Knowledge Discovery from Agricultural Survey Data: The Case of Teff Production in Ethiopia

Bruk Legesse
brucklyn21@gmail.com

Berhanu Borena
PhD Candidate, Addis Ababa University, Ethiopia
berhanuborena@gmail.com

Abstract

This paper examines classification and association pattern of agricultural productivity survey data of Teff in Ethiopia. Although different approaches of Statistic, Technology, Metrology and Geology were applied to identify factors contributing for improvement of Teff productivity, there remains a lot of work to bring overall change on productivity of Teff. This research focused on identifying relationships between attributes of agricultural productivity survey data of Teff with input mechanisms and techniques to clearly understand the nature of production of Teff in Ethiopia. In order to conduct this research we followed the approach called CRoss-Industry Standard Process for Data Mining (CRISP-DM). The results of this research using classification and association rule have discovered that the technique of data mining is applicable to generate knowledge from agricultural productivity survey data of Teff in agricultural production. Association algorithms: Apriori, Tertius, and FilteredAssociator were used to select the features/attributes that have direct and indirect relationship with the target class in addition to the expert judgment. Subsequently the classification algorithms namely J48, RandomForest, REPTree used to generate rules. The generated model is represented in terms of IF THEN RULE.

Keywords: Teff Productivity; Data mining; Classification; Decision Tree

1. Introduction

Agriculture is the foundation of Ethiopia's economy, accounting for half of the gross domestic product (GDP), 83.9% of exports, and 80% of total employment [1]. Ethiopia's agriculture is plagued by periodic drought and soil degradation caused by inappropriate agricultural practices and overgrazing, deforestation, high population density, undeveloped water resources, and poor transport infrastructure [2].

Teff is one of the cereals believed to have originated in Ethiopia between 4000 BCE and 1000 BCE. Teff accounts for about a quarter of the total cereal production in Ethiopia [4]. Teff has been widely cultivated and used in Ethiopia, which is used for making Injera. It is now raised in the U.S., in Idaho in particular, with experimental plots in Kansas, American customers including people from traditional Teff consuming countries and also people desiring to eat Teff to avoid glutens that irritate celiac disease [5]. Teff is native to Ethiopia where it accounts for one quarter of the total cereal production.

With the growth of population consuming Teff the need to improve its production is felt and given due attention. In this effort Although different approaches of statistic, technology, metrology and geology were applied to identify factors contributing for improvement of Teff productivity, there remains a lot of work to bring overall change on productivity of Teff Teff productivity were not thoroughly investigated using data mining approaches.

In this research we employed data mining approach to uncover previously unknown patterns related census data and Teff productivity.

2. Related Work

The literatures covered below are taken from the perspective of what kind of data input used, amount of instance, method or approach used, tool and algorithm used, output data and deployment.

Revathi and Hemalatha [9] presented a brief idea of some of the widely used data mining techniques over cotton growth and development. In this work some of the data mining techniques are discussed and presented such as Naïve Bayes, Decision tree, Multilayer Perception. The cotton seed dataset consists of 500 instances, with 24 attributes of varying quality - Good, Average and Bad. Their experiment shows that J48 decision tree predicts the best performance from other classifiers used for experiments. They also found the time taken to build the J48 and Random Tree take less seconds from other classifiers.

Similarly Peyakunta and Singaraju [6] compared the effectiveness of the classification algorithms Genetic algorithm, Fuzzy classification and Fuzzy clustering. They found that the accuracy rate of Fuzzy Classification rules is higher than Fuzzy C-means algorithm; Fuzzy C-means algorithm is more suitable for Unsupervised Fuzzy soil data.

The work of Rao and Das [7] was also on the classification of herbal garden to classify the herbal gardens information. The data was collected from different herbal gardens and developed a database and uploaded online at www.herbalgardenindia.org. This website at present has a total of 71 herbal gardens from all over the country that are registered as members in this network. A total of 1024 species are available in these herbal gardens from which the majority of the species present in the garden are herbs of plant type. It is noted that the total number of quantity of planting material available in the entire registered herbal garden is 9,060,426 cuttings [7]. The initiative behind this experiment was the question that what type of habit of species is present in which location. They modeled these questions to find the relationship between location of species and their habit.

Another work by Gholap *et al.* [8], proposed an analysis of soil data using different algorithms and prediction techniques such as Naïve Bayes, J48, and JRip with the help of the WEKA data mining tool. They found that J48 is a very simple classifier to make a decision tree, and it gives the best result in the experiment.

3. Method

3.1 Data Collection and Preparation

More than 80% of researchers working on data mining projects spend 40%-60% of their time on cleaning and preparation of data [10].

Before data preprocessing data understanding is needed. In order to accomplish the study the data is collected from FDRE Ministry of Agriculture and CSA (Central Statistical Agency) database and library. The initial target dataset contains a total of 135,524 records with originally 44 attributes.

After the initial data collection, the dataset produced is converted into CSV (Comma Separated Variable) file format to allow them to be used in WEKA.

The activities during the data preparation phase included data cleaning, data selection, attribute or feature selection, transformation and aggregation, integration and formatting.

Data cleaning is the process of examining data and determining the existence of incorrect characters and mis-transmitted information [3]. The activities during this phase includes deleting attributes and data, filling missing values as advised by a domain expert, identifying and removing outliers and resolve inconsistencies that occur in all field attributes.

Irrelevant or unneeded data are usually eliminated from the data mining database before starting the actual data mining function. Other criteria for excluding data may include resource constraints, cost, restrictions on data use, or quality problems [3].

We deleted 15 attributes including attributes that don't have direct relationship with the productivity of Teff as the experts judge.

3.2 Attribute/feature selection

Instances are evaluated and classified based on the values of their attributes. Theoretically, decision tree could determine relevant attributes for classification automatically using the concept of

information gain or entropy without manual efforts. Of course, some of the attributes of an instance may be irrelevant to the process of classification as well as association and thus should be excluded.

Missing data is a common problem in statistical analysis and Data Mining. Rates of less than 1% missing data are generally considered trivial, 1-5% manageable. However, 5-15% requires sophisticated methods to handle, and more than 15% may severely impact any kind of interpretation [11].

Accordingly the recommended data size we can deal with is attributes having up to 20% of missing values. Attributes more than 20% missing value were removed. These include:

1. Weight of improved seed (WTNISEED)➔ Number of deleted instances=1288
2. Source of water for irrigation (SIRRG)➔ Number of deleted instances=2275
3. Number of trees (TREES) ➔ Number of deleted instances=93
4. Number of trees of bearing age (TREESBA) ➔ Number of deleted instances=91
5. Improved seed cost (COSTIMPS) ➔ Number of deleted instances=1281
6. Type of natural fertilizer (D23)➔ Number of deleted instances=14,193

Table 1 shows the selected attributes among the original and total dataset.

Table 1: List of selected attributes and their description

Field name	*Data type*	*Description*
HHSEX	Nominal	Head Sex
FLDTYPE	Nominal	Field Type
OWNTYPE	Nominal	Owner Type
EXT	Nominal	Is the field under Extension program?
IRRG	Nominal	Is the field used Irrigated?
SERRO	Nominal	Is the field affected by soil erosion?
MERRO	Nominal	Measure taken for Soil erosion
SEEDTYPE	Nominal	Seed type used
WTNISEED	Numeric	Weight of non improved seed(Kg)
DAMAGE	Nominal	Is the Field damaged?
DREASON	Nominal	Damage reason if field damaged
DPERCENT	Numeric	Damage percent if the field was damaged.
DMEASURE	Nominal	Measure taken to prevent damage
DMTYPE	Nominal	Type of damage prevention
DMCHEM	Nominal	Chemical Used?
FERT	Nominal	Fertilizer Used?
FERTTYPE	Nominal	Type of fertilizer used?
D22A	Nominal	Type of chemical fertilizer used?
D22B	Numeric	If chemical fertilizer used, quantity in Kg
D24	Nominal	How many times do the farmers produce crops?
D26	Nominal	What was the field used for?
AREAH	Numeric	Area measured in hectare (Hectare)
PRODQ	Nominal	Production in Quintal per Hectare Area

4. Results and Findings

4.1 Feature/Attribute Selection by the Association Rule

We used expert's advice to select the features/attributes that have direct and indirect relation with the target class which is production of Teff in quintal per hectare. In addition to this commonly known procedure, we also tried to select and categorize those attributes as directly related or not by applying Association rule.

As we can see from the test runs conducted for the selected association rules, the result shows that there are attributes which have direct and indirect relationship or have association with the target class which is yield of Teff in quintal.

Though there are some differences the outcome of the entire three association rule test runs share most common points with the experts' advice.

The Apriori and Tertius algorithms in the above test runs indicate that most attributes has association with the productivity. Such as

1. Irrigation used (IRRG)
2. Extension (EXT)➔
3. Seed type (SEEDTYPE)
4. Any damage? (DAMAGE)
5. Damage reason (DREASON)
6. Any measure to prevent damage (DMEASURE)
7. Damage percent (DPERCENT)
8. Fertilizer used (FERT)
9. How many times do you produce crops (D24)
10. What was the field used for? (D26)
11. Field type (FLDTYPE)

Those attributes listed in above are indicated by the association rules assuming that they have a relation with the productivity as confirmed by the expert. Even though there are also other directly related attributes confirmed by the expert but not by the experiment such as

1. Type of chemical fertilizer used (D22A)
2. If chemical fertilizer, quantity in kg (D22B)
3. Area in hectare (AREAH)
4. Type of damage prevention (DMTYPE)
5. Chemical used (DMCHEM)
6. Fertilizer type (FERTTYPE)
7. Soil erosion (SERRO)
8. Measure taken to soil erosion (MERRO)
9. Weight of non improved seed (WTIMSEED)

The time taken to generate 100 rules by Apriori is 2.38 second, Tertius is 3.09 second, and FilteredAssociator is 1.25 second. Tertius algorithm is the slowest of all compared with all rest algorithms testes in this research.

4.2 Build and Test the Modeling for Classification using Decision Tree

The experimentation is performed through decision tree algorithms. The algorithms selected are J48, REPTree, and Random forest. The decision tree software employed for the purpose of this research was the Weka software package, which contains several classifiers, clusters and association algorithms.

4.2.1 Generating Rules from Decision Tree (J48)

Table 2: Confusion matrix for J48

Actual	*Predicted*		*Total*
	High	*Low*	
High	11714	3430	*15144*
Low	3845	17990	*21835*
Total	*15559*	*21420*	*36979*

The confusion matrix in Table 2 depicts that out of the total records provided to the program, 11,714 (77.35%) and 17,990 (82.39%) records were classified correctly in the class of High and Low respectively.

4.2.2 Generating Rules from Decision Tree (REPTree)

Table 3: Confusion matrix for REPTree

Actual	*Predicted*		*Total*
	High	*Low*	
High	11722	3422	*15144*
Low	4061	17774	*21835*
Total	*15783*	*21196*	*36979*

The confusion matrix in Table 3 depicts that out of the total records provided to the program, 11,722 (77.40%) and 17,774 (81.40%) records were classified correctly in the class of High and Low respectively.

4.2.3 Generating Rules from Decision Tree (Random Forest)

Table 4: Confusion matrix for Random Forest

Actual	*Predicted*		*Total*
	High	*Low*	
High	11324	3820	*15144*
Low	3663	18172	*21835*
Total	*14987*	*21992*	*36979*

The confusion matrix in Table 4 depicts that out of the total records provided to the program, 11,324

(74.77%) and 18,172 (83.22%) records were classified correctly in the class of High and Low respectively.

4.3 Evaluating the classification model

Evaluating the performance of a data mining technique is a fundamental aspect of machine learning. Evaluation method is the yardstick to examine the efficiency and performance of any model. The evaluation is important for understanding the quality of the model or technique, for refining parameters in the iterative process of learning and for selecting the most acceptable model or technique from a given set of models or techniques.

We used the available ones which was present on Weka tool such as k-fold cross validation, F-Measure, ROC and Confusion Matrix (precision and recall).

Table 5: Comparison of Algorithms for their Efficiency

Algorithm Efficiency Evaluator	*J48*	*REPTree*	*Random Forest*
F-Measure	0.804	0.798	0.797
Precision	0.804	0.799	0.797
Recall	0.803	0.798	0.798
ROC	0.857	0.858	0.87

As shown in Table 5, the three algorithms have shown to be very closely efficient with the efficiency evaluations. But there is a slight difference as shown in the result. ROC curve shows that REPTree and J48 are almost the same and RandomForest has a slight difference. Precision, F-Measure and Recall shows that J48 is more efficient.

Table 6: Comparison of Classification Algorithms for their performance

	J48	*REPTree*	*Random Forest*
Total Record	36,979	36,979	36,979
Number of Correctly classified instances	29704	29,496	29,496
Accuracy in Percentage	80.32%	79.76%	79.76%
Time taken to build the model	1.33 sec	1.51 sec	3.06 sec

As shown in Table 6 each algorithm takes the same numbers of instances which is 36,979 (100%) and gave us the knowledge. Though they take equal number of instances, the correctly classified numbers of instances were different due to the difference in algorithms.

5. Recommendation and Conclusion

This research attempted to study the application of classification rule mining to find relationship and interesting patterns between attributes of census or survey data and Teff productivity. Data collection, selection and cleaning were major tasks which took most of this research.

We attempted to select relevant records and attributes based on the objectives of this research and practices in data mining. The total number of attributes in the original target dataset was 44. We attempted to select the relevant attributes for the data mining task using statistical and agricultural experts' advice and data mining algorithm.

A dataset totaling 36,979 records was used in generating classification and association rules. Initially, all of the records were given with 24 attributes to Apriori, Tertius, and FilteredAssociator, which are association rule mining algorithms in order to identify directly related attributes with Teff productivity. Subsequently the records were submitted to J48, RandomForest, and REPTree for classification rule mining.

The result from the association rule indicates that from the total dataset used in this research 95.45% of Teff production was cultivated on pure field type

which is dedicated only for Teff production using damage prevention mechanisms and this has an association with the use of fertilizer which is about 94.71% of the total dataset.

The result from the classification rule also indicates that from the total dataset used in this research high productivity of Teff is associated with area in hectare, damage preventive measure, use of extension service, fertilizer used, , type of seed, sample weight of seed and type of head sex.

On the basis of the findings of the study and the experience gained from the research, the following recommendations are suggested. More attributes should be added to allow complete analysis of the productivity of Teff. Possible values of an attribute should be less ambiguous.

There are attributes which have directly and indirectly related attributes and not included in this research due to different reasons during data preprocessing. There are also attributes having missing values more than the accepted range value. Thus to enhance such studies we recommend collecting and using data that have a better quality and size.

Although the findings of the research are not conclusive, they, however, can be considered as insight-giving to the bigger picture of the phenomena of Teff Production against inputs and methodology being used. Accordingly, the findings can be incorporated for the advocacy and awareness raising efforts of CSA and the Ministry of Agriculture, preferably after being substantiated and supplemented by qualitative research.

References

[1] Wikipedia, the free encyclopedia, Agriculture in Ethiopia, 2012, Retrieved from: http://en.wikipedia.org/wiki/Agriculture_in_Ethiopia, Last accessed on June 5,2012

[2] Background note: Ethiopia, 2012, Retrieved from http://www.state.gov/r/pa/ei/bgn/2859.htm, Last accessed on June 1,2012

[3] Fayyad, Usma, Piatetsky-shapiro, G. Smyth, and Padharic, "From Data Mining to Knowledge Discovery in Database", 1996, Retrieved from: http://citeseer.nj.nec.com.fayyad96from.html, Last accessed on June 8, 2012

[4] Gabre-Madhin and Eleni Zaude, "Market Institutions, Transaction Costs, and Social Capital in the Ethiopian Grain Market", Washington, DC: International Food Policy Research Institute, 2001.

[5] Teff for Gluten Intolerance, 2012, Retrieved from: http://www.matr.net/article-6172.html. Last accessed on June 6, 2012.

[6] Bhargavi Peyakunta and Jyothi Singaraju, "Soil Classification Using Data Mining Techniques: A Comparative Study", International Journal of Engineering Trends and Technology, July to Aug, 2011.

[7] Nukella Srinivasa Rao and Susanta Kumar Das, "Classification of herbal gardens in India using data mining", Journal of Theoretical and Applied Information Technology, March 31st, 2011.

[8] J. Gholap, A. Ingole, J. Gohil, S. Gargade, and V. Attar, "Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction", 2011.

[9] P. Revathi and M. Hemalatha, "Efficient Classification Mining Approach for Agriculture", International Journal of Research and Reviews in Information Science, June 2011.

[10] DV Kalashnikov, S Mehrotra, and Z Chen (2005), "Exploiting Relationships for Domain-Independent Data Cleaning", SIAM Data Mining (SDM) Conf.

[11] Edgar Acuna1 and Caroline Rodriguez (2004), "The Treatment of Missing Values and its Effect in the Classifier Accuracy".