A STUDY OF EMBEDDED FEATURE SELECTION METHODS FOR HOUSEHOLD FOOD INSECURITY CLASSIFICATION ON HICE DATA

Mersha Nigus[1], Dorsewamy[2]
[1]Research scholar, Department of Computer Science Mangalore University, Karnataka, India,
mbezadtu@gmail.com
[2]Professor, Department of Computer Science Mangalore University, Karnataka, India,
dore@mangaloreuniversity.ac.in

Abstract

Feature Selection is the method of selecting the most appropriate features of a given task while discarding the messy, unnecessary and redundant features of the data set. The selection of features is important for a number of reasons, such as simplification, high accuracy, computational efficiency and interpret-ability. In order to validate the accuracy of selected features, classifiers such as K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB) are used. The experimental result shows that recursive feature elimination with RF classifier which scores an accuracy of 99.9% followed by KNN, SVM, LR and NB with an accuracy of 96.6%, 0.90.9%, 0.89% and 0.85% respectively.

Key words: Feature Selection, Preprocessing, Calorie, Food insecurity, Classification, HICE

1. INTRODUCTION

Food insecurity is characterized as lack of access to adequate food for all for an active and safe life by all persons at all times. Almost all countries in the world haven't achieved food security. And in sub-Saharan Africa, the question is more pronounced. It has been estimated that around 204 million people were undernourished in Africa between 2000 and 2002, 86 million from the east side of the continent, including Somalia, Ethiopia, Sudan, Kenya and Tanzania [1].

Nutrition is the most important human requirement for survival, well-being and productivity. This is the cornerstone of social and economic development. As we know, enough food is needed to meet the needs of all people in the world today [2]. Food insecurity is the total cost of food which would satisfy the average nutritional needs of families of different sizes and composition. A family will likely suffer from food insecurity if its average income is lower than the poverty line for food. Food insecurity can also be defined as a situation where people lack secure access to sufficient amounts of safe and nutritious food for average growth and development and for active and healthy living. Food unavailability, lack of purchasing power and inadequate distribution of food at household level may be affected [3].

Households are considered to be food insecure if they lack economic access to the food they need, no access to cash food processing, and their dietary energy intake (kilo/ calories) is below the appropriate level. One of the biggest challenges for household food insecurity prediction is forecasting food insecurity with high accuracy and less computational cost from data volume. Using data mining and machine learning algorithms, choosing the key features from the original data set would reduce the computational cost and eventually improve the prediction with greater accuracy.

In the field of classification, feature selection is very important for improving the performance of classification, particularly in the case of high-dimensional data. Through removing unnecessary and redundant features, helps to boost learning speed, predictive accuracy, usability and interpretation of learned results [4].

Feature selection is one of the most common preprocessing strategies for data and has become an essential part of the learning method. It is the mechanism whereby relevant features are detected and obsolete, redundant, or noisy data removed. This method speeds up algorithms to increase predictive accuracy, and improves understand-ability. Irrelevant features are those which do not provide useful information and redundant features do not provide more information than the features currently selected [5].

Feature selection is a supervised learning algorithm aimed at finding the appropriate features and deleting unnecessary features or redundant features for greater accuracy. Feature selection improve performance (in terms of speed, predictive power, model simplicity), visualize data for model selection and reduce dimensionality and eliminate noise. [6].

The main objective of this paper is to apply embedded feature selection approach on Ethiopian household income, consumption, and expenditure survey data collected in 11 regions from 2011 to 2016, covering both all urban and rural parts of the country. There is no previous study utilizing this data to identify and forecast food insecurity in the household level. Poverty in Ethiopia also remains a big challenge to resolve. Accordingly, there is persistent and severe nutritional scarcity. Approximately 10 percent of Ethiopian citizens are chronically food insecure and this figure rises to over 15 percent during frequent drought years, 2.7 million people in 2014 need emergency food assistance and 238,761 children in 2014 need treatment for severe and acute malnutrition [7].

1. RELATED WORKS

Articles related to food insecurity and feature selection methods are reviewed below.

Endalew et al. [7] identifies factors affecting food security for households in different developing countries. The findings of the study indicate that features such as gender, educational level, age and household income have a positive effect on food security, whereas household size has a negative impact on household food security.

Selection of feature subsets is an important topic when it comes to problems with training classifiers in Machine Learning (ML). Many input features in the ML may contribute to the "curse of dimensionality" which explains the fact that the difficulty of changing the parameters of the classifier during training increases exponentially with the number of features. Consequently, ML algorithms are considered to suffer from a significant reduction in predictive precision when faced with many unnecessary features [8].

The feature search process is integrated into the classification algorithm in the embedded method and impossible to separate the learning process and the feature selection process. Compared to the wrapper method, the embedded method involves interacting with the classifier while at the same time saving large amounts of computing costs than the wrapper method [4].

Machine learning is the method of constructing a scientific model based on knowledge learned from the sample training data set. It is also a complex computational method to automatically identify patterns and make reasonable, sample-based decision. Selection of features is the method of choosing a subset of the most important features for model construction [9].

A critical problem in machine learning is the identification of the sample features on which to construct a classification model for a particular task. Good feature sets are strongly class-related but not interrelated [9].

Machine learning provides a tool through which, large amounts of data can be analyzed automatically and feature selection is fundamental. Based on different search strategies, the selection of features can also be classified into three techniques, such as filter-based methods, wrapper-based methods and embedded methods.

There has been a considerable increase in need to apply feature selection methods in large data set. This is because most data sets have a large number of high-dimensional feature samples. This makes it impractical, computationally expensive and reduces classification accuracy when using a complete set of inputs [10].

Report on food insecurity by Kakwani et al. [11] revealing that 842 million people worldwide have suffered from chronic hunger, which is 12 per cent in 2011-2013. The Food and Agriculture Organization (FAO) measures food insecurity based on the prevalence of under nutrition and compares the usual consumption of food expressed in terms of dietary energy (kilo/calories) with certain energy requirements. It also calculates food insecurity by the percentage of the population whose consumption of dietary energy is below the nutrition requirement standard.

The food insecurity proposed by W.Okori et al. [12] is influential in leading stakeholders where early intervention relief may be focused. Consequently, the effect of food insecurity in the process may be controlled or minimized. The study of predicting food insecurity has produced positive results in some parts of the world. If accurate monitoring of food insecurity is carried out successful implementation of federal programs, food assistance, and other government initiatives will reduce food insecurity.

Alisha et al. [13] examined food malnutrition as a global issue impacting millions of citizens facing extreme hunger in developed countries. People have suffered loss of life due to food insecurity both in the past and in recent years. The situation is too serious in Ethiopia and the problem of food deprivation devastates millions of people, some of whom risk their lives. Most people across the world are facing food insecurity and demand various local and global food aid agencies to provide nutritious assistance.

P .Yildirim and Pinar [14] have suggested a variety of approaches for choosing features in machine learning. For a number of reasons, feature selection strategies were introduced, such as reducing computational costs , reducing complexity of the model, reducing over-fitting and making it easy to interpret by model.

E. C. Blessie and E. Karthikeyan [15] suggested methods for selecting features to calculate how useful it is in making an ideal choice. The authors classify feature selection approaches into the wrapper-based method, filter-based method, and hybrid. To determine goodness, the wrapper-based approach uses the predictive precision of a predetermined learning algorithm. Wrapper-based approach is cost efficient with a high range of features for data sets. The filter-based approach selects important features independent from any learning algorithm, and uses distance, knowledge and dependence.

M. S. Srivastava et al. [16] introduced methodologies for selecting features to accomplish the shared objective of maximizing classifier accuracy, decreasing related costs of calculation, increasing accuracy by deleting irrelevant and possibly redundant features, reducing complexity and related computational costs, and enhancing the likelihood of an intelligible and realistic solution.

R. M. Barbosa and D. R. Nelson [17] applied the SVM algorithm to identify farm households' food security status in a safe and insecure food. The authors were chosen as the top 14 features out of 75 features in total and selected features score 77 percent accuracy and 84 percent recall.

B. Endalew et al [7] researched the triggers, determinants and food security situation in Ethiopia. As a result, increased household size, lower levels of household head education and an increase in household head age are significantly associated with household food insecurity. Household food insecurity is associated with increased household size. There is an inverse relationship between the level of education household head has attained and the likelihood of food insecurity falling.

2. METHODOLOGY

Feature Selection is a method widely used in machine learning, in which subsets of features to implement learning algorithms are selected from the available data. So we prefer the model with the fewest possible parameters that properly represent the data [18]. It improves efficiency, model classification accuracy that was developed to solve the problem. The goal of selecting feature algorithms is to find features that provide the best accuracy in classification and eliminate other irrelevant features.

2.1 Feature selection Methods

Selecting a feature is a process that selects a reduced number of explanatory variables to describe the variable response. The main reasons for selecting a feature are:

- Making the model easier to understand, eliminating redundant variables
- Reduce the problem to allow algorithms to work faster and allow the handling of high-dimensional data
- Reduce over fitting.

Feature selection is the process of finding the most relevant variables for a predictive model. Such strategies are used to recognize and delete unnecessary, irrelevant and redundant features that do not contribute to the accuracy of the predictive model. There are many built-in feature selection methods for selecting features that can select the best features from the original data set, such as Lasso, Tree-based and recursive feature elimination.

Lasso is a built-in feature selection method that uses selectFromModel for a meta-transformer and an estimator that has a coefficient or significance attribute after fitting. Attributes shall be considered unimportant and omitted if the respective coefficients are below the defined threshold point. The aim of lasso regression is to obtain a subset of predictors which minimize prediction error. Lasso regression is achieved by placing a restriction on the model's parameters which causes the regression coefficients for certain variables to shrink to zero. After shrinking, variables with a regression coefficient equal to zero shall be omitted from the model. Variables with non-zero coefficient of regression variables are more strongly related to the response variable. Therefore it might be helpful to do a lasso regression while running a regression analysis to decide how many variables the analysis must contain.

- **Tree-Based Feature Selection (Random Forest**) is one of the most common algorithms in machine learning. It is so effective because it provides good predictive efficiency, low over-fitting and simple interpretability in general. This interpretability is based on the fact that the value of each variable in tree decisions is easily extracted. In other words, the calculation of how much each variable contributes to the decision is simple.

- **Recursive feature elimination** is based on the principle of repeatedly constructing a layout and choosing either the best or worst performing function to set the feature aside and then replicate the procedure with the rest of the features. The process will be expanded until all the features in the data set have been exhausted. The attributes are then rated according to when deleted.

Classification algorithms such as KNN, RF, SVM, LR and NB are used to evaluate the validity of the selected features.

- **Logistic Regression**: is the model of classification, where test data probabilities are estimated. Uses conditional probabilities for the classification of food-secure and food-insecure problems with two possible outcomes 1 and 0.

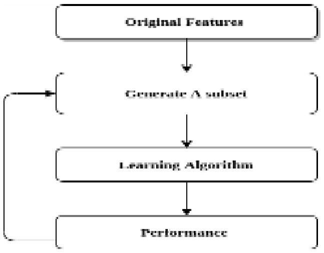

Fig. 1. Embedded Feature Selection method

K-Nearest Neighbor (KNN): is a classification algorithm which resides with all available features and classifies new cases based on distance functions or measure of similarity. KNN is the basic classification algorithm that takes into account all the data set points to be categorized within their class belongingness. It considers k-nearest points from the data point selected and ranks them in ascending order.

Support Vector Machine (SVM): Is a classification algorithm in which we have several kernel options depending on the fashion of the distribution of data. It can classify data in a number of linear ways but SVM gives us the best choice among all the available options. Forms of the kernels, linear, rbf, poly, and sigmoid.

Random Forest (RF): It is part of the ensemble learning and integrates several algorithms to obtain optimized efficiency. We combine decision tree classification algorithms multiple times in random forest classification.

Naive Bayes (NB): This is a Bayes Theorem-based mathematical classification methodology. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the algorithm that is fast, accurate and reliable and has high precision and speed on large data sets.

3. PERFORMANCE METRICS

Various assessment criteria are used to evaluate the efficiency of the selected features Accuracy, recall, accuracy, f1-measure and confusion matrix are the most commonly used evaluation metrics.

Classification Accuracy (CA): is the number of correct predictions divided by the total number of predictions multiplied by 100 to make it a percentage.

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

Recall: is measured on the basis of the number of True Positives (TP) divided by the number of True Positives (TP) and the number of False Negatives (FN).

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

Precision: is determined on the basis of the number of True Positives (TP) divided by the number of Negative Positives (TP) and False Positives (FP).

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (3)$$

F1 measure: is calculated based on precision and recall.

$$\text{F1 measure} = 2\,\frac{\mathrm{precision}*\mathrm{recall}}{\mathrm{precison}+\mathrm{recall}} \qquad (4)$$

Confusion Matrix is a metric shows correctly classified and misclassified samples from a given test data, as shown in the following table.

Table 1: confusion matrix

	Predicted		
Actual		Positive	Negative
	Positive	True positive	False Positive
	Negative	False Negative	True Negative

4. DESCRIPTION OF DATA SET

The data set used for this work was collected from the Central Statistics Agency (CSA) of the Federal Democratic Republic of Ethiopia. The data dealt with survey data on income consumption and expenditure from 2011 to 2016 at national level, and the survey covered both rural and urban areas of the country. The data set comprises 21 features including the class label (Zone, Woreda , Region, Town, K/ketema, Kebele, Household serial No, Household size, adult equivalent scale, Area of residence, Report level, Ecology, Household head sex, Household head age, Marital status of household head, Education attend, Weight, Annual Expenditure, Year, Net calorie and Food security status). The class label shows whether the household is food secure or insecure.

5. RESULT AND DISCUSSION

In this section, the experimental result of the proposed feature selection methods namely Lasso, Tree based and recursive feature elimination using the classifiers KNN, LR, SVM, RF and NB are presented. As Table II shows that recursive feature elimination feature selection method Random Forest classification Algorithm scores better with an accuracy of 99.9%. Table III show Lasso feature selection method and under the selected features are evaluated with different algorithm, from these algorithm, RF score better result which accuracy value of 96.3% accuracy and 96% ROC. Whereas under Tree-based feature selection RF score better result which is 96.7% and 97% of ROC followed by KNN, SVM ,LR and NB with accuracy result of 96.6%, 90.6%, 89% and 85% respectively.

Table II: Accuracy of recursive feature elimination method

Algorithm	FS-Type	No-F	Accuracy	Precision	Recall	F1 score	Roc
KNN	RFE	10	0.92	0.91	0.91	0.91	0.92
RF			0.9998	0.9996	1.00	0.9998	1.00
SVM			0.93	0.9557	0.884	0.9186	0.93
LR			0.80	0.796	0.739	0.766	0.79
NB			0.77	0.71	0.81	0.75	0.78

Table III: Accuracy of lasso method

Algorithm	FS-Type	No-F	Accuracy	Precision	Recall	F1 score	Roc
KNN	Lasso	4	0.946	0.961	0.915	0.937	0.94
RF			0.963	0.979	0.935	0.957	0.96
SVM			0.93	0.9557	0.884	0.9186	0.93
LR			0.907	0.893	0.897	0.895	0.91
NB			0.786	0.71	0.86	0.78	0.79

Table IV: Accuracy of Tree based feature selection

Algorithm	FS-Type	No-F	Accuracy	Precision	Recall	F1 score	Roc
KNN	Tree-Based	2	0.966	0.959	0.962	0.961	0.97
RF			0.967	0.960	0.965	0.963	0.97
SVM			0.909	0.920	0.867	0.893	0.90
LR			0.890	0.867	0.867	0.893	0.90
NB			0.85	0.767	0.954	0.85	0.85

6. CONCLUSION AND FUTURE WORK

In the machine learning domain, several feature selection methods were implemented with the goal of eliminating unnecessary or redundant features from the data set and getting better classification accuracy. Classifiers such as KNN, LR, SVM, RF, and, NB are used to validate accuracy of the selected features. The experimental results show that lasso feature selection approach outperforms with 96% accuracy using random forest classifier, followed by KNN, SVM, LR and NB with 95%, 93%, 90.7% and 78.6% respectively. While using tree-based approach with random forest classification performs better compared to others with 96.7% accuracy. Finally in recursive feature elimination method using random forest classifier score better result compared to others with an accuracy of 99.98%. Finally we conclude that feature selection plays a major role in increasing classification accuracy. Based on the experiment result random forest classifier is the best approach relative to others and recursive feature elimination is the best feature selection method

REFERENCES

[1] J. Mbukwa, "A model for predicting food security status among households in developing countries," *International Journal of Development and Sustainability*, vol. 2, no. 2, 2013

[2] L. C. Smith, H. Alderman, and D. Aduayom, Food insecurity in sub-Saharan Africa: new estimates from household expenditure surveys. Intl Food Policy Res Inst, 2006, vol. 146.

[3] Food and A. O. F. for Agricultural Development/World Food Programme, "The state of food insecurity in the world. The multiple dimensions of food security," 2013.

[4] M. Zhu and J. Song, "An embedded backward feature selection method for mclp classification algorithm," Procedia Computer Science, vol. 17, pp. 1047–1054, 2013.

[5] V. Kumar and S. Minz, "Feature selection: a literature review," Smart CR, vol. 4, no. 3, pp. 211–229, 2014.

[6] L. P. Dhyaram and B. Vishnuvardhan, "Random subset feature selection for classification." *International Journal of Advanced Research in Computer Science*, vol. 9, no. 2, 2018.

[7] B. Endalew, M. Muche, and S. Tadesse, "Assessment of food security situation in Ethiopia," *World Journal of Dairy and Food Sciences*, vol. 10, no. 1, pp. 37–43, 2015.

[8] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Esfs: A new embedded feature selection method based on sfs," 2008.

[9] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[10] M. W. Mwadulo, "A review on feature selection methods for classification tasks," *International Journal of Computer Applications Technology and Research*, vol. 5, no. 6, pp. 395–402, 2016.

[11] N. Kakwani and H. H. Son, "Measuring food insecurity: Global estimates," in *Social welfare functions and development*. Springer, 2016, pp. 253–294.

[12] W. Okori and J. Obua, "Machine learning classification technique for famine prediction," in *Proceedings of the world congress on engineering*, vol. 2, 2011, pp. 991–996.

[13] C.-J. Alisha, M. P. Rabbitt, C. A. Gregory, and A. Singh, "Household food security in the united states in 2016," United States Department of Agriculture, Economic Research Service, Tech. Rep., 2017.

[14] P. Yildirim, "Filter based feature selection methods for prediction of risks in hepatitis disease," *International*

[15] E. C. Blessie and E. Karthikeyan, “Sigmis: A feature selection algorithm using correlation based method,” *Journal of Algorithms & Computational Technology*, vol. 6, no. 3, pp. 385–394, 2012.

[16] M. S. Srivastava, M. N. Joshi, and M. Gaur, “A review paper on feature selection methodologies and their applications,” *IJCSNS*, vol. 14, no. 5, p. 78, 2014. p. 258, 2015.

[17] R. M. Barbosa and D. R. Nelson, “The use of support vector machine to analyze food security in a region of brazil,” *Applied Artificial Intelligence*, vol. 30, no. 4, pp. 318–330, 2016.

[18] B. Kumari and T. Swarnkar, “Filter versus wrapper feature subset selection in large dimensionality micro array: A review,” 2011.