A Parallel Weighted Decision Tree Classifier for Complex Spatial Landslide Analysis: Big Data Computation Approach

P. Anbalagan
Assistant Professor,
Department of Computer Science & Engineering,
Annamalai University
Annamalainagar – 608 002,
Tamil Nadu, India.

R.M. Chandrasekaran
Professor,
Department of Computer Science & Engineering,
Annamalai University
Annamalainagar – 608 002,
Tamil Nadu, India.

ABSTRACT

Effective and efficient strategies to acquire manage and analyze data leads to better decision making and competitive advantage. The development of cloud computing and the big data era, brings up challenges to traditional data mining algorithms. The processing capacity, architecture and algorithms of traditional database system are not coping with big data analysis. Big Data are now rapidly growing in all science and engineering domains, including biological, biomedical sciences and disaster management. The characteristics of complexity formulate an extreme challenge for discovering useful knowledge from the big data. Spatial data is complex big data. The aim of this paper is to propose Parallel Weighted Decision Tree Classifier to handle complex spatial landslide big data using Map Reduce programming model. The Proposed Classifier performance is validated with massive dataset. The results indicate that our classifier exhibits both time efficiency and scalability.

Keywords

Big Data, Classifier, Spatial Data, Map Reduce, Landslide..

1. INTRODUCTION

Very large amount of Geo-spatial data leads to definition of complex relationship, which creates challenges in today data mining research. Current scientific advancement has led to a flood of data from distinctive domains such as healthcare and scientific sensors, user-generated data, internet and disaster management. Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. For instance, big data is commonly unstructured and require more real-time analysis. This development calls forms system architectures for data acquisition, transmission, storage, and large-scale data processing mechanisms. Hadoop is a platform for distributing computing problems across a number of servers. First developed and released as open source by Yahoo, it implements the MapReduce approach pioneered by Google in compiling its search indexes. Hadoop's MapReduce involves distributing a dataset among multiple servers and operating on the data: the "map" stage. The partial results are then recombined: the "reduce" stage. To store data, Hadoop utilizes its own distributed file system, HDFS, which makes data available to multiple computing nodes.

Natural disasters like hurricanes, earthquakes, erosion, tsunamis and landslides cause countless deaths and fearsome damage to infrastructure and the environment. Landslide is the one of the major problem in hilly areas. Landslide Risk can be identified using different methods based on the GIS technology. In Ooty, Nilgiri district, landslide was happened due to the heavy rainfall and frequent modification of land use features. Landslide disaster could have been reduced, if more had been known about forecasting and mitigation. So far, few attempts have been made to predict these landslides or prevent the damages caused by them. In the previous studies, various approaches were applied to such problems which show that it is difficult to understand and tricky to predict accurately. In order to analyze these landslides, various factors, such as Rainfall, Geology, Slope, land use/land cover, soil and Geomorphology are considered and the relevant thematic layers are prepared in GIS for landslide susceptibility mapping. The data collected from various research institutes related to landslide helped to predict and analyze the landslide susceptibility. The spatial landslide data is one of the complex big data. To handle such as large amount of landslide data, the previous study weighted decision tree approach is improvised and parallel weighted decision classifier is proposed using map reduce programming model.

2. RELATED WORK

Decision trees are one of the most accepted methods for classification in diverse data mining applications [1-2] and help the development of decision making [3].

One of the well known decision tree algorithms is C4.5 [4-5], an expansion of basic ID3 algorithm [6]. However, with the growing improvement of cloud computing [7] as well as the big challenge [8,9], traditional decision tree algorithms reveal numerous restriction. First and foremost, building a decision tree can be very time consuming when the volume of dataset is extremely big, and new computing paradigm should be applied for clusters. Second, although parallel computing [10] in clusters can be leveraged in decision tree based classification algorithms [11,12], the strategy of data distribution should be optimized, so that required data for building one node is localized and mean while the communication cost to be minimized. Weighted classifications are well-suited for many real-world binary classification problems. Weighted classification [14] assigns different importance degrees to different attributes. Many different splitting criteria for attribute selection have been proposed in the literature and they all tend to provide similar results [13].

HACE theorem [15] has been presented which characterizes the features of the Big Data revolution, and proposes a Big

Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

An integration of remote sensing, GIS and Data mining techniques has been used to predicting the landslide risk. The probabilistic and statistical approaches were applied for estimating the landslide susceptibility area. Landslide susceptibility map is reduced the landslide hazard and is used for land cover planning. The frequency ratio model has better than logistic regression model. Fuzzy membership functions and factor analysis were used to assess the landslide susceptibility using various factors. The spatial data were collected and processed and create a spatial database using GIS and Image processing techniques. The landslide occurrence factor was identified and processed. Each factor weight was determined and calculated the training using back-propagation. Improvised Bayesian Classification approach [16] and decision tree approach [17] have been applied to predict the landslide susceptibility in Nilgiri district.

3. PARALLEL WEIGHTED DECISION TREE CLASSIFIER

Classification is the process to predict the unknown class label using training data set. Classification approaches are categorized into Decision Tree, Back propagation Neural Network, Support Vector machine(SVM),Rule based Classification and Bayesian Classification. In the present scenario, landslide analysis study was done by using Neural Network and Bayesian but these approaches are difficult to understand and tricky to predict. In this paper, Parallel Weighted Decision Tree Classifier to handle complex spatial landslide big data using MapReduce programming model is proposed for landslide Risk Analysis. The performance of the proposed approach is measured with various parameters.

Decision Tree (DT) approach is used to analyze the data in the form of tree. The Tree is constructed using the top-down and recursive splitting technique. A tree structure consists of a root node, internal nodes, and leaf nodes. Weighted classification techniques give simpler models for the important classes. Weighted classification assigns different importance degrees to different landslide factor. In this paper, we assign weights to the different landslide factors in order to represent the relative importance of each landslide factor. We represent the weight corresponding to landslide factor. In a distributing computing environment, the large data sets are handled by an open source framework called Hadoop. It consists of Mapreduce, Hadoop Distribution File System (HDFS) and number of related projects Apache Hive, HBase and Zookeeper.

The Hadoop Distributed File Systems (HDFS) architecture is illustrated in Fig. 1 NameNode is the master node of HDFS handling metadata, and DataNode is slave node with data storage in terms of blocks. Similarly, the Master node of Hadoop MapReduce is called JobTracker, which is in charge of managing and scheduling several tasks, and the slave node is called TaskTracker, where Map and Reduce procedures are actually performed.

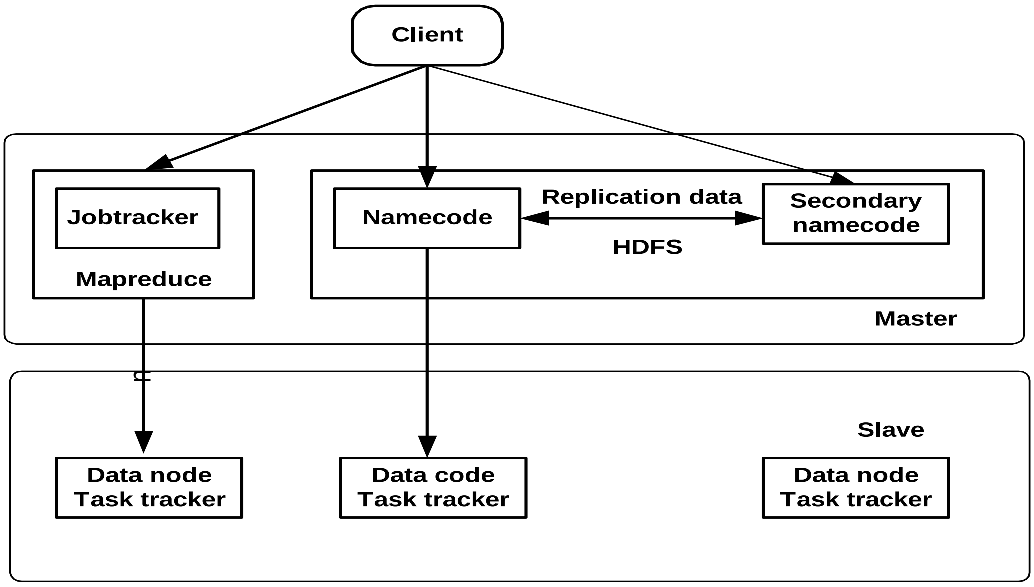

Fig. 1 HDFS Architecture

MapReduce programming model is used for parallel and distri buted processing of large datasets on clusters [16]. There are two basic procedures in MapReduce: Map and Redu ce.

In general, the input and output are both in the form of key value pairs. The Fig. 2 shows MapReduce programming model architecture. The input data is divided in to block in the size of 68MB or 128 MB. The mapper input will be supplied as key/value pairs and it produces the relative output in the form of key/pairs. Partitioner and combiner are used in between mapper and reducer to perform sorting and shuffling. The Reducer iterates through the values that are associated with specific key and produces zero or more outputs.

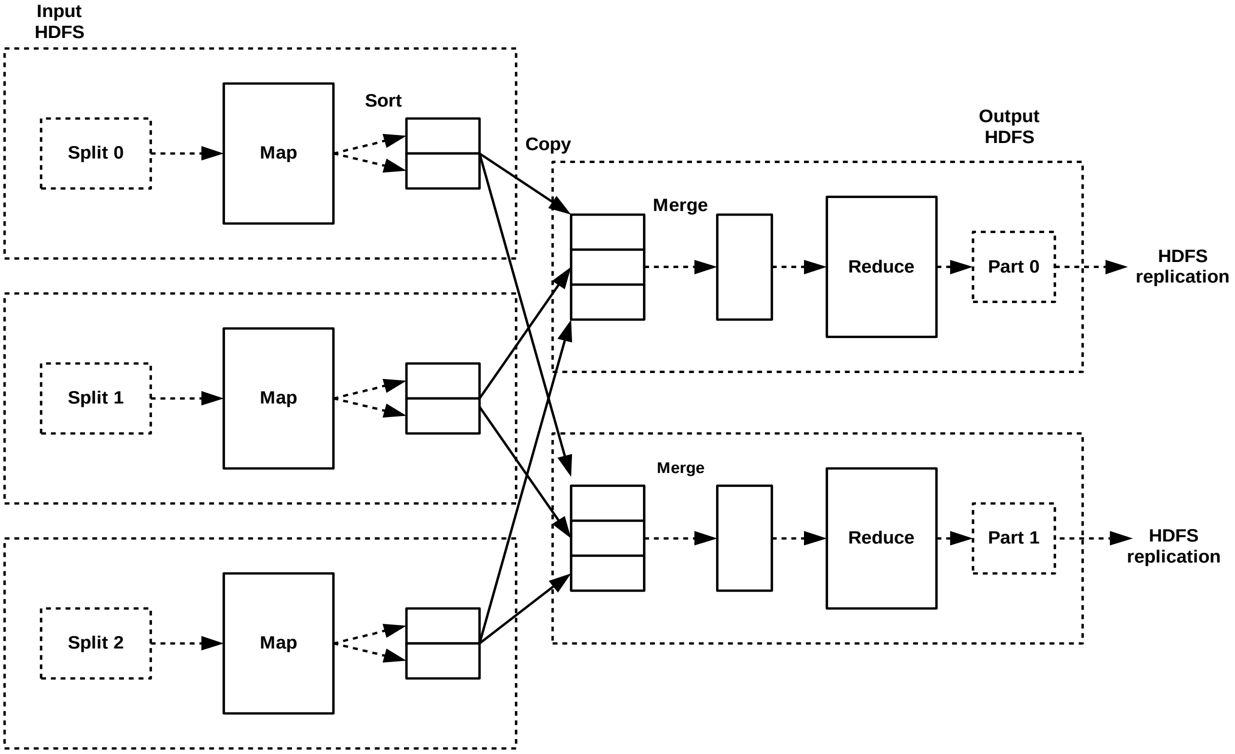

Fig. 2 Map Reduce Architecture

The dataset is relatively huge in a big data atmosphere, designing appropriate data structures for parallel programming is very much important. Three data structures such as attribute table, count table, hash table are used to build parallel decision tree classifier.

Basic information of attribute "a", the row identifier of instance "row_id", values of attribute "values(a)" and class labels of instances "c" are stored in attribute table. Count table computes the count of instances with specific class labels if split by attribute a. That is, two fields are included: class label c and a count. The last one is hash table, which stores the link information between tree nodes node_id and row_id, as well as the link between parent node node_id and its branches.

```
Algorithm -I: Data Conversion
Procedure Map_Attribute (tuple_id,(A1,A2,…A3,C))
        emit (Aj,(row_id,C))
end procedure
Procedure Reduce_Attribute ((Aj,(row_id,C))
emit (Aj,(C,Cnt))
end Procedure
```

In Decision Tree Classifier, selecting best splitting attribute abest is important task. The algorithm - II shows that, mapper performs the computation of information and split information of Aj. The reducer computes the information gain ratio. The attribute Aj which has maximum value of GainRatio is selected as splitting attribute.

```
Algorithm-II: Splitting Attribute Selection
Procedure Reduce_Population((Aj,(C,Cnt))
        emit (Aj,all)
end Procedure
Procedure Map_Computation((Aj,(C,Cnt,all)))
        $Compute\ Entrophy\ (A_j\ )$
$Compute\ Info\ (A_j\ ) = \frac{Cnt}{all} Entrophy(A_j)$
$Compute\ Split\ Info\ (A_j\ ) = \frac{Cnt}{all} log\ \frac{Cnt}{all}$
end Procedure
Procedure                                    Reduce_Computation
( $(A_j, Info\ (A_j\ ), Split\ Infor\ (A_j\ ))$ )
 $emit\ (A_j, Gain\ Ratio\ (A_j\ ))$
end Procedure
```

As shown in algorithm - III, the records are read from attribute table with key value equals to abest and emit the count of class labels.

```
Algorithm-III: Hash Table updation

Procedure Map_Update_Count((Abest,(row_id,C)))
emit (Abest, (C,Cnt'))
end Procedure
Procedure Map_Hash((Abest, row_id))
        compute node_id= hash(Abest )
        emit (row_id, node_id)
end Procedure
```

Algorithm –IV shows the procedure to grow the decision tree by building linkages between nodes.

```
Algorithm-IV: Building Tree
Procedure Map((Abest,row_id))
Compute node_id=hash(Abest)
If node_id is same with the old value then
```

```
        emit(row_id, node_id)
end if
add a new subnode
emit(row_id, node_id, subnode_id)
end Procedure
```

Weighted classification techniques give simpler models for the important classes. Weighted classification assigns different importance degrees to different landslide factor. In this paper, we assign weights to the different landslide factors in order to represent the relative importance of each landslide factor. Weighted decision tree classification algorithm is improved as parallel weighted decision tree classifier using map reduce programming model as shown in the above algorithms. The developed classifier is used to analyze the landslide risk in the Ooty region of Nilgiri district. The proposed classifier scalability is improved and performance is compared with the existing classification methods.

4. EXPERIMENT AND RESULT ANALYSIS

The proposed parallel decision tree classifier is implemented in Hadoop cluster. We have HPC cluster with 6 nodes. We let one of them as HDFS NameNode and MapReduce JobTracker (i.e., master), and the remaining nodes act as HDFS DataNode and MapReduce Task Tracker (i.e., slave). The efficiency of weighted decision tree classification algorithm is theoretically and empirically proved in our previous study. In this paper, we are concerned with the time efficiency of parallel version of weighted decision tree classification algorithm in big data environment. This paper focuses landslide risk analysis using big data computational techniques. The needed toposheets and required maps are collected from the geological survey of India. Many number of factors causes landslide in the hill region, but four factors are very important for landslide study such as rainfall, slope, geology, and land use/land cover. The above said factors thematic layers are prepared from the LISS III+ PAN images using ArcGIS Tool. Ooty, Nilgiri district is considered as study area .We have applied the proposed weighted decision tree classifier on Ooty landslide data as shown in Table1.

Land use	Geology	Rainfall	Slope	Zone
Agriculture	Ultrabasic rocks	135.63-150.82	11.76-19.79	Low
Agriculture	Gneiss	135.63-150.82	8.02-11.76	Low
Agriculture	Ultrabasic rocks	135.63-150.82	8.02-11.76	Very Low
Scrub Forest	Gneiss	135.63-150.82	0-8.02	Very Low
Scrub Forest	Ultrabasic rocks	135.63-150.82	0-8.02	Very Low
Scrub Forest	Gneiss	135.63-150.82	11.76-19.79	Low
Scrub Forest	Ultrabasic rocks	135.63-150.82	11.76-19.79	Very Low

Table 1 Sample Landslide Data

The proposed parallel weighted decision tree classifier is applied on ooty landslide data and the landslide risk level is analyzed and it is shown in Fig. 3.

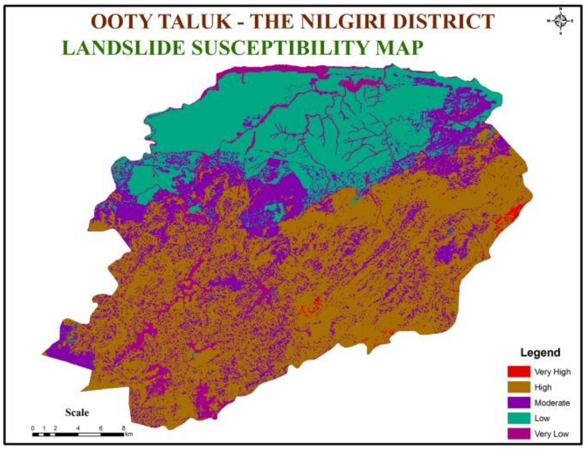
OOTY TALUK - THE NILGIRI DISTRICT
LANDSLIDE SUSCEPTIBILITY MAP
Legend
Very High
High
Moderate
Low
Very Low
Scale
0 1 2 4 6 8
km

Fig. 3 Landslide Risk analysis using parallel Weighted Decsion Tree Classifer

The performance of proposed parallel weighted decision tree classifier is compared with the weighted decsion tree classifier and decision tree classifier on single node. Fig. 4 illustrates the following observations.

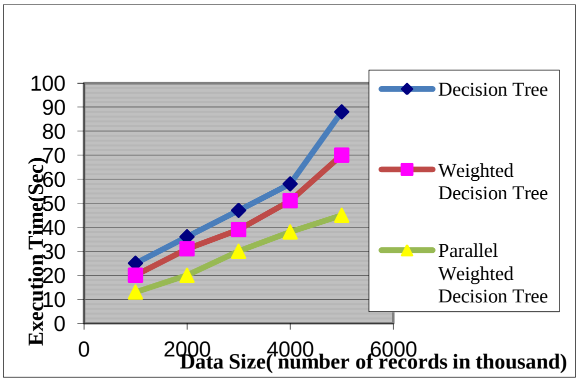

Fig. 4 Performance of Parallel Weighted Decsion Tree Classifier

First, the larger the dataset is, the more time consuming it is to build the normal decision tree approach. Second, the execution time of weighted decsion tree classification takes more time than proposed parallel weighted decsion tree classifier. The proposed MapReduce based parallel weighted decsion tree classifier algorithm takes less time the original decision tree as the size of dataset increases. Therefore, it is proved that the proposed parallel decision tree classifier outperforms the sequential version even on a single node environment.

The scalability of the proposed weighted decsion tree classification is also tested in distributed parallel domain. The scalability evaluation includes two aspects: (1) performance with different numbers of nodes, and (2) performance with different size of training datasets.

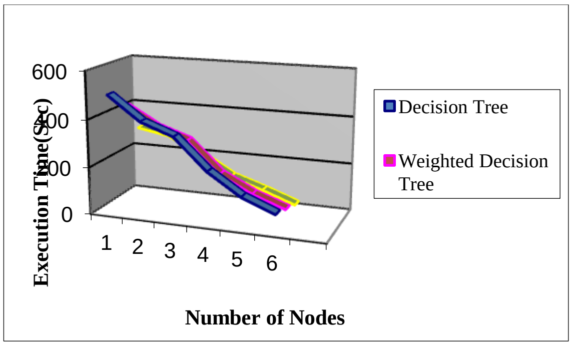

Fig. 5 Performance of Parallel Weighted Decsion Tree Classifier based on number of nodes

Fig. 5 illustrates the execution time of our proposed weighted decsion tree classification algorithm with different numbers of nodes when the number of record is 1, 2 and 3 lakhs respectively, We have observed that the overall execution time decreases when the number of nodes increases. This indicates that the more nodes are involved for computing increases the efficiency of the algorithm.

5. CONCLUSION

Predicting and analyzing disaster is complex task. In this paper, landslide risk is analyzed using Parallel Weighted Decision Classifier approach. Disaster management domain generates huge amount of data.Traditional sequential decision tree algorithms cannot fit to handle such huge data sets. For example, as the size of training data grows, the process of building decision trees can be very time consuming. To solve the above challenges, parallel weighted decision classifier approach is proposed to improve the scalability of the model. We have compared the performance of the proposed approach with existing approach with respect to number of nodes and number of record. The empirical results shows that the proposed algorithm exhibit both time efficiency and scalability. In future works, the rainfall induced landslide risk analysis will be studied using big data computational approaches.

6. REFERENCES

[1] H. I. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2005.

[2] M. J. Berry and G. S. Linoff, "Data mining techniques : For marketing, sales, and customer support", John Wiley & Sons, Inc., 1997.

[3] J.R. Quinlan, "Decision trees and decision-making", IEE Transactions Systems, Man and Cybernetics, Vol. 20, No.2, pp. 336-346, 1990.

[4] J. R. Quinlan, "C4.5: programs for machine learning", Morgan Kaufmann, 1993.

[5] J. R. Quinlan, "Improved use of continuous attributes in C4.5", arXiv preprint cs/9603103, 1996.

[6] J. R. Quinlan, "Induction of decision trees", Machine Learning, vol. 1, no. 1, pp. 81-106,1986.

[7] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I.Stoica and M. Zaharia, "A view of cloud computing", Communications of the ACM, vol. 53, no. 4, pp.50-58, 2010.

[8] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D.P. Hill, R. Kania, M. Schaeffer, S.S.Pierre, S. Twigger, O. White and S.Y. Rhee. "Big data: The future of biocuration", Nature, vol.455, no.7209, pp.47-50, 2008.

[9] P. Zikopoulos and C. Eaton, "Understanding big data: Analytics for enterprise class hadoop and streaming data", McGraw-Hill Osborne Media, 2011.

[10] V. Kumar, A. Grama, A. Gupta and G. Karypis, "Introduction to parallel computing"Redwood City: Benjamin/Cummings, vol. 110, 1994.

[11] K. W. Bowyer, L. O. Hall, T. Moore, N. Chawla and W. P. Kegelmeyer, "A parallel decision tree builder for mining very large visualization datasets", IEEE International Conference on Systems, Man, and Cybernetics, vol. 3, pp. 1888-1893, 2000.

[12] J. Shafer, R. Agrawal and M. Mehta, "SPRINT: A scalable parallel classifier for data mining", Proc. 1996 Int.Conf. Very Large Data Bases, 1996.

[13] F.Berzal, J.C.Cubero, F.Cuenca, and M.J.Martín-Bautista, "On the quest for easy-to-understand splitting rules", Data and Knowledge Engineering, Vol. 44, No. 1, pp. 31–48, 2003.

[14] J.L.Polo, F.Berzal, and J.C.Cubero, "Taking class importance into account", Lecture Notes in Computer Science, 2007.

[15] Xindong Wu,Xingquan Zhu,Gong-Qing Wu,and WeiDing.S, "Data Mining with BigData", IEEE transactions on knowledge and data engineering, Vol.26, No.1, january2014.

[16] Venkatesan M, Rajawat A S, Arunkumar T, Anbarasi M, Malarvizhi K, "GIS Based Data Mining Classification Approaches for Landslide Susceptibility Analysis", International Journal of Applied Environmental Sciences, Volume 9, Number 5, pp. 2345-2357, 2014.

[17] Venkatesan.M, Arunkumar .Thangavelu, and Prabhavathy.P, "An Improved Bayesian Classification Data mining Method for Early Warning Landslide Susceptibility Model Using GIS", Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications 9BICTA, Advances in Intelligent Systems and Computing Springer India 2013.