IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 2 (Sep. - Oct. 2013), PP 01-05
www.iosrjournals.org

Using Data-Mining Technique for Census Analysis to Give Geo-Spatial Distribution of Nigeria.

*Ogochukwu C.Okeke And **Boniface C,Ekechukwu
**Computer Science Department,Anambra State University Uli,Nigeria*
***Computer Science Department.Nnamdi Azikiwe University,Awka,Nigeria*

Abstract: *There are patterns buried within the mass of data in the various editions of population census figures in this country. These are patterns that will be impossible for humans working with bare eyes and hands, to uncover without computer system to give geo-spatial distribution of population in that area. This paper is an effort towards harnessing the power of data-mining technique to develop mining model applicable to the analysis of census data that could uncover some hidden patterns to get their geo-spatial distribution. This could help better-informed business decisions and provide government with the intelligence for strategic planning, tactical decision-making and better policy formulation.*
Decision tree learning is a method for approximating discrete-valued target function, in which the leaned function is represented by a decision tree.
Decision tree algorithm was used to predict some basic attributes of population in the census database. Structured System Analysis and Design Methodology were used.
Key words*: Census, Data-mining and GIS*

I. Introduction

Census analysis is often not critically analyzed to bring out some of the basic and important attributes of census data information to give geo-spatial distribution of population. This is due to non-availability of the required tools for carrying out such analysis. This paper suggests the use of data-mining technique(Decision tree algorithm technique) to extract hidden information from large census data warehouse and geographical information system (GIS) as an integrating technology that gives geo-spatial distribution of the population.

Data-mining is the process of discovering previously unknown, actionable and profitable information from large consolidated databases and using it to support tactical and strategic decisions (Gajendra, 2008). It is also extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies, industries, institutions, government e.t.c focus on the most important information in their data warehouses. Data-mining tools predict future trends and behaviors, allowing business to make proactive, knowledge –driven decisions. The automated, prospective analyses offered by data-mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data-mining tools can answer business questions that traditionally were time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data-mining is the exploration of historical data (usually large in size) in search of a consistent pattern and / or a systematic relationship between variable. It is then used to volute the findings by applying the detected pattern to new subsets of data. The roots of data-mining originate in three areas, Classical statistics, artificial intelligence and machine learning. Pregiborn (1997) described data-mining as a blend of statistics, artificial intelligence, and database research and noted that it was not field of interest of many until recently.
Mena (2005) has asserted data-mining is the process of discovering actionable and meaningful patterns, profiles and trends by sifting through your data using pattern recognition technologies such as neural networks machine learning and genetic algorithms. Also Foloruns & Ogunde (2004) have asserted that data-mining is a technologic for knowledge management in business process redesign (BPR). It helps in rethinking a process in order to enhance its performance academics and business practitioners have been developing methodologies to support the application of BPR principles. However, most methodologies generally lack actual guidance on deriving a process design thereby threatening the success of BPR (Selma & Hago 2003). Indeed a survey has proved that 85% of business process redesign projects fail or experience problems (Crow& Giudici 2002).

Moreover, Data-mining takes evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery (Koh, Hian & Low 2004). Data mining is ready for application in business community because it is supported by tree technologies that are now sufficiently mature which are massive data technologies that are now sufficiently mature which are massive data collection, powerful processor computers and data mining algorithm. Most companies already collect and refine massive quantities of data. Data-mining techniques can be implemented rapidly on existing software and hardware platforms to

enhance the value of existing information resources, can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data-mining tools can analyze massive databases to deliver answers to questions.

The United Nation (UN) defines census as total as total process of collecting compiling, analyzing, evaluating, publishing and disseminating demographic economic, social and housing data pertaining at a specified time to all persons and all building in a country or in a well delineated part. A population and housing census is of great relevance to the economics, political and socio-cultural planning of a country. Reliable and detailed data on the size, structure, distribution and socio-economic and demographic characteristics of a country's population is required for planning, policy intervention and monitoring of development goals. Within the masses of information in the census database lays hidden information of strategic importance. Data-mining is a key element in finding the particular pattern and relationship that can help governments, organizations and businesses. Data-mining find those patterns and relationships using sophisticated data analysis tool and technique to build models. Data-mining model will predict attributes of the population like youths of a given age limit, number of males, number of females ,sex, employment e.t.c Then geographical information system (GIS) will integrates, edits, analyzes, shares, and display geo-spatial distribution of the population . Census can do all these with data-mining; the statistical techniques of data-mining are familiar. They include linear and logistic regressions, multivariate analysis, principal component analysis, decision trees and neural networks. Traditional approaches to statistical inference fail with large databases. Decision- tree is a tree-shaped structure that represents set of decision. This decision generates rules for the classification of a dataset (Gajendra, 2008).

However, because with thousands or millions of cases and hundreds or thousand of variable there will be spurious relationships which will be highly significant by any statistical test. The objective is to build a model with significant predictive power that would give geo-spatial distribution of the population. It is not enough just to find which relationships are statistically significant. There are two main kinds of models in data-mining; the first kind is predictive models which use data with known results to develop values. For examples, based on marital status, gender, age, employment etc. in a census database the model will predict wealth. The strength of the predictive model lies in learning (self- training teaching it how to predict the outcome of a given process).The second kind of model is descriptive models which may be used to guide decisions as opposed to making explicitly predictions. For example, the model might identify different ethnicity in a database.

The data-mining algorithm is the mechanism that creates mining models. To create a model, an algorithm first analyzes a set of data, looking for specific patterns and trends. The algorithm then uses the result of this analysis to define the parameters of the mining model to give geo-spatial distribution.

The data-mining model that an algorithm creates takes various forms including; number of males, number of females, sex, literacy, employment and illiteracy to give geo-spatial distribution of population. The data-mining extract these attributes out from pool of census database and give geo-spatial distribution. In its simplest form, a Geographic Information System (GIS) is a computer-based data management system for storing, editing, manipulating, analyzing, and displaying geographically referenced information. However, effective use of GIS also requires good quality data, skilled personnel, and institutional arrangements to collect, share, and disseminate the data. Geographically referenced, or geospatial, data describes anything that can be located in physical space, most typically with respect to earth's surface, and therefore can be displayed on a map.

Process of Data-Mining Technique; this process consists of three stages: the initial exploration, model building or pattern identification with validation and verification and deployment (i.e. the application of the model to new data in order to generate predictions)

Stage 1: Exploration, this stage usually starts with data preparation which may involve cleaning data, data transformation, selecting subsets of record and in case of data set with large numbers of variable (fields)- Performing some preliminary feature, selection operations to bring the number of variables to a manageable range (depending on statistical methods which are being considered). Then, depending on the nature of the analytic problem, this first stage of the process of data- mining may involve anywhere between a simple choice of straight forward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods.

Stage 2: Model building and validation: Berry & Linoff (2000) asserted that this stage involves considering various models and choosing the best one based on their predictive performance (i.e. explaining the variability in question and producing results across samples). This may sound like a simple operation, but in fact it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve that goal. Many of which are based on so called" competitive evaluation of model", that is applying different models to the same dataset and then comparing their performance to choose the best. These techniques which are often considered the core of predictive data-mining –include: Bagging (Voting, Averaging), Boosting, Stacking (Stacked Generalizations), and Meta-Learning.

Bagging (Voting, Averaging) applies to the area of predictive data-mining to combine the predicted classifications (prediction) from multiple models, or from the same type of model for different learning data. It is also used to address the inherent instability of results when applying complex models to relatively small data sets .Suppose your data-mining task is to build a model for predictive classification, and the datasets from which to train the model (learning data set, which contains observed classifications) is relatively small. You could repeatedly sub-sample (with replacement) from the dataset, and apply, for example a tree classifier (e.g,C&RT and CHAID) to the successive samples. In practice, very different trees will often be grown for the different samples, illustrating the instability of models often evident with small datasets. One method of deriving a single prediction (for new observations) is to use all trees found in the different samples, and to apply some simple voting. The final classification is the one most often predicted by the different trees. Some weighted combination of predictions (weighted vote, weighted average) is also possible, and commonly used. A sophisticated (machine learning) algorithm for generating weights for weighted prediction or voting is the boosting procedure.

Boosting applies to the area of predictive data-mining, to generate multiple models or classifiers (for prediction or classification), and to derive weights to combine the prediction from those models into a single prediction or predicted classification. Boosting will generate a sequence of classifiers , where each consecutive classifier in the sequence is an expert in classifying observations that were not well classified by those preceding it .Boosting can also be applied to learning methods that do not explicitly support weights or misclassification costs .In that case, random sub –sampling can be applied to the learning data in the successive steps of the iterative boosting procedure, where the probability for selection of an observation into the sub sample is inversely proportional to the accuracy of the prediction for that observation in the previous iteration.

A simple algorithm for boosting works like this: Start by applying some methods (e.g, a tree classifier such as a C&RT or CHAID) to the learning data, where each observation is assigned an equal weight .Compute the predicted classification ,and apply weight to the observation in the learning sample that are inversely proportional to the accuracy of the classification .

Meta –Learning .The concept of meta-learning applies to the area of predictive data-mining, to combine the predictions from multiple models. It is particularly useful when the types of models included in the project are very different. Stacking (Stacked Generalization) is use to combine the predictions from multiple models. It is particularly useful when the types of models included in the project are very different. Experience has shown that combining the prediction from multiple methods often yields more accurate predictions (Witten & Frank , 2000).In stacking ,the prediction from different classifier are used as input into a meta-learner, which attempts to combine the predictions to create a final best classification. So, for example, the predicted classifications from the tree classifier, linear model, and the neural network classifier(s) can be used as input variables into a neural network meta-classifier, which will attempt to learn from the data how to combine the predictions from different models to yield maximum accuracy.

Stage 3: Deployment: The final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate prediction or estimates of expected outcome. It is also application of a model for prediction or classification to new data. After a satisfactory model or set of models has been identified (trained) for a particular application, one usually wants to deploy these models so that prediction or predicted classifications can quickly be obtained for new data. For example, a credit card company may want to deploy a trained model or set of models (e.g. neural network, meta- leaner) to quickly identify transactions which have a high probability of being fraudulent. Census data are cleaned and reduced to give a good output (Okeke, 2013).

Decision trees are one of the powerful tools for classification and prediction. The strength of decision trees is due to the fact that, decision trees represent rules. Rules can readily be expressed so that human can understand then or even directly used in a database access language like SQL so that records falling into a particular category may be retrieved. A decision tree is a predictive modeling technique used to classify, cluster and predict tasks (Gajendra, 2008). It uses “divide-and-conquer” technique to split the problem search space into subsets.

For example, in marketing one has describe the customer segments to marketing professionals, so that they can utilize this knowledge in launching a successful marketing campaign. These domain experts must recognize and approve this discovered knowledge, and for this we need good descriptions. There are a variety of algorithms for building decision trees that share the desirable quality of interpretability. Decision tree is a classifier in the form of a tree structure, where root node and each internal node are labeled with question (kwedlo, 2001). The arcs emanating from each node represent each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem under construction.

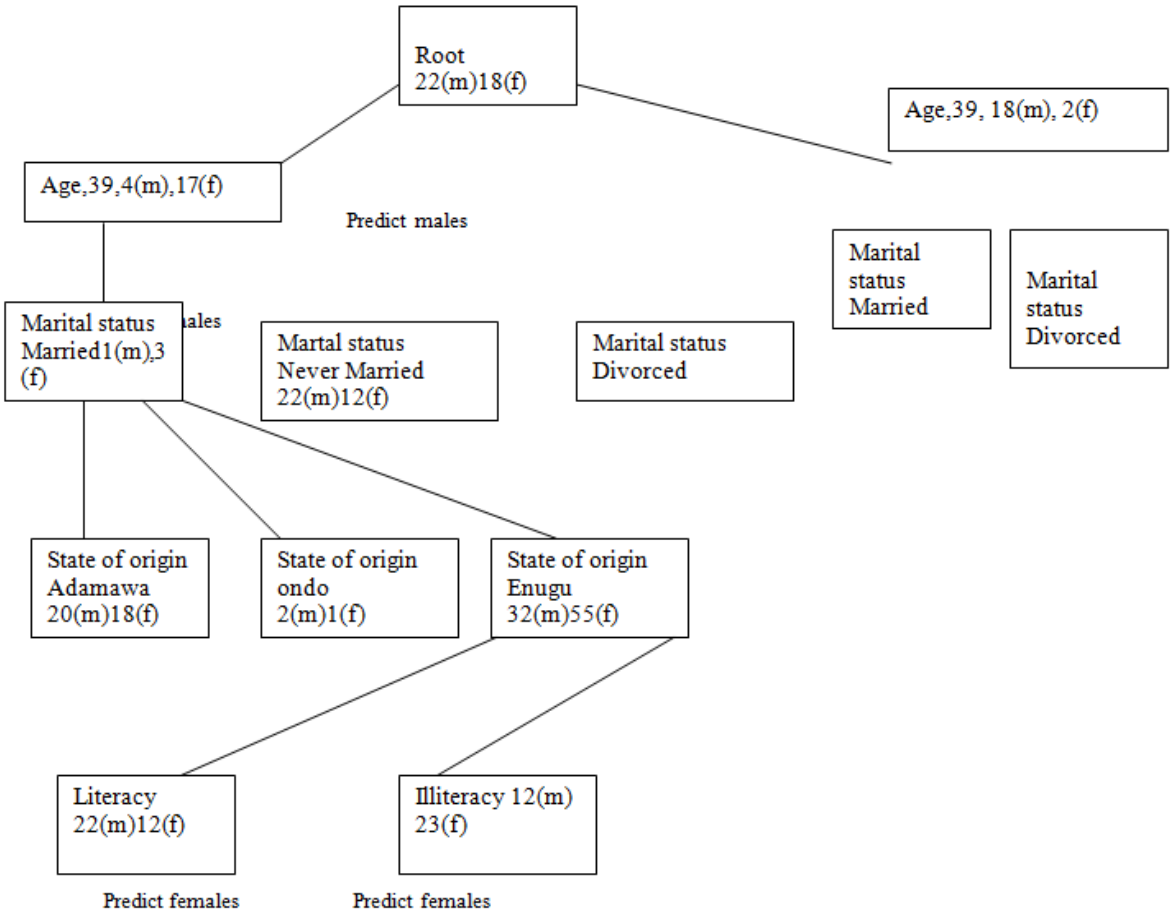

Figure 1.1 Decision tree algorithm predicting males and females

Results; this work achieved the goal of applying data-mining technique to the analysis of census data. The result of this paper is a predictive attributes of a population to give geo-spatial distribution in Nigeria. For instance based on marital status, sex, employment and unemployment etc in census database the model will predict wealth of a nation. The effort yielded the possibility of implementing the IDE3 decision algorithm in building decision trees from which attributes of a population can be predicted to give geo-spatial distribution.

II. Conclusion

Decision tree starts with a root node on which it is for users to take actions. From this node users split each node recursively according to decision tree learning algorithm. Data-mining helps governments, individuals, companies to uncover hidden patterns in large database which is used for development and Geographic Information System (GIS) captures the data from IDE3 algorithm using census data as source of input in development, stores it, manages it and give geo-spatial distribution needed.

Recommendations

Due to the possibility that the IDE3 algorithm pays attention to parts of the data that are irrelevant (what is called over- fitting), it may perform less well on test set data. This work did not consider noise compensation into consideration in the implementation of algorithm. Further work on this subject should bring in techniques that avoids over-fitting .The initial definition of IDE3 is restricted in dealing with discrete sets of values .It handles symbolic attributes effectively. In this work it was extended to handle numeric attributes (age in this case).What was done in this work was discretize the attributes age to Boolean value. Further work should do better by basing continuous attributes (numeric attributes) based on proper computation of information gain threshold.

References

[1]. Berry, M.J.A & Linoff (2000). Mastering Data-mining .Wiley Press: New York.
[2]. Crow, M.C &Giudici. (2003).Applied Data-mining :Statistical Method For Business And Industry. John Wiley and Sons .West Sussex, England.
[3]. Folorunso, O. & Ogunde, A.O. (2004). Data-mining as a Technique for Knowledge in Business Process Redesign. The Electronic Journal of Knowledge Management Volume 2 issue 1, pp, 33-44, available on line at www.ejkm.com .
[4]. Gajendra, S. (2008). Data-mining, Data-ware housing and OLAP. Kataria & sons: New Delhi.
[5]. Koh, Chye, H. & Kee, C.L (2004). Going concern prediction using Data-mining Techniques Managerial Auditing Journal, 19:3.
[6]. Kwedlo, W. & Kretowski, M. (2001). Learning Decision Rules using a Distributed Evolutionary Algorithm. Gdansk Press: Poland.
[7]. Mena, K.C (2005). Data mining and Statistics: Guild Form press: New York.
[8]. Pregibon,D.(1997).Data-mining "Statistical Computing and Graphics, pp 7-8."
[9]. RedLands,C.A.(1990).Understanding GIS. Environmental System Research Institute Oxford University Press: New York.
[10]. Rambaldi, G., and J. Callosa (2000). Manual on Participatory 3-Dimensional Modeling fo Natural Resource Management (Volume 7). NIPAP, PAWB-DENR: Philippines Department of Environment and Natural Resources.
[11]. Rhind's (2001): review of activities at the Experimental Cartographic Unit in the United Kingdom.
[12]. Sieber, R. (2000). 'Conforming (to) the Opposition: the Social Construction of Geographical Information Systems in Social Movements.' International Journal of Geographical Information Science, 14(8): 775–793.