ASIAN JOURNAL OF INFORMATION TECHNOLOGY
MEDWELL PUBLICATIONS

Feature Selection for Household Insecurity Classification: Wrapper Approach

Mersha Nigus Alemayehu and Doreswamy
Department of Computer Science, Mangalore University, Konaje, India

Key words: Food insecurity, feature selection, classification, HICE, ROC

Corresponding Author:
Mersha Nigus Alemayehu
Department of Computer Science, Mangalore University, Konaje, India

Page No.: 146-151
Volume: 20, Issue 5, 2021
ISSN: 1682-3915
Asian Journal of Information Technology

Abstract: Feature selection will become crucial, specifically in facts units with a huge variety of variables and features. It's going to cast off irrelevant variables and boom classification accuracy and overall performance. With the intention to decrease the model's computational cost and growth its efficiency, it is a good concept to reduce the variety of input variables. This study employs a wrapper approach to discover a subset of features most relevant to the classification problem. Sequential backward series, sequential forward choice and recursive feature exclusion are the 3 forms of feature selection that Wrapper procedures help. Machine learning classifiers inclusive of k-Nearest Neighbor, Logistic Regression, support vector machine and random forest are used to determine the classification accuracy of selected attributes. The findings reveal that the random forest classifier is the excellent and sequential backward selection with seven attributes is the great filtering approach with 99.97% accuracy and a 100% ROC. Finally, the experiment result of the paper inform to government, policy makers and humanitarian organizations to take an emergency action to fix the problems of household who are food insecure and needs emergency action to survive their lives.

INTRODUCTION

High dimensional data, especially data with many features, is increasingly being used in machine learning problems these days. To solve these problems, many researchers concentrate on experiments. Furthermore, essential features must be extracted from these high-dimensional variables and data. To reduce noise and redundant data, statistical techniques were used.

We live in the twenty-first century which has shown remarkable results in terms of performance, technology and financial advancement. We've made big strides in medicine, improving people's health and lengthening their lives. As a result of this expansion, almost one-sixth of the world's population now suffers from chronic poverty and malnutrition caused by food shortages.

Families are food shaky, when they need monetary admittance to cash for food and its dietary energies are underneath the necessary norm. Choosing significant highlights from the first informational collection with the assistance of AI calculations will lessen computational expense and will support the expectation with better exactness. AI is the way toward building a logical model dependent on information found from the example

preparing informational index. It is additionally a mind boggling calculation interaction to perceive designs consequently and to settle on a clever choice dependent on the example information[1].

Quite possibly the main subjects in AI and related fields is feature selection[2, 3]. True informational indexes likewise contain countless old or repetitive highlights that, if not sufficiently discarded will drastically lessen model consistency and learning speed. Highlight choice involves recognizing a subset of highlights that can be utilized to build forecast precision or diminish the size of the construction without influencing the expectation exactness of a classifier planned with just the picked highlights[4].

Feature selection methods are used for dimensionality reduction strategy with the goal of selecting important features from the original data set by eliminating redundant and noisy features that may not have a prediction accuracy[4]. Based on searching strategies, feature selection can be classified into three techniques, namely filter approach, wrapper approach and embedded approach. Wrapper approaches are used greedy search algorithms which evaluate all possible combinations of features and select the combination which produces the best results[5].

The discovery of a sample features from which to construct a classification model for a given task is a key problem for machine learning. Good features are highly associated with class but are uncorrelated each other[1, 6]. Applying feature selection is increasing because there are a large number of high-dimensional feature samples in the real world data set. This makes it impractical, computationally expensive and leads to less classification accuracy when a whole set of inputs is used[7].

In this study, we are using Ethiopian household income , consumption and expenditure survey data set which was collected from 2011-2016 and the data set contains 21 attributes including the class label and 58064 rows. The main objective of this study is selecting the best features from the original data set and predicts household food insecurity. This enables policy makers can used this data for further decision making regarding the countries food security status.

Literature review: Several related research efforts have been conducted related to food insecurity and feature selection. Yildirim[8], identified factors that affect household food security in many developing countries. The research result tells that attributes such as d gender of household head, literacy level , age of household head and household's income have a positive influence on food security of the family where as family size has a negative influence on household food security.

According to Okori and Obua[9], conducted research on Prediction of food insecurity and they conclude that predicting food insecurity drives leading actors where to direct relief from early intervention. The study predicts food insecurity has been producing promising results in certain parts of the world. When effective household food insecurity monitoring is implemented, efficient implementation of government programs, nutritional assistance and other policy measures would reduce household food insecurity globally.

According to Okori and Obua[9], prediction of household food insecurity is important in directing stakeholders where to target early intervention aids. Therefore the effect of food insecurity in the process can be controlled or eliminated. There are areas of the world where the Positive results were obtained in science of predicting food insecurity. When adequate analysis of food insecurity is conducted, good outcomes have been obtained, government programs would implemented successfully and assistance and other policy measures would reduce food insecurity.

Srivastava *et al*.[10] proposed various feature selection approaches to eliminate redundant features and features that do not have good prediction accuracy. Feature selection techniques used for many reasons such as reducing computational costs, reducing a model's complexity and reduce over fitting.

Blessie and Karthikeyan[11] suggested different feature selection approaches to choose the most important features that have good prediction accuracy from the original data set. Authors identified feature selection approach as wrapper, filter and hybrid method. To select features which are said to be good, Wrapper approach used predictive accuracy of a fixed learning algorithm. It is computationally expensive for high dimensional data. Filter approach select features independent any learning algorithm and use criteria such as distance, knowledge and dependency.

Srivastava *et al*.[10] implemented different feature selection approaches to maximize classifier performance minimize calculation costs and enhancing performance by eliminating features which are redundant and noisy.

Support vector machine were implemented by the authors to predict food insecurity status of farm households into food secure and insecure. Authors select top 14 features out of 75 features. After feature selection 77% accuracy and recall of 84% were scored[12].

Sanchez-Maro *et al*.[13] implemented wrapper and filter approach for feature selection .To select the right features Wrapper approach use classifier where as filter approach select features independent of the classifier.

Yildirim[8] conducted research to determine household food insecurity condition in Ethiopia. Based on the authors finding, higher family size, lower educational level and household old household head age are positive correlation with household food insecurity. Kumari and Swarnkar[14] implemented wrapper-based

technique for selecting features in supervised learning algorithms. Based on evaluation methodology for the subset. Correlations and interactions between the features are regarded when choosing the features. Prediction bias helps to optimize the algorithm's performance. In Support Vector Machine (SVM), during SVM learning, weight is assigned to each feature. The key downside of the wrapper strategy is computational expensiveness because of the quest for the ideal collection from broad computational complexity space.

MATERIALS AND METHODS

In the feature subset selection problem, a learning algorithm must choose a relevant subset of features on which to concentrate its attention while ignoring the rest. Within the feature selection problem, a learning algorithm must pick a relevant subset of features on which to focus its attention while ignoring the remainder. To obtain the best possible results for a specific learning algorithm on a specific training set, a feature subset selection method should consider how the algorithm as well as the training set perform[15]. Feature selection is a machine learning method that uses a learning algorithm to select a subset of features from a data set. Choose the model with the fewest parameters that correctly reflects the data[16]. Feature selection increases the performance and precision of learning algorithm's models. The primary goal of feature selection is to pick features of the best classification accuracy and to exclude any features that are deemed redundant. We emphasize the wrapper-based feature selection approach in this article. This study relied on information from the Ethiopian Household Income, Consumption and Expenditure (HICE) report. Wrapper approaches make use of greedy search algorithms which test all possible attribute combinations before choosing the one that best matches a given machine learning algorithm (Fig. 1).

Wrapper method: Wrappers need an assured technique on exploring the area of entire viable subsets of options, assessing their quality via learning or evaluating a classifier with that subset on features. The strategy on selecting features is primarily based on a specific machine learning algorithmic rule as we tend to attempt in accordance with match between a given data set. By examination all the viable combinations of options con to the analysis criteria. Wrapper methods check variety of models the usage of techniques as embody and/or cite predictors to find the foremost acceptable combination therefore maximizes model efficiency. This methodology is usually based mostly over the Greedy Search algorithmic program. The subsequent Fig. 2 demonstrates wrapper based approach.

Forward Feature Selection (FFS): Starts with a single predictor and adds more attractively. The best of the remaining original predictors will be added at each subsequent iteration on the basis of performance criteria.

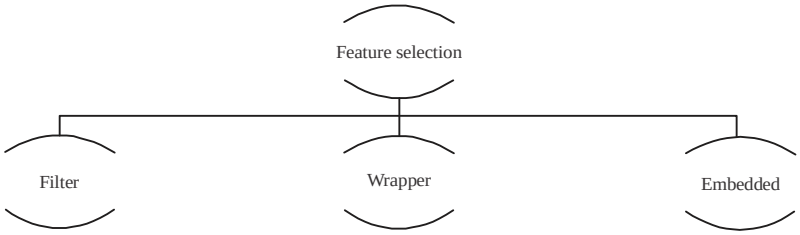

Fig. 1: Machine learning feature selection

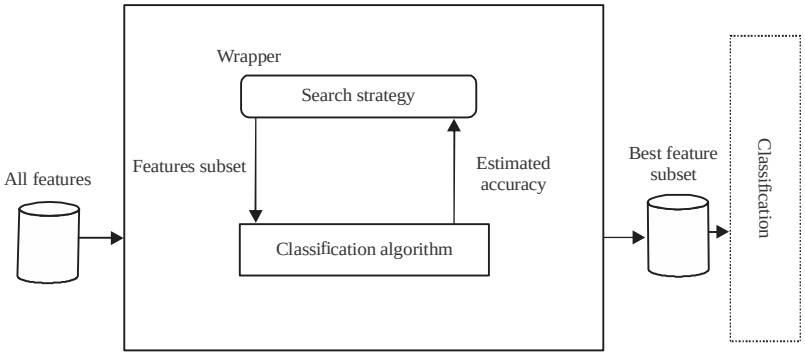

Fig. 2: Wrapper feature selection method

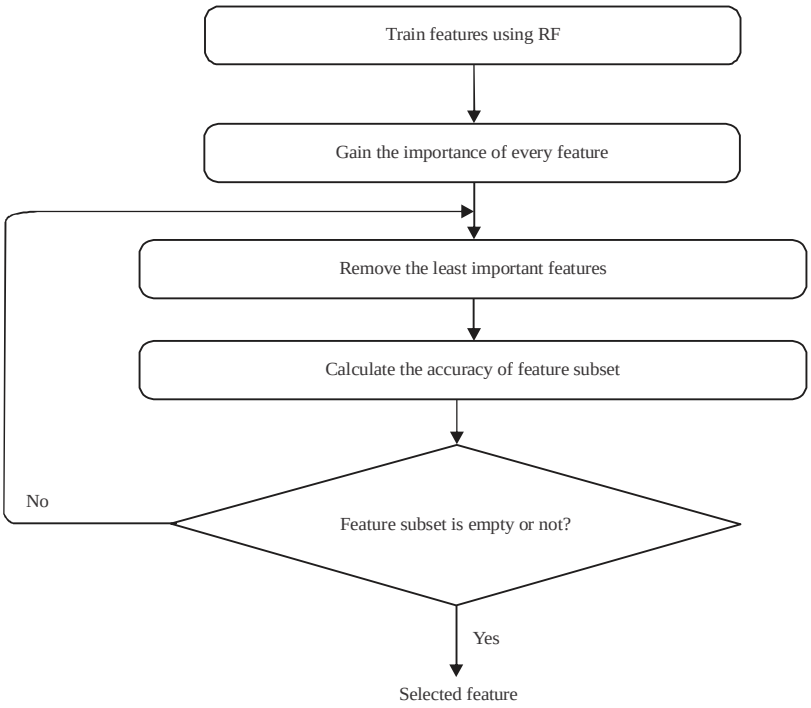

Fig. 3: Recursive feature elimination process

Backward Feature Selection (BFS): All features are included at the beginning and eliminate unnecessary and redundant features. In each iteration a feature is omitted from the full data. Each validation method tests the resulting selection. If the performance rate of the new subset feature is higher than that of the previous subset then it will replace the current best subset of features. The cycle continues until each feature is removed from the data sets and an empty set is reached[14].

Recursive Feature Elimination (RFE): Starts with all the predictors and eliminates one-by-one. It is one of the most popular algorithms, eliminating less important predictors based on a feature importance ranking. Figure 3 shows that how recursive feature elimination works for feature selection.

Classification: Data from eleven regions of Ethiopia were classified using KNN, LR, RF and SVM to establish the influence of regional disparity. in addition to this different performance metrics are used.

LR was representations of the probability of labeling issues for two potential results 1 and 0 which are food. Secure and insecure foods. Or logistic regression is a classification technique, where test data probabilities are expected. This is linear in essence such that the quantitative response is plotted straight.

kNN is a classification algorithm that resides all available features and classify new cases based on Distance functions or similarity measure. kNN is the simple classification algorithm with considers all the data set points to classify them. It considers k nearest points and lists them ascending from the chosen data level.

SVM is a classification algorithm in which we have multiple kernels option depending on the fashion of the distribution of data. it can classify data in multiple linear ways but SVM gives us the optimal among all the possible option. Types of kernel: linear, RBF, poly, sigmoid.

RF is a part of ensemble learning. Ensemble learning is when we combine same or different algorithm multiple Times to get the optimized output. In random forest classification we combine decision tree classification algorithms multiple times.

Performance evaluation: Performance measures such as Accuracy, recall, precision, F1-measure and Operating Receiver Characteristic (ROC) are used to evaluate the results of the proposed model. Accuracy (CA) is the ratio of accurate predictions and total number of predictions:

$$CA = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

Recall is the proportion of real positive cases that were accurately estimated positive by the model is referred to as recall:

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

Precision is the proportion of estimated positive/negative cases that were ultimately positive or negative is referred to as precision:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

F1-measure: The accuracy and recall are used to determine the F1-measure:

$$\text{F1 measure} = \frac{\text{Precision*Recall}}{\text{Precision+recall}} \quad (4)$$

ROC: When agreeing on observable outcomes, the ROC is commonly used. There are the scenarios where the false positive figure on the X and Y axes has a true positive situation.

Data set description: The HICE data under consideration consists of four distinct data sets compiled in 2000, 2005, 2011 and 2016. Ethiopia Household Income, Consumption and Expenditure (HICE) were conducted by Central Statistical Agency (CSA) and National Planning Commission together with the then Ministry of Finance and Economic Development (MoFED) (CSA). This research used HICE as a data source, especially regarding household food security status to identify whether a Household is Food secure or food Insecure. We use a nationwide sample of 58.064 households from around the country with 32.209 instances belonging to one class and 23.668 instances belonging to another. Cases are described by 21 characteristics, some of which are numerical and some of which are nominal and the predicted output class is either food secure or food insecure.

RESULTS AND DISCUSSION

This study addresses the experimental effects of the proposed methods on the data set household, income, consumption and spending. This study introduces wrapper-based feature selection techniques for determining whether a household is food safe or not. Techniques such as sequential forward selection, sequential backward selection and recursive attribute elimination are used. Classifiers such as kNN, LR, SVM and RF are used to verify the classification accuracy of selected features. The outcome of the experiment reveals that sequential backward selection with five features performs better which is 96% accuracy and 96% ROC, followed by sequential backward feature with seven features and recursive feature elimination with ten features whose accuracy score is 93 and 89%, respectively using kNN classifier. In SVM classifier forward feature selection with seven features and sequential backward selection with five features score the same which is an accuracy of 91 and 91% of ROC followed by recursive Feature elimination with five features with an accuracy of 89%. In the case of RF sequential backward selection with seven features score better than sequential forward selection and recursive feature elimination with an accuracy of 99 and 100% ROC. Finally, sequential backward selection method with ten features score better result than others which is an accuracy of 89% and ROC of 89% using LR classifier (Table 1-4).

Table 1: Performance of kNN classifier

Algorithm	Feature selection technique	No. of features	Accuracy	Precision	Recall	F1 score	ROC
kNN	Forward feature selection	10	0.90	0.87	0.92	0.89	0.91
		7	0.92	0.91	0.92	0.91	0.93
		5	0.95	0.95	0.95	0.95	0.96
	Backward feature selection	10	0.89	0.88	0.88	0.88	0.90
		7	0.93	0.91	0.92	0.91	0.93
		5	0.96	0.95	0.95	0.95	0.96
	Recursive feature elimination	5	0.96	0.96	0.95	0.95	0.96

Table 2: Performance of SVM classifier

Algorithm	Feature selection technique	No. of features	Accuracy	Precision	Recall	F1 score	ROC
SVM	Forward feature selection	10	0.88	0.90	0.82	0.85	0.87
		7	0.91	0.93	0.85	0.89	0.91
		5	0.88	0.90	0.82	0.86	0.88
	Backward feature selection	10	0.88	0.88	0.83	0.86	0.88
		7	0.88	0.90	0.83	0.86	0.88
		5	0.91	0.92	0.87	0.89	0.91
	Recursive feature elimination	5	0.89	0.90	0.84	0.87	0.89

Table 3: Performance of random forest classifier

Algorithm	Feature selection technique	No. of features	Accuracy	Precision	Recall	F1-score	ROC
RF	Backward	10	0.9997	0.9994	1	0.9997	1
		7	0.9994	0.9990	0.9998	0.9994	1
		5	0.9996	0.9992	1	0.9996	1
	Forward	10	0.9995	0.9994	0.9996	0.9995	1
		7	0.9997	0.9996	0.9998	0.9997	1
		5	0.9996	0.9992	1	0.9996	1
	Recursive	5	0.9994	0.9996	0.9992	0.9994	1

Table 4: Performance of LR classifier

Algorithm	Feature selection technique	No. of features	Accuracy	Precision	Recall	F1 score	ROC
LR	Forward feature selection	10	0.87	0.86	0.84	0.85	0.87
		7	0.88	0.86	0.88	0.87	0.89
		5	0.88	0.86	0.87	0.87	0.89
	Backward feature selection	10	0.89	0.86	0.88	0.87	0.89
		7	0.83	0.82	0.77	0.80	0.83
		5	0.88	0.86	0.88	0.87	0.89
	Recursive feature elimination	5	0.86	0.85	0.83	0.84	0.86

CONCLUSION

This research introduces wrapper-based filtering strategies such as sequential backward, sequential forward and recursive feature elimination to pick the best features from the original data set that yield good predictive precision once features are selected. Classification algorithms such as random forest, k-neighbor, logistic regression and SVC are used to validate the chosen features. For feature selection, the HICE survey data set is included in this study. The validity of the given algorithms is assessed using performance measures such as accuracy, recall, precision, F1-score, AUC/ROC and uncertainty matrix. The experimental results of the proposed feature selection method reveal that sequential backward selection of ten features outperforms others, with an accuracy of 99.97% and a ROC of 100%. Finally, we recommend government of policy makers and humanitarian organizations to take an emergency action to fix the problems of household who are food insecure and needs emergency action to survive their lives.

REFERENCES

01. Hall, M.A., 1999. Correlation-Based Feature Selection for Machine Learning. University of Waikato Press, New Zealand, Pages: 178.
02. Guyon, I. and A. Elisseeff, 2003. An introduction to variable and feature selection. J. Machine Learn. Res., 3: 1157-1182.
03. Dash, M. and H. Liu, 1997. Feature selection for classification. Intell. Data Anal., 1: 131-156.
04. Miao, J. and L. Niu, 2016. A survey on feature selection. Procedia Comput. Sci., 91: 919-926.
05. Das, S., 2001. Filters, wrappers and a boosting-based hybrid for feature selection. Proceeding of the 18th International Conference on Machine Learning, 28 June-July 1, 2001, San Francisco, CA., USA., pp: 74-81.
06. Mwadulo, M.W., 2016. A review on feature selection methods for classification tasks. Int. J. Comput. Appl. Technol. Res., 5: 395-402.
07. Birara, E., M. Mequanent and T. Samuel, 2015. Assessment of food security situation in Ethiopia. World J. Dairy Food Sci., 10: 37-43.
08. Yildirim, P., 2015. Filter based feature selection methods for prediction of risks in hepatitis disease. Intl. J. Mach. Learn. Comput., 5: 258-263.
09. Okori, W. and J. Obua, 2011. Machine learning classification technique for famine prediction. Proceedings of the World Congress on Engineering Vol. 2, July 6-8, 2011, IAENG, London, UK., pp: 4-9.
10. Srivastava, M.S., M.N. Joshi and M. Gaur, 2014. A review paper on feature selection methodologies and their applications. IJCSNS. Int. J. Comput. Sci. Network Secur., 14: 78-81.
11. Blessie, E.C. and E. Karthikeyan, 2012. Sigmis: A feature selection algorithm using correlation based method. J. Algorithms Comput. Technol., 6: 385-394.
12. Khaire, U.M. and R. Dhanalakshmi, 2019. Stability of feature selection algorithm: A review. J. King Saud Univ. Comput. Inf. Sci., Vol. 1, 10.1016/j.jksuci.2019.06.012
13. Sanchez-Marono, N., A. Alonso-Betanzos and M. Tombilla-Sanroman, 2007. Filter methods for feature selection-a comparative study. Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning, December 16-19, 2007, Springer, Birmingham, UK., pp: 178-187.
14. Kumari, B. and T. Swarnkar, 2011. Filter versus wrapper feature subset selection in large dimensionality micro array: A review. Int. J. Comput. Sci. Inf. Technol., 2: 1048-1053.
15. Kohavi, R. and G.H. John, 1997. Wrappers for feature subset selection. Artif. Intell., 97: 273-324.
16. Barbosa, R.M. and D.R. Nelson, 2016. The use of support vector machine to analyze food security in a region of Brazil. Applied Artif. Intell., 30: 318-330.