Journal of Applied Artificial Intelligence
https://jaai.sabapub.com
ISSN: 2709-5908

2023 Volume 4, Issue 1 : 1–8
DOI : https://doi.org/10.48185/jaai.v4i1.692

Feature Subset Generation for Ensemble Learning Using Feature Clustering and Mutual Information

Hana Amar [1, *]

[1] Higher Institute of Science and Technology /souk- Al Juma/ Tripoli, Libya

Received: 24.01.2023 • Accepted: 15.03.2023 • Published: 20.04.2023 • Final Version: 25.04.2023

Abstract: Ensemble learning is a powerful technique for constructing accurate predictive models. Feature subset generation is an important step for ensemble learning. This paper proposes a new feature subset generation technique for ensemble learning using feature clustering and mutual information. The proposed feature subset generation technique clusters the features using a hierarchical clustering algorithm. Mutual information is used to compute the similarity between the features within each cluster. Feature subset generation is then performed by selecting the most informative features from each cluster. Experiments are conducted on a real-world dataset to compare the proposed feature subset generation technique to other existing feature subset generation techniques. The experimental results show that the proposed technique outperforms other existing feature subset generation techniques. In other words, at the end of my study, the required achievements were reached successfully between 79% and 90% as it shown in the table1, table2, table3 with most valuable subsets and effective features.

Keywords: Ensemble Learning (EL), Feature Clustering, Feature Subset Generation (VG), Minimum Redundancy-Maximum Relevance Algorithm, Support Vector Machine (SVM)

1. Introduction

Finding the most important features in a dataset is done using the feature subset generation technique for ensemble learning [1], which employs feature clustering and mutual information. To group related features together and determine which are most significant, this strategy makes use of mutual information and clustering algorithms. Mutual information is utilized to gauge how relevant each item is to the larger problem, while clustering methods are used to group related characteristics. The chosen characteristics can then be combined to provide a smaller set of features for ensemble learning. The data's dimensionality may be decreased with this technique, and the accuracy of ensemble learning models can be increased by removing pointless features. In many situations, ensemble learning models' performance has been enhanced by using this feature selection technique. While this feature selection technique has been shown to increase the accuracy of ensemble learning models, it can also lead to improved interpretability and a reduction in model complexity The goal of this feature selection technique is to identify important features in the data that can provide insight into the underlying structure of the dataset and improve the overall performance of ensemble learning models This feature selection technique works by recursively eliminating features of the dataset that are not relevant to the model's task or do not provide any useful information This process is repeated

[*] Corresponding Author: hanaomarkanan1@gmail.com

until all the most important features of the dataset have been identified Thus, using this feature selection technique[2], it is possible to identify important features of the dataset that are useful for model training and can provide a better understanding of the underlying structure of the data.

2. Summary of Some EL Studies:

Some experts specified four approaches of multiple algorithms combination. Firstly, setting for weak classifiers. Then, the outputs of bagging, boosting, and the random subspace method was compared. So, they reached to bagging was useful for weak and unsteady algorithms. Boosting is helpful only for weak, simple learners which were built on a big size of training instances [3]. The random subspace method is advantageous for weakened and unsteady algorithms that were applied in a few numbers of examples. Other researchers achieved another result. That was the averaging [4] versus the voting measure with multiple models. Where averaging often outperforms voting for Gaussian error of appreciation whereas a heavy tail function vote could be a winner. This way is used in economic issues. besides, new methods were invented by other experts. Those approaches are (stacking by extending this technique with probability distribution), and (multi-response linear regression). Thus, other researchers suggested a framework to construct hundreds or thousands of algorithms on small data sets. Their results showed that the new approach is scalable, fast, and accurate. According to all surveys done.

The general flowchart of this paper's methods and implements is shown in Figure 1, along with the following steps to take in order to achieve the improved results shown in each data table's results:

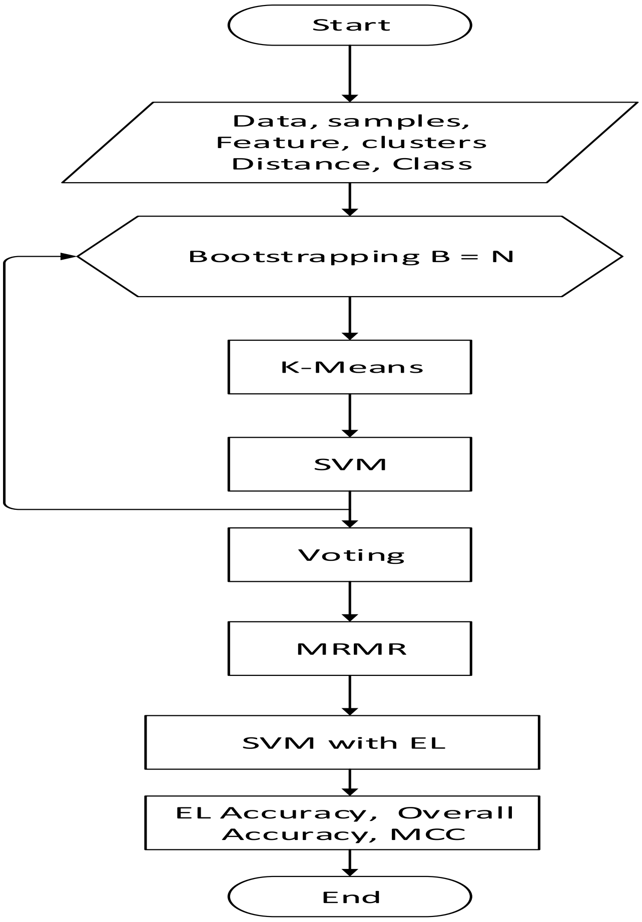
Start
Data, samples, Feature, clusters Distance, Class
Bootstrapping B = N
K-Means
SVM
Voting
MRMR
SVM with EL
EL Accuracy, Overall Accuracy, MCC
End

Figure 1. The General Flowchart of All Implements and Methods

1) Required data were got from UCI Machine Learning Repository and discard unwanted ones "first rows and columns ".

2) cluster data1, data2, and data3, and divided all the samples to create different and relevant features.

3) specify all the standards and feed our data to the code step by step to generate subsets, then, implement selection algorithms in order to pick up the best features and cluster them according to the MRMR, SVM, K- Means algorithms [4], and finalize our algorithm with voting step and see the results of MRMR, EL.

4) train and test a model until the obtained accuracies became improved unless repeat all that from Bootstrapping to voting with EL [6] and accuracies.

2.1. MRMR ALGORITHM

maximum significance Using the minimal redundancy algorithm (MRMR), features may be chosen based on mutual information (its ordinary job). Most characteristics that significantly affect class are those with the highest degree of feature-class relevancy [5]. Additionally, minimal redundancy denotes a reduction in the number of repeated variables. Therefore, it was thought that feature selection was a crucial problem for classification applications. As a result, many professionals used the technique and carried out extensive research to find the best characteristics that could be chosen based on the maximum dependency [6] and mutual information. However, because it was challenging to implement the maximal dependence situation, the researchers came up with another method that did the same thing and was known as MRMR [7]. Therefore, one of the most often used approaches to comprehend max dependence was maximum.

3. DATA SETS

3.1. DATA 1

The National Centre for Voice and Speech and Oxford University [8] collaborated to create the Data1 (PD) Parkinson's Disease Detection data collection. In order to obtain the properties (features) that represented the columns, experts and professionals captured the patients' voice signals. The 192 samples that made up the data represented rows. Each row represents one of the 192 voices recorded by 31 people, each of whom contributed six recordings. Additionally, 23 features were offered, of which 22 were recordings. Additionally, the class label's 23rd characteristic was set to 1 for Parkinson's disease and 0 for healthy (i.e., who have disease).

3.2. RESULTS OF PARKINSON'S DISEASE DATA SET

The number of instances, the number of divided sets, the number of clusters, the distance used in k-means, the optimization of the SVM [9] parameters (box constraint C with linear kernel function [10], C and Sigma with rbf function), the number of features, the number of class groups, and the data type itself are just a few variables that could influence the results (integer, real e.g.). The best training performance with a lot of instances was therefore achieved. The SVM parameters were optimized to provide better results. When the linear kernel function with the k-means [11] parameters

(K=5, correlation, and Euclidean distance) was applied (test many values of C as 2e-1, 4e-1, till C=9e-1). With correlation distance, the classification [12] on the training set was more reasonable and accurate. In comparison, the rbf kernel function produced more accurate results (e.g., C=8e-1, sigma=0.7).

Table 1. data1 outcomes:

Subset No	Features Number	Features No	Individual Accuracy	Combined subsets Num	EL Accuracy	TN Rate	TP Rate	MCC
1	6	4, 5, 6, 7,8, 15	79%	1	79%	91.3043%	8.6956%	0.2894
2	4	1, 2, 3, 16	48%	2	86%	84.0000%	16.0000%	0.5797
4	6	9, 10, 11, 12, 13, 14	69%	3	79%	91.3043%	8.6956%	0.2894
6	3	18, 19, 22	62%	4	83%	83.3333%	16.6666%	0.5076
16	10	4, 5, 6, 7,8, 15, 17, 19, 20, 22	79%	9	79%	91.3043%	8.6956%	0.2894
21	3	1, 2 , 3	48%	10	79%	91.3043%	8.6956%	0.2894
19	2	16, 18	62%	11	79%	91.3043%	8.6956%	0.2894
Overall Accuracy						79%		

The outcome showed the value of using EL [13] to Parkinson's disease data. Finally, by combining an EL accuracy improvement with the best accuracy (1st and 2nd subsets). According to the individual accuracies of their subsets, the characteristics (4th, 5th, 6th, 7th, 8th, 9th, 10th, 11th, 12th, 13th, 14th, and 15th) were the most effective and influential on the class among all qualities. Additionally, according to the data itself, the right classification [14] was on the negative samples being greater than positive ones based on the TP, TN rates in the prior table.

3.3. DATA 2

The Breast Cancer Wisconsin (Prognostic) Data Set was developed by physicians from Wisconsin University and is known as Data2 (BC). wherein the first 30 characteristics were extracted from a digital picture of a tiny needle aspirate (FNA). The final four characteristics were determined through medical testing [15]. There are 198 instances in the data, each of which represents a row. The columns correspond to the characteristics (variables), and each row designates one of the 198 recordings. 34 qualities were therefore offered. 34th for the class label, which was set to 0 if the sickness didn't reoccur and 1 if it did (recur).

3.4. RESULTS OF BREAST CANCER DATA SET

With this data, many results have been reached. As a result, the linear kernel function (test various values such as 2e-1, 4e-1...until C=9e-1) and the k-means [16] parameters (K=5, correlation, and Euclidean distance) were used. With k-means correlation distance, the classification [17] on the training set was more reasonable and accurate. When the first three perspectives were combined, the ensemble result was ideal.

Table 2. data2 outcomes:

Subset No	Feature Num	Features No	Individual Accuracy	Combined subsets Num	EL Accuracy	TP Rate	TN Rate	MCC
1	2	8, 10	13%	1	23%	42.8571%	57.1428%	0.1307
4	1	7, 9, 22	67%	2	77%	0%	100%	-0.1307
6	7	1, 5, 6 , 16 17, 20, 21	80%	3	90%	7.4074%	92.5925%	0.5229
11	5	1, 6, 16, 17, 21	80%	4	83%	8%	92%	0.3888
21	8	2, 3, 4, 11 12, 15, 18 19	37%	6	80%	8.3333%	91.6666%	0.3415
26	5	5, 6, 17, 20, 21	80%	7	83%	8%	92%	0.3888
31	6	5, 6, 17 , 18, 20, 21	77%	8	77%	8.6956%	91.3043%	0.3015
41	4	2, 3, 4, 19	43%	11	70%	9.5238%	90.4761%	0.2357
Overall Accuracy						78%		

3.5. DATA 3

The Madelon data collection was Data3. It was a fake dataset that had nothing to do with identifying cancer. The data was divided into a predetermined number of clusters and randomly assigned the labels 1 or -1. It was extracted from the variable selection [18] benchmark report for the NIPS 2003 experiments' design. Four thousand, four hundred occurrences, five hundred, and one characteristic were also included in the data. The last one was for the class label, and they were all features. The data underwent some preprocessing in order to facilitate quick and simple operations. In order to prevent them, we changed each -1 in the class value to 0. The algorithm [19] was then used on several examples with different characteristics to compare how well it performed.

3.6. RESULTS OF MADELON DATA SET

Consequently, data3's findings demonstrate that even though Madelon data contained a lot of characteristics, employing EL on it was adequate. Additionally, by merging the first five subgroups, we were able to get the greatest accuracy when using EL [20]. because after presenting one data set to the classifier [21], the EL accuracy was greater than the individual accuracy [22]. The preceding table showed that accuracy was generally accurate. According to individual accuracy, the views (4th, 22nd, 9th, 12th, and 7th) were the strongest subsets and most pertinent to their class among all subsets.

Table 3. Table3: data3 outcomes:

Subset No	Feature Num	Individual Accuracy	Combined Subsets Num	EL Accuracy	TP Rate	TN Rate	MCC
2	103	33%	1	51%	65.2173%	34.7826%	0.0165
5	97	44%	2	56%	60%	40%	0.1365
4	99	51%	3	58%	73.0769%	26.9230%	0.0915
7	90	62%	4	62%	71.4285%	28.5714%	0.1872
12	100	58%	5	73%	75.7575%	24.2424%	0.3919
17	103	42%	6	71%	75%	25%	0.3460
10	129	44%	7	62%	75%	25%	0.1641
9	111	56%	8	64%	72.4137%	27.5862%	0.2241
22	94	53%	9	64%	72.4137%	27.5862%	0.2241
Overall Accuracy					62%		

4. CONCLUSION

To summarize in a few sentences, using more has numerous advantages. In a nutshell, by referring to earlier findings like the data1, data2, and data3 tables, all advantages from employing several approaches were realized at the conclusion of this study. Furthermore, we discovered that ensemble learning consistently produces adequate and flawless results, especially when there are a large number of characteristics and various created subsets. However, EL processes with several characteristics may need complex computations. Consequently, EL was successful and helpful when m was not a large variable, such as in the case of Parkinson's disease and breast cancer data, yet it was highly valuable when applied to data that had a large number of characteristics, such as Madelon data. Finally, with the MRMR algorithm, SVM classifier, one approach was reached at the end of this research by looking at the previous results, as shown in the table1, table2, and table3. Also, we have noticed that ensemble learning most of the time gave us sufficient and perfect outputs, especially with a large number of features **m** where all generated subsets were diverse. Although, EL process with a large number of attributes could require complicated calculations, as a result, EL was effective and beneficial when the variables were small, such as in Parkinson's or Breast Cancer data, but it was ineffective and ineffective when applied to data with a large number of features, such as Madelon data.

Therefore, using clustering, the SVM classifier, the MRMR algorithm, and EL helped us obtain accurate and diverse subsets. The details of the algorithm steps were simply demonstrated in the methods chapter. Thus, the major standpoint was completely carried out thanks to a flexible approach and efficient algorithms that were jointly worked on. Consequently, diverse, accurate, and sufficient subsets were produced, as they should be. So, they were chosen. In other words, this achievement was the thesis target. As an engineering perspective, a real data set could be used in a future work with some modifications to my next studies. The methods used may be improved. This work could also be expanded into a PhD dissertation with additional additions.

References

[1] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing 300 (2018) 70–79.

[2] https://en.wikipedia.org/wiki/Active_learning_(machine_learning)

[3] Steven, C. H., Michael, R., & Zhu, J. The Chinese University of Hong Kong., & Jin ,R. Michigan State University, USA. *Batch Mode Active Learning and Its Application to Medical Image Classification.*

[4] Alpaydin, E., 2010. *Introduction to machine learning*- 2nd ed. Cambridge, London.

[5] https://en. Wikipedia. org/ wiki/ Statistical_ classification

[6] Phyu, T., N., 2009. *Survey of Classification Techniques in Data Mining*. Hong Kong.

[7]] A. Khoder, F. Dornaika, An enhanced approach to the robust discriminant analysis and class sparsity based embedding, Neural Netw. (2021)

[8] Wei, Di., & Melba, M. Crawford, *Active Learning via Multi-View and Local Proximity Co-Regularization for Hyper spectral Image Classification.*

[9] R. Wyss, S. Schneeweiss, M. van der Laan, S.D. Lendle, C. Ju, J.M. Franklin, Using super learner prediction modeling to improve high-dimensional propensity score estimation, Epidemiology 29 (1) (2018) 96–106.

[10] Abello, J., & Cormode, G., 2006. *Report on DIMACS Tutorial on Data Mining and Epidemiology*. Rutgers University. America.

[11] https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-choice/

[12] D. Michie, D.J. Spiegelhalter, & C.C. Taylor, 1994. *Machine Learning, Neural and Statistical Classification*.

[13]] V.H.A. Ribeiro, G. Reynoso-Meza, Ensemble learning by means of a multiobjective optimization design approach for dealing with imbalanced datasets, Expert Syst. Appl. 147 (2020) 113232.

[14] Srinet, A., & Snyder, D. *Bagging and Boosting Slides*. A. Krogh and J. Vedelsby, 1995. *Ensembles, Cross Validation and Active Learning*.

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., 2011. *Scikit- learn: Machine Learning in Python*. Journal of Machine Learning Research.

[16] Y.-H. Hung, Improved ensemble-learning algorithm for predictive maintenance in the manufacturing process, Appl. Sci. 11 (15) (2021) 6832.

[17] Holmes, P. *IC1 of Simple Decisions*. Princeton University., Bevers, M. *IC2 of Optimization Models*. USDA. & Ping Lo, Y. *MS3 Two- Dimensional*, Loughborough University.

[18] David*, M., 2003. Chapter 20. An Example Inference Task Clustering PDF. Information Theory, Inference and Learning Algorithms. Cambridge University.*

[19] https://en.wikipedia.org/wiki/K-means_clustering.

[20] Osmar, R. Z., 1999. Chapter 8. Data Clustering book. Alberta University. from source jiawei han data mining book.

[21] S. Hijazi, Semi-Supervised Margin-Based Feature Selection for Classification (Ph.D. thesis), Université du Littoral Côte d’Opale; Université Libanaise, école doctorale, 2019.

[22] A. Mujib, T. Djatna, Ensemble learning for predictive maintenance on wafer stick machine using IoT sensor data, in: 2020 International Conference on Computer Science and its Application in Agriculture (ICOSICA), IEEE, 2020, pp. 1–5..