Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach

Umesh D. R.
Assistant Professor
Department of Computer Science & Engineering
PESCE, Mandya, Karnataka, India

B. Ramachandra, PhD
Professor
Department of Electrical & Electronics Engineering
PESCE, Mandya, Karnataka, India

ABSTRACT
The traditional data analytic might not have the capacity to handle enormous amount of data. Due to the rapid growth of information, solutions need to be contemplated and provided in order to handle and extract value and knowledge from these data sets. Moreover, decision makers should have the capacity to increase significant bits of knowledge from such fluctuated and quickly evolving information. Such esteem can be given utilizing big data analytic, which is the utilization of advanced analytic techniques on big data using MapReduce approach. This paper examines to develop a high performance platform to efficiently analyse big SEER (Surveillance, Epidemiology, and End Results) breast cancer data set using MapReduce to find the recurrence of breast cancer.

Keywords
Breast cancer; Big data, Classification; Data analytics, MapReduce, SEER.

1. INTRODUCTION
The volume of data generated in the fields of science & technology is growing extremely fast [1-4]. Fortunately, with the support of the MapReduce [5–8] paradigm, researchers gear up to take on a simple programming interface for parallel scaling up of many data mining algorithms on larger data sets. It was shown [9] that algorithms which accommodate the Statistical Query Model [10] can be written in a certain "summation form". They illustrated 10 variant algorithms that can be reasonable parallelized on multi-core computers applying the MapReduce paradigm. In 2009, Google represented PLANET: a framework for large-scale tree erudition using a MapReduce cluster [11]. Their goal in building PLANET was to build up an adaptable tree learner which could accomplish similar exactness execution as the conventional in-memory calculations furthermore have the capacity to manage bigger datasets. PLANET is utilized to develop versatile order and relapse trees, and also groups of these models. It understands parallelization by isolating tree learning into numerous circulated calculations, each executed with MapReduce. There are two main steps in the supervised classification process. The first is the training step where the classification model is built. The second is the classification itself, which applies the trained model to assign obscure information to one out of a given set of class labels. In spite of the fact that the training step is the one that draws more exploratory consideration [12-15], it generally depends on a little illustrative information set that does not speak to an issue for big data applications. Accordingly, the big data challenge affects mostly the classification step.

This paper is organized as follows: section "Related work" introduces work that has previously been proposed for solving the problem in Hadoop MapReduce; section "Proposed Algorithm" presents the proposed algorithm; section "Experimental Result" demonstrates the performance of MapReduce in terms of accuracy; section "Conclusion" concludes the paper.

2. RELATED WORK
Santi Wulan Purnami et al. in their research work used bolster vector machine for feature selection and classification of breast cancer [16]. They focused on how 1-norm SVM can be used as a part of feature selection and smooth SVM (SSVM) for classification. Wisconsin breast cancer dataset was used for breast cancer analysis. The basic attributes were at first recognized and the finding was done based on nine chosen attributes. Then again, the study can't be termed precise because of the limited number of attributes.

Farzaneh Keivanfard et al.in their work, have connected feature selection and classification methods in perspective of artificial neural network to characterize breast cancer on dynamic Magnetic Resonance Imaging (MRI) [17]. A forward selection method was applied to find the best elements for characterization. Likewise, artificial neural networks such as Multilayer Preceptron (MLP) neural network, Probabilistic Neural Network (PNN) and Generalized Regression Neural Network (GRNN) were connected to classify breast cancer into two groups; benign and malignant tumor. An accuracy of 100% was accomplished utilizing GRNN and PNN. Then again, specificity procured in this study can't be termed precise in light of the way that the quantity of circumspect cases in the database was not respectably high.

Lambrou et al. exhibited a Conformal Predictor in light of Genetic Algorithms, and connected to Wisconsin Breast Cancer Diagnosis (WBCD) problem [18]. A standard based Genetic Algorithms (GAs) was used as a procedure for building a Conformal Prediction (CP). The resulting algorithm was connected to the problem of breast cancer diagnosis for 683 records without missing values from WDBC dataset. The error rates insisted the authenticity of their CP for any given confidence level 1-e, where e is the error rate.

Liu Ya-Qin et al proposed predictive models for breast cancer survivability utilizing SEER data [19]. C5.0 decision tree algorithm was initially utilized on the imbalanced data and afterward under testing was applied to the models to defeat the impediment of imbalanced data. Bagging algorithm was then used to build the characterization's execution for predicting breast cancer survivability. The results procured showed an accuracy of 0.7678.

Ankit Agrawal et al. in their work examined the lung cancer data available from the SEER database for making survival forecast models utilizing data mining techniques [20]. SEER

data attributes were classified as demographic attributes, diagnosis attributes, treatment attributes and outcome attributes. A few classification techniques were applied to model the five outcomes of survival after 6 months, 9 months, 1 year, 2 years and 5 years. Attribute selection techniques were applied to recognize a small non-redundant set of attributes to develop a model mortality risk calculator. It was found that the way of forecast was held even with small number of non-redundant attributes.

Delen et al, in their work, have created models for predicting the survivability of analyzed cases utilizing SEER breast cancer dataset [21]. Two algorithms artificial neural network (ANN) and C5.0 decision tree were utilized to create prediction models. C5.0 gave an accuracy of 93.6% while ANN gave an accuracy of 91.2%. Bellaachia et al. took the investigation of Delen et al. as the basis of their research [22]. They have reported that the pre-classification method of Delen et al was not accurate in deciding the records of "not survived" class as the reason for death and survivability rate were not taken into consideration. They investigated three data mining techniques: the Naïve Bayes, the back propagated neural network, and the C4.5 decision tree algorithms. They have reported that C4.5 algorithm gave the best execution of 86.7% accuracy.

Umesh et al, in their work have used Association rule mining for predicting breast cancer recurrence on SEER dataset [23] using 17 attributes with a limited random dataset among the three best samples. The result procured an accuracy of 87.72% with a limited set of data records (i.e. 2143).

Pregel [24] is a concept like MapReduce. The distinction is that it gives a characteristic API to distributed programming system aimed for graph algorithms. It likewise supports iterative computations over the graph. This is a property which MapReduce needs. In Pregel computations, super steps, a sequence of iterations is adopted. With super steps, a vertex can get data from the past iteration furthermore send data to different vertices that will be received at a next super step. In any case, Pregel concentrates on graph mining algorithms, while we are keen on more general applications.

3. PROPOSED ALGORITHM

In this paper, an algorithm is proposed for predicting the recurrence of breast cancer for a breast cancer patient in SEER (Surveillance, Epidemiology, and End Results) dataset of Program of the National Cancer Institute (NCI). This dataset contained population characteristics and included 17 input variables. The data were pre-processed to evacuate inadmissible cases. After using data cleansing and data preparation strategies, the final dataset was constructed. Finally, SEER dataset were analyzed for breast cancer recurrences happen in the initial 5 years after breast cancer treatment. The independent variables that were utilized are demonstrated as a part of Table 1. The dataset were cleaned by handling missing values, noise, identifying and correcting inconsistencies by using Expectation maximization (EM) method [25].

Table 1: Variables Used For Breast Cancer Recurrance Modeling

Sl. No.	Variable Name
1.	Race
2.	Marital Status
3.	Primary site code
4.	Histological type
5.	Behavior code
6.	Grade
7.	Extension of Tumor
8.	Lymph node involvement
9.	Site specific surgery code
10.	Radiation
11.	Stage of cancer
12.	Age
13.	Tumor size
14.	Number of positive nodes
15.	Number of nodes
16.	Number of primaries
17.	Menopause

The proposed algorithm has entry to just a particular subset of training data. The algorithm generates a set of hypotheses and they are combined through weighted majority voting of the classes predicted by the individual hypotheses. To generate the hypotheses by training a weak classifier, instances drawn from an iteratively updated distribution of the training data are used. This distribution is updated so that instances misclassified by the previous hypothesis are more likely to be included in the training data of the next classifier. The pseudocode for the algorithm is said underneath:

Algorithm (D_n, T)

Input: Consider SEER dataset of n records $(x_1,y_1), \ldots, (x_n, y_n)$ with label classifications $y_i \in Y$ = {Recurrence (R), Non-Recurrence (NR)}; $x_i \in X$ is the object or instance; Base learner *B; and* Number of iterations *T*

Output: The final classifier $H_{final}(x)$

1. Initialize all the records with weight, so that $D_1(i) = \frac{1}{n}$
2. for t ← 1 to T do
3. Create distribution D_t on {1,…..,n} from the selected training subset S_t
4. Call base learner B, train B with S_t
5. Select weak classifier with smallest error rate (ε_t) on D_t

$$\varepsilon_t = Pr_{D_t}[h_t(x_i) \neq y_i]$$
$$h_t: x \rightarrow \{R, NR\}$$

6. if $\varepsilon_t > 0.5$, then set T = t – 1 and exit from loop.
7. Update distribution $D_{t+1}(i) = \frac{D_t(i)}{Z_t}\ C(x)$

$$C(x) = \begin{cases} \frac{\varepsilon_t}{1-\varepsilon_t} : y_i = h_t(x_i) \\ 1 : y_i \neq h_t(x_i) \end{cases}$$
$$\alpha_t = \log \frac{1-\varepsilon_t}{\varepsilon_t} > 0$$
$$Z_t \rightarrow \text{Normalization constant} \leq 1$$

8. Output: The final classifier $H_{final}(x) = argmax_{y i \in Y} \sum_{t:h_t(x)=y} \alpha_t$

Let the data set D_n={ (x_1, y_1) , (x_2, y_2) , ….., (x_m, y_m)} *with label classification* $y_i \in$ {Recurrence (R), *Non-Recurrence (NR)};* $x_i \in X$ is the object or instance; The algorithm initialize all the records with weight, so that $D_1(i) = \frac{1}{n}$ for all the examples in D_n, where $t \in [1, T]$ and T is the total number of iterations. Before beginning the first iteration these weights are uniformly initialized (line 1) also, they are updated in every consecutive iteration. At each iteration, a base learner function is applied to the weighted form of the data which then returns an optimal weak hypothesis h_t (line 5). This weak hypothesis minimizes the weighted error. At each iteration, a weight (α_t) is assigned to the weak classifier (line 7). At the end of T iterations, the algorithm returns the final classifier H which is a weighted average of all the weak classifiers.

Computational Complexity of Algorithm depends on the base learner algorithm in line 4. Rest of the operations can be performed in $\Theta(n)$. Let's consider decision trees with only two leaf nodes as base learners. Then the cost is $\Theta(dn)$ if the data examples are sorted in each attribute. Sorting all the attributes will take $\Theta(dn \log n)$ time and this has to be done only once before starting the first iteration. So, the overall cost of the T iterations is $\Theta(dn(T + \log n))$.

In order to implement this algorithm using MapReduce, for T iterations T MapReduce jobs should be put together by the driver program. This driver program additionally needs to decide for each iteration t whether this abortion condition $\varepsilon_t > 1/2$ is met. For this situation, the number of MapReduce jobs is smaller than T.

The experiments are deployed on Amazon EC2 and have utilized Weka software tool to experiment with this algorithm. Expectation maximization (EM) algorithm were adopted for an efficient estimation from incomplete data. In any inadequate dataset, there is indirect evidence about the probable estimations of the unobserved values. This evidence, when joined with a few suppositions, comprises a predictive probability distribution for the missing values that should be averaged in the statistical analysis. The EM algorithm is a typical strategy for coordinating models to deficient information. EM is vital on the relationship between missing information and obscure parameters of a model. At the point when the parameters are known, then it is conceivable to acquire unprejudiced expectations for the missing values [26].

4. EXPERIMENTAL RESULT

In this section, the performance of our proposed algorithms were demonstrated in terms of classification accuracy. The task is to learn a model that predicts whether breast cancer will recur for the breast cancer patient on a SEER dataset. All our experiments were performed on *Amazon EC2* cloud computing environment and the computing hubs used were of type *M3 instance configured* with Latest Intel Xeon Processor and SSD-backed instance storage that delivers higher I/O performance.

For experimenting 2,20,811 instances and 17 attributes were used for determining the classification accuracy. It can be seen from the Table 2: confusion matrix, that 25,291 of 2,20,811 records are characterized vaguely. 11,021 of the "RECURRENCE" cases have been classified as "NON-RECURRENCE" (False Negatives). 14,270 of the "NON-RECURRENCE" cases have been classified as "RECURRENCE" (False Positives) and also represented graphically in Figure 1 and Figure 2. Table 3: demonstrates the examination of execution as for sensitivity, specificity and accuracy.

Table 2: Confusion Matrix For Seer Dataset

	R	NR	TOTAL
RECURRENCE (R)	14,352 (TP)	11,021 (FN)	25,373
NONRECURRENCE (NR)	14,270 (FP)	1,81,168 (TN)	1,95,438
TOTAL	28,622	1,92,189	2,20,811

Figure 1: Result Of Recurrence (R) For Seer Breast Cancer Dataset

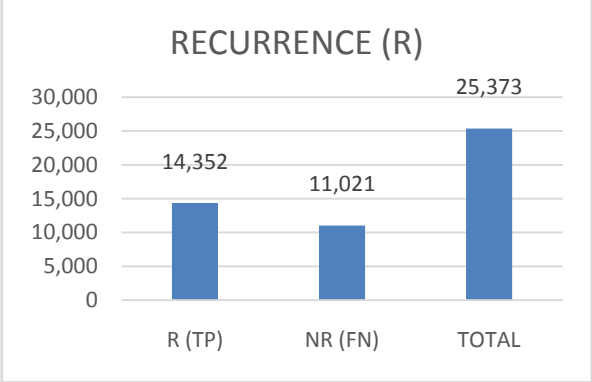

Figure 2: Result Of Non-Recurrence (Nr) For Seer Breast Cancer Dataset

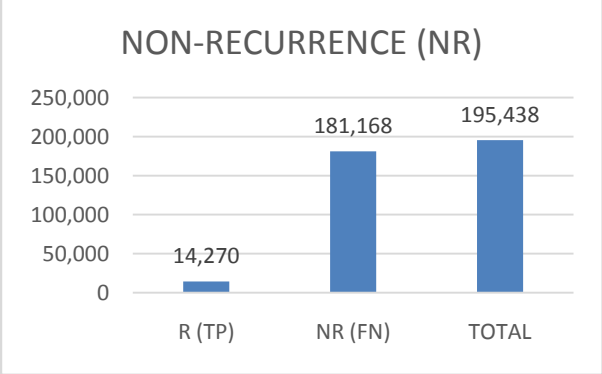

Table 3: Performance Analysis

SENSITIVITY	56.56%
SPECIFICITY	92.69%
ACCURACY	**88.54%**

There are couple of limitations with this usage. To start with, the training data sent to each MapReduce job dependent on each other as each training data subset S_t is drawn from the distribution D_t (Algorithm - line 3) and this distribution is updated based on the results of the previous MapReduce job. This implies these MapReduce jobs can't be executed in parallel as they need to sit tight for the distribution D_t from the past MapReduce job. Second, every time a MapReduce work begins it needs to read data from the HDFS where the past MapReduce job has stored the distribution D_t which will choose the training subset S_t. After this MapReduce job completes, it again composes its outcomes into the HDFS. For T iterations, the correspondence overhead is considerable as data are re-stacked, re-spared and re-processed for T times.

Therefore, a lot of CPU resources, network bandwidth and I/O are squandered. For smaller datasets, it turns into a main consideration which decreases the exhibitions. Third, as said some time recently, a driver project is required for each MapReduce job to check the end condition: $\varepsilon_t > 1/2$. This driver program is an additional MapReduce job and causes overheads as additional assignments should be booked, additional information need to peruse and spare to HDFS, additional networks resources are requested to move these data.

5. CONCLUSION

Our proposed algorithm implemented with MapReduce. The experimental results show that the error rates are more accurate and smaller in predicting the recurrence of breast cancer. For the proposed algorithm, since the base learners which process part of the original datasets work in one single machine sequentially.

In the future, parallelize and distributing the computation to more computing hubs for the sake of increasing the computational efficiency is planned and intend to use parallelized machine learning algorithms which will also improve the scalability to larger datasets.

6. REFERENCES

[1] Sagiroglu, S., and Sinanc, D., 2013. Big Data: A Review. *International Conference on Collaboration Technologies and Systems (CTS)*, pp. 42-47.

[2] Zaslavsky, A., Perera, C., and Georgakopoulos, D., 2012. Sensing as a Service and Big Data. *Proceedings of the International Conference on Advances in Cloud Computing (ACC)*, pp. 21-29.

[3] Suthaharan, S., 2014. Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning. *ACM SIGMETRICS Performance Evaluation Review*, 41(4), pp. 70-73.

[4] Kishor, D., 2013. Big Data: The New Challenges in Data Mining. *International Journal of Innovative Research in Computer Science & Technology*, 1(2), pp. 39-42.

[5] Dean J, Ghemawat S (2008) Mapreduce: simplified data processing on large clusters. Commun ACM 51(1):107–113

[6] White T (2012) Hadoop: The Definitive Guide. " O'Reilly Media, Inc.", California

[7] Venner J, Cyrus S (2009) Pro Hadoop. vol. 1. Springer, New York

[8] Lam C (2010) Hadoop in Action. Manning Publications Co., New York

[9] Chu C, Kim SK, Lin YA, Yu Y, Bradski G, Ng AY, Olukotun K (2007) Map-reduce for machine learning on multicore. Adv neural Info processing systems 19:281

[10] Kearns M (1998) Efficient noise-tolerant learning from statistical queries. J ACM (JACM) 45(6):983–1006

[11] Panda B, Herbach JS, Basu S, Bayardo RJ (2009) Planet: massively parallel learning of tree ensembles with mapreduce. Proc. VLDB Endowment 2(2):1426–1437

[12] Liu, B., Blasch, E., Chen, Y., Shen, D., and Chen, G., 2013. Scalable Sentiment Classification for Big Data Analysis Using Naïve Bayes Classifier. *IEEE International Conference on Big Data*, pp. 99-104.

[13] Dai, W., and Ji, W., 2014. A MapReduce Implementation of C4.5 Decision Tree Algorithm. *International Journal of Database Theory and Application*, 7(1), pp. 49-60.

[14] Kiran, M., Kumar, A., Mukherjee, S., and Prakash, R., 2013. Verification and Validation of MapReduce Program Model for Parallel Support Vector Machine. *International Journal of Computer Science Issues*, 10(3), pp. 317-325.

[15] Han, J., Liu, Y., and Sun, X., 2013. A Scalable Random Forest Algorithm Based on MapReduce. *4th IEEE International Conference on Software Engineering and Service Science*, pp. 849-852.

[16] Santi Wulan Purnami, S.P. Rahayu and Abdullah Embong, "Feature selection and classification of breast cancer diagnosis based on support vector machine", IEEE 2008.

[17] Farzaneh Keivanfard , Mohammad Teshnehlab , Mahdi Aliyari Shoorehdeli , "Feature Selection and Classification of Breast Cancer on Dynamic Magnetic Resonance Imaging by Using Artificial Neural Networks", Proceedings of the 17th Iranian Conference of Biomedical Engineering (ICBME2010), 3-4 November 2010.

[18] A. Lambrou, H. Papadopoulos, A. Gammerman, "Evolutionary Conformal Prediction for Breast Cancer diagnosis", Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine, ITAB 2009, Larnaca, Cyprus, 5-7 November 2009.

[19] Liu Ya-Qin, Wang Cheng, Zhang Lu, "Decision tree based predictive models for breast cancer survivability on imbalanced data ", IEEE 2009.

[20] Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, Alok Choudhary, "A Lung Cancer Mortality Risk Calculator Based on SEER Data", IEEE 2011.

[21] D. Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: comparison of three data mining methods," Artificial Intelligence in Medicine, vol. 34, pp. 113-127, 2005

[22] A.Bellachia and E.Guvan,"Predicting breast cancer survivability using data mining techniques", Scientific Data Mining Workshop, in conjunction with the 2006 SIAM Conference on Data Mining, 2006

[23] Umesh D R and B Ramachandra, "Association Rule Mining Based Predicting Breast Cancer recurrence on SEER Breast Cancer Data" IEEE 2015

[24] Malewicz G, Austern MH, Bik AJ, Dehnert JC, Horn I, Leiser N, Czajkowski G (2010) Pregel: a system for large-scale graph processing. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. ACM, New York, NY, USA. pp 135–146

[25] Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Series B 39:1-38.

[26] Erika Laranjeira & Filipe Grilo, The impact of innovation on Healthcare costs: A multiple imputation

approach. 2nd Portuguese Stata User Group Meeting Olisipo.

7. BIOGRAPHY

Umesh D R completed his Engineering from PES College of Engineering Mandya, Masters from NIE Mysore, presently pursuing Ph.D. from University of Mysore, Mysore. Working in PES College of Engineering Mandya from 2005.

Dr.B.Ramachandra working as Professor and Head in Department of Electrical & Electronics, PES College of Engineering Mandya. He had his Ph.D. From Indian Institute of Science, Bangalore, Master's from Indian Institute of Technology, Bombay.