Undergraduate Student Retention Using Wavelet Decomposition

Ji-Wu Jia, *Member IAENG*, Manohar Mareboyana, *Member IAENG*

***Abstract*---In this paper, we have presented some results on undergraduate student retention using signal processing techniques for classification of the student data. The experiments revealed that the main factor that influences student retention in the Historically Black Colleges and Universities (HBCU) is the cumulative grade point average (GPA). The linear smoothing of the data helped remove the noise spikes in data thereby improving the retention results. The data is decomposed into Haar coefficients that helped accurate classification. The results showed that the HBCU undergraduate student retention corresponds to an average GPA of 2.8597 and the difference of -0.023307. Using this approach, we obtained more accurate retention results on training data.**

***Index Terms*---Haar Transform, Linear Smoothing, Machine Learning, Signal Processing, Student Retention**

I. INTRODUCTION

In this paper we study the HBCU undergraduate student retention using signal processing techniques to obtain the HBCU undergraduate student retention criterion for average GPA [20]-[23].

We started collecting data from the HBCU Fall 2006 full-time and first-time undergraduate students. We tracked these students' records in the following six years from Fall 2006 to Fall 2011. The data was queried from the Campus Solution database. The six-year training data set size is 771 instances with two attributes shown in Table I. The HBCU undergraduate six years retention rate 44.9% was derived from the six-year training data set [5]. The HBCU six-year training data set numeric attributes and statistics are shown in Table II [21].

We classified the data under two groups – "Retention" – students who were retained in the HBCU and "No Retention" – students who were not retained in the HBCU [1]-[19].

TABLE I
LIST OF DATA SET ATTRIBUTES

Number	Name	Description	Type
1	GPA	The last cumulative GPA while student enrolled	Number
2	Retention	If student graduated or enrolled in Fall 2011 then yes, else no	Text

Manuscript received February 25, 2014; revised March 22, 2014.

Ji-Wu Jia is a senior PeopleSoft developer, Bowie State University, 14000 Jericho Park Road, Bowie, Maryland 20715, USA. jjia@bowiestate.edu

Manohar Mareboyana is Professor in Department of Computer Science, Bowie State University, 14000 Jericho Park Road, Bowie, Maryland 20715, USA. MMareboyana@bowiestate.edu

TABLE II
TRAINING DATA SET NUMERIC ATTRIBUTES

Naive Bayes	No Retention		Retention	
Attribute Name	Mean	Std. Dev.	Mean	Std. Dev.
GPA	1.9371	±0.8913	2.8864	±0.4276

The most basic wavelet transform is the Haar transform described by Alfred Haar in 1910. It serves as the prototypical wavelet transform. We will describe the (discrete) Haar transform, as it encapsulates the basic concepts of wavelet transforms used today. We will also see its limitation, which the newer wavelet transform (in 1998 by Ingrid Daubechies) resolves [23].

The algorithm to calculate the Haar transform of an array of *n* (number of years) samples is below [23]:

1. Treat the array as n/2 pairs called (a, b)
2. Calculate (a + b) / sqrt(2) for each pair, these values will be the first half of the output array.
3. Calculate (a - b) / sqrt(2) for each pair, these values will be the second half.
4. Repeat the process on the first half of the array. (the array length should be a power of two)

First, we pre-processed the student data, added missing data, and grouped the student data into two files. They are retention and no-retention files.

Second, we applied linear smoothing to the discrete GPA signals for removing noise.

Finally, we applied Haar transform to the GPA data, and calculated the average and difference from the retention data, and discussed the average GPA for the HBCU undergraduate student retention. The framework of the study is shown as Figure 1.

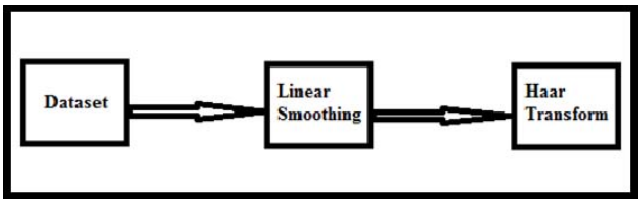

Fig. 1. The framework for the study

In the following sections, we describe the methodology and algorithms.

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. METHODOLOGY

A. Student Data

- Retention Student Data

The retention student data are shown in Figure 2 as star. The data are concentrative distributed.

```
clf
hold
b=[1 2 3 4 5 6]
r=load('F:\Octa\used\retention.txt')
plot(b,r,"*")
```

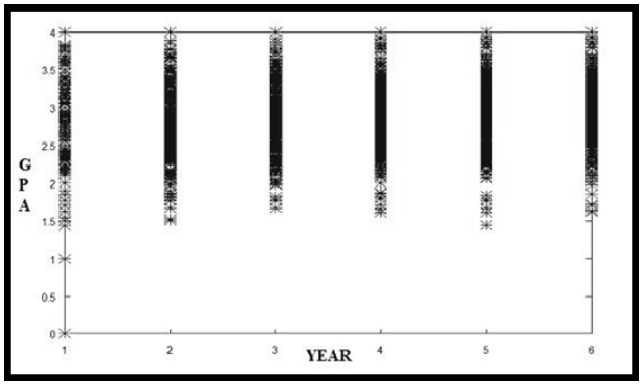

Fig. 2. Retention student data

- No-retention Student Data

No-retention student data are shown in Figure 3 as star. The data are not concentrative distributed.

```
clf
hold
b=[1 2 3 4 5 6]
nr=load('F:\Octa\used\no-retention.txt')
plot(b,nr,"*")
```

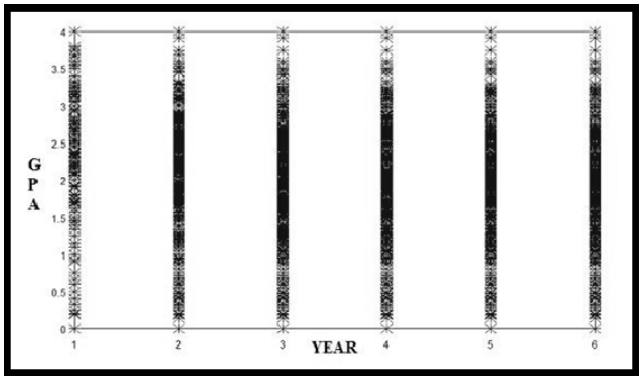

Fig. 3. No-retention student data

B. Linear Smoothing Retention GPA Data

We applied linear smoothing to the retention GPA data. The linear smoothing is applied as shown below.

$$y_1 = \frac{x_1 + x_2 + x_3}{3}$$

$$y_2 = \frac{x_2 + x_3 + x_4}{3} \qquad (1)$$

.....

Where $x_1, x_2, x_3 \ldots$ are the original GPA data, and $y_1, y_2, y_3 \ldots$ are the new GPA data.

The new retention student data are shown in Figure 4 as star. Compared the smoothing retention data to the original data (Figure 2), we can see the lowest and highest values are removed, and six years data have been smoothed into three periods (four data points).

```
clf
hold
a=[1 2 3 4}
r3=r(:,1:4)
[m n]=size(r)
for i=1:m
r3(i, 1)=(r(i, 1)+r(i, 2)+r(i, 3))/3;
r3(i, 2)=(r(i, 2)+r(i, 3)+r(i, 4))/3;
r3(i, 3)=(r(i, 3)+r(i, 4)+r(i, 5))/3;
r3(i, 4)=(r(i, 4)+r(i, 5)+r(i, 6))/3;
end
for k=1 m
plot(a, r3(k, :),"*");
end
```

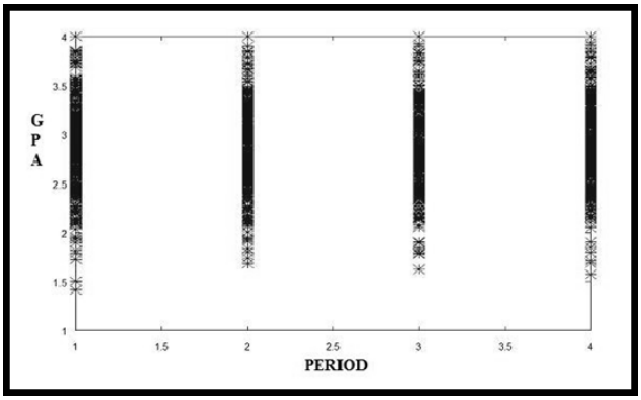

Fig. 4. Linear smoothing retention discrete GPA data

C. Linear Smoothing No-retention GPA Data

We have used the same algorithms for no-retention data for smoothing and compared the results with the original data (Figure 3). We can see the lowest and highest values are removed, and six years data have been smoothed into three periods (four data points).

The no-retention data dimension is bigger than retention data, and the no-retention linear smoothing graph is shown in Figure 5 as star.

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

```
clf
hold
a=[1 2 3 4}
[m n]=size(nr)
s=r(:,1:4)
for i=1:m
s(i, 1)=(nr(i, 1)+nr(i, 2)+nr(i, 3))/3;
s(i, 2)=(nr(i, 2)+nr(i, 3)+nr(i, 4))/3;
s(i, 3)=(nr(i, 3)+nr(i, 4)+nr(i, 5))/3;
s(i, 4)=(nr(i, 4)+nr(i, 5)+nr(i, 6))/3;
end
for k=1:m
plot(a, s(k, :),"*");
end
```

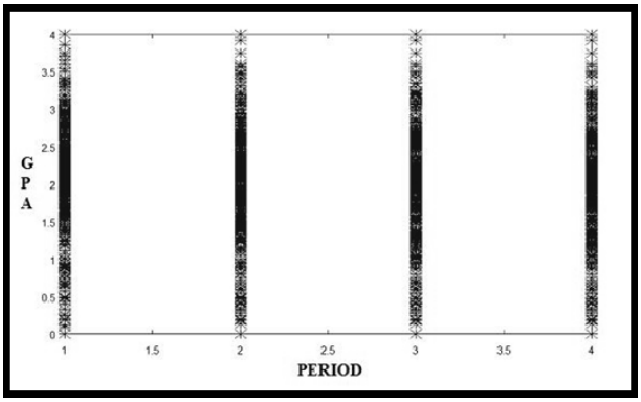

Fig. 5. Linear smoothing no-retention data

D. Wavelet Transform

After the retention data have been smoothed by linear filters, we used Haar transform processing. Haar transform's decompositions can be written as below [23].

$$c(n) = 0.5 * y(2n) + 0.5 * y(2n+1)$$
$$d(n) = 0.5 * y(2n) - 0.5 * y(2n+1) \quad (2)$$

Where c(n) is average of the pairs of data, and d(n) is their differences.

1. The First Level Decomposition

We have used the Haar algorithm as given in equation 2 to smooth the data. The first level decomposition of the average GPA is shown as star in Figure 6.

We have used the Haar algorithm as given in equation 2 to smooth the data. The first level decomposition of the difference GPA is shown as rhombus in Figure 7.

```
hc1= r3(:,1:2)
[m n]=size(r)
for i=1:m
hc1(i,1)=0.5*r3(i,1)+0.5*r3(i,2);
hc1(i,2)=0.5*r3(i,3)+0.5*r3(i,4);
end
clf
hold
p2=[1 2]
for i=1:m
plot(p2,hc1(i,:),"*");
end
```

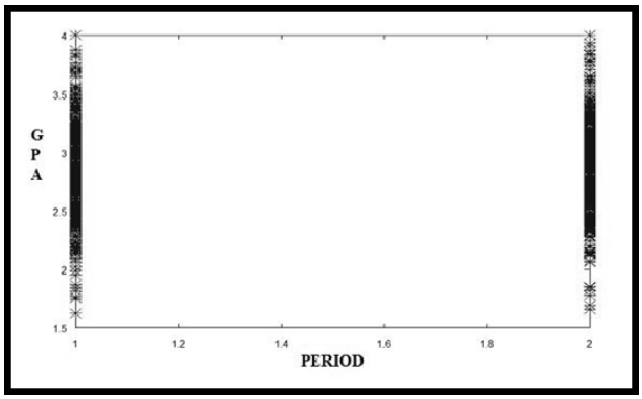

Fig. 6. Haar average 1

```
hd1= r3(:,1:2)
for i=1:m
hd1(i,1)=0.5*r3(i,1)-0.5*r3(i,2);
hd1(i,2)=0.5*r3(i,3)-0.5*r3(i,4);
end
clf
hold
p2=[1 2]
for i=1:m
plot(p2,hd1(i,:),"d");
end
```

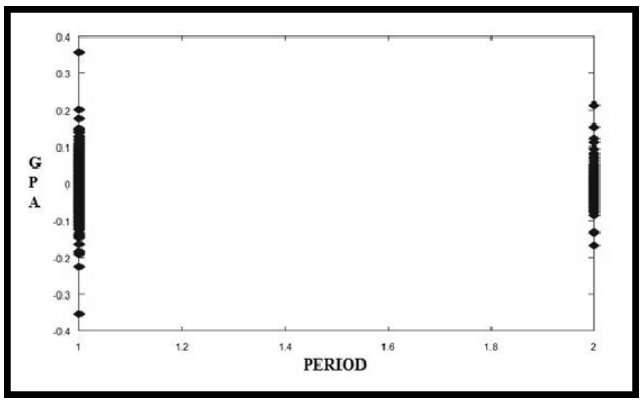

Fig. 7. Haar difference 1

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

2. The Second Level Decomposition

We again have applied the Haar transform to the first level decomposition of retention data, and the second level decomposition retention average, and difference. The results are shown below.

The average points are shown as star on the top, and the difference points are shown as rhombus on the bottom in Figure 8.

```
hc2=r3(:,1:1)
hd2= r3(:,1:1)
for i=1:m
hc2(i)=0.5*hc1(i,1)+0.5*hc1(i,2);
hd2(i)=0.5*hc1(i,1)-0.5*hc1(i,2);
end
clf
hold
for i=1:m
plot(1,hc2(i),"*");
plot(1,hd2(i),"d");
end
```

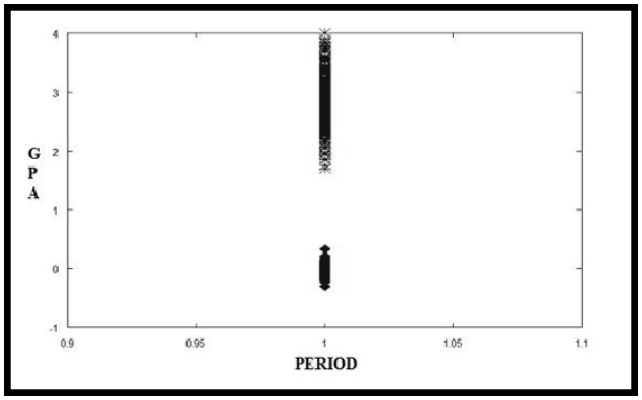

Fig. 8. Haar average and difference 2

3. The Retention Data Representation

The retention data Haar processes are shown in Figure 9.

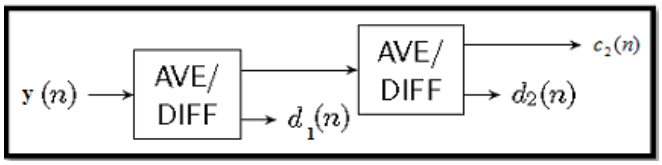

Fig. 9. The Haar retention processes

Where $d_1(n)$ is the first level decomposition GPA difference, $d_2(n)$ is the second level decomposition GPA difference, and $c_2(n)$ is the second level decomposition GPA average, and the retention representation is shown in Figure 10. The average points are shown as star on the top, and the difference points are shown as rhombus on the bottom [23].

```
clf
hold
for i=1:m
plot(p2,hd1(i,:),"d");
end
for i=1:m
plot(3,hc2(i),"*");
plot(3,hd2(i),"d");
end
```

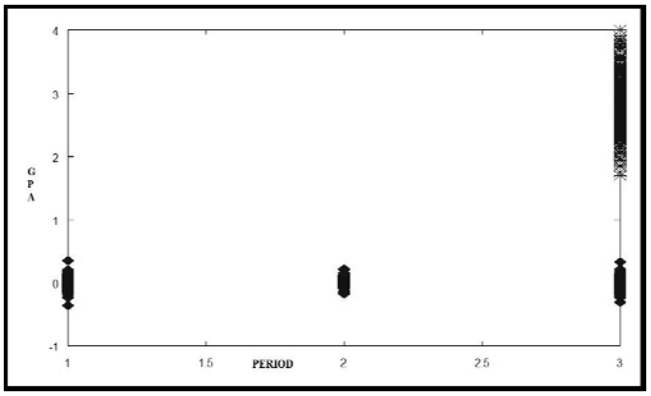

Fig. 10. The Haar retention representation

III. RESULTS

By using the results from the second level Haar transform over the entire student population, we computed the average GPA and the difference for retention students.

In Figure 11, the average point is shown as star on the top, and the difference point is shown as rhombus on the bottom.

```
tgpa=0
tdiff=0
for i=1:m
tgpa=tgpa+hc2(i);
tdiff=tdiff+hd2(i);
end
atgpa=0
atdiff=0
clf
hold
atgpa=tgpa/m
atdiff=tdiff/m
plot(1,atgpa,"*")
plot(1,atdiff,"d")
atgpa
atdiff
```

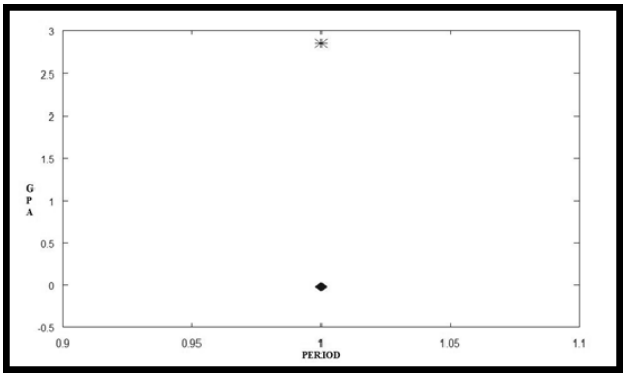

Fig. 11. The average GPA and difference

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

We tested the Haar average retention GPA using test data set (the HBCU Fall 2007 to Fall 2012 student test data set with 820 instances) and compared the results to Naïve Bayes mean value. The results are given in Table III. Haar based classification is better than Naïve Bayes.

TABLE III
TESTED RESULTS

	Average GPA	Retention Accuracy (%)
Naïve Bayes	2.8864	74.8
Haar	2.8597	75.6

IV. CONCLUSIONS

Smoothing the data removed the highest and lowest GPA values for both of retention and no-retention data. The algorithm filtered out the noise and made the data more pure.

From the Haar transform's results, we can say that the average GPA for the HBCU undergraduate student retention should be 2.8597, and the average difference should be -0.023307.

REFERENCES

[1] M. H. Dunham, "Data mining introductory and advanced," ISBN 0-13-088892-3, Prentice Hall, 2003.
[2] T. Mitchell, "Machine learning," ISBN 0070428077, McGraw Hill, 1997.
[3] S. Russell and P. Norvig, "Artificial intelligence, a modern approach," Third Edition, Pearson, ISBN-13: 978-0-13-604259-4, ISBN-10: 0-13-604259-7, 2010.
[4] Y. S. Abu-Mostafa, M. Magdon-Lsmail, and H. Lin, "Learning from data," ISBN 10:1-60049-006-9, AMLbook, 2012.
[5] S. L. Hagedorn, "How to define retention: a new look at an old problem," In Alan Seidman (Ed), College student retention: Formula for student Success, Westport, CT: Praeger Publishers, 2005.
[6] R. Alkhasawneh and R. Hobson, "Modeling student retention in science and engineering disciplines using neural networks," IEEE Global Engineering Education Conference (EDUCON), 660-663, 2011.
[7] D. B. Stone, "African-American males in computer science – examining the pipeline for clogs," The School of Engineering and Applied Science of the George Washington University, Thesis, 2008.
[8] E. Frank, "Pruning Decision Trees and Lists," Department of Computer Science, University of Waikato, Thesis, 2000.
[9] C. H. Yu, S. Digangi, A. Jannasch-pennell, and C. Kaprolet, "A data mining approach for identifying predictors of student retention from sophomore to junior year," Journal of Data Science 8, 307-325, 2010.
[10] S. K. Yadav, B. Bharadwaj, and S. Pal, "Mining educational data to predict student's retention: a comparative study," International Journal of Computer Science and Information Security, 10(2), 113-117, 2012.
[11] A. Nandeshwara, T. Menziesb, and A. Nelson, "Learning patterns of university student retention," *Expert Systems with Applications*, *38*(12), 14984–14996, 2011.
[12] S. A. Kumar and M. V. N., "Implication of classification techniques in predicting student's recital**,"** International Journal of Data Mining & Knowledge Management Process (IJDKP), 1(5), 2011.
[13] D. Kabakchieva, "Student performance prediction by using data mining classification algorithms," International Journal of Computer Science and Management Research, 1(4), 686-690, 2012.
[14] S. Lin, "Data mining for student retention management," The Consortium for Computing Sciences in Colleges, 2012.
[15] S. Singh and V. Kumar, "Classification of student's data using data mining techniques for training & placement department in technical education," International Journal of Computer Science and Network (IJCSN) 1(4), 2012.
[16] F. Esposito, D. Malerba and G. Semeraro, "A Comparative Analysis of Methods for Pruning Decision Tree," IEEE, Transaction on Pattern Analysis and Machine Intelligence, Vol 19, No 5, 1997.
[17] D. D. Patil, V. M. Wadhai and J. A. Gokhale, "Evaluation of Decision Tree Pruning Algorithms for Complexity and Classification Accuracy," International Journal of Computer Applications (0975-8887), Vol 11, No 2, 2010.
[18] N. Stanevski and D. Tsvetkov, "Using Support Vector Machines as a Binary Classifier," International Conference on Computer Systems and Technologies – CompSys Tech' 2005.
[19] S. Sembiring, M. Zarlis, D. Hartama, and E. Wani, "Prediction of student academic performance by an application of data mining techniques," International Conference on Management and Artificial Intelligence IPEDR, 6, 2011.
[20] S. D. Stearns and D. R. Hush, "Digital Signal Processing with Examples in MATLAB," Second Edition, CRC Press, 2011.
[21] J. Jia and M. Mareboyana, "Machine Learning Algorithms and Predictive Models for Undergraduate Student Retention," World Congress on Engineering & Computer Science 2013, Volume I, 222-227.
[22] R. X. Gao and R. Yan, "Wavelets: Theory and Applications for Manufacturing," ISBN 978-1-4419-1544-3, Springer, 2011.
[23] I. W. Selesnick, "Wavelet Transforms – A Quick Study," Physics Today magazine, 2007.

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)