Predicting Microfinance Participation in Indian Villages

Govind Manian and Karen Shen

December 15, 2012

Abstract

Using data from a microfinance organization operating in southern Indian villages, we use a logistic regression classifier to predict the probability that an individual will join the microfinance program in her village. We perform our classification in two initial settings: one where we use only demographic and graph features of an individual and one where we also use labellings on the other nodes in the graph. As expected, the latter is able to improve our predictions. We therefore try an iterative classification algorithm to attempt to achieve better prediction accuracy than the base model without being given knowledge of the other labellings.

1 Introduction

Between 2006 and 2010, the microfinance organization Bharatha Swamuki Samsthe (BSS) identified 75 rural villages in southern India in which they hoped to start operations and which previously had almost no exposure to microfinance institutions or access to formal credit. When they begin work in a village, they do so by seeking out people in the village to serve as "injection points" who they expect to be well-connected within the village–teachers, leaders of self-help groups, shopkeepers, etc.–and holds a private meeting with these village "leaders" to teach them about the program and to ask them to help spread the word about microfinance in the village.

In conjunction with BSS, the authors of Banerjee, et al. collected social network and demographic data in these villages in order to show that the "influence" of these leaders (specifically, their eigenvalue centrality) in the social network graph of a village was positively correlated with the village's overall participation level in the microfinance program. [1]

Using this same dataset, we attempt to make similar conclusions at the node-level. Specifically, we seek to answer the following questions:

- Given an individual's characteristics and relationships, what is the probability that she will hear about and decide to participate in microfinance?

- Can we use the relationship data and limited data on who else in the village decided to participate in microfinance to update these prior probabilities and find the posterior probability that an individual participated in microfinance when we know that certain people in her village did (or didn't)?

2 Data and Preprocessing

The data consists of three sources:

- Household Census: Each household was asked for its religion, caste/sub-caste, roof type, number of rooms and beds, latrine type, and whether their home is owned or rented.

- Individual Survey: About half of BSS-eligible villagers in each village and asked for **demographic** information such as their age, religion, caste, subcaste, mother tongue and other languages spoken, whether or not they are a native of the village or how long ago and why they moved, where they work, whether they belong to a savings group, have an election card or a ration card. They were also asked to provide information on their social **network** by listing other people in the village who they interacted with along thirteen different dimensions: who they lend/borrow money or material goods (kersone, rice) to/from, give/seek advice from, go to temple with, people they visit and who visit them.

- Participation Data: The IDs of the village leaders are recorded, as well as who ultimately joined the program.

By the time data collection ended, BSS had entered 43 of the 75 villages they initially identified. Due to keying and other errors in the data (e.g. nonunique person IDs), we restricted our sample to 36 of these 43 villages. Because we only have data for the surveyed individuals, we also chose to restrict our social network graphs to the graphs between surveyed individuals. This eliminates the potential for asymmetries in the network, e.g. where a surveyed individual is not listed as many times as she lists other people, solely because the others did not have the opportunity to list her because they were not surveyed. Across the 36 villages, a little less than 10,000 people were surveyed. After cleaning, we were left with 9,431 clean observations. Of these, 551 were microfinance participants, and between 5 and 22 in each village were identified as village leaders.

3 Methodology

3.1 Features

We use three types of features:

- Demographic: We use the features from the individual surveys. Because many of these features are categorical, we do some initial dimensionality reduction by restricting to features possessed by 15 or more villagers. Removing these relatively uninformative variables reduces the sparsity of the matrix and leaves us with 73 features.
- Graph Measures: We construct various measures of well-connectedness to use as additional features. Intuitively, a more connected person should be more likely to hear about microfinance, and possibly be convinced to join it. To that end, we use a person's clustering coefficient, betweenness centrality, eigenvector centrality, degree, and closeness. Our adjacency matrix is initially defined so that all edges are undirected and represent any type of relationship between two nodes.
- Participation: If we are given information on who else in the village is participating in the microfinance program, we can improve upon the graph measures used above by using the fact that an individual is probably more likely to participate in microfinance if she has a relationship with someone else in the program. We therefore add features such as whether or not an individual has a microfinance participant "neighbor" in the social network graph, how many neighbors she has, and her minimum and average distance to the leaders in the village.

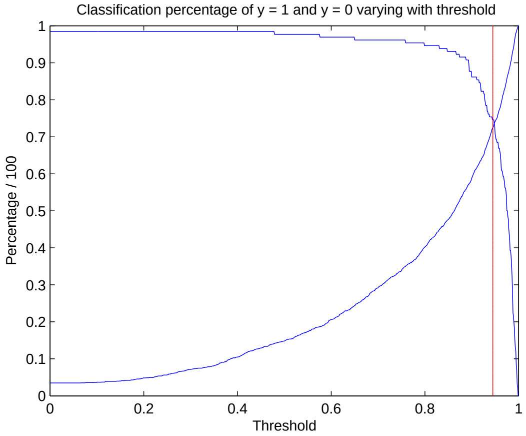
3.2 Learning

We chose to classify using logistic regression. Because we split most of our features into categorical "dummy" variables, we decided that an assumption of linearity was unlikely to be too harmful. We confirmed by trying principal components regression and SVM with various kernels and saw little to no improvement in accuracy.

To get a sense of the importance and predictive power of each of the three feature classes above, we fit each feature class separately before combining them into two larger models, one with both the demographic and graph features, and another with all three feature classes. In fitting the demographic model, we used forward feature selection to identify the most relevant predictors.

One significant obstacle we faced was the problem of class imbalance. Because less than 6 percent of our sample participates in microfinance, a model with a good prediction accuracy, by the usual standards, would be one that just classifies everyone as a nonparticipant. Thus, it was necessary to adjust our measure of model accuracy from classification error and the probability threshold above which we classified a node as a participant. As a preliminary attempt, we picked the probability threshold at what we called the "crossover" point, where the percent of correctly classified participants was equal to the percent of correctly classified nonparticipants (see Figure 1). We chose this threshold because we do not have any information to suggest that false positives are more problematic than false negatives, discussed further in Future Work. This measure is advantageous because it captures both types of error in one assessment of model accuracy. Class imbalance can also distort the fit of a model since many learning algorithms assume a balanced dataset [6]. A common solution is to either undersample the majority class or oversample the minority class in order to construct a balanced dataset. In these cases, a correction should be applied to the logistic regression coefficients [7]. We tried both undersampling and oversampling, using the correction suggested in [7].

Figure 1: "Crossover" Accuracy of Full Model



3.3 Prediction

There are essentially no villagers with links between the villages so we treat them as separate graphs. We furthermore assume that villages are homogenous when working with demographic data. Given that we had over 20 examples per feature, as in [3], we split our data into a training set of 18 villages, a test set of 9 villages, and then a "vaulted" set of 9 villages. We trained on 18 of these villages, and given our models, we seek to evaluate them on a test set of 9 villages. We choose to do so under different informational (no information, limited information, and full information) settings and using different models.

In the no information setting, we are not given any information about who participated in the microfinance program in the village. Thus, we classify using our model that was fit using only the demographic and graph features. On the other hand, in the full information setting, we are given the label of every node in the graph except the one we are making a prediction for. Thus, we can use our full model to make predictions.

To make use of the training data (where we do have participation data) without requiring the same of the test data, we implement an Iterative Classification Algorithm (ICA) due to Lu and Getoor [5]. The basic structure of the algorithm is as follows:

- "Bootstrap" stage: Assign an initial classification to every node without using the participation data.

- "Iteration" stage: Iteratively apply the full model to classify every node until a termination condition is reached, i.e. the number of classifications doesn't change by more than τ from one iteration to the next.

Clearly, the more accurate the bootstrap stage is, the more efficient the algorithm and the more likely we are to get the right classifications. We thus tried two different versions of the bootstrap stage. In one, we assigned initial classifications using the partial model (with demographic and graph features). In another, we used our prior knowledge of the village leaders to label just the leaders as microfinance participants.

After applying this algorithm, we found that we generally overpredicted the number of microfinance participants, because there was too much "diffusion" from the initial nodes. We found that due to the nature of our participation features, it was unlikely for a node to go from being labeled as a participant in one iteration to a nonparticipant in the next. This is because having a neighbor do microfinance is a strong predictor in our model of microfinance participation since participants are so rare, and so once a microfinance participant is labeled it becomes likely that almost all of her neighbors will also be classified as participants. Our solution was to "slow" the diffusion by using a weighted average of the participation features in the current iteration and in the past iteration to make predictions. We can select this weight by cross validation.

4 Results

4.1 Feature Selection

Feature selection using the demographic data selected six vocations (painter, teacher, carpenter, electronics, waterman, mechanic) and four subcastes (Acharya, Alamatha Gowda, Totigaru, Besthru) as its top ten features. We should be careful of making overly optimistic claims about the importance of these vocations and subcastes, since all of these categories as well as microfinance participation itself appear in relatively small numbers and so a "highly" predictive feature may just indicate a category of about the right size and with some overlap with the set of microfinance participants. Nevertheless, this result does suggest some differences that future programs may be able to take advantage of.

Figure 2: ICA vs Baseline, weight=.01

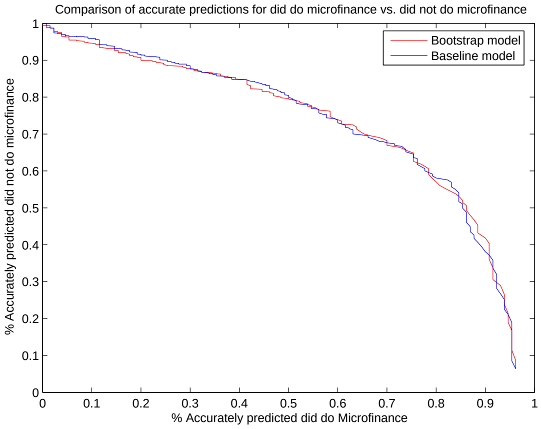
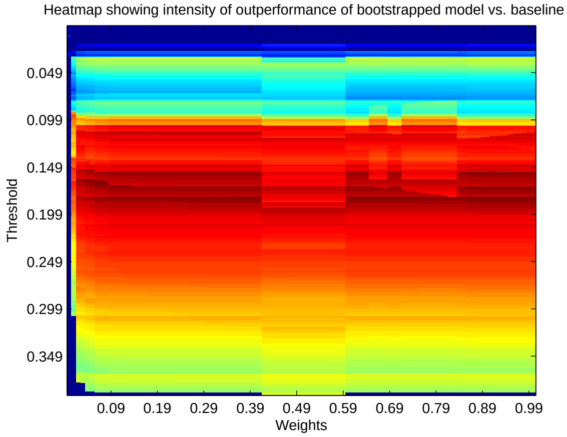


Figure 3: Heat Map of Outperformance



4.2 Classification

Model	TrTh.	TeTh.	Oversampling
Demographic	.68	.76	.69
Graph	.62	.55	.68
Participation	.60	.60	.64
Partial (D, G)	.68	.76	.69
Full	.71	.79	.74
ICA Leaders	.59	.61	-
ICA	.59	.69	-

Above we have recorded the different "crossover" accuracies for our various models and algorithms. The first column refers to the accuracy when we make classifications based on the threshold chosen on the train set. The second column refers to the accuracy if we were able to choose the threshold given the test set. The third shows the effect of oversampling and applying the necessary coefficient corrections, again using the training threshold. We generally see a small improvement from the imbalanced model.

The demographic model uses all demographic features; the graph model uses only network features; the particpation model uses average distance to local leaders, minimum distance to local leaders; and the partial model uses demographic and graph data and the full uses demographc, graph, and participation. ICA is, as before, the Iterative Classification Algorithm. As expected, the individual and partial models do worse than the full model. It also appears that our ICA was not able to achieve the same crossover accuracy as the no-information model, with the leaders version doing particularly badly. Upon further review of the data, we found that many of the leaders did not themselves end up participating in microfinance. It seems like the algorithm is unable to overcome the error of labelling them as participants in the first iteration in order to start the diffusion process.

However, for the ICA where we used the partial model to make initial classifications, we found that while the ICA did not succeed in improving on the crossover accuracy, it does dominate the partial demographic and graph model at some points of the sensitivity-specificity curve. That is, for a fixed positive ICA model classification accuracy, the corresponding negative classification accuracy is strictly superior to that of the partial demographic and graph model. This can be seen in in Figure 4.2, which shows the sensitivity-specficty curve for the baseline model (blue) and the ICA model (red) with a diffusion rate of 1%.

In Figure 4.2, a heatmap shows the space and intensity in which the ICA outperforms the partial model. The x-axis in this plot is the weight, or diffusion rate, in the ICA model. As the weight increases, information diffuses more quickly, so that when the weight is one, the diffusion is immediate. The y-axis shows the classification threshold above which response predictions are labelled positive. We find that, for diffusion rates less than 40%, and typically above 5%, there exists some range of thresholds such that the ICA model strictly dominates the partial model. Furthermore, in general, there is a band of thresholds between 0.2 and 0.3 with greatest outperformance over all diffusion rates.

5 Future Work

One of our initial reasons for using logistic regression was for its interpretability. Identifying features that are particularly predictive of microfinance take-up are of particular interest in guiding real world microfinance programs. However, we have learned recently of methods to interpret algorithmic model structures from Breiman [2]. Given that we are not to the functional form of logistic regression, using algorithmic models may prove worthwhile.

We believe that Random Forests may be a good candidate for two reasons. The first is that BSS currently chooses new microfinance injection points using a decision tree and tree structures determined algorithmically would inform the current selection process. The second is that our data set is relatively noise-free and therefore a good candidate for boosting methods such as boosted trees or random forests [3]. This method - and others - can be extended by determining from BSS whether it their resources constrain them to a maximum ratio of acceptable false positives to positive results. This would allow us to construct a loss matrix modeling actual costs of false positives and negatives to modify the Gini coefficient used when pruning trees.

It may be also possible to model the transmission of information about microfinance from one person to another using a dynamic Bayes network and improve classifications by finding the most probable paths for transmission to take. The participation data was actually collected at different periods of time, but this panel data has yet to be released, but would serve as a useful comparison for such a model.

There are also many aspects of the dataset that we were not able to fully explore, but that could be helpful in improving our classification results. For example, we construct our adjacency matrices as undirected adjacencies that simply indicate that either individual listed the other on at least one of the thirteen dimensions of relationship they were surveyed on. It is possible that for our problem, certain types of relationships matter more than others (e.g. giving/receiving advice) and that we could take advantage of these differences to improve classification. As with many graph problems, there is always the possibility that better measures of the underlying property that we are trying to capture with our model exist.

6 Conclusion

Using logistic regression and an iterative classification algorithm, we are able to train a model of microfinance participation in Indian Villages using both demographic and network features, and to classify entire village graphs at once. We show that this model is able to improve on the partial model that doesn't use features such as how many of an individual's neighbors are participating in microfinance in certain situations. Our full model is also able to attain a 79 percent classification accuracy.

References

[1] Abhijit Banerjee, Arun Chandrasekhar, Esther Duflo, Matthew O. Jackson (2011) “The Diffusion of Microfinance.”

[2] Leo Breiman (2001). ”Statistical Modeling: The Two Cultures.”

[3] Trevor Hastie, Robert Tibshirani and Jerome Friedman (2009). ”The Elements of Statistical Learning”

[4] Matthew O. Jackson, Tomas Rodriguez-Barraquer, Xu Tan (2012) “Social Capital and Social Quilts: Network Patterns of Favor Exchange,”

[5] Qing Lu and Lise Getoor, (2003) “Link-based Classification.”

[6] Haibo He and Edwardo A. Garcia, (2009) “Learning from Imbalanced Data.”

[7] Gary King and Langche Zeng (2001) ”Logistic Regression in Rare Events Data.”

[8] Jennifer Neville and David Jensen, (2003) “Collective Classification with Relational Dependency Networks.”