The International Journal Of Engineering And Science (IJES)
|| Volume || 4 || Issue || 8 || Pages || PP -05-11|| 2015 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Application of Data-mining Technique and Intelligent System on Demography Analysis in Nigeria

[1]Okeke Ogochukwu C. , [2]Ezenwegbu Nnamdi C.
[1]*Computer Science Department,Chukwuemeka Odumegwu Ojukwu University Uli,Nigeria*
[2]*Computer Science Department, Chukwuemeka Odumegwu Ojukwu University Uli,Nigeria*

----------ABSTRACT----------

Controversy over Nigeria's census figures is nothing new. Accusations that the country's official population figures had been rigged date back to the 1950s and have continued unabated under military and civilian regimes. Furthermore, the demographics of Nigeria have undergone several changes over the past few decades as a result of migration and settlement from the far off countries. At the present, demographic figures in Nigeria mires in controversy. This paper is an attempt to straighten Nigerian demographic analysis through data mining techniques and intelligent system to properly identify patterns and trends in Nigerian demographic figures. The study will first focus on Uli community and subsequently applied for Nigeria at large. We use Decision-Tree-Based classification model to extract from rich demographic data hidden information that can be used for the investigation of national conditions and national power. We also utilized the power of intelligent system for complex processing and data analysis. We propose an intelligent system developed on Microsoft .NET platform. Since the intelligent system is an object oriented system, we used the Object Oriented Analysis and Design Methodology (OOADM).

Keywords: *demographic analysis, data mining, Tree-Based classification, intelligent systems.*

Date of Submission: 14 July 2015 Date of Accepted: 15 August 2015

I. Introduction

Demographic analysis is a technique used to develop an understanding of the age, sex, and racial composition of a population and how it has changed over time through the basic demographic processes of birth, death, and migration. Demographic Analysis (usually abbreviated as DA) also refers to a specific set of techniques for developing national population estimates by age, sex, and race from administrative records to be used to assess the quality of the decennial census.

This paper focuses of constructing population estimates using data mining techniques as well as intelligent system. This involves data mining statistics, estimates of net international migration, and for the population aged 65 and over, data from an intelligent system. Traditionally, the DA estimates have been disaggregated by sex and single year of age. New data sources and changes in the racial and ethnic make-up of the nation form patterns that data mining can reveal when harnessed. (2010 Demographic Analysis Estimates News Conference)

Population demographics are of great relevance to the economic, political cultural and social development of a country. It is the major source of the bench-mark data on the size, structure and distribution of the country's population required for both planning revenue allocation, distribution of public utilities and research purposes. The 1991 Population demographics is one of the cardinal items on the Transition to Civil Rule Programme of Nigeria's Federal Military Government (1988-1992). In pursuance of the conduct of population demographics, the Government promulgated Decree No. 23 of 1989 establishing the National Population Commission (NPC). The same Decree also gave legal backing to the conduct of the 1991 population demographics. (http://www.123independenceday.com/ nigeria/ demography.html)

The primary objective of that census is to provide information on the number, distribution and social-demographic characteristics of the people. But sadly, the purpose of all these efforts was defeated as demographic figure remains a controversy in Nigeria. To apply data mining techniques and intelligent system in demographic analysis is promising, but before we do that it is important that we establish a good understanding of the data mining technique and intelligent system we used.

II. Data mining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. (Han *et al* 2001)

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps. In our data mining, we used Decision-Tree-Based classification technique. (Witten *et al* 2011)

2.1 Decision trees:

These are tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

III. Intelligent system

An intelligent system is a machine with an embedded, Internet-connected computer that has the capacity to gather and analyse data and communicate with other systems. Requirements for an intelligent system include security, connectivity, the ability to adapt according to current data and the capacity for remote monitoring and management. Essentially, an intelligent system is anything that contains a functional, although not usually general-purpose, computer with Internet connectivity. An embedded system may be powerful and capable of complex processing and data analysis, but it is usually specialized for tasks relevant to the host machine.

Intelligent systems exist all around us in point-of-sale (POS) terminals, digital televisions, traffic lights, smart meters, automobiles, digital signage and airplane controls, among a great number of other possibilities. As this on-going trend continues, many foresee a scenario known as the Internet of Things (IoT), in which objects, animals and people can all be provided with unique identifiers and the ability to automatically transfer data over a network without requiring human-to-human or human-to-computer interaction. (Margaret Rouse *et al* 2013)

IV. Uli Town

As noted earlier, our system is applied on a town in Anambra state called Uli. Uli is a town of historic importance situated at the extreme southeast corner of Ihiala local government area of Anambra state in Nigeria. Its closest neighbouring towns are Amaofuo (formerly a village in Uli town), Ihiala, Amorka, Ubulu, Ozara, Egbuoma and Ohakpu. Uli town extends westward to the confluence of the rivers of Atamiri and Enyinja, and across Usham Lake down to the lower Niger region. The history, life, culture and custom of Uli people can be found in many texts with the most elaborate one being the book titled URI History, Life, Culture and Custom of a People and written by Ichie (Sir) G.C. Okonkwo. The Anambra state University of technology is located in Uli. During the Biafran Civil War, the Biafran Airport code named Annabel Airport was also located in a land strip at Umuchima village, Uli. This landing strip was used extensively to bring in relief supplies during the Biafran airlift.

Uli has also produced many eminent people in Nigeria among whom are Igwe Eze Udoka II Damian Onyekaonwu, Engr. Fort Dike, Dr ABC Orjiako a business mogul, Dr. Chinwoke Mbadinuju a former Governor of Anambra state also came from Uli, Ichie (Sir) G.C. Okonkwo an elder statesman and the author of the most comprehensive book on the history of Uli people 'URI', and many other great men and women. Uli is a small

progressive town, with many of its sons and daughter all over the world. The people of Uli are predominantly Christians of different denominations though there are some adherents of African traditional beliefs popularly called "Ndi Odinana" scattered across the four quarters.

V. Main Issue - Application of data mining technique and intelligent system on demographic analysis in Nigeria.

5.1 Role of data mining:

We proposed the Decision Tree Classification data mining technique. Decision tree builds classification or regression models in the form of a tree structure. It breaks down Nigerian demographic dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Let's look at a household in Uli for example. A decision node (e.g., Household) has two or more branches (e.g.household_id, street_id, ward_id, town_id, lg_id, state_id, zone_id, house_description, head_of_household). Leaf node (e.g. age) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

You can use classification to build up a population distribution by age and sex, employment and age group, marital status and age group, household and age group etc. by describing multiple attributes to identify a particular class. For example, you can easily classify person into different types (head, husband, daughter, and son) by identifying different attributes (appearance, shape). Given a new person, you might apply it into a particular class by comparing the attributes with our known definition. You can apply the same principles to other people, for example by classifying them by age and social group.

Additionally, you can use classification as a feeder to, or the result of, other techniques. For example, you can use decision trees to determine a classification. Clustering allows you to use common attributes in different classifications to identify clusters. Carrying out population census tabulation program involves a determination of the number of different levels of geographic details. Data are presented for administrative divisions and subdivisions of the country, Federal government, state government and local government in various categories. For small geographic area, such as villages, the results as a rule are limited to a report of the number of inhabitants or perhaps the population by sex only. At another level, tabulation may provide only inventory statistics that is simply a count of persons in the various categories of age, marital status, economic activity and so on, with but little cross-classification with other characteristics. In some cases, most subjects are cross-tabulated by age and sex and often there are cross-classification with social and economic characteristics, such as educational attainment by economic activity or employment status by occupation.

5.2 Role of intelligent system

The intelligent system is a system that will enable government have first-hand information on the population of the country and tends to arrest the manual form of census taking by converting the paper work into a database which can be updated. The census database is the most fundamental part of the census system. It houses all information needed. The system at a click will be able to display first hand population results. This will involve recording demographic information in a database as well as keeping track of births and death certificates to update the census figure. The intelligent system was programmed on the Microsoft .Net framework using visual basic 9.0 on the asp.net technology. Remember that we noted in section 3 that intelligent system is a machine with an embedded, Internet-connected computer that has the capacity to gather and analyse data and communicate with other systems. The intelligent system does the job with the assisted data mining technique to collect, store and retrieve data faster. Because the intelligent system is connected to the internet, the system gathers mined information from related sources and insert or update them in the central distributed database. The intelligent system maintains accuracy, security and integrity of data collected. It undertakes periodic enumeration of population through sample survey, Censuses or otherwise. The intelligent system establishes and maintains machinery for continuous and universal registration of births and deaths throughout the federation. The intelligent system further publishes and provides information data on population for the purpose of facilitating economic and development planning.

5.2.1 Database specification for the intelligent system

The Database is composed of the quantitative (structured) and qualitative (unstructured) knowledge of population dynamic acquired by the human. The database is a network of semantically related static and dynamic objects, each of which is modelled, in a relational form. The structured knowledge is concerned with facts, rule and events of human population dynamics, which are commonly agreed upon by experts in demography. The unstructured knowledge is that knowledge which is acquired by demographic expert from experience and population census survey. Data was organized using a relational, hierarchical network or object orientated database model. The databases are accessed via networks using technologies like client-server. The

prominent form of database organization described as relation allows the user to think in the form of two dimensional tables which is the way many people see data reports. It takes its name from the mathematical theory of relations. The data in the database are stored together with minimum of redundancy to serve multiple applications, so the database is independent of the computer program that uses it and the type of hardware where it is stored. The general description of a relation is given in (Dates, 1986; Navathe, 2000), a relation (or relation state) r of the relation schema R(A1, A2, ..., An), also denoted by r(R), is a set of n-tuples r = {t1, t2, ..., tm}. Each n-tuple t is an ordered list of n values t = <v1, v2, ..., vn>, where each value vi, is an element of dom(Ai) or is a special null value. The i^{th} value in tuple t, which corresponds to the attribute Ai, is referred to as t(Ai). Some of the relations in this database are as presented below:
ZONE (zone_id, zone_name, zone_head_quarter)
STATE (state_zone_id, state_id, state_name, state_capital)
LOCAL GOVERNMENT (lg_state_id, lg_id, lg_name, lg_headquarter)
TOWN (town_lg_id, town_id, town_name, Head_of_community)
WARD (ward_town_id, ward_id, ward_name)
STREET (street_ward_id, street_id, street_name)
HOUSEHOLD (household_id, street_id, ward_id, town_id, lg_id, state_id, zone_id, house_description, head_of_household)
BIO (respondent_id, house_id, respondent_name, relation_to_household_head, sex, age, tribe, nationality, disability, duration of residence, previous_residential_address, present_residential_address)
EDUCATION (respondent_id, literacy, highest_educational_qualification)
ECONOMIC (respondent_id, work_status, type_of_employment, sector_of_employment)
MARITAL (respondent_ id, marital_status, age_at_marriage, no_of_children_born, n_death_child_the_last_1year).

VI. Result

The Data Mining algorithm is implemented fully in Visual Basic.net. Before we dive into the code part, let's set up our database of people. In the database (called population), we'll have two other important tables: GivenPopulation and NewPopulation. GivenPopulation has the population that have already been entered, and NewPopulation will have the population size that we are going to predict. The system uses the ActiveX Data Objects (ADO) provider for a relational database to access a data mining database. It fully implements the data mining algorithm in Visual Basic code, completely eliminating any database specific SQL syntax, stored procedures and queries. It needs only a simple SQL SELECT statement to access a relational database. The only part of the data mining application that changes is the SQL syntax used to select data from a specific table and relational database. Hence the data mining application is truly portable to any relational database where an ADO provider exists (most major database vendors have ADO providers for their databases). To proceed we:

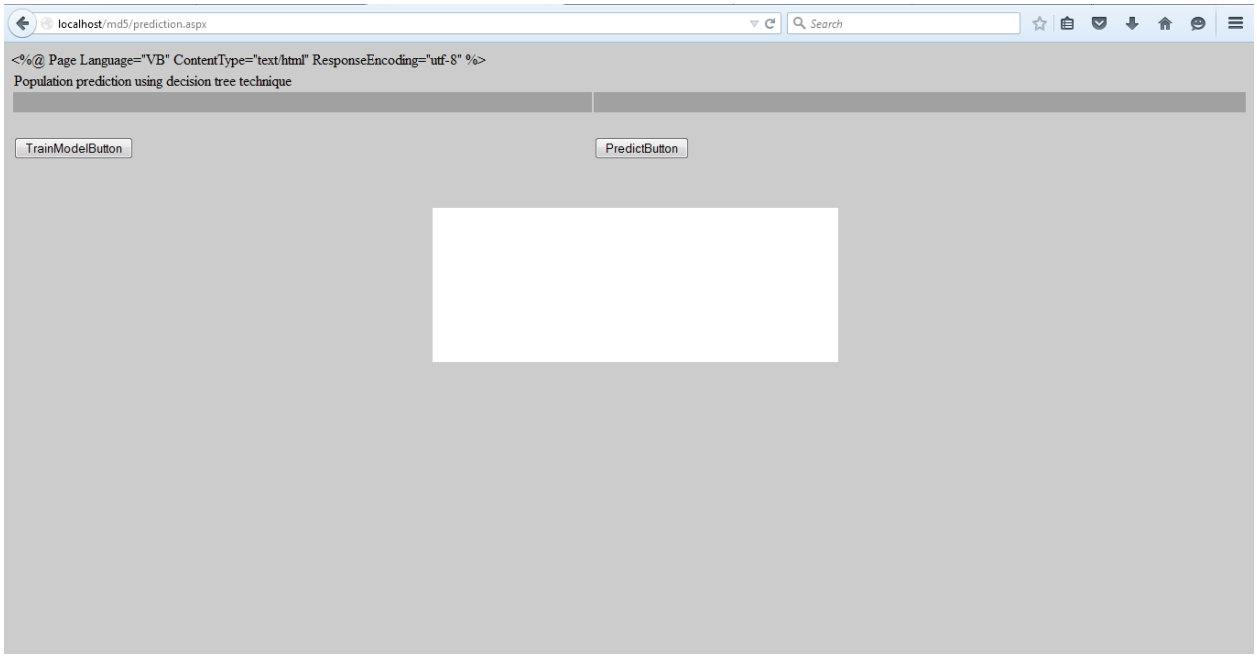
1. Create a new Analysis Services project using SQL Server Business Intelligence Development Studio.
2. Add a new Data Source, call it Population_size, and have it point to the newly created population database
3. Add a new Data Source View employing the new data source (from the last step) and name it population, too.
4. Add a new Mining Structure and have it use the Microsoft Decision Trees algorithm. When prompted, the person_id field is the key, and all remaining fields are inputs. gender should be marked as Boolean, age as Long, and the remaining fields as Text.

Now, publish the project to the database. You will need to open the project's properties, click the Deployment options, and enter ‘population’ in the name of the database you want to publish in the Analysis Services. We are now ready to utilize the data mining on our NewPopulation table. To do so, we'll use a very simple ASP.NET page with two buttons, TrainModelButton and PredictButton, which do exactly what we expect. Look at the codes:

```
Imports System.Data.OleDb

Partial Class _Default
  Inherits System.Web.UI.Page

  Private cs As String = "Provider=MSOLAP.3;Data Source=localhost;Initial Catalog=population"
  Private conn As New OleDbConnection()

  Protected Sub Page_Load(ByVal sender As Object, ByVal e As System.EventArgs)
    Handles Me.Load
    conn.Open()
```

```
  End Sub

  Protected Sub Page_Unload(ByVal sender As Object, ByVal e As System.EventArgs)
    Handles Me.Unload
    conn.Close()
  End Sub

  Protected Sub TrainModelButton_Click(ByVal sender As Object, _
    BVal e As System.EventArgs) Handles TrainModelButton.Click

    ' Delete the mining structure so we can re-train it.
    Dim SQL As String = "DELETE FROM MINING STRUCTURE PostedGrades"

    Dim CMDDelete As OleDbCommand = New OleDbCommand(SQL, conn)
    CMDDelete.ExecuteNonQuery()
    CMDDelete.Dispose()

    ' Train the mining structure with data from the GivenPopulation table.
    Dim PipeDataToModel As String = "INSERT INTO MINING STRUCTURE GivenPopulation " _
    & "(person_id, gender, age, name, zonal_id) " _
    & "OPENQUERY (population, 'SELECT person_id, gender, age, name, zonal_id " _
    & FROM GivenPopulation)"

    Dim CMD As New OleDbCommand(PipeDataToModel, conn)
    CMD.ExecuteNonQuery()
    CMD.Dispose()

  End Sub

  Protected Sub PredictButton_Click(ByVal sender As Object, _
    BVal e As System.EventArgs) Handles PredictButton.Click

    ' Make our query to predict population for the persons in the NewPopulation table.
    Dim Query As String = "SELECT T.person_id, GivenPopulation.grade, " _
    & PredictProbability(population_size) FROM " _
    & " GivenPopulation NATURAL PREDICTION JOIN OPENQUERY(population, " _
    & 'SELECT * FROM NewPopulation) AS T"

    Dim CMD As New OleDbCommand(Query, conn)

    Dim myReader As OleDbDataReader = CMD.ExecuteReader()

    If myReader.HasRows = True Then
    While myReader.Read()
      ' Just output the results of the query.
     Response.Write(myReader(0).ToString & " " & _
       & myReader(1).ToString & " " & _
       myReader(2).ToString & "<BR>")
    End While
    End If

    myReader.Close()
    CMD.Dispose()

  End Sub

End Class
```

The front looks like:



The purpose of the code in TrainModelButton_Click() is to read the data from the GivenPopulation table into the mining structure (train the classifier). Once that is done, PredictButton_Click() will predict the population size for Uli town in the NewPopulation table and write the results to the page.

The mining structure GivenPopulation is first dropped in TrainModelButton_Click() because we cannot incrementally train a classifier - we just have to delete it and make it again. For the possible exception of the OpenQuery function, the SQL used is pretty straight-forward. The OpenQuery function lets you perform a query on a linked server. In our case, the linked server is just the database server.

VII. Conclusion

The primary source of information about the population of a country is the population census. This has to deal with a process of collecting, compiling and publishing demographic, economic, and social data pertaining at a specified time or times, to all persons in a country or delimited territory. There have been many developments in demographic analyses of census results in recent years. Most of them fall within the same broad direction of orientation, namely the provision of more information about the social and economic characteristics of

populations and about the pattern of social and economic organization of communities. Until recent times, no noticeable achievement has been made in Nigerian demographics. In this study, a decision-tree-based classification technique was used to mine the population pattern in Uli town Anambra state in Nigeria. We carried out a demographic survey on the target population. Population census was administered; hidden demographic patterns were uncovered, coded and entered into our intelligent system to construct the required relations. The intelligent system is designed and implemented using Visual Basic 9.0 programming language and Microsoft SQL Server. Demographic models are also embedded, in the system, which help to estimate some population statistics. The case study was carried out to demonstrate the functionality of the system. The system being reported in this study is capable of, accepting input data from population census for the purpose of demographic analysis.

REFERENCES

[1] Dates, C. J. (1986). An introduction to database system (4nd ed.). California: Addison-Wesley Publishing Company, p. 639. ISBN:0-201-14201-5.

[2] Demographic Analysis Estimates News Conference (2010) https://www. census.gov/coverage_ measurement/ demographic_analysis/

[3] Han, Jiawei; Kamber, Micheline (2001). Data mining: concepts and techniques. Morgan Kaufmann. p. 5. ISBN 9781558604896.

[4] http://www.123independenceday.com/ nigeria/ demography.html

[5] King, D. (1993). Intelligent support systems: art, augmentation, and agents, in R. H. Sprague, Jr and H. J. Watson (eds), Decision Support Systems: Putting Theory into Practice (3nd ed.). Englewood Cliffs, NJ: Prentice Hall.

[6] Margaret Rouse, Ivy Wigmore 2013 http://whatis.techtarget.com/definition/ intelligent-system

[7] Navathe, E. (2000). Fundamental of database systems (3nd ed.). Teturo Sawada, Exclusive publisher and Distributor.

[8] Olabode, O. (1999). Computer aided system for demography in Nigeria. Unpublished M.Tech Thesis, Federal University of Technology, Akure.

[9] Witten, Ian H.; Frank, Eibe; Hall, Mark A. (30 January 2011). Data Mining: Practical Machine Learning Tools and Techniques (3 ed.). Elsevier. ISBN 978-0-12-374856-0.