University of Reading

Improved filter-based feature selection using correlation and clustering techniques

Book or Report Section

Accepted Version

improved fitering.pdf

Atmakuru, A., Di Fatta, G., Nicosia, G. and Badii, A. (2024) Improved filter-based feature selection using correlation and clustering techniques. In: Nicosia, G., Ojha, V., La Malfa, E., La Malfa, G., Pardalos, P. M. and Umeton, R. (eds.) Machine Learning, Optimization, and Data Science: 9th International Conference, LOD 2023, Grasmere, UK, September 22–26, 2023, Revised Selected Papers, Part I:. Lecture Notes in Computer Science. Springer, pp. 379-389. ISBN 9783031539688 doi: 10.1007/978-3-031-53969-5_28 Available at https://reading-pure-test.eprints-hosting.org/120660/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1007/978-3-031-53969-5_28

Publisher: Springer

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in

the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Improved Filter-Based Feature Selection using Correlation and Clustering Techniques

Akhila Atmakuru[1] Giuseppe Di Fatta[2] Giuseppe Nicosia[3] and Atta Badii[4]

[1] University of Reading, UK, akhila.atmakurupt@gmail.com
[2] Free University of Bozen-Bolzano, Italy, giuseppe.difatta@unibz.it
[3] University of Catania, Italy, giuseppe.nicosia.1@gmail.com
[4] University of Reading, UK, atta.badii@reading.ac.uk

Abstract. Feature engineering and feature selection are essential techniques to most data science and machine learning applications, in which, respectively, raw data are transformed into features and features are selected to provide the most effective subset of features for the application. Feature selection techniques are particularly useful when dealing with high-dimensional datasets that contain noisy and redundant data. An optimised feature subset could enhance the performance as well as the interpretability of the model. There are three types of feature selection methods, namely filter, wrapper and embedded techniques. Amongst these methods, the filter method is more efficient than the others as it is computationally less expensive and more generalised. This work presents two improved filter-based feature selection methods based on a correlation coefficient and clustering techniques. The first approach is based on feature correlation where the feature subset consists of features above a similarity threshold to identify a kind of neighbourhood for each feature. The second method uses clustering analysis on the correlation data to identify features that can be used to represent the entire cluster. The obtained feature subsets have been applied as pre-processing step for logistic regression and artificial neural networks. The performance of the proposed methods has been compared against the popular ReliefF feature selection method. The experimental analysis shows that the proposed feature selection methods provide an observable improvement in accuracy by choosing the most effective features.

Keywords: Feature Selection, Correlation, Clustering, Principal Coordinate analysis, Neural Network and High DimensionalDataset.

1 Introduction

Machine learning models can identify patterns in data and make predictions accurately based on identified patterns. However, raw data are typically noisy and contain redundant and irrelevant information: machine learning models can be quite inaccurate and ineffective when applied directly to raw data.

Feature selection is the process of choosing the most effective subset of features from raw data. The performance of a machine learning model is improved by minimising the number of the less effective, redun-

dant, or noisy features in the dataset. Identifying and ranking the most effective features, enables researchers to have a better grasp of the relationships between the features and the target variable, making the output of a machine learning model more robust. Feature ranking could also aid in reducing overfitting, increasing model accuracy, reducing model complexity and training time. There are three different feature selection methods, namely filter, wrapper and embedded methods [1]. The filter-based methods rank each individual feature or an entire subset of features based on measures such as information, distance, consistency, similarity (correlation), and other statistical measures. Feature correlation is typically computed as a matrix representation of the relationship between each possible pair for all the variables and is useful to summarise a large dataset and visualise similarity patterns within the features. The performance of filter-based methods does not depend on to the data modelling approach chosen for the application. Thus, compared to the wrapper and embedded techniques, the filter- based approaches are more general and computationally less expensive since they are independent of the learning algorithm. ReliefF, Correlation based Feature Selection (CFS), and the max-relevancy min-redundancy (mRMR) feature selection algorithm are a few examples of filter-based techniques.

The wrapper methods use searching algorithms to identify the significant feature subsets, which are ranked based on their performance when applied to a specified modelling method. The creation of each subset is reliant on the search technique, and the evaluation is repeated for each subset. Since they are dependent on the resource requirements of the modelling process, wrappers are substantially slower than filters at identifying the more effective subsets. The feature subsets thus identified, are skewed in favour of the modelling approach that was used to rank them.

The embedded methods perform feature selection during the execution of the modelling algorithm. By maximising a performance parameter, such as accuracy or error rate, these techniques seek to identify and choose the more effective features during training. These methods are embedded in the algorithm either as its normal or extended functionality. Due to their iterative process, embedded-based approaches can be computationally costly. Moreover, they require more domain expertise than filter or wrapper approaches since they change an existing learning algorithm rather than just rank the efficacy of feature subsets independently of it, as is the case with other feature selection methods.

The ReliefF algorithm assumes that the variables are linear in nature and suitable for a binary classification problem. It calculates the significance of each feature relative to its performance target. Features with high correlation are retained. and so, some redundant features remain making the algorithm inefficient. The Correlation based Feature Selection (CFS) algorithm computes correlation for a subset of features. The method is then repeated for several feature subsets until a subset with a lower average of feature-to-feature correlation and a higher average of feature to target correlation is determined. The CFS algorithm based on iterative heuristic search strategies is computationally expensive as it requires repeated computation of correlation between features and target. The max-relevancy min-redundancy (mRMR) feature selection algorithm is based on mutual information. The feature relevancy, i.e., the relative effectiveness of a feature, is determined by mutual information computed between individual features and the target class. The redundant features are identified by applying mutually exclusive conditions. The algorithm is computationally expensive as many mutual information computations are required.

Another prominent strategy for identifying a subset of the relatively more significant features in large datasets is clustering-based feature selection. The characteristics features are clustered based on their closeness, and the best representative feature from each cluster is then selected. The disadvantage of the clustering methods arises from the loss of performance due to overlap of clusters and influence of outliers.

There are mainly four types of clustering techniques. The correlation-based clustering involves clustering features according to their correlation coefficient. The most representative feature from each cluster is chosen after the features with strong correlation are clustered together. K-Means clustering algorithm groups the features depending on their distance from the cluster centroid. The most representative feature of each

cluster is chosen as a subset of features. Hierarchical clustering approaches group the features into hierarchical structures based on how similar these are. At various levels of the hierarchical structure, the most representative feature is chosen from each cluster. However, this approach, is inefficient while handling missing data or large datasets. The spectral clustering approach groups the features using a graph representation. The most representative feature of each cluster is chosen as a subset of features. computing eigenvectors for a high dimensional dataset can cause bottlenecks in this approach.

This paper proposes two distinct strategies for implementing feature selection. These are based on a correlation matrix generated from a raw dataset. The first technique analyses the correlation matrix to find key attributes that will be used to represent the neighbourhood. The second technique utilises an unsupervised clustering method on the correlation matrix to identify the key features that represent the complete cluster.

The findings showed that the features obtained from these techniques improved the performance interpretability by identifying key features and the corresponding performance.

The remainder of this paper is structured as follows: Section 2 presents an overview of the literature on current feature selection approaches. Section 3 provides a description of the dataset that was used in this work. Section 4 discusses the approach of the feature selection strategies that have been developed. Section 5 includes the Results and Discussion which: summarises the methodology outcomes utilising suggested feature selection methods. Section 6 concludes this paper by discussing the future trajectory of the work.

2 Literature Review

While dealing with a high dimensional dataset, it is beneficial and efficient to use filter-based feature selection methods as they do not need repeated training and evaluation and are independent of the model being used for training. The literature study will focus on methods based on correlation and clustering methods.

The widely used feature selection method known as "Relief" assumes the features are linear in nature and only works for two-class situations. This method is based on a distance-based metric function that weights each feature according to how relevant (correlation) it is to the target class [2]. The features with higher correlation to the target and lower correlation to other features were retained. The ReliefF [3], a Relief variant, can deal with multi-class problems and noisy datasets, but it is inefficient due to redundant features. ReliefF has been augmented by various other algorithms to handle the redundant features issue.

Another popular method for correlation-based feature selection (CFS) [4] is based on a heuristic best first search approach. In this algorithm, correlation between features, and features to the target is computed. The algorithm then determines which features could be used together in a subset in a heuristic manner. The algorithm repeats the process for several feature subsets until a subset containing a lower average of feature-to-feature correlation and higher average of feature-to-the-target correlation is identified. The algorithm is computationally expensive for a high dimensional dataset.

Another Fast Correlation-Based Filter (FCBF) [5] algorithm removes both irrelevant and redundant features by calculating the symmetrical uncertainty which is the measure of relevance between the feature and the target. Beginning with the whole feature set, FCBF adopts a heuristic backward selection approach combined with a sequential search strategy to eliminate unnecessary and duplicate features. The algorithm terminates once there are no more characteristics to discard. The algorithm is able to reduce the features, while still maintaining high accuracy. However, the algorithm is computationally expensive.

The algorithm maximum Relevancy Minimum Redundancy (mRMR) [6] is based on Mutual Information (MI), which is computed between the individual feature and target class for identifying feature relevancy. To handle the redundant features, mutually exclusive conditions are applied.

For clustering methods, one of the algorithms [7] is based on graph clustering to identify similar features for removing redundant features. Each cluster is ranked based on similarity measures, thereafter the top-ranked clusters are selected and the representative features retained. Another method [8] clusters the features based on the clustering chi-square statistical measure and then selects the most representative feature for each cluster. This approach utilises a modified K-Means clustering algorithm to form clusters. The method Multi Cluster feature selection [9] is based on a spectral clustering algorithm on the correlation similarity matrix between the features. The importance of each feature is assessed within each cluster using a non-parametric Wilcoxon rank-sum test, and the most relevant features are identified.

A clustering method based on mutual information [10] starts with splitting the dataset into training and testing sets and then groups the features in the training set according to how relevant they are to the classification task using a supervised clustering algorithm. The clustering technique uses the mutual information between each feature and the class label as the distance metric. The most representative characteristic from each cluster is then chosen as the final set of features after the clusters are sorted according to their mutual information with the class label. A classifier is then trained using the selected features on the training set, and its performance is assessed using the testing set.

A similar approach [11] uses a graph partitioning clustering algorithm. The data is represented as a graph and partitioned into clusters to select the most relevant feature from each cluster. Another method [12] is based on similarity measures to form clusters using hierarchical clustering algorithms. At each hierarchical level, both the intra-cluster and inter-cluster similarity score is computed based on which the features with each cluster are retained. This process continues until stopping criteria are met or the specified number of clusters are obtained.

An approach for arrhythmia classification using a feature selection schema for creating an ensemble of classifiers [13] improves accuracy in high dimensional datasets by identifying relevant feature sets that affect classification. Multiple feature subsets are extracted, and classification models are built based on each subset. The models are then combined using a voting approach to calculate each classifier score in the ensemble. In an experimental study, the proposed method generated three top distinct feature sets, and three classifiers were constructed based on these subsets. The classifier ensemble using the voting approach significantly improved classification accuracy in high dimensional datasets and led to a more stable classification model. The performance of each classifier and the ensemble was compared to a classifier using the entire feature space of the dataset.

Different feature selection methods have been compared [14] on the arrhythmia dataset to improve the accuracy of machine learning models for diagnosing heart disease. The two methods used are filters and wrappers, both of which perform well, but filters are faster. The paper finds that random forest with the wrapper method has the highest accuracy. The paper also suggests further investigation into the correlation of features and optimization for learning machines targeted on the arrhythmia dataset. A Gaussian naive Bayes classifier is also trained, and the paper finds that filters greatly improve the accuracy of this classifier. Support Vector Machines (SVMs) are also used, and the paper finds that an SVM classifier with feature selection methods scores higher on the test set than that on the train set.

3 Dataset Description

The Arrhythmia Dataset is a widely used dataset for research in the field of cardiac arrhythmia detection and classification. The dataset was initially created and made publicly available by the University of California, Irvine (UCI) Machine Learning Repository. The specific source reference for the dataset is "Arrhythmia Data Set" in the "UCI Machine Learning Repository" [15].

This dataset was collected to aid in the identification of various types of cardiac arrhythmias from regular 12-lead ECG recordings. It consists of records from 452 patients, with each record containing 280 features. The features encompass clinical measurements derived from ECG signals, such as QRS length, R-R interval, P-R interval, Q-T interval, and other demographic information including sex, age, weight, and a cardiologist's recommendation.

The primary objective of using this dataset is to classify patients into two categories: normal patients and those with one of the 15 different types of arrhythmias. The dataset poses challenges due to its high dimensionality and the presence of missing values, which require appropriate pre-processing techniques.

Researchers and machine learning practitioners often utilise this dataset as a benchmark for developing and evaluating arrhythmia detection algorithms, feature selection methods, and classification models.

The dataset contains features that need pre-processing, some of which have missing values. In order to address this, the missing values were replaced with 0. Furthermore, certain features were removed from the dataset due to their numerous NaN values. Despite this, there are still 262 features available for analysis. In this paper, target variables are divided into two classes: normal patients and arrhythmia patients.

4 Methodology

This section describes the feature selection methods developed using Python. The feature selection methods are based on the correlation between the input features. An Arrhythmia dataset was used to perform the feature selection. The developed methods were further tested on machine learning algorithms and neural network models to compare their performance.

4.1 CGN-FS: Correlation-based Greedy Neighbourhood Feature Selection method

The method involves computation of feature-to-feature correlation values, thresholding the values and selecting the features having the lowest correlation. This approach will hereafter be referred to as the Correlation-based Greedy Neighbourhood Feature Selection method (CGN-FS). The anticipated output of the algorithm is a subset of features from the main dataset.

To pick the features, first a correlation matrix is generated. The correlation matrix would include the correlation coefficient for each feature in relation to each of the other features in all possible pairs. The diagonal value of the correlation matrix, for each feature, represents the correlation coefficient value of the feature with itself which is equal to 1. The diagonal values are therefore excluded from the analysis. The absolute values of the element of the correlation matrix are then obtained. The correlation magnitude is represented by an absolute value in the matrix; the higher the value of the element, the more strongly the variables are related to one another.

The absolute correlation values of the matrix are utilised to calculate various evaluation metrics, such as 'sum' and 'count.' The cumulative total of the correlation values of each feature concerning all other features determines the ‘sum’ for each feature. The cumulative total for each feature is appended to the absolute correlation matrix under the column ‘sum’.

Next, the count is determined by applying a threshold to the correlation values, which can range between 0 and 1.00. Further analysis excludes the correlation values of the feature with respect to all other features, while the features that exceed the threshold are identified as the neighbours of the feature. The count value, which corresponds to the number of these neighbours, is then added to a new column called 'count' in the absolute correlation matrix.

The above metrics are computed for different threshold values. At each threshold value the selected features are evaluated by two Machine Learning classifiers. The threshold value with best performance is finalised as the optimum threshold value. Starting with 0.50, the various threshold values used for testing are increase by 0.05. After deriving the necessary metrices, the matrix is organised in descending order using the primary and secondary columns of "Count" and "Sum". Additionally, a flag labelled "Keep" is initialised for each feature. The neighbouring features are then marked as "Remove," while the flag for each feature remains as ‘Keep’. Ultimately, the list of features with the "Keep" flag exclusively assigned is retrieved. This final list represents the anticipated subset of features selected from the entire range. The process can be replicated on various datasets to obtain the chosen subset of features.

Two models have been evaluated to assess their performance for each threshold value. With the help of a 10- fold stratified cross-validation approach, these models are trained using a final subset of features that were generated using the CGN-FS methodology. For each of the two models, the accuracy mean and standard deviation were computed. The best performance is the one where the accuracy is maximum while the standard deviation of the accuracy is at a minimum.

After calculating the needed components, the matrix is sorted using the primary and secondary columns of "Count" and "Sum" in decreasing order. A flag labelled "Keep" is initialised for each feature as well. The neighbours of each feature are then marked as "Remove," while the flag for each feature itself remains as ‘Keep’.

Finally, the list of features that only have the flag "Keep" assigned is retrieved. The expected subset of features chosen from all the features is represented through this final list. This process could be repeated on different datasets to obtain the selected subset of features.

The proposed method, Correlation-based Greedy Neighbourhood Feature Selection (CGN-FS), presents various advantages and limitations. On the positive side, CGN-FS efficiently detects pertinent features by means of feature-to-feature correlation computation. This results in an enhanced classification performance through the selection of informative features. The flexibility of the method enables customisation of the feature selection process by adjusting threshold values, making it applicable to various datasets and classification tasks. Nevertheless, the effectiveness of the approach is heavily dependent on the choice of the optimal threshold value, which can be challenging and reliant on the dataset. To maximise the benefits of CGN-FS, careful threshold selection and consideration of alternative evaluation metrics are crucial.

Algorithm 01: Correlation-based Greedy Neighbourhood Feature Selection Method (CGN-FS)

Input: High Dimensional Dataset

Threshold: To be chosen after thorough analysis

Output: Subset of features

Procedure:

1. Calculate the feature correlation using Pearson Method.
2. Obtain the absolute values of the feature correlation matrix.
3. Calculate the sum of each features correlation values w.r.t all other features.
4. For each feature i, count the number of correlation values above threshold w.r.t to all other features and identify these features as neighbours(i). The number of neighbours is 'count' value.
5. Sort by decreasing order of count (primary) and sum (secondary).
6. Initialize flag as 'Keep' for all features.
7. For each feature i, if Flag(i) is "Keep" mark the features in the neighbours(i) as 'Removed".
8. Return the features with flag "Keep".

Fig. 1. Algorithm for CGN-FS: Correlation-based Greedy Neighbourhood Feature Selection method

4.2 RCH-FSC: Region and Correlation based Heuristic Feature Selection with Clustering Analysis

The clustering analysis uses an arrhythmia dataset, which is a high-dimensional dataset, as input for the analysis. The objective of this method is to carry out clustering of the input features and feature selection by selecting the features that represent the complete clusters produced by the clustering algorithm. The analysis used K-medoids clustering, which is an improved version of the conventional K-means clustering and is more resistant to noise and outliers. A medoid indicates the centre of a cluster as a data-point in the cluster rather than the mean point. Medoid is the data-point in the cluster that is nearest to the centre and has the least total distances from other locations. This developed method will be termed as Region and Correlation-based Heuristic Feature Selection with Clustering Analysis (RCH-FSC) method. This method is expected to produce a lower number of final subset features while not compromising on the performance. Initially, a correlation matrix is created for the dataset. The correlation matrix will include the correlation coefficients for each feature in relation to each other feature in all conceivable pairs. Next, the absolute correlation matrix is determined by taking absolute values.

The distance matrix for the absolute correlation matrix is calculated. Knowledge of the linear or non- linear dependency between the two variables might be gained from the distance matrix.

Next, the distance matrix is subjected to dimensionality reduction. For this purpose, a Principal Coordinate Analysis (PCoA) method, an improved version of the classical Principal Component Analysis, variant Multi-dimensional Scaling (MDS) is used. PCoA is a graphical depiction of the matrix in a low dimensional space. This approach emphasises the differences between the features by visualising similarity of features. In PCoA with MDS, the issues arising due to unstable eigen decomposition and non-linear dissimilarities are handled by the algorithm. The number of principal components is determined by using a cumulative explained variance with a threshold of 50%. The threshold used for an explanation was 50% and the resultant number of principal components were 219 as a result.

The K-medoids clustering method is applied to the resultant PCoA data that has been generated. The elbow technique and silhouette score are used to calculate the optimal number of clusters. The outcomes of the silhouette score depend on how well each feature fits into the cluster, which is crucial to the goal of the

developing method. Thus, the silhouette score determines the K-value which is the number of clusters. K-medoids attempt to lower the total sum of dissimilarities between the data-points of a cluster and its medoid. The datapoint in the cluster with the most central location is a medoid. As the medoids reflect the complete cluster, these medoids represent the final subset of the selected features.

The RCH-FSC Method, which stands for Region and Correlation-based Heuristic Feature Selection with Clustering Analysis, presents various benefits for selecting features in datasets with high dimensions. Through the utilisation of the K-medoids clustering algorithm, it can identify features that are representative and form complete clusters. This makes for a robust selection process, particularly when dealing with noise and anomalies. To lower dimensionality and capture both linear and non-linear relationships among characteristics, the principal coordinate analysis (PCoA) using Multi-dimensional Scaling (MDS) is additionally implemented. By means of silhouette scores, this approach can efficiently determine the optimal number of clusters, thereby ensuring a more effective feature subset. Furthermore, the algorithm strives to attain a compromise between reducing the feature set and classification effectiveness, resulting in a smaller feature selection without any significant decrease in overall accuracy.

Algorithm 02: Region and Correlation based Heuristic Feature Selection with Clustering Analysis Method (RCH-FSC)
Input: High Dimensional Dataset **Output:** Subset of features
Procedure: 1. Calculate the feature correlation using the Pearson Method. 2. Obtain the distance matrix from the correlation matrix. 3. Perform Principal Co-Ordinate Analysis with Multi Dimensional Scaling. 4. Perform K-Medoids Clustering analysis. 5. Identify one feature to represent each and entire cluster. 6. Identified features are the final subset.

Fig. 2. Algorithm for RCH-FSC: Region and Correlation based Heuristic Feature Selection with Clustering Analysis

5 Results and Discussion

5.1 Correlation Method

For clustering-based method, the arrhythmia dataset was utilised to perform the analysis. The final number of features used for feature selection were 262.

Starting with 0.50, the various threshold values used for testing increased by 0.05. Three models have been evaluated to assess performance for each threshold value. With the help of a repeated 10-fold stratified cross -validation approach, these models were trained using a final subset of features that was generated using the CGN- FS methodology. For each of the two models, the accuracy mean, and standard deviation were computed. The best performance was taken as the one where the accuracy was maximum while the standard deviation of the accuracy was at minimum which occurred at 0.70 threshold. The number of keep values at the optimum threshold was 101 features.

The algorithm was evaluated on two different machine learning models namely, logistic regression, and neural networks. The threshold 0.80 was selected as optimum and the logistic regression accuracy was

68.57% with a standard deviation of 3.87. For the neural network the best threshold was at 0.85 and number of features selected were 200 features and accuracy was 92.10% with standard deviation of 2.77.

ReliefF was implemented for the same dataset used by the CGNFS algorithm. For ReliefF the model used for evaluation was logistic regression. The parameters required were the number of features to be selected which was 101 similar to the threshold value at .80. The obtained accuracy for the ReliefF method was 70.36%.

5.2 Clustering Method

The clustering-based method was used to carry out the analysis of the arrhythmia dataset. During the feature selection process, a total of 262 features were utilised. Furthermore, the PCoA with MDS method was implemented for dimensionality reduction. This resulted in a total of 219 dimensions for 50% cumulative explained variance.

Upon using the developed algorithm, 15 features were obtained for this particular dataset. Upon evaluating these 15 features with logistic regression, using the repeated (10 times) stratified 10 cross-validation method, an accuracy of 70.60% was achieved. The standard deviation was calculated to be 6.57. For comparison purposes, testing was conducted using all features with the repeated (10 times) stratified 10 cross-validation method. The accuracy achieved was 72.36%, with a standard deviation of 7.25. Additionally, using the neural network for 15 features with repeated (10 times) stratified 10 cross-validation method resulted in an accuracy of 67.39%, with a standard deviation of 3.90.

6 Conclusions

The application of correlation and clustering-based feature selection methods on the arrhythmia dataset has proven to be a valuable approach for enhancing the efficiency and interpretability of the predictive models. The correlation-based approach has successfully identified features with low inter-feature correlation, ultimately leading to a reduced feature set without compromising accuracy. This reduction in features has not only streamlined the modelling process but has also contributed to improved interpretability and robustness, consequently allowing for more meaningful insights into the underlying relationships between variables.

On the other hand, the clustering-based approach has demonstrated its efficacy in selecting a compact set of 15 relevant features, with minimal trade-offs in accuracy (approximately 1.5%). By leveraging the patterns of data points in a 2D space and determining cluster centroids, this method has effectively highlighted key features that capture essential information for classification tasks. Thus, the ability of the clustering algorithm to reduce feature dimensionality has led to a noticeable enhancement in model performance and interpretability.

The combination of these two feature selection techniques offers a promising avenue for optimising machine learning models in various domains, especially when faced with high-dimensional datasets such as the arrhythmia dataset. Not only do they contribute to improve predictive accuracy, but they also provide a clearer understanding of the underlying data patterns, leading to more reliable and transparent decision-making processes. As the field of feature selection continues to evolve, incorporating these methods can be

significantly beneficial to practitioners seeking to develop efficient and effective machine learning models for real-world applications.

References

1. Gnana, D.A.A., Balamurugan, S.A.A. and Leavline, E.J., 2016. Literature review on feature selection methods for high-dimensional data. International Journal of Computer Applications, 136(1), pp.9-17.
2. Kira, K. and Rendell, L.A., 1992. A practical approach to feature selection. In Machine learning proceedings 1992 (pp. 249-256). Morgan Kaufmann.
3. Kononenko, I., 1994, April. Estimating attributes: Analysis and extensions of RELIEF. In ECML (Vol. 94, pp. 171- 182).
4. Hall, M.A., 2000. Correlation-based feature selection of discrete and numeric class machine learning.
5. Yu, L. and Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research, 5, pp.1205-1224.
6. Peng, H., Long, F. and Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), pp.1226-1238.
7. Song, Q., Ni, J. and Wang, G., 2011. A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE transactions on knowledge and data engineering, 25(1), pp.1-14.
8. Li, Y., Luo, C. and Chung, S.M., 2008. Text clustering with feature selection by using statistical data. IEEE Transactions on knowledge and Data Engineering, 20(5), pp.641-652.
9. Cai, D., Zhang, C. and He, X., 2010, July. Unsupervised feature selection for multi-cluster data. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 333-342).
10. Chow, T.W. and Huang, D., 2005. Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. IEEE Transactions on Neural networks, 16(1), pp.213-224.
11. Mitra, S., Acharya, T. and Luo, J., 2006. Data Mining: Multimedia, Soft Computing, and Bioinformatics. J. Electronic Imaging, 15(1), p.019901.
12. Sotoca, J.M. and Pla, F., 2010. Supervised feature selection by clustering using conditional mutual information- based distances. Pattern Recognition, 43(6), pp.2068-2081.
13. Erdenetuya, Namsrai., Tsendsuren, Munkhdalai., Meijing, Li., Jung-Hoon, Shin., Oyun-Erdene, Namsrai., Keun, Ho, Ryu. (2013). A Feature Selection-based Ensemble Method for Arrhythmia Classification. Journal of Information Processing Systems, 9(1):31-40. doi: 10.3745/JIPS.2013.9.1.031
14. Liu Ziheng. 2021. Comparison of Feature Selection Methods on Arrhythmia Dataset. In Proceedings of the 2021 3rd International Conference on Image Processing and Machine Vision (IPMV '21). Association for Computing Machinery, New York, NY, USA, 66–71. https://doi.org/10.1145/3469951.3469963
15. Guvenir, H, Acar Burak, Muderrisoglu Haldun, and Quinlan R. (1998). Arrhythmia. UCI Machine Learning Repository. https://doi.org/10.24432/C5BS32.

****