Poverty Level Characterization via Feature Selection and Machine Learning

Jama Hussein Mohamud
Department of Electrical and Electronics Engineering
Anadolu University
Eskisehir, Turkey
jamahusseinmohamud@eskisehir.edu.tr

Omer Nazih Gerek
Department of Electrical and Electronics Engineering
Eskisehir Technical University
Eskisehir, Turkey
ongerek@eskisehir.edu.tr

***Abstract*—A persistent socio-cultural problem of mankind is "poverty", which requires accurate characterization in order to construct well designed policies for intervention. Unfortunately, the categorization along the poverty - wealthiness scale is not simply determined by applying surveys. Population is large, subjective opinions are usually biased, and available data are only indirectly related. In this paper, we attempt to identify poverty levels using feature selections from these indirect observations and machine learning techniques. In poverty assessment, similar to many other classification problems, it is crucial to know how any feature contributes to the classification of each class of poverty. We designed an approach that (1) extracts a subset of features that best characterize each poverty class, (2) examines how this subset affect the chosen class and finally (3) employ ensemble models. In this research, we adopt the Proxy Means Test (PMT) for labeling the data that was obtained from the Inter-American Development Bank of Costa Rica. Through this approach we analyze poverty classes within a multidimensional feature space perspective, contrary to the classically used single dimensional perspective defined as "living on a consumption expenditure of less than the predefined income threshold". The application and usefulness of our proposed framework is tested on the mentioned dataset using 85-15 data folding.**

***Keywords*—poverty characterization, poverty measurement, poverty identification, multidimensional poverty, feature extraction, machine learning.**

I. Introduction

Despite its obvious importance, poverty classification or prediction is time-consuming, expensive and tough in developing countries. Data scarcity and security complications are reasons that avoid accurate assessment in some countries. Even when various different data are collected from households, it may still be hard to define poverty. Besides, poverty is a heterogeneous problem and has many aspects varying according to the geographical location and time. For example, being poor in America is quite different from being poor in Asia or Africa.

According to Sen, measurement of poverty has two separate complications, (I) Poverty identification (ii) Creation of an index to measure poverty [1]. Income is classically used to overcome the first problem, but the second part is long debated by researchers and practitioners [3]. Researchers have proposed several poverty-measurement indices to solve the second complication, and one of them is the multidimensional poverty index (MPI) by [2-6]. Luckily, machine learning (ML) models can help us target poverty by learning from datasets that are labeled using MPI's. Unfortunately, ML models are designed only according to the data, and they don't help understanding reasons behind predictions.

Although there is little research on ML application to poverty estimation, thanks to recent advances in data obtainability, scholars have started to use big data and ML to predict poverty levels in emerging countries. The concept of proxy means test (PMT) becomes a common tool for targeting poverty by using visible characteristics of the household when income data is not presented [10]. In [7], ML models were employed to improve the accuracy of PMT poverty targeting tools, where stochastic ensemble methods were exploited to boost out-of-sample performance. In another study, ML models that best identify B40 - Bottom 40 % - household population in Malaysia were selected [8]. Their study showed that decision tree models perform well. In that work, they assumed that different variables explain "falling into" and "escaping from" poverty, which enables utilization of ML methods to categorize the respective strength of these variables [9]. Unlike these studies, our framework focuses on obtaining the features that best characterize each class.

The proposed approach employs i) a method of poverty prediction based on the multidimensional poverty concept that takes into account various household characteristics, contrary to the conventional measure that is based on a single dimension of poverty; ii) a novel feature extraction framework to find features that put a household in that specific class of poverty; and iii) four classes of poverty levels, instead of traditional two-class scheme (poor/non-poor). Even though modeling is an important aspect of our work, the more emphasizing part is mining the right features that characterizes each poverty class.

The rest of the paper Section 2, where the utilized data is reviewed. Section 3 focuses on our approach to Poverty Level Characterization, Section 4 reviews the classification success measures, and Section 5 discusses applied models and results.

II. Data

The utilized dataset is based on Costa Rican data, provided by Inter-American Development Bank (available from Kaggle). It consists of four classes (Extreme poverty, Moderate, Vulnerable and Non-vulnerable), and various household characteristics. The PMT (Proxy Means Test) model was

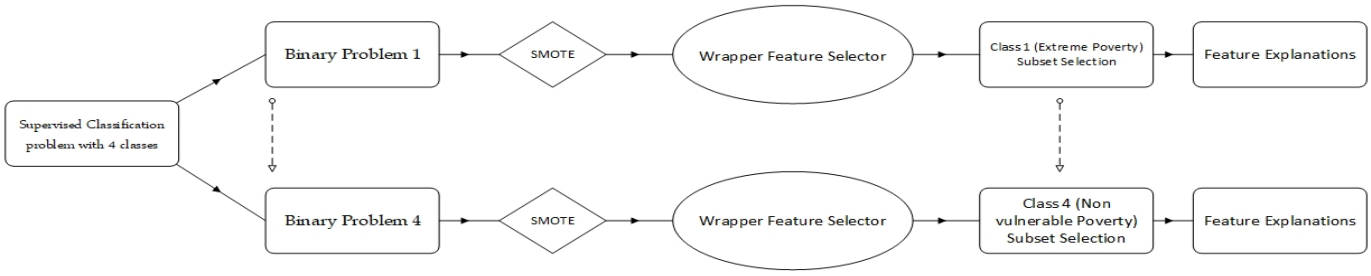

Fig. 1. General frame-work for each poverty class feature subset selection

used to label the data. The PMT uses observable household characteristics as a proxy to estimate the household socio-economic status [10]. Verification of the accuracy of PMT was left beyond the scope of this work, and assumed as the ground truth for ML training to learn and predict unseen instances. We explore the contributions and impacts of the variables to the target, which first requires certain time-consuming pre-processing stages such as missing value interpolation, aggregation of household characteristics, computing feature interactions, performing feature transformations, etc.

In most cases, the non-monetary measure of poverty is composed of different forms of dimensions (e.g. current assets, education, natural resources) that all contribute to an individuals welfare. Unfortunately, the disproportion (Fig. 2) of the classes in the dataset and the complexity of the class memberships (Fig. 3) makes it tough for several linear regression models to achieve high accuracy. ML models are known to better handle such complicated datasets.

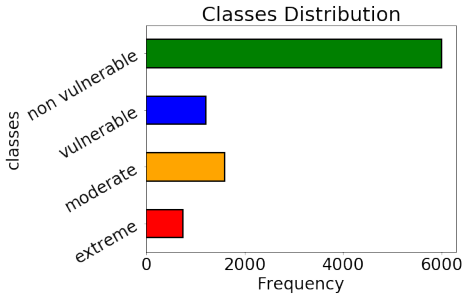

Fig. 2. Disproportion of the classes

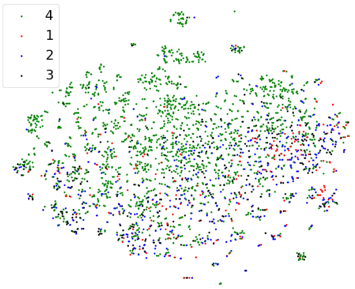

Fig. 3. Visualization of poverty classes by t-SNE. (1 = Extreme, 2 = Moderate, 3 = Vulnerable, 4 = Non Vulnerable)

III. General Frame-work for Feature Contribution Extraction

The non-discriminate representation of poverty status of an individual or household imposes that the considered problem should be considered as a multi-class problem, since different people may have differences in the features that represent their poverty level. Owing to this, we proposed a frame-work to extract feature contributions to each class, using a four-stage strategy: binarization, Synthetic Minority Over-sampling Technique (SMOTE), wrapper feature selector, and feature explanation (Fig. 1). A number of similar approaches have been introduced in the literature [11-13]. These methods are generally called class-specific feature selection, but our purpose in this study is not simply ”feature selection for better accuracy”; rather, we need to visualize features with high discriminative power, which are able to differentiate within a sub-class of the problem.

- *Binarization:* The first stage is class binarization, which transforms a k-class problem into several binary problems. This allows each class to be compared against all others, rendering the process suitable for ML [14].
- *SMOTE:* After the binarization phase for each class, an ”imbalance” problem appears. We see that the number of elements in ”non-vulnarable” class is 4-6 times more than any other class, causing the recall optimization unfavourably for the within-poverty classifications. To overcome this issue, we proposed to use a sampling method, SMOTE, which over-samples the minority class by creating synthetic minority class instances. The synthetic instance are created along the line segments linking all of k-minority class nearest neighbors [15].
- *Class Specific Subset Selection:* At this stage we had 4 different binary classes and our target was to find a discrete subset for each class which best isolates this class from all the other classes. Thus, we considered common feature selectors and obtained attributes that prompted someone to be under this class of poverty (see Table 1). In our case we employed Wrapper feature selectors. These selectors are based on greedy search algorithms and they select a set of variables that produces best results for a given ML algorithm [16].

In the literature, there are several wrapper feature selection methods, including sequential backward selection (SBS), sequential forward selection (SFS) and Exhaustive feature selection. Since the dimensionality of our problem is high (about 209) with several attributes for each feature, we preferred to

TABLE I
CONTRIBUTION OF FEATURES TO EACH POVERTY CLASS

Levels of Poverty	Features	Descriptions	Dimension	LIME Explanations (for couple of Examples)
Extreme poverty	v18q1	Number of Tables	Current Assets	Deprived
	epared3	IF Walls good	standard of living	Deprived
	instlevel1	No Level of Education	Education	Yes
	rent_per_room	Rent Per Room	-	0 (mostly don't pay rent)
	instlevel9	Postgraduate H.education	Education	Mostly lack Postgraduate higher Education
	pisonatur	IF Floor is natural Material	standard of living	Not Natural Matrial (Deprived)
Moderate	hogar_adul	Number of Adults in the Household	-	Adults (1,2,3,4,5) are indicative to fall in to this class
	hogar_mayor	Indivuduals Age > 65	-	Almost 30 % of this class, Age >65
	age_std	Age Standard deviation	-	Probability of moderate class increases if std >22
	paredmad	Material on the outside wall is wood	Physical capital	About 80% in this class walls are not wood
	paredzinc	Material on the outside wall is zinc	Physical capital	Deprived (Almost 98% wall is not zinc)
	Escolari_mean	Average years Education	Education	[6.75 - 8]
Vulnerable	hogar_nin	Number of Children (0-19) in household	-	1
	instlevel2	Incomplete Primary Education	Education	Yes
	pisocemento	Material on the floor is cement	Physical capital	(around 17 % have cement on the floor)
	meaneduc	Average years of Education for Adults	Education	6
	sanitario1	IF no Toilet In the Dwelling	Physical capital	0 (Almost Every HH has Toilet in the dwellig)
	epared3	IF Walls good	Physical capital	Walls good (not deprived)
Non Vulnerable	phone_per_person_household	Phone per person in household	Physical capital	1 (indicates almost every person has telephone)
	television	Television	Physical capital	1 (indicates HH has televion)
	etecho3	IF roof is good	Physical capital	1 (HH has a good roof)
	dependence	Dependence Rate	Social capical	Mostly dependence rate is 0
	escolari_mean	Average years of schooling	Education	>= 13 (higher the more likely to be this class)
	eviv3	IF floor is good	Physical capital	1 (indicating floor is good)

TABLE II
RANDOM FOREST CLASSIFIER RESULTS

Classifier-Metric	Train F1 macro score	Validation F1 macro score	Test F1 macro score
Random Forest Classifier	0.57	0.414	0.425

use step forward feature selector (SFFS). The method is as follows: In the first step, the classifier is evaluated against each feature, one at a time, then the feature that performs the best is retained. Next, the selected feature is combined one-by-one with all other features and the combination of two features that performs the best is reserved. The evolution continues until a specified success rate is reached or the feature set success starts to deteriorate [16]. Finally, the features which best describe each poverty class are obtained as in Table I. To further strengthen our argument and understand how such features really trigger a household to be poor, we used the Local Interpretable Model-agnostic Explanations (LIME) model explanation technique (last column of Table I). This technique provides the outcome of any model in a decipherable and truthful manner, by learning an interpretable model locally around the prediction [17]. Such determination of the factors behind predictions is necessary, especially when the model is used for policy-making. In a humans poverty status, predictions cannot be acted upon isolatedly as the penalties may be intolerable. Here, LIME data assists us to demonstrate the effect of the features we retrieved from our frame-work on the observations we are investigating. It provides qualitative interpretation of the relationship between the instance's variables and the model's prediction [17].

IV. CLASSIFICATION

Following the selection of representative features, we had gone through further aggregation of household characteristics,

computing feature interactions, and performing feature transformations in order to build a consistent model. Then, we applied the data to a random forest classifier. The smallness and skewness of the data made it tough for the classifier to perform well in separation of less populated classes. The non-vulnerable class (class 4) is over-represented compared to other classes. If we are to exactly separate class 4 from all other classes, our decision boundary would fail to identify a fair decision boundary to isolate between the other three classes (Extreme, Moderate, Vulnerable). Therefore, an overall F1-macro score optimization is adopted and presented (Table II). Eventually, we observed that the combination of our framework and LIME method seems to be a promising move to depict each poverty class.

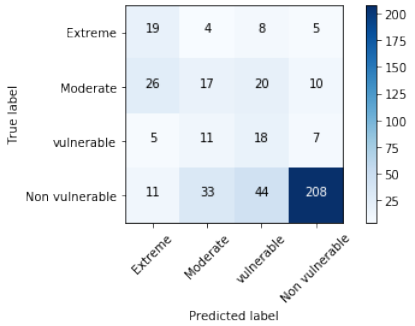

Fig. 4. Confusion Matrix

V. Discussion of Results

The analysis of this paper yields interesting results on the non-orthodox problem of classification of poverty levels from indirect data. Instead of a combined selection of features for all classes, we chose 6 features from the combined pool of features that explains *each* class the best. These features have different influence on different classes. For example, we see how dimensions like current assets (Table 1), and standard of living (housing related feature) best contribute to class 1 (Extreme Poverty), whereas they are not among the features to separate class 3 from other classes. Similarly, it can be observed that different levels of education contribute differently to the classes. For instance, deprivation of basic education is a characteristic of extreme poverty, while good education is an indicator of the non-vulnerable class. Bearing in mind that we don't have any features that are symptomatic of unitary variables (like income); this further reinforces the theory that poverty is a multidimensional concept that depends on the cause of many characteristics.

The overall classification performance is evaluated using the F1-macro score and the confusion matrix. We were forced to utilize F1-macro score on the *test* data, as we don't have access to the actual labels of the test data. The F1-macro score is automatically provided by the international web challenge of the Inter-American World Bank, however they do not provide the labels of the individual instances. Yet, we do have the labels for the training data, so we have divided the train data into 85% train and 15% validation portions and provided a confusion matrix (see Fig.4). The results show that the class members are strongly mixed, only with an exception of the "non vulnerable" class.

VI. Conclusion

In this study, we proposed a general framework to retrieve a subset of features that mostly characterize each class of poverty. Then, we used the LIME explanation technique to further illustrate the effect of the chosen features on the outcome. Finally we fitted Random Forest classifier to our data and evaluated on unseen validation and test data. We have observed that, instead of a unified ensamble of features, different set of features are required to describe different levels of poverty. Unbalancedness, skewness, missing data, and too much mixing of poverty class item points all contribute adversely to the classification performances. Yet, this work makes a proof that feature selection and classification are reasonable tools that can be used for poverty categorization.

References

[1] A. Sen, Poverty: An Ordinal Approach to Measurement, Econometrica, vol. 44, no. 2, p. 219, 1976.

[2] S. Alkire and M. E. Santos, "Measuring Acute Poverty in the Developing World: Robustness and Scope of the Multidimensional Poverty Index," World Dev., vol. 59, pp. 251274, 2014.

[3] F. Bourguignon and S. R. Chakravarty, "The Measurement of Multidimensional Poverty," J. Econ. Inequal., vol. 1225, no. February, pp. 4142, 2003.

[4] S. Alkire and M. E. Santos, "Multidimensional Poverty Index," Oxford Poverty Hum. Dev. Initiat., no. July, pp. 18, 2010.

[5] S. Alkire and S. Seth, "Multidimensional Poverty Reduction in India between 1999 and 2006: Where and How?," World Dev., vol. 72, pp. 93108, 2015.

[6] N. Nari and N. Quinn, "Alkire-Foster Method The Global MPI Policy Use Public Communication The Global Multidimensional Poverty Index," no. November, 2017.

[7] L. Mcbride and A. Nichols, "Improved poverty targeting through machine learning: An application to the USAID Poverty Assessment Tools," p. 24, 2015.

[8] N. S. Sani, M. A. Rahman, A. A. Bakar, S. Sahran, and H. Mohd, "Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification," vol. 8, no. 4, pp. 16981705, 2018.

[9] S. Narendranath, S. Khare, D. Gupta, and A. Jyotishi, "Characteristics of Escaping and Falling into Poverty in India: An Analysis of IHDS Panel Data using machine learning approach," 2018 Int. Conf. Adv. Comput. Commun. Informatics, pp. 13911397, 2018.

[10] World bank, "Measuring income and poverty using Proxy Means Tests."

[11] B. B. Pineda-Bautista, J. A. Carrasco-Ochoa, and J. F. Martnez-Trinidad, "General framework for class-specific feature selection," Expert Systems with Applications, vol. 38, no. 8. pp. 1001810024, 2011.

[12] A. Roy, P. D. Mackin, and S. Mukhopadhyay, "Methods for pattern selection, class-specific feature selection and classification for automated learning," Neural Networks, vol. 41. Elsevier Ltd, pp. 113129, 2013.

[13] A. M. P. Canuto, K. M. O. Vale, A. Feitos, and A. Signoretti, "ReinSel: A class-based mechanism for feature selection in ensemble of classifiers," Applied Soft Computing Journal, vol. 12, no. 8. Elsevier B.V., pp. 25172529, 2012.

[14] Y. S. Erin L. Allwein, Robert E. Schapire, Reducing Multiclass to Binary, J. Mach. Learn. Res., vol. 1, pp. 113141, 2000.

[15] Ott SM., "W. P. K. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, Synthetic Minority Over-sampling Technique (SMOTE), J. Artif. Intell. Res., vol. 16, pp. 321357, 2002.

[16] L. A. Z. (eds. . Isabelle Guyon, Steve Gunn, Masoud Nikravesh, Feature Extraction: Foundations and Applications, vol. 91. 2017.

[17] C. G. M. T. Ribeiro, S. Singh, "Why should i trust you?: Explaining the predictions of any classifier," in International Conference on Knowledge Discovery and Data Mining, 2016, pp. 11351144.