Performance Evaluation of Entropy and Gini using Threaded and Non Threaded ID3 on Anaemia Dataset

Conference Paper · April 2015

DOI: 10.1109/CSNT.2015.112

3 authors:

Ravi Kishore
Indian Institute of Technology Madras

Karu Prasada Rao
GITAM University

G.R.S. Murthy
Lendi Institute of Engineering and Technology

2015 Fifth International Conference on Communication Systems and Network Technologies

Performance Evaluation of Entorpy and Gini using Threaded and Non Threaded ID3 on Anaemia Dataset

Ch.Ravi Kishore[1],K.Prasada Rao[2], Dr.G.R.S.Murthy[3]
Department of Information Technology[1, 3], Department of Computer Science and Engineering[2]
Aditya Institute of Technology and Management,
Tekkali, Srikakulam, Dist, A.P., India
cauchy9@gmail.com, prasadrao.karu@gmail.com, murthy.grs@gmail.com

***Abstract*— Classification is an important data mining task, and decision trees have emerged as a popular classifier due to their simplicity and relatively low computational complexity. Time required to build a decision tree becomes intractable, as datasets get extremely large. To overcome this problem we proposed a parallel mode of ID3 algorithm. Decision tree building is well-suited for thread-level parallelism as it requires a large number of independent computations. In this paper, we present the analysis and parallel implementation of the ID3 algorithm using Entropy and Gini as heuristics, along with experimental results conducted on the anaemic patient's data set.**

Keywords- Decision tree; Entropy; Gini; ID3;Parallel data mining;

I. INTRODUCTION

Classification is one of the most important data mining tasks. Decision trees are emerged as the first and most fundamental classification technique and still continue to be a subject of research today.

Decision trees are popular because, they are simple in concept, construction is computationally feasible and they are easily interpretable by the user. When we want to create a decision tree from a given data set, we may have problem of choosing which heuristic measure is suitable for impurity/purity checking, and also we have to decide which measure creates decision tree in less time.

Today the number of cores on chip is increasing rapidly, so we can take advantage of thread level parallelism to construct decision tree quickly. According to Han Xiao [1] researchers are realizing that parallel processing is a novel technique for scaling up the classification algorithms. Although there are various reasons for performing data mining algorithms in a distributed manner, the most immediate and practical motivation is that developing learning algorithms that are able to take advantage of the increasing availability of multi-processor and grid computing technology.

Srivastava *et.al*. [2] stated that reasonable accuracy in reasonable time can be achieved by using parallel algorithms. Many researchers felt that parallelism may be a solution to reduce the amount of time spent in building decision trees using datasets while keeping high classification accuracy levels [3-6].

Anemia is a condition in which the number of red blood cells in human body is reduced. The red blood cells are elements that plays vital role, as they carry oxygen from the lungs to all other tissues in the body. Some of the most common symptoms of anemia are weakness, fatigue, poor concentration, pale skin, mild depression, and increased risk of infection. This disease occurs because of lack of three essential substances iron, vitamin B12, and folic acid.

Bentley and Griffiths in their research [7] hypothesized that rural women would have a higher prevalence of anaemia compared with urban women, particularly among the lower income groups.

In North Costal Region of Andhra Pradesh many people suffers with anemia due to poverty and lack of awareness. Medical Authorities feel that technology should help them in diagnosis of anaemia. With this motivation we proposed a thread level ID3 approach for constructing a decision tree based classification model for anaemia data set obtained from patient records of north Andhra region.

Proposed model helps in analyzing two types of anaemia such as Iron deficiency anaemia (ID) and B12 deficiency anaemia (B12).We used heuristics such as Entropy and Gini, both threaded and non-threaded, and analyzed the performance of the proposed algorithm.

The rest of the paper proceeds as follows. Section II discusses the Literature Survey. Section III describes the proposed method. Section IV highlights results and section V is the conclusion.

II. LITERATURE SURVEY

In the technical report on Parallel algorithms in Data mining Mahesh et.al [8] stressed the need for developing effective parallel algorithms for various data mining techniques such as decision trees. Ruoming Jin and Gagan Agrawal in their work [9] presented and evaluated a new approach for decision tree construction, with a particular focus on parallel efficiency and they offered high-level interfaces for parallel data mining. Decision Tree model has been applied in very diverse areas like medicine, data analytics, retail marketing, fraud detection and security.

Kissia and Ramdanib in their work [10] utilized decision tree algorithm to select the most important variables in QSAR modelling and then these variables were used as inputs of ANFIS to predict the anti-HIV activity. Guh et.al [11] in their work presented a hybrid intelligence method which integrates genetic algorithm and decision learning techniques for knowledge mining of an IVF medical database.

978-1-4799-1797-6/15 $31.00 © 2015 IEEE
DOI 10.1109/CSNT.2015.112

Their proposed model assist the IVF physician in predicting the IVF outcome and also find useful knowledge, which helps the IVF physician to modify the IVF treatment to the individual patient with the aim of improving the pregnancy success rate. The proposed method identified twenty-eight most significant attributes for determining the pregnancy rate (e.g., patient's age, number of embryo transferred, number of frozen embryos, and culture days of embryo) and their combinative relationships (represented by if–then rules).

Ruben [12] suggested that data mining in healthcare is an emerging field of high importance for providing diagnosis and a deeper understanding of medical data. Data mining applications in healthcare includes, prevention and early detection of diseases, prevention of hospital errors and preventable hospital deaths, analysis of health care centres for purchasing better health care polices and also used in finding false insurance claims.

Various Researchers are using data mining techniques in diagnosis of several diseases [13-21], which shows the usefulness of these techniques in medical field.

III. Proposed Work

The data we are considering in this case is related to records of anaemia patients we have collected in North Costal Andhra region. We have collected Complete Blood Count (CBC) reports of the patients. Even though there are many parameters in those reports under the guidance of doctors who are domain experts, we have considered only six of those parameters as attributes to our dataset. The attributes we are considering here are, Age, Gender, Hemoglobin(Hb), Mean Corpuscular volume (MCV), Mean Corpuscular Hemoglobin(MCH), Hematocrit (HCT).

TABLE I. Sample records of Anemia data

Age	Sex	Hb	MCV	MCH	HCT	class
child	F	severe	micro cyitic mcv	micro cytic mch	low	ID
old	M	normal	micro cyitic mcv	macro mocytic mch	medium	ID
adult	M	moderate	macro cyitic mcv	normo cytic mch	low	ID
old	M	normal	macro cyitic mcv	macro cyitic mch	medium	B12
child	M	normal	macro cyitic mcv	macro cyitic mch	low	B12
adult	F	moderate	macro cyiticmcv	macro cyitic mch	low	B12

We have discretized all these continuous attributes into categorical values by considering permissible values on the suggestion of domain experts (doctors) for each attribute. Table II shows our selected parameters and permissible ranges for each value as suggested by doctors.

Using these categorical attribute values we have categorized anaemia in to two classes

a) Iron Deficiency Anemia(ID)

b) B12 Deficiency Anemia(B12).

TABLE II. Discritization of parameters

Attributes	Attribute ranges	Categorical values
Age	0-12 12-40 >40	Child Adult Old
Gender	Male Female	M F
Hb	<6.5 6.5-10 10-12 >12	Life threatening Severe Moderate Hb normal
MCV	<97 97-100 >100	MicrocyticMCV NormocyticMCV MacrocyticMCV
MCH	<27 27-33 >33	MicrocyticMCH NormocyticMCH MacrocyticMCH
HCT	<33 33-50 >50	Low Medium High

A. *Original ID3 Decision Tree Algorithm*

Decision trees classify examples according to the values of their attributes. They are constructed by recursively partitioning training examples based on the remaining attribute that has the highest information gain. Attributes become nodes in the constructed tree and their possible values determine the paths of the tree. The process of partitioning the data continues until the data is divided into subsets that contain a single class, or until some stopping condition is met (this corresponds to a leaf in the tree). The two heuristics to measure the purity or impurity of the training data are Entropy and Gini. The gain can be computed using either Entropy or Gini.

$$Gini(t) = 1 - \sum_j [p(j \mid t)]^2 \quad (1)$$

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t) \quad (2)$$

$$Gain_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right) \quad (3)$$

where p(j | t) is the relative frequency of class j at node t Parent Node, p is split into k partitions; n_i is number of records in partition i, n is the total number of records in the data set.

B. *Non Threaded ID3(Examples,Attributes,TargetAttribute)*

Create Root Node for the tree
if all members of Examples are belongs to the same class C
then Root Node = single-node tree with label = C
else if Attributes is empty
then Root Node = single-node tree with
label = most common value of Target_attribute in Examples;
else
A = element in Attributes that maximizes Information Gain(Examples, A)

```
A is decision attribute for RootNode
for each possible value v of A
     add a Branch below RootNode, testing for A = v
     Examples_v = subset of Examples with A = v
     if Examples_v is empty
       then below Branch add Leaf with label = most
       common value of Target_attribute in Examples;
 else
       below Branch add Sub tree
        Non Threaded ID3(Examples_v,Attributes - {A},
                              TargetAttribute);
return RootNode;
```

C. Our contribution

A single processor using all the training set starts the construction phase at each node. The attribute test divides the data into independent partitions where we can create a new thread and assign the data partition to that thread. During the evaluation of the possible splits each thread is responsible only for the evaluation of its attributes. In the ***for*** loop of original algorithm we implemented thread level parallelism as shown in the Pseudocode below:

Threaded ID3(Examples, Attributes, TargetAttribute)

```
Create Root Node for the tree
if all members of Examples belongs to the same class C
       then Root Node = single-node tree with label = C
else if Attributes is empty
       then Root Node = single-node tree with
        label = most  common value of Target_attribute in
                  Examples;
else
A = element in Attributes that maximizes Information
       Gain(Examples, A)

A is decision attribute for RootNode

for each possible value v of A
     create a thread and assign
       Examples_v = subset of  Examples with A = v
if Examples_v is empty
    then below Branch add Leaf with label = most
    common  value of   Target_attribute in Examples;
else
     below Branch add Sub tree
    ThreadedID3(Examples_v, Attributes - {A},
                    TargetAttribute);
return RootNode;
```

IV. RESULTS AND ANALYSIS

The total number of records collected in our Anemia database is 480. We applied our threaded parallelism ID3 method on this database, with different heuristics such as

a) Comparison between Non threaded and threaded entropy
b) Comparison between Non threaded and threaded gini
c) Comparison between Non threaded entropy and non threaded gini
d) Comparison between threaded entropy and threaded Gini.

We conducted experiments by incrementing the size of records by 40. The decision tree model for the Anemia data set is as show below.

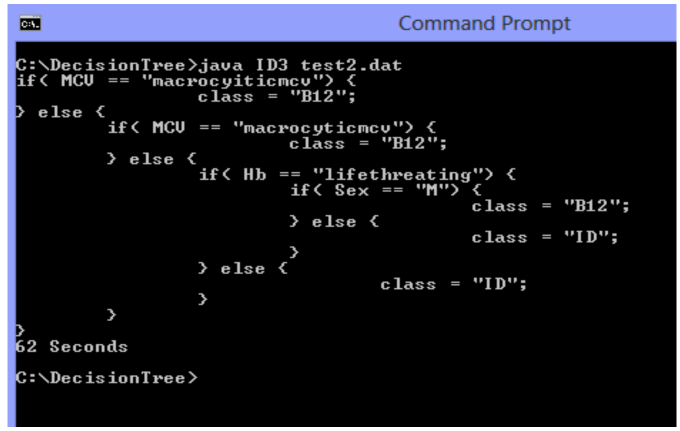

Figure 1. Decision Tree model for Anemia data set

The comparative study of heuristic measures of entropy and Gini for both threaded and non threaded are given in the tables as shown below.

TABLE III. NON THREADED & THREADED ENTROPY

Time in Seconds		
Dataset Size	Non Threaded Entropy	Threaded Entropy
40	20	16
80	47	44
120	78	62
160	80	78
200	105	93
240	116	109
280	133	125
320	158	146
360	162	156
400	188	166
440	192	176
480	206	189

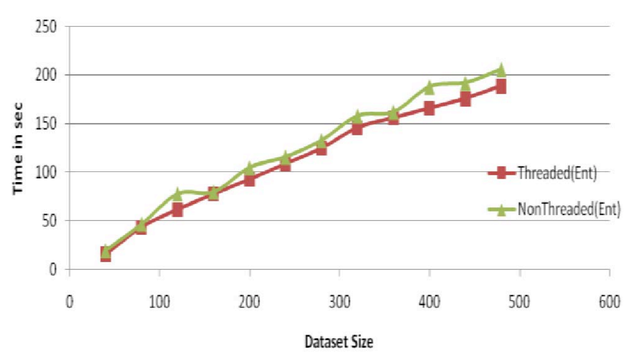

Figure 2. Performance analysis between Non Threaded and Threaded Entropy

TABLE IV. NON THREADED & THREADED GINI

Time in Seconds		
Dataset Size	Non Threaded Gini	Threaded Gini
40	31	15
80	62	47
120	76	62
160	93	78
200	109	94
240	125	116
280	166	125
320	172	156
360	199	188
400	205	197
440	216	209
480	238	218

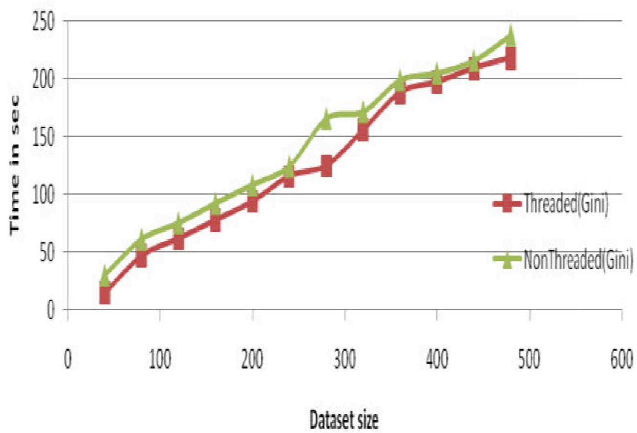

Figure 3. Performance analysis between NonThreaded and Threaded Gini

TABLE V. NON THREADED ENTROPY & NON THREADED GINI

Time in Seconds		
Dataset Size	Non Threaded Entropy	Non Threaded Gini
40	20	31
80	47	62
120	78	76
160	80	93
200	105	109
240	116	125
280	133	166
320	158	172
360	162	199
400	188	205
440	192	216
480	206	238

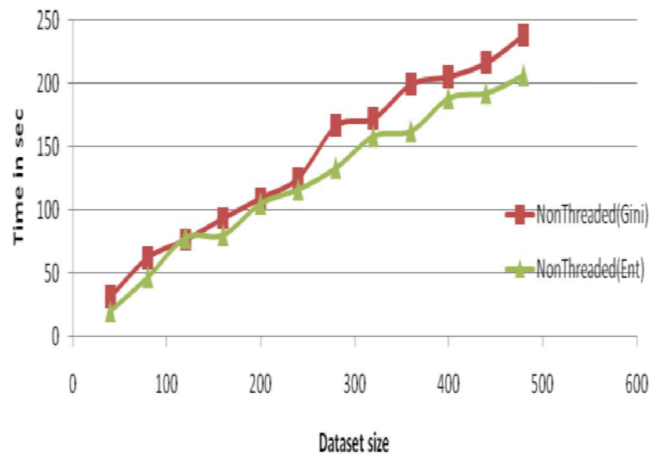

Figure 4. Performance analysis between NonThreaded and NonThreaded Gini

TABLE VI. THREADED ENTROPY & THREADED GINI

Time in Seconds		
Dataset Size	Threaded Entropy	Threaded Gini
40	16	15
80	44	47
120	62	62
160	78	78
200	93	94
240	109	116
280	125	125
320	146	156
360	156	188
400	166	197
440	176	209
480	189	218

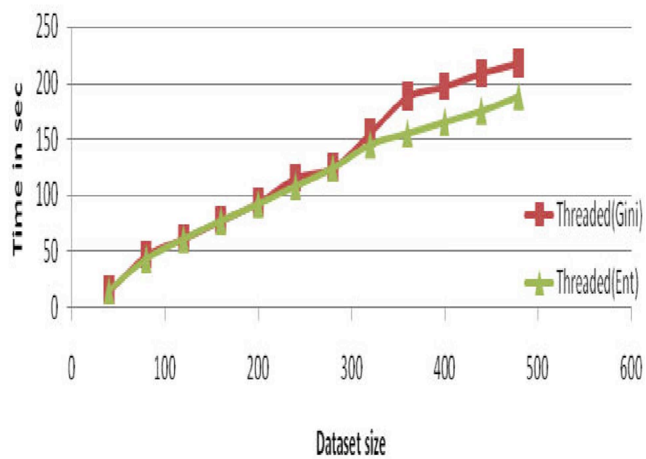

Figure 5. Performance analysis between Threaded Entropy and Gini

From fig 2 and fig. 5 we can found that even with small dataset (480 records) threaded ID3 takes less time to create the decision tree with both Entropy and Gini. From Fig.3 and

Fig 4 we can found that with the same data set Entropy heuristic measure performs well for both non threaded and threaded ID3. Up to the record size of 280 both performed similarly where as from 320 records onwards Entropy heuristic measure performed well.

V. CONCLUSIONS

We have developed a decision tree for anemia data base. With the given moderate dataset for decision tree construction, threaded ID3 with Entropy heuristic is proved as better option when compared with Gini as a heuristic. Even though we reduced communication between threads, parallelism still suffers from load imbalance.

REFERENCES

[1] Han Xiao, Towards Parallel and Distributed Computing in Large-Scale Data Mining: A Survey. Technical Report, Technical University of Munich,TR 01-001, Feb 2010, pp.1-30. Can be downloaded from http://home.in.tum.de/~xiaoh/

[2] Srivastava, A.; Eui-Hong Han; Singh, V.; Kumar, V., "Parallel formulations of decision-tree classification algorithms," Parallel Processing, 1998. Proceedings. 1998 International Conference on , vol., no., pp.237,244, 10-14 Aug 1998. doi: 10.1109/ICPP.1998.708491

[3] J. Darlington, Y. Guo, J. Sutiwaraphun, H.W. To, Parallel induction algorithms for data mining, in: Advances in Intelligent Data Analysis: Reasoning About Data IDA '97, vol. 1280 of LNCS, 1997, pp. 437-445.

[4] A. A. Freitas and S. H. Lavington. Mining Very Large Databases with Parallel Processing. Kluwer Academic Publishers, 1998.

[5] J. C. Shafer, R. Agrawal, and M. Mehta. SPRINT: A scalable parallel classifier for data mining. In Proc. 22nd Int. Conf. Very Large Databases, pages 544–555. Morgan Kaufmann, 3–6 1996.

[6] A. Srivastava, E. Han, V. Kumar, and V. Singh. Parallel formulations of decision-tree classification algorithms. Data Mining and Knowledge Discovery, 3, 1999

[7] M E Bentley and P L Griffiths, The burden of anemia among women in India, European Journal of Clinical Nutrition, vol 57, 2003 , pp 52–60. doi:10.1038/sj.ejcn.1601504

[8] Technical Report: Parallel Algorithms in Data Mining by Mahesh Joshi, Euihong (sam) Han, George Karypis, and Vipin Kumar. http://www.cs.umn.edu/tech_reports_upload/tr2001/01-001.pdf

[9] Ruoming Jin and Gagan Agrawal, "Communcation and Memory Efficient Parallel Decision Tree Construction", in the Third SIAM International Conference on Data Mining (SDM), 2003.

[10] Mohamed Kissia, Mohammed Ramdanib, "A hybrid decision trees – adaptive neuro fuzzy inference system in prediction of anti-HIV molecules", Expert Systems with Applications Volume 38, Issue 5, May 2011, Pages 6376–6380.

[11] Ruey-Shiang Guh, Tsung-Chieh Jackson Wu, Shao-Ping Weng,"Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes"Expert Syst. Appl 01/2011; 38:pp 4437-4449. DOI: 10.1016/j.eswa.2010.09.112.

[12] Ruben, D. C. J. (2009). "Data Mining in Healthcare:Current Applications and Issues." MSIT Thesis submitted to CMU, Australia.

[13] Panzarasa S, Quaglini S, Sacchi L, Cavallini A, Micieli G, Stefanelli M. "Data mining techniques for analyzing stroke care processes." Studies in health technology and informatics, 2010;160(Pt 2),pp 939-43.

[14] T. Porter, B. Green, "Identifying Diabetic Patients: A Data Mining Approach, Americas Conference on Information Systems," Proceedings of the Fifteenth Americas Conference on Information Systems, San Francisco, 2009.

[15] Li L, T. H., Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA (2004). "Data mining techniques for cancer detection using serum proteomic profiling." Artificial Intelligence in Medicine,. Artificial Intelligence in Medicine, Elsevier Volume 32, Issue 2, October 2004, Pages 71–83

[16] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles." Expert Systems with Applications, Elsevier vol. 36-4, May, 2009 Pages 7675-7680

[17] Andreeva, P.: Data Modelling and Specific Rule Generation via Data Mining Techniques. International Conference on Computer Systems and Technologies - CompSysTech' IIIA.,17, pp.1-62006.

[18] Hongmei Yan ; Bioeng. Inst., Chongqing Univ., China ; Jun Zheng ; Yingtao Jiang ; Chenglin Peng . "Development of a decision support system for heart disease diagnosis using multilayer perceptron." Proceedings, Circuits and Systems, 2003. ISCAS '03. Proceedings of the 2003 International Symposium on Volume:5, pp.709-712,

[19] Sitar-Taut, V. A., D. Zdrenghea, Pop. D, D.A. Sitar Taut . "Using machine learning algorithms in cardiovascular disease risk evaluation." Journal of Applied Computer Science & Mathematics 5-3, pp 29-32, 2009.

[20] Srinivas, K., B. K. Rani, Dr. A.Govardhan . "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks." International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 02: pp 250- 255,2010.

[21] Rajkumar, A. and G. S. Reena . "Diagnosis Of Heart Disease Using Datamining Algorithm." Global Journal of Computer Science and Technology Vol. 10 2010.