Literacy Rate Analysis

Tarun Verma, Sweety Raj, Mohammad Asif Khan, Palak Modi

Abstract— This paper provides the information that literacy is one of the main important issues that every country wants to fulfill. UNESCO, UNDP, ASER are some organizations that work towards this goal and collected data regarding literacy rate and dropout rate. Some solution approaches have been studied for analyzing the data associated with literacy. Weka tool is used along with a pre-existing solution to define the literacy rates region wise, on an international and national level. Data to be analyzed is given as input to the tool and output is obtained in the form of a decision tree. The decision tree, which is the final output, is very easy to understand and comprehend.

Index Terms— Data Mining, Dropout Rate, Knowledge Discovery, Literacy Rate, Out-of-school, Primary Education, Primary Education.

1 Introduction

LITERACY is defined as the ability to read, write and think rationally. It represents the lifelong, intellectual process of gaining meaning from print. Key to all literacy is reading development, which involves a progression of skills that begins with the ability to understand spoken words and decode written words, and culminates in the deep understanding of text. Reading development involves a range of complex language underpinnings including awareness of speech sounds (phonology), spelling patterns (orthography), word meaning (semantics), grammar (syntax) and patterns of word formation (morphology), all of which provide a necessary platform for reading fluency and comprehension. Once these skills are acquired the reader can attain full language literacy, which includes the abilities to approach printed material with critical analysis, inference and synthesis; to write with accuracy and coherence; and to use information and insights from text as the basis for informed decisions and creative thought [1].

Literacy has always been an issue for the world. Every country aims to achieve full literacy rate. Although literacy rate has increased upto a great extent now but still there is a need to know the areas that are still lagging behind. So, study and analysis of literacy data of the world is required to provide a timely and informed basis for helping planning and management of education services and to establish or contribute to an education system for collection, organization and utilization of education data.

2 Background

This report extends the work done by United Nations Development Programme (UNDP) [2], ASER (India) [3] and India Census 2011 [4].

3 Solution Approaches

Following are the five different algorithms that can be4 used for analyzing the data [5].

3.1 IDE3

IDE3 (Iterative Dichotomiser 3) decision tree algorithm was introduced in 1986 by Quinlan Ross (Quinlan, 1986 and 1987). It is based on Hunt's algorithm and it is serially implemented. Like other decision tree algorithms the tree is constructed in two phases; tree growth and tree pruning. Data is sorted at every node during the tree building phase in-order to select the best splitting single attribute (Shafer et al, 1996). IDE3 uses information gain measure in choosing the splitting attribute. It only accepts categorical attributes in building a tree model (Quinlan, 1986 and 1987). IDE3 does not give accurate result when there is too-much noise or details in the training data set, thus a an intensive pre-processing of data is carried out before building a decision tree model with IDE3.

3.2 C4.5

C4.5 algorithm is an improvement of IDE3 algorithm, developed by Quinlan Ross (1993). It is based on Hunt's algorithm and also like IDE3, it is serially implemented. Pruning takes place in C4.5 by replacing the internal node with a leaf node thereby reducing the error rate (Podgorelec et al, 2002). Unlike IDE3, C4.5 accepts both continuous and categorical attributes in building the Decision tree. It has an enhanced method of tree pruning that reduces misclassification errors due noise or too-much details in the training data set. Like IDE3 the data is sorted at every node of the tree in order to determine the best splitting attribute. It uses gain ratio impurity method to evaluate the splitting attribute (Quinlan, 1993).

- *Tarun Verma is currently pursuing Bachelors degree program in Information Technology from JSS Academy of Technical Educ ation, India, PH-(+919582695085). E-mail:tarun.verma25d@gmail.com:*
- *Sweety Raj is currently pursuing Bachelors degree program in Information Technology from JSS Academy of Technical Educ ation, India, PH-(+919015525780). E-mail:sweetyraj.raj@gmail.com*
- *Mohammad Asif Khan is currently pursuing Bachelors degree program in Computer Science Engineering in JSS Academy of Technical Educ ation, India, PH-(+919718651204). E-mail:mohammad.asif.khan13@gmail.com*
- *Palak Modi is currently pursuing Bachelors degree program in Computer Science Engineering in JSS Academy of Technical Educ ation, India, PH-(+91)9837074908. E-mail palakmodi25@gmail.com*

3.3 CART

CART (Classification and regression trees) was introduced by Breiman, (1984). It builds both classifications and regressions trees. The classification tree construction by CART is based on binary splitting of the attributes. It is also based on Hunt's model of decision tree construction and can be implemented serially (Breiman, 1984). It uses gini index splitting measure in selecting the splitting attribute. Pruning is done in CART by using a portion of the training data set (Podgorelec et al, 2002). CART uses both numeric and categorical attributes for building the decision tree and has in-built features that deal with missing attributes (Lewis, 200). CART is unique from other Hunt's based algorithm as it is also use for regression analysis with the help of the regression trees. The regression analysis feature is used in forecasting a dependent variable (result) given a set of predictor variables over a given period of time (Breiman, 1984). It uses many single variable splitting criteria like gini index, symgini etc and one multi-variable (linear combinations) in determining the best split point and data is sorted at every node to determine the best splitting point. The linear combination splitting criteria is used during regression analysis. SALFORD SYSTEMS implemented a version of CART called CART using the original code of Breiman (1984). CART has enhanced features and capabilities that address the short-comings of CART giving rise to a modern decision tree classifier with high classification and prediction accuracy.

3.4 SLIQ

SLIQ (Supervised Learning In Ques) was introduced by Mehta et al, (1996). It is a fast, scalable decision tree algorithm that can be implemented in serial and parallel pattern. It is not based on Hunt's algorithm for decision tree classification. It partitions a training data set recursively using breadth-first greedy strategy that is integrated with pre-sorting technique during the tree building phase (Mehta et al, 1996). With the pre-sorting technique sorting at decision tree nodes is eliminated and replaced with one-time sort, with the use of list data structure for each attribute to determine the best split point (Mehta et al, 1996 and Shafer et al, 1996). In building a decision tree model SLIQ handles both numeric and categorical attributes. One of the disadvantages of SLIQ is that it uses a class list data structure that is memory resident thereby imposing memory restrictions on the data (Shafer et al, 1996). It uses Minimum Description length Principle (MDL) in pruning the tree after constructing it MDL is an inexpensive technique in tree pruning that uses the least amount of coding in producing tree that are small in size using bottom-up technique (Anyanwu et al, 2009 and Mehta et al, 1996).

3.5 SPRINT

SPRINT (Scalable Parallelizable Induction of decision Tree algorithm) was introduced by Shafer et al, 1996. It is a fast, scalable decision tree classifier. It is not based on Hunt's algorithm in constructing the decision tree, rather it partitions the training data set recursively using breadth first greedy technique until each partition belong to the same leaf node or class (Anyanwu et al, 2009 and Shafer et al, 1996). It is an enhancement of SLIQ as it can be implemented in both serial and parallel pattern for good data placement and load balancing (Shafer et al, 1996). In this paper we will focus on the serial implementation of SPRINT. Like SLIQ it uses one time sort of the data items and it has no restriction on the input data size. Unlike SLIQ it uses two data structures: attribute list and histogram which is not memory resident making SPRINT suitable for large data set, thus it removes all the data memory restrictions on data (Shafer et al, 1996). It handles both continuous and categorical attributes.

4 Proposed Approach

The algorithm that was used C4.5 because of the following advantages:

- Handling both continuous and discrete attributes - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Handling training data with missing attribute values - C4.5 allows attribute values to be marked as '?' for missing. Missing attribute values are simply not used in gain and entropy calculations.
- Handling attributes with differing costs.
- Pruning trees after creation - C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

Since weka 3.7 tool uses the java version of C4.5 as j48 algorithm. So this tool has been used for analyzing the data [6].

5 Analysis

As per the data, according to United Nations Development Programme (UNDP) report 2011. Georgia ranks 1 and has 100% literacy rate. Estonia, Latvia, Slovenia, Ukraine, Russia, Hungary are some of the countries including 94 other countries that show highest literacy rate. 65 countries show moderate literacy rate from 80%-50%. Some of the countries as shown in figure 1 and 13 countries like Mali, South Sudan, Niger and Guinea show the lowest literacy rate among all.

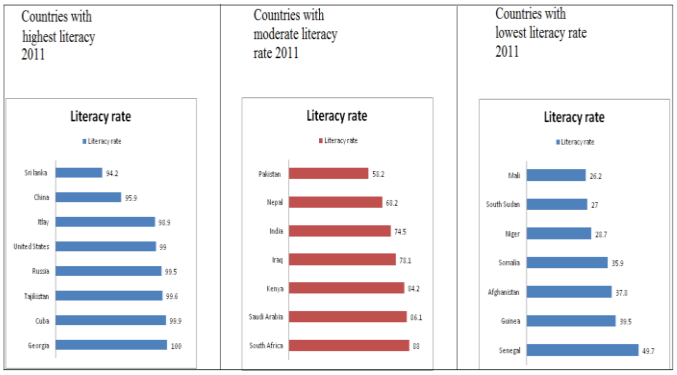

Figure 1 List of Countries according to their literacy rate

When the data was applied to weka 3.7 tool, the following decision tree was inferred as shown in figure 2.

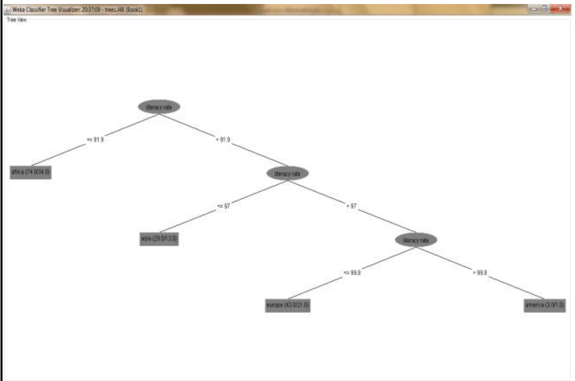

Figure 2 Decision tree for world literacy rate

From the decision tree the following information was obtained

1. Africa has a literacy rate less than or equal to 91.9%.

2. Asia has a literacy rate between 91.9% and 97%.

3. Europe has a literacy rate between 97% and 99.8%.

4. America has a literacy rate greater than 99.8%.

As per the data published by the 2011 census (INDIA), it was found that Kerala ranks top in the Indian states for Literacy rates among girls and overall literacy rate with 92% and 96% respectively and Rajasthan at the bottom of the table for female literacy rate with 52.7%. Refer figure 3.

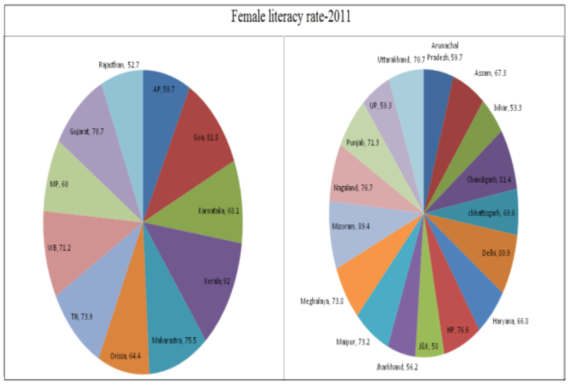

Figure 3 Female Literacy rate, 2011

When the data was applied to weka 3.7 tool, the following decision tree was inferred. Refer figure 4.

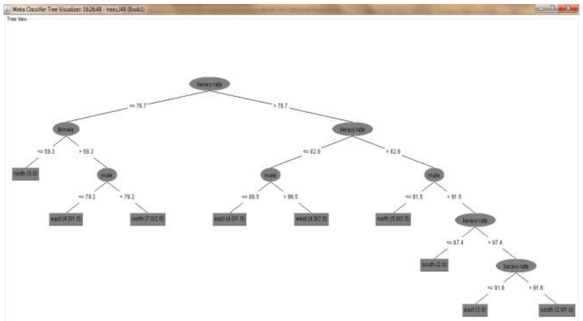

Figure 4 Decision tree of Indian Literacy rate

From the decision tree the following information was inferred

1. In India, females have a literacy rate less than 76.7% of total literacy rate and less than 59.3% of female literacy rate is found in north India.

2. Males in India have a literacy rate less than 76.7% of total literacy rate, greater than 59.3% of female literacy rate and males with a literacy rate less than 79.2% of male literacy rate are found in east India. Whereas males with a literacy rate greater than 79.2% of male literacy rate are found in North India.

3. Males with more than 82.9% of total literacy rate with less than 86.5% of male literacy rate, are in east India and those with literacy rate greater than 86.5% of male literacy rate are in north India.

4. Males with literacy rate more than 82.9% of total literacy rate and less than 91.5% of male literacy rate are found in north India.

5. Males with literacy rate greater than 82.9% of total literacy rate, more than 91.5% of male literacy rate and less than the 87.4% of total literacy rate is found in south India.

6. Males with literacy rate greater than 82.9% of total literacy rate and more than 91.5% of male literacy rate, greater than 87.4% of total literacy rate and less than 91.6% of total literacy rate are found in east India.

7. Males with literacy rate greater than 82.9% of total literacy rate and more than 91.5% of male literacy rate, greater than 87.4% of total literacy rate and greater than 91.6% of total literacy rate are found in south India.
Only enrolment to education system is not necessary but also completing education till 5th grade is also important. So analysis of such data is important.

As per the data published by the ASER report 2011 (INDIA) for ages 6-14, it was found that Goa, Pondicherry and Daman & Diu have zero dropouts and Uttar Pradesh has highest dropouts with 6.1%. Again when the data was applied to weka 3.7 tool, following decision tree was obtained. Refer figure 5.dcdcndncxdjncxdj

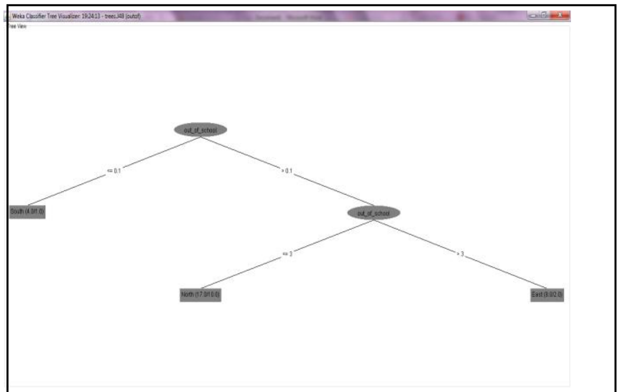

Figure 5 Decision tree for dropouts of India

What can be derived from the above decision tree is that south has less out of school children then north and then east.

1. South has dropout rate less than and equal to 0.1%.

2. North has dropout rate greater than 0.1% and less than 3%
.
3. East has dropout rate greater than 3%.

4. Tree is not showing west region as Goa, Maharashtra and Gujarat have different ranges, Goa has 0%, Maharashtra has 1.1% and Gujarat has 2.7%. Such fluctuating ranges cannot be shown in the form of decision tree.

6 Conclusion

In the world, five countries with highest literacy rate are Georgia, Cube, Estonia, Latvia and Barbados and five countries with lowest literacy rate are Mali, South Sudan, Ethiopia, Niger and Burkina Faso. If categorized continent wise America has the highest literacy rate followed by Europe, Asia and Africa respectively.

1. **America ranks 1st**
2. **Europe ranks 2nd**
3. **Africa ranks 3rd**
4. **Asia ranks 4th**

In India, Kerala has the highest literacy rate of 93.9% and Bihar has the lowest literacy rate of 63.8%. If categorized region wise South has the highest literacy rate followed by East, West and North respectively.

1. **South ranks 1st**
2. **East ranks 2nd**
3. **West ranks 3rd**
4. **North ranks 4th**

In India, it was found that Goa, Pondicherry and Daman & Diu have zero dropouts and Uttar Pradesh has highest dropouts with 6.1%. If categorized region wise East has the highest dropouts followed by North and South respectively.

1. **East ranks 1st**
2. **North ranks 2nd**
3. **South ranks 3rd**

Acknowledgment

We owe special debt of gratitude to **Professor S.S. Prasad** Principal of JSS Academy of Technical Education, Noida for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

References

[1] Wikipedia, “Literacy”, http://wikipedia.org/wiki/literacy, May 1, 2012.

[2] Wikipedia, “List of countries by literacy rate”, http://en.wikipedia.org/wiki/List_of_countries_by_literacy_rate.htm, April 2, 2012.

[3] Pratham, Annual Status of Education Report (Rural) 2011, http://www.pratham.org/M-20-3-ASER.aspx, January 16, 2012.

[4] MapsofIndia, “Literacy Rate in India”, http://www.mapsofindia.com/census2011/literacy-rate.html, March 20, 2012.

[5] Matthew N. Anyanwu, Sajjan G. Shiva, “Comparative Analysis of Serial Decision Tree Classification Algorithms”, *International Journal of Computer Science and Security (IJCSS),* Volume 3, Issue 3, June 2009, ISSN (Online): 1985-1553

[6] Weka the University of Waikato, “Weka 3: Data Mining Software in Java”, http://www.cs.waikato.ac.nz/ml/weka/, March 15, 2012.