C.P., Chikkamannur, A.A. (2021). Feature selection: An empirical study. International Journal of Engineering Trends and Technology, 69(2): 165-170. https://doi.org/10.14445/22315381/ijett-v69i2p223

[11] Homsapaya, K., Sornil, O. (2018). Modified floating search feature selection based on genetic algorithm. In MATEC Web of Conferences, 164: 01023. https://doi.org/10.1051/matecconf/201816401023

[12] Raghavendra, G.S., Mahesh, S. and Rao, M.V.P.C.S. (2021) Prediction of accuracy in emergency health records using hybrid machine learning model. Journal of Pharmaceutical Research International. 33(58A): pp. 206-212. https://doi.org/10.9734/jpri/2021/v33i58A34107

[13] Liu, W., Wang, J. (2021). Recursive elimination–election algorithms for wrapper feature selection. Applied Soft Computing, 113: 107956. https://doi.org/10.1016/j.asoc.2021.107956

[14] Padmashree, A., Krishnamoorthi, M. (2022). Decision Tree with Pearson Correlation-based Recursive Feature Elimination Model for Attack Detection in IoT Environment. Information Technology and Control, 51(4): 771-785. https://doi.org/10.5755/j01.itc.51.4.31818

[15] Chen, X.W., Jeong, J.C. (2007). Enhanced recursive feature elimination. In Sixth International Conference on Machine Learning and Applications (ICMLA 2007), pp. 429-435. https://doi.org/10.1109/icmla.2007.35

[16] Gumma, L.N., Thiruvengatanadhan, R., Lakshmi, P.D., LakshmiNadh, K. (2022). A Binary multi class and multi level classification with dual priority labelling model for COVID-19 and other thorax disease detection. Revue d'Intelligence Artificielle, 36(5): 657-664. https://doi.org/10.18280/ria.360501

[17] Karunanithi, A., Singh, A.S., Kannapiran, T. (2022). Enhanced hybrid neural networks (CoAtNet) for paddy crops disease detection and classification. Revue d'Intelligence Artificielle, 36(5): 671-679. https://doi.org/10.18280/ria.360503

[18] Basysyar, F.M., Dwilestari, G. (2022). House price prediction using exploratory data analysis and machine learning with feature selection. Acadlore Transactions on AI and Machine Learning, 1(1): 11-21. https://doi.org/10.56578/ataiml010103

[19] Wawage, P., Deshpande, Y. Real-time prediction of car driver's emotions using facial expression with a convolutional neural network-based intelligent system. Acadlore Transactions on AI and Machine Learning, 1(4): 22-29. https://doi.org/10.56578/ataiml010104