Hybrid Approaches Using Decision Tree, Naïve Bayes, Means and Euclidean Distances for Childhood Obesity Prediction

Muhamad Hariz Muhamad Adnan, Wahidah Husain and Nur'Aini Abdul Rashid

School of Computer Sciences, Universiti Sains Malaysia
11800 USM, Penang, Malaysia
mhma.com08@student.usm.my, wahidah@cs.usm.my, nuraini@cs.usm.my

Abstract

Even by using the data mining, many weaknesses still existed in childhood obesity prediction and it is still far from achieving perfect prediction. This paper studies previous steps involved in childhood obesity prediction using different data mining techniques and proposed hybrid approaches to improve the accuracy of the prediction. The steps taken in this study were a review of childhood obesity, data collections, data cleaning and preprocessing, implementation of the hybrid approach, and evaluation of the proposed approach. The hybrid approach consists of the classification and regression tree, Naïve Bayes, mean value identification and Euclidean distances classification. The results from the evaluation have shown that the proposed approach has 60% sensitivity for childhood obesity prediction and 95% sensitivity for childhood overweight prediction.

Keywords: *Hybrid approach, Classification and regression tree, Naïve Bayes, Mean, Euclidean distances, Childhood obesity*

1. Introduction

A child with a Body Mass Index (BMI) over than 95th percentile is categorized as obese while between 85th to 95th percentiles is categorized as overweight [1]. Regarding the childhood obesity prediction, one of the most extensive work was done by Zhang et al. in which six data mining techniques were compared with the logistic regression [2]. The techniques are; decision tree (C4.5), association rules, Neural Network, NB, Bayesian networks, linear SVM and the Radial Basis Function (RBF) SVM. Based on the results for overweight predictions at 3 years old, the linear SVM and RBF SVM have the highest sensitivity with 59.6% and 60% respectively but showed the lowest specificity compared to others. The Bayesian techniques overall accuracy were the highest at 91.9 %. Meanwhile, based on the results of the childhood obesity predictions, all the techniques showed poor sensitivities except for the Bayesian classifiers at 62%. Therefore, this study shows that the Naïve Bayes is a suitable classifier for childhood obesity prediction. Beside from that, the Naïve Bayes has showed good performances in other predictions which are quite similar to the childhood obesity predictions such as the coronary heart diseases, breast cancer, diabetes, and the in-vitro fertilization [3-6]. Based from these, the Naïve Bayes was used as a classifier in presented this paper.

From a comparative study made using 11 data mining techniques for childhood obesity predictions, the classification and regression tree (CART) was proven to be the most consistent [7]. This study also used the CART for classification and variable selections. The paper is organized as follows. In Section 2, the data mining techniques used in this paper are discussed briefly. Section 3 presents the proposed hybrid approaches while Section 4 presents

the materials and methods used. Section 5 presents the results and discussion. The conclusions are summarized in Section 6.

2. Data Mining Techniques

2.1. Classification and Regression Tree

The Classification and Regression Tree (CART) classifies the data by using binary recursive partitioning and build decision trees [8]. The CART has the capability to uncover hidden relationships from the data and is advantageous over traditional regression procedures in which it is not limited by postulations, impervious to outliers, better identification of complex interactions than the logistic regression, and it does not need risk specification [8].

2.2. Naive Bayes

The Naïve Bayes has been identified as a suitable classifier in the medical domains. The advantages of Naïve Bayes include: computational simplicity, easy to understand, extremely fast, and only requires a single pass through the data if the attributes are discrete [9]. The Naïve Bayes is a simple classifier with a postulation of independence among attributes and often give better classification accuracy compared to other classifiers on real life data [10].

2.3. Euclidean Distances

The Euclidean distance is used to calculate the distance or similarity between two points, which in this study the two points are output value from the Naïve Bayes and the mean value. The mean refers to the sum of the values divided by the number of values. The mean calculation was widely used in statistics and by the K-Means clustering technique for many purposes such as in the determination of an algorithm accuracy, finding optimal solutions, clustering and classification [11,12].

3. Proposed Hybrid Approaches

3.1. The First Approach

In this approach, the CART was used to select important variables based on its relative importance. The variables selected by CART are shown in Figure 1.

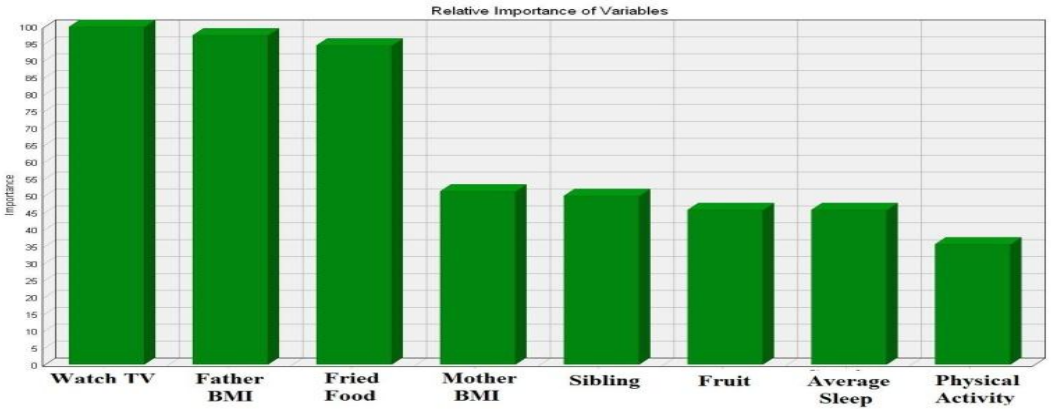

Figure 1. The Variables Selected using CART

The CART has selected eight important variables that are watching-TV, father-BMI, fried-food, mother-BMI, sibling, fruit, average-sleep, and physical-activity. The identified variables will be classified using the Na ïve Bayes.

3.2. The Second Approach

In addition to the first approach, the means value identification and the Euclidean Distance classifications were used. The mean value was used because similar patterns were identified from the Na ïve Bayes output where a more similar output values were identified from the same groups compared to a different group. The architecture of this approach is shown in Figure 2. The outputs of the Na ïve Bayes were clustered into the positive and negative groups. For training, two clusters were created in the positive and the negative groups to keep the outputs from the Na ïve Bayes classification.

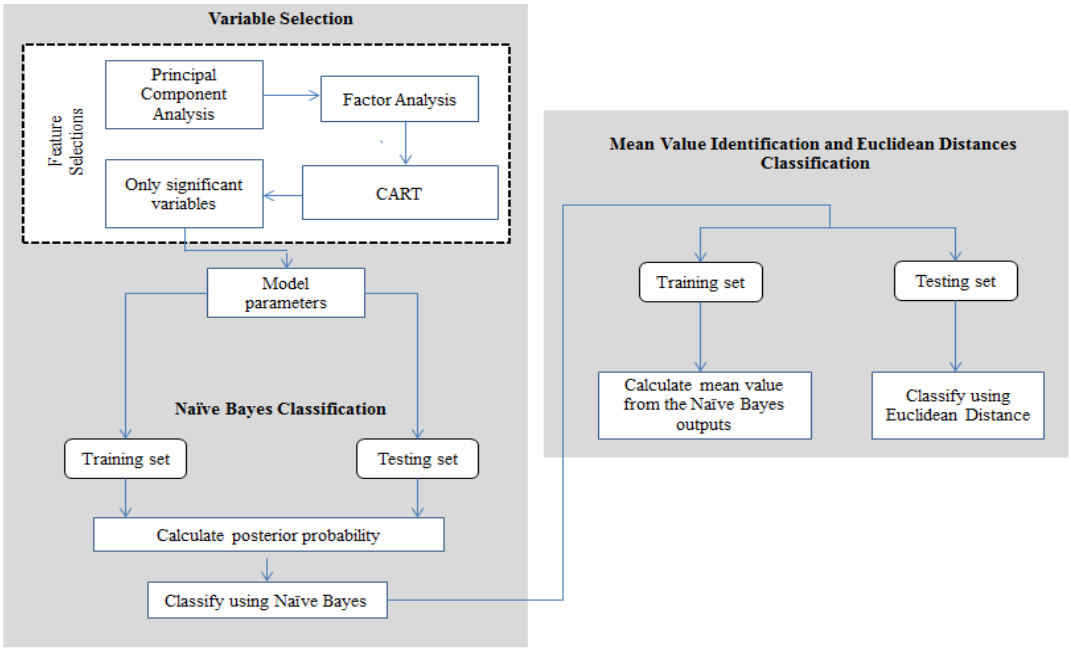

Figure 2. The Second Approach Architecture

The clustering was made based on the child real BMI class (positive or negative). An important step that has to be taken is to filter and remove highly deviated value so that the final mean value would not be noised by these high deviation values. In the testing phase, the sample is classified based on its P (C | X) value closest to the means of either the positive and negative cluster. The Euclidean distance was used to calculate the distance or similarity between the sample output value and the means for both clusters. The flows of the processes starting from the Na ïve Bayes classification are as shown in Figure 3.

4. Experimental

4.1. Materials and Methods

The journal articles, conference proceedings, and online databases are the reliable sources to gain information and knowledge of childhood obesity. Datasets consisted of 320 children

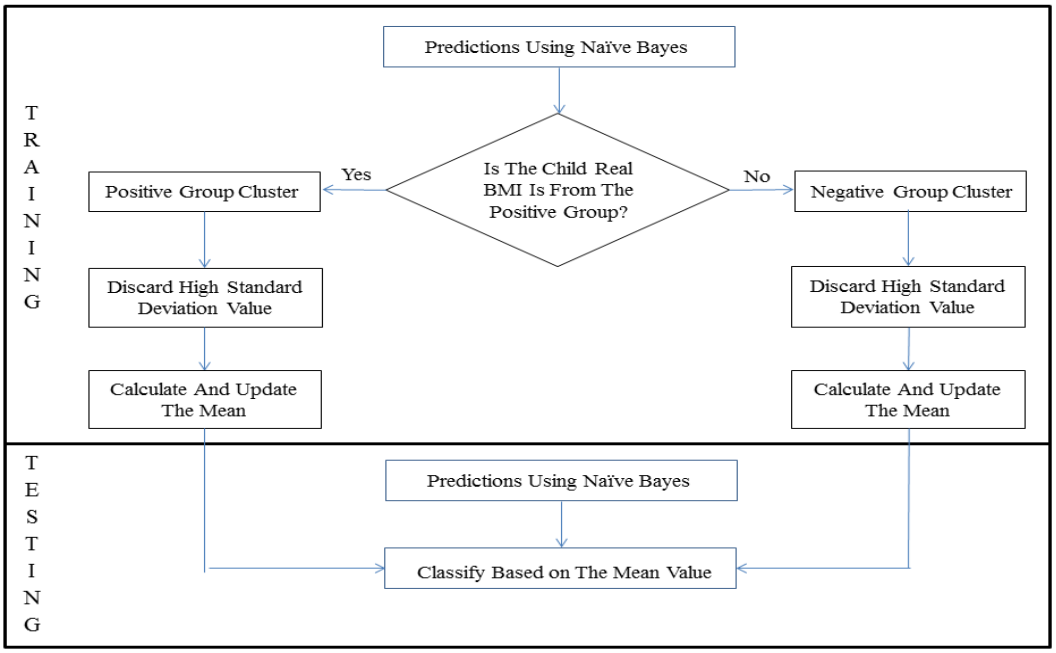

Figure 3. The Flow of Processes Starting from the Naïve Bayes Classification

samples were used for training and testing with a ratio of 61:23:16 of normal, obese, and overweight samples. The SPSS statistical software was used to analyze the data. The methods used were the risk factor study where 39 factors were identified, data collection, data pre-processing, feature selection, hybrid approach implementation and evaluations [13]. The evaluation parameters are the specificity, sensitivity, and overall accuracy. The sensitivity or the true positive rate (TPR) is defined by TP / (TP + FN); while the specificity or the true negative rate (TNR) is defined by TN / (TN + FP); and the accuracy is defined by (TP + TN) / (TP + FP + TN + FN) [2, 4, 5].

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

5. Results and Discussion

5.1. The First Approach

The results of normal, obesity, and overweight predictions using the first approach are shown in Table 1. In comparison to the CART performances, this approach has 2.31% lower sensitivity for normal prediction, 26.67% lower sensitivity for obesity prediction, and 75% higher sensitivity for obesity prediction. This approach cannot be declared as better than the CART because it did improve the sensitivity of overweight predictions greatly compared to the CART that failed to predict any overweight cases but it also shows reduced sensitivity for obesity prediction by 26.67% compared to the CART.

Table 1. The Results of the First Approach

Prediction	Sensitivity (%)	Specificity (%)	Accuracy (%)
Normal	90	0	37.5
Obesity	40	100	75
Overweight	75	100	95.83

The number of overweight and obese cases correctly classified by both techniques was compared. For the CART, the number of obese and overweight cases correctly predicted is 49 cases and zero cases respectively. For this approach, the number of obese cases and overweight cases correctly classified is 30 and 38 respectively with the total of 68. Therefore, this approach was better for both class classifications. In order to clarify that the CART did improve the classification of the Naïve Bayes with the variable selections, the same predictions were made using the Naïve Bayes using 13 independent variables (FA variables). The results are shown in Table 2.

Table 2. The Results of Naïve Bayes Classification using 13 Independent Variables

Prediction	Sensitivity (%)	Specificity (%)	Accuracy (%)
Normal	70	14	37.5
Obesity	0	100	83.3
Overweight	0	100	83.3

Based on the results, it shows that the CART variable selections have greatly improved the performance of the Naive Bayes classifier. The accuracy of the Naïve Bayes was reported to be greatly improved when the input parameters were reduced [5]. If the variables were selected using CART, the sensitivity in normal, obesity, and overweight prediction have been increased by 20%, 40%, and 75% respectively. It shows that a Naïve Bayes classifier with a good variable selection technique produced better results.

5.2. The Second Approach

The results of normal, obesity, and overweight predictions using the second approach are shown in Table 3.

Table 3. The Results of the Second Approach

Prediction	Sensitivity (%)	Specificity (%)	Accuracy (%)
Normal	60	100	83
Obesity	60	78.6	70.8
Overweight	95	50	58.3

Based on the results, the sensitivity for normal prediction was worse than the first approach by 30%. But, the accuracy has been improved by 45.5%. For obesity and overweight prediction, the sensitivity in comparison with the first approach has been increased by 20%

and 15% respectively. The results show that the addition of the means value identification and the Euclidean distance had increased the second approach performances for obesity and overweight prediction. In comparison with the CART, this approach sensitivity is 6.67% worse than the CART for obesity prediction which means it was still comparable with the CART. For overweight predictions, this approach was better at overweight predictions by 95%.

To compare with the CART, the numbers of obese and overweight cases correctly predicted were identified. There were 44 obese cases and 48 overweight cases correctly predicted with the total of 92 cases. Therefore, this approach correctly predicted more obese and overweight cases compared with the CART (49 cases) and the second approach (68 cases). Meanwhile, this approach sensitivity in obese and overweight predictions was better than the predictions by Zhang et al. by 5.3% and 32% respectively. This approach shows a good potential for better performances if the datasets are larger.

5.3. A Comparison using ROC Curve

The overall performance of both approaches can be determined from the area under the ROC curve or known as the AUC which measures the average value of sensitivity for all possible values of specificity. The true positive rate (TPR) is plotted along the y axis while the false positive rate (FPR) is shown on the x axis. Better classifier should be located closer to the upper left corner of the ROC curve diagram. The comparisons were made for these two approaches. The ROC will help to visualize which approach is the best for obesity and overweight predictions. The ROC curve for obesity predictions are shown in Figure 4.

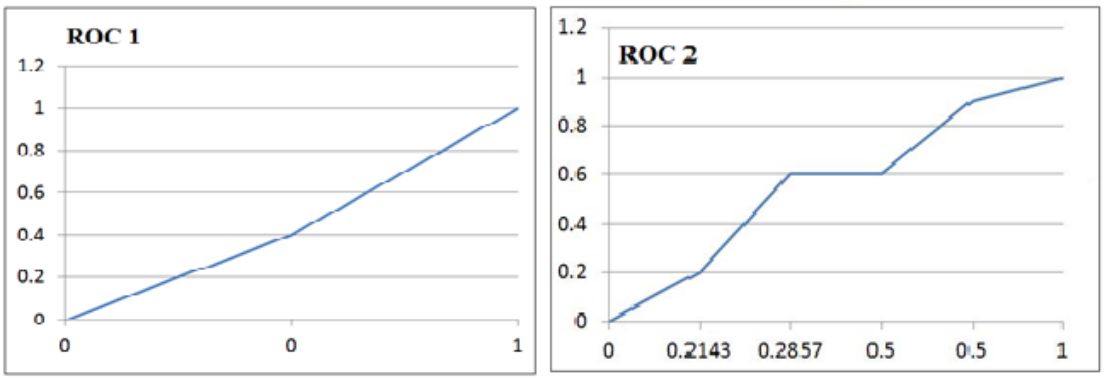

Figure 4. The ROC Curve for Obesity Predictions

Based from the figure, ROC 2 that represents the the second approach has a closer line to the upper left corner and the widest AUC. This shows that the second approach is better for the obesity prediction. Next, the ROC curves for overweight predictions are shown in Figure 5. Based from the figure, ROC 2 that represents the second approach has a closer line to the upper left corner and the widest AUC. This shows that the second approach is better for the overweight prediction.

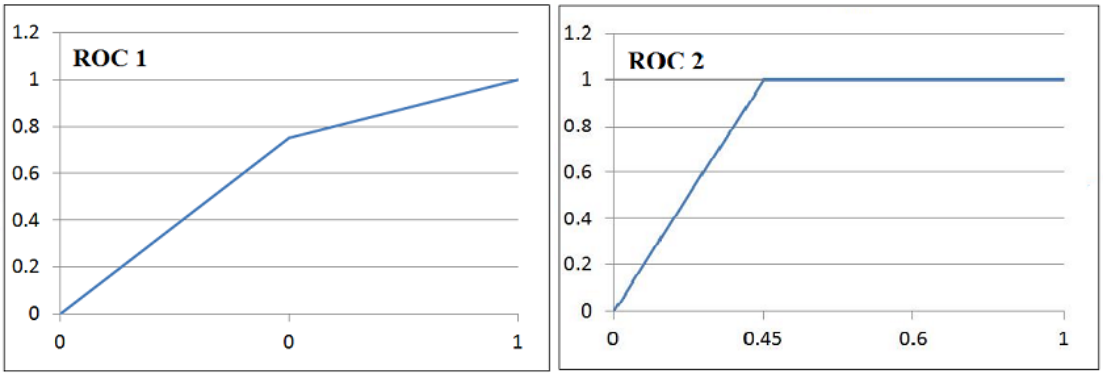

Figure 5. The ROC Curve for Overweight Predictions

6. Conclusion

This paper presents two hybrid approaches using the data mining techniques. The first approach used the CART for variable selection and the Naïve Bayes for classification. The second approach added a mean value identifications and Euclidean distances to the first approach. From the results, comparisons were made using the ROC curves. The conclusion of this paper is that the implementation of the first approach has increased the performance of CART, while the second approach has increased the performance of the first approach.

Acknowledgements

The authors would like to thank the Ministry of Higher Education (MOHE), Malaysia (Grant No. 203/PKOMP/6730002) and University Sains Malaysia for supporting this study.

References

[1] Centers for Disease Control and Prevention, http://www.cdc.gov/growthcharts/.

[2] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan and J. Keane, "Comparing Data Mining Methods with Logistic Regression in Childhood Obesity Prediction", Information Systems Frontiers, vol. 11, **(2009)**, pp. 51.

[3] E. Strumbelj, Z. Bosnic, I. Kononenko, B. Zakotnik, C. Grasic and Kuhar, "Explanation and Reliability of Prediction Models: The Case of Breast Cancer Recurrence", Knowl. Inf. Syst., vol. 24, **(2010)**, pp. 305-324.

[4] J. Chen, Y. Xing, G. Xi, J. Chen, J. Yi, D. Zhao and J. Wang, "A Comparison of Four Data Mining Models: Bayes, Neural Network, SVM and Decision Trees in Identifying Syndromes in Coronary Heart Disease", **(2007)**, pp. 4491.

[5] Y. Huang, P. McCullagh, N. Black and R. Harper, "Evaluation of Outcome Prediction for a Clinical Diabetes Database, Knowledge Exploration in Life Science Informatics", vol. 3303, J. López, et al., Eds., ed: Springer Berlin / Heidelberg, **(2004)**, pp. 181-190.

[6] A. Uyar, A. Bener, H. N.Ciray and M. Bahceci, "ROC Based Evaluation and Comparison of Classifiers for IVF Implantation Prediction", Electronic Healthcare, vol. 27, P. Kostkova, Ed., ed: Springer Berlin Heidelberg, **(2010)**, pp.108-111.

[7] M. H. M. Adnan, W. Husain and N. A. Rashid, "Comparative experiments of data mining methods for childhood obesity and overweight predictions", Universiti Sains Malaysia, **(2012)**, unpublished.

[8] L. J. Scheetz, J. Zhang and J. Kolassa, "Classification Tree Modeling to Identify Severe and Moderate Vehicular Injuries in Young and Middle-aged Adults", Artificial Intelligence in Medicine, vol. 45, **(2009)**, pp. 1-10.

[9] R. Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", in Proceedings of The Second International Conference on Knowledge Discovery and Data Mining, **(1996)**.

[10] B. Chandra and M. Gupta, "Robust Approach for Estimating Probabilities in Na ïve–Bayes Classifier for Gene Expression Data", Expert Systems with Applications, vol. 38, **(2011)**, pp. 1293-1298.

[11] P.-N. Tan, M.Steinbach and V. Kumar, "Introduction to Data Mining", Addison-Wesley, **(2006)**.

[12] J. Huaiyu, J. Hongjie, L. Yannan and S. Wei, "Symmetry of Translating Solutions to Mean Curvature Flows", Acta Mathematica Scientia, vol. 30, **(2010)**, pp. 2006-2016.

[13] M. H. M. Adnan, W. Husain and N. A. Rashid, "Parameter identification and selection for childhood obesity prediction using data mining", in 2012 2nd International Conference on Management and Artificial Intelligence (ICMAI), Bangkok, Thailand, **(2012)**.

Authors

Muhamad Hariz B. Muhamad Adnan

He received BsC in Universiti Sains Malaysia in 2008. He is a research Officer and a postgraduate in Universiti Sains Malaysia. His research interests are in the area of Software Engineering, Data Mining, Artificial Intelligence, and Computer Vision.

Wahidah Husain

She received M. Sc. Degree in Computer Science from Northrop University, California in 1988. She is a senior lecturer at the School of Computer Sciences, Universiti Sains Malaysia, Penang. Her research interests are in the area of Bioinformatics and Health Informatics.

Nur`Aini Abdul Rashid

She received B. Sc. Degree in Computer Science from Mississipi State, U.S.A. She received M.Sc. and PhD in Universiti Sains Malaysia. She is an associate professor at the School of Computer Sciences, Universiti Sains Malaysia, Penang. Her research interests are in the area of Parallel and Distributed Processing, Genomic Information Processing, and String and Pattern Matching