DOI 10.1007/s10796-009-9157-0

Comparing data mining methods with logistic regression in childhood obesity prediction

Shaoyan Zhang · Christos Tjortjis · Xiaojun Zeng ·
Hong Qiao · Iain Buchan · John Keane

Published online: 24 February 2009

Abstract The epidemiological question of concern here is “can young children at risk of obesity be identified from their early growth records?” Pilot work using logistic regression to predict overweight and obese children demonstrated relatively limited success. Hence we investigate the incorporation of non-linear interactions to help improve accuracy of prediction; by comparing the result of logistic regression with those of six mature data mining techniques.

The contributions of this paper are as follows: a) a comparison of logistic regression with six data mining techniques: specifically, for the prediction of overweight and obese children at 3 years using data recorded at birth, 6 weeks, 8 months and 2 years respectively; b) improved accuracy of prediction: prediction at 8 months accuracy is improved very slightly, in this case by using neural networks, whereas for prediction at 2 years obtained accuracy is improved by over 10%, in this case by using Bayesian methods. It has also been shown that incorporation of non-linear interactions could be important in epidemiological prediction, and that data mining techniques are becoming sufficiently well established to offer the medical research community a valid alternative to logistic regression.

Keywords Medical data mining · Machine learning · Public health · Prediction · Accuracy

S. Zhang · X. Zeng · H. Qiao · J. Keane
School of Computer Science, University of Manchester,
Manchester M60 1QD, UK

S. Zhang
e-mail: s.zhang-3@manchester.ac.uk

X. Zeng
e-mail: x.zeng@manchester.ac.uk

H. Qiao
e-mail: hong.qiao@manchester.ac.uk

J. Keane
e-mail: john.keane@manchester.ac.uk

C. Tjortjis (✉)
Department Engineering Informatics and Telecommunications,
University of Western Macedonia,
Vermiou & Ligeris,
Kozani 50100, Greece
e-mail: christos.tjortjis@manchester.ac.uk

C. Tjortjis
Department of Computer Science, University of Ioannina,
P.O. 1186, 45110 Ioannina, Greece

I. Buchan
School of Medicine, University of Manchester,
Manchester M13 9PT, UK
e-mail: Buchan@manchester.ac.uk

1 Introduction

There is a growing epidemic of obesity affecting all age groups, with the prevalence of obesity in the UK rising rapidly in children as young as 3 years (Bouchard et al. 1990). It has been reported that among two to four-year-olds, obesity has doubled since the early 1990s, while the rate has trebled for six to 15-year-olds (BBC, Editor 2005). In the UK, among those under 11, obesity increased from 9.6% in 1995, to 13.7% in 2003 (From Health Surveys for England reported in The Times, February 28th 2006). The increase in childhood obesity is causing concern in other countries, as well as the UK. Being fat as a child causes immediate harm, such as low self-esteem, and has consequences for adult health including life-long risk of obesity and an increased risk of type 2 diabetes (Bhargava et al. 2004).

Several ways have been suggested to treat obesity in children, such as physical exercise combined with nutrition education or behaviour modification (Wolf et al. 1985). However, there would be a greater public health impact from preventing obesity than treating it. There may be targets for obesity prevention in early childhood, to be identified through a combination of biomedical and epidemiological research. The epidemiological question here is "can young children at risk of obesity be identified from their early growth records?" This prediction can be regarded as a general classification/prediction problem. Classification means a procedure of finding a set of models/functions in a given database with given classes; using the models/functions obtained, the class of objects where class labels are unknown can be predicted. Classification has been used extensively in the medical domain (Tjortjis et al. 2007; Abu-Hanna and de Keizer 2003).

The Wirral child database has been built from data collected by health visitors in Wirral, England. This collection covers the 16-year period from 1988 to 2003, and comprises data from a total of 16,653 samples. It records parameters of children from their birth to around 3 years old, with 56 attributes for each sample. These mainly include weight, height and Body Mass Index (BMI) at each of five visits, and their standard deviation scores (SDS) adjusted for age and sex (British 1990 revised reference) (Cole et al. 1998).

The aim of this work is to evaluate the importance of non-linear information on childhood obesity prediction. Pilot work using logistic regression was used to predict overweight and obese children obesity with relatively limited success (Buchan 2005). It is therefore postulated that incorporation of non-linear interactions may improve the accuracy of this prediction; hence, we investigate the incorporation of non-linear interactions to help improve accuracy of prediction; by comparing the result of logistic regression with those of six mature data mining techniques (the techniques were selected based on significant experience in the area).

The contributions of this paper are as follows:

1. a comparison of logistic regression with six data mining techniques (decision trees (C4.5), association rules, Neural Networks (NNs), naïve Bayes, Bayesian networks and Support Vector Machines (SVMs)) to this prediction problem: specifically, for the prediction of overweight and obese children at 3 years of age using data recorded at birth, 6 weeks, 8 months and 2 years respectively;
2. improved accuracy of prediction by using data mining techniques: prediction at 8 months accuracy is improved very slightly, in this case by using NNs, whereas for prediction at 2 years old obtained accuracy is improved by over 10%, in this case by using Bayesian methods.

The work has achieved improvement in prediction of an important real-world problem. Associated with this, it has been shown that incorporation of non-linear interactions is likely to be of importance in epidemiological prediction, and that data mining techniques are becoming sufficiently well established to offer the medical research community a valid alternative to logistic regression.

The structure of this paper is as follows: in Section 2 the methods and algorithms to be used are presented; Section 3 focuses on experimentation with the recorded childhood data to compare and analyse the performance of different algorithms for childhood overweight/obesity prediction; finally, conclusions are presented in Section 4.

2 Methods and algorithms

2.1 Problem modelling

To achieve the aim of this work, we evaluate a number of well-known data mining algorithms, using non-linear information, on childhood obesity prediction. In data mining, prediction can be viewed as the construction and use of a model to assess the class of an unlabeled sample, or to assess the value or value ranges of an attribute that a given sample is likely to have (Han and Kamber 2006).

In the Wirral database, the value of the attribute "overweight" divides the database into two classes, where a label of "+1" means that the child is overweight by 3 years old, whilst a label of "−1" means that the child belongs to the lean class. Determining whether a child is lean, obese or overweight is decided by the BMI at 3 years old. BMI can be expressed as:

$$BMI = \frac{weight}{height^2} \tag{1}$$

An experience threshold BMI_{cutoff} is given, where:

$$BMI \begin{cases} \geq BMI_{cutoff} \Rightarrow overweight = +1 \Rightarrow overweight \\ < BMI_{cutoff} \Rightarrow overweight = -1 \Rightarrow lean \end{cases}. \tag{2}$$

The parameter "BMI" is normalized to be the parameter "overweight" to define whether a child is overweight, using a transformation (f), the standard deviation score which adjusts the child's measurement for their expected growth in terms of age and sex (Cole et al. 1998). With this "labelled" database, it is possible to construct a model first, and then use the model to predict the class of unlabeled samples.

Classification and regression are two major types of prediction algorithms: classification is used to predict discrete or nominal values, while regression is used to predict continuous or ordered value. The main objective is to identify and predict the group of children who are at risk of becoming overweight and thus require preventative action, rather than to predict the BMI of individual children. To predict child overweight/obesity by year 3, data recorded on the weight of the child during the first 2 years of a child's life is used (i.e. recorded at 6 weeks, 8 months and 2 years respectively).

2.2 Accuracy measurement

To evaluate the prediction rate, the following related parameters are to be studied: sensitivity $P_r(+|D)$, specificity $P_r(-|{\sim}D)$, positive predictive value $P_r(D|+)$, negative predictive value $P_r({\sim}D|-)$, and overall accuracy. "D" means the overweight or obese case and "~D" means the lean case.

$$P_r(+|D) = \frac{\text{number of correctly classified overweight or obeses cases}}{\text{number of total overweight or obese cases}} \tag{3}$$

$$P_r(-|\sim D) = \frac{\text{number of correctly classified lean cases}}{\text{number of total lean cases}} \tag{4}$$

$$P_r(D|+) = \frac{\text{number of correctly classified overweight or obeses cases}}{\text{total number of cases classified as overweight or obese}} \tag{5}$$

$$P_r(\sim D|-) = \frac{\text{number of correctly classified lean cases}}{\text{total number of cases classified as lean}} \tag{6}$$

If every sample is correctly classified, then the values of all the above four parameters are 1.0. In many applications, specificity or overall accuracy is more important than sensitivity; however, this is not the case here for the following reasons:

1. As the majority of the children are not overweight or obese, it is simple to have a classifier with perfect specificity and very good overall accuracy but without any disclosure of the risk group of potential overweight or obese children. For example, for classifying children who are at risk of obesity, a trivial classification that all children are not overweight can achieve 100% specificity and approximately 96.7% overall accuracy. However, such a classifier does not provide any useful information about which group of children is at risk of becoming obese. In other words, specificity and overall accuracy are much less important and useful here.
2. Further, the higher the sensitivity, the more accurate the identification of the overweight/obese risk group, and the better the chance that prevention can be achieved. Even in the case that specificity or overall accuracy is a little lower, which means some lean cases would be mis-classified as overweight and thus preventative methods may be applied to lean children. This may be of no harm as, if handled sensitively, such preventative methods are likely to have positive effects on all children. This is very different to many medical diagnoses where lower specification means that more "unaffected" patients are misdiagnosed and have to receive potentially harmful or uncomfortable treatments unnecessarily. For this reason, sensitivity in equation (3) is by far the most important parameter in this context.

2.3 The database

Let **S** to be a database consisting s samples, assume that the class label attribute (for example, the attribute "overweight" in the Wirral Database) has m distinctive values defining m classes $C_i, i = 1, \cdots m$. Define s_i to be the number of samples that belong to class C_i.

Each data sample X has n attributes, which can be expressed in the form of a vector as $X = (x_1,\ x_2, \ldots, x_n)$. In vector X, each element x_i is a nominal or continuous value that corresponds with the attribute $A_i = (i = 1, \ldots, n)$. Suppose that the attribute A_i has q distinct values $\{a_{i1}, a_{i2}, \cdots, a_{iq}\}$, and then A_i can partition the original dataset S into q subsets, $\{S_1, S_2, \cdots, S_q\}$, where $S_j$contains samples that have value of a_{ij} in A_i. Then let s_{ij} be the number of samples of class C_i in a subset S_j.

2.4 Prediction algorithms

There are many kinds of prediction algorithm; based on collective experience in the field of data mining, we have selected six prominent data mining algorithms that have potential to yield good results and are accessible to a wider audience, these are decision trees (C4.5), association rules, Neural Networks, naïve Bayes, Bayesian networks, and SVM. Another fact in the selection of the methods is the maturity of the techniques and the related software tools (Witten and Frank 2005).

1. Decision Trees. A decision tree is a tree-like structure, which starts from root attributes, and ends with leaf nodes. Generally a decision tree has several branches

consisting of different attributes, the leaf node on each branch representing a class or a kind of class distribution. The tree is generated according to the information-gain measure, the procedures are briefly as:

a. Calculate:

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \ \log_2(p_i) \tag{7}$$

$$P_i = s_i/s \tag{8}$$

where p_i is the probability that an arbitrary sample belongs to class C_i.

b. Calculate the entropy $E(A_i)$, which is the expected information based on the partitioning by attribute A_i:

$$E(A_i) = \sum_{j=1}^{q} \frac{s_{1j} + s_{2j} + \cdots + s_{mj}}{s} I\left(s_{1j}, \cdots s_{mj}\right) \tag{9}$$

$$I\left(s_{1j}, s_{2j} \cdots s_{mj}\right) = \sum_{i=1}^{m} p_{ij} \log_2\left(p_{ij}\right) \tag{10}$$

where $p_{ij} = s_{ij}/\left|S_j\right|$, and $\left|S_j\right|$ is the number of samples in subset S_j.

c. Then the encoding information that would be gained by branching on A_i is:

$$Gain(A_i) = I(s_1, s_2, \cdots, s_m) - E(A_i) \tag{11}$$

The attribute A_i with the highest information gain is selected as the root node, the branches of the root node is formed according to different distinctive values of $a_{ij}, j = 1, \cdots, q$. The tree grows like this until if all the samples are all of the same class, and then the node becomes a leaf and is labelled with that class.

Decision tree algorithms describe the relationship among attributes, and the relative importance of attributes. In addition, human-understandable rules can be extracted from the tree. Both the learning and classification steps of decision tree induction are generally fast. The work described here uses the well-known C4.5 algorithm (Quinlan 1993); a recent discussion of the area is given in (Rokach and Maimon 2005).

2. Bayesian Classifiers. These are statistical classifiers that predict class membership by probabilities, such as the probability that a given sample belongs to a particular class. Several Bayes' algorithms have been developed, among which Bayesian networks and naïve Bayes are the two fundamental methods. Naïve Bayes algorithms assume that the effect that an attribute plays on a given class is independent of the values of other attributes. In practice, dependencies often exist among attributes; hence Bayesian networks are graphical models, which, unlike naïve Bayesian classifiers, can describe joint conditional probability distributions.

Let *X* to be a sample whose class is to be determined. Let *H* be some hypothesis, such as the sample *X* belongs to a specified class *C*. For classification problems, we want to determine *P*(*H*|*X*), the probability that the hypothesis *H* holds given the observed data sample *X*.

Bayes theorem provides an efficient algorithm to calculate the posterior probability, *P*(*H*|*X*); the formula is as follows

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{12}$$

where *P*(*H*) is the prior probability of H. *P*(*H*|*X*) is the posterior probability of X conditioned on H. *P*(*X*)is the prior probability of X.

In formula (12), *P*(*X*),*P*(*H*) and *P*(*X*|*H*)may be estimated from the given data for training, if all three values can be determined, the probability for sample X to be in hypothesis H can be determined.

Given an unknown data sample *X*, the naïve Bayesian classifier works as follows:

a. Calculate the posterior probability $P(C_i|X), 1 \leq \mathrm{i} \leq \mathrm{m}$, conditioned on X. The naïve Bayesian classifier assigns sample *X* to the class C_i if and only if

$$P(C_i|X) > P\left(C_j|X\right) \text{ for } 1 \leq \mathrm{j} \leq \mathrm{m}, \neq \mathrm{j} \tag{13}$$

where, *P*(C_i|*X*) can be determined by Bayes theorem,

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{14}$$

b. As *P*(*X*) is constant for all classes, only the maximum *P* (*X*|C_i)*P*(C_i) need to be sorted out. If there are no samples given, then it is commonly assumed that the classes are equally likely, and we would therefore maximize *P*(*X*|C_i). Otherwise, we maximize *P*(*X*|C_i)*P* (C_i).

P(C_i) can be determined by:

$$P(C_i) = \frac{NU_i}{NU} \tag{15}$$

where NU_i is the number of training samples that belong to class C_i, and *NU* is the total number of training samples.

c. Assume that all the attributes are conditionally independent of one another, which means that there are no dependence relationships among the attribute. Thus,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k, C_i) \tag{16}$$

The probabilities $P(x_1|C_i)$, $P(x_2|C_i)$,…,$P(x_n|C_i)$ can be estimated from the training samples, where:

if A_k is categorical, then $P(x_k|C_i) = \frac{NU_{ik}}{NU_i}$, where NU_{ik} is the number of training samples of class C_i having the nominal value x_k for A_k, and NU_i is the number of training samples belonging to C_i.

d. Substitute formulae (14–16) into (13), sample X is then assigned to the class C_i with the highest probability $P(X|C_i)P(C_i)$.

Compared with Bayesian networks, the naïve Bayesian classifier is much easier to use, and if the attributes are independent, the naïve Bayesian can be the most accurate when compared with other classifiers. Bayesian classifiers have exhibited high accuracy and speed when applied to large databases and are especially popular in medical domains, for example, using Bayesian networks to analyse DNA hybridization arrays, and in medical diagnosis; related material can be seen in (Kononenko 1993; Fayyad and Irani 1993; John and Langley 1995). Due to its advantages, especially its high performance in medical domains, Bayesian classifiers are selected for the prediction of child overweight/obesity. Further details on Bayesian approaches can be found in (Ross and Pate 1987; Gortmaker et al. 1987).

3. Association Rule Classifiers. Association rules describe relationships between attributes in a database, and are widely used in market basket analysis. Two key concepts of association rules are *support* and *confidence:*

$$\text{support}(A \Rightarrow B) = P(A \cup B) \tag{17}$$

$$\text{confidence}(A \Rightarrow B) = P(B|A) \tag{18}$$

where $P(A \cup B)$ is the probability that A and B occur together in the database; and $P(B|A)$ is the conditional probability, which means that if A occurs, how many times does B occur at the same time. Rules that satisfy both a minimum support threshold and minimum confidence are called *strong* rules.

The basic steps of finding association rules include:

a. Find frequent item-sets. An item-set means a set of attributes; an item-set that contains k attributes is called a k-item-set. The *occurrence frequency* of an item-set is the frequency of samples in the database that contains the item-set. The item-set is called a frequent item-set, if it occurs at least as frequently as a pre-determined minimum support count.

b. Generate strong association rules from the frequent item-sets.

Based on the above basic ideas, many association rule algorithms have been developed, such as Apriori (Agrawal and Srikant 1994). Recently, the concepts in association rule mining have been developed to address classification (Ross and Pate 1987; Gortmaker et al. 1987). For example, suppose that after training, the following three rules are obtained as shown in Table 1. Each rule in Table 1 is related with an associated output—overweight or not. Then for classification, if the child's attribute values coincide with rule 1 or 2, then the child is predicted to be overweight by 3 years old; if the attribute values do not accord with rule 1 or 2, the child it predicted to be not overweight. Note that the rules given in Table 1 are simplified examples; the real case rules are far more complicated.

4. Neural Networks. A neural network is a set of connected input/output units where each connection has a weight associated with it. A neural network has an input layer, an output layer, and hidden layers. Unlike decision trees, which have only one input node (root node) for the input layer, a neural network has one input node for each attribute value to be examined. In contrast to decision trees, a neural network adjusts the weights during the learning process, in order to satisfy all the input and output relations. Neural networks have hidden layers with arbitrary number of nodes, which make it easier to regulate the weight of each node, to satisfy the input and output relationships.

Instead of illustrating neural networks by formulae, a simple structure is given in Fig. 1 to show how a neural network works for the Wirral database. The structure is much simplified, where only two attributes "height" and "weight" are selected as inputs. Neural networks are popularly applied in classification and prediction, as they

Table 1 Examples of association rules obtained for child overweight after training

1	If (sdsbmiv4 = '(2.2445-inf)') and (sdshtv4 = '(1.13-inf)') and (sdswtv4 = '(3.0555-inf)') => then, overweight = yes
2	Else If (sdsbmiv4 = '(2.2445-inf)') and (sdswtv1 = '(0.6365-inf)') and (sdswtv3 = '(0.8305–2.0905)') and (sdslenv1 = '(-inf-1.9755)') and (sdsgainb2v1 = '(−0.2775–0.5555)') => then, overweight = yes
3	Else, overweight = no

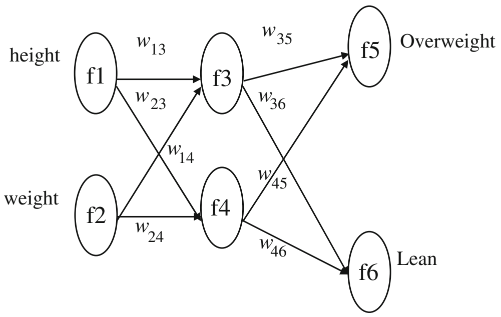

Fig. 1 A simplified neural network for child overweight analysis

have advantages such as high tolerance to noise, and the ability to classify unseen patterns**.**

5. Support vector machines (SVMs): SVMs are pattern classification algorithms developed by Vapnik (1995, 1998). For a training set of l samples, the learning procedure is represented as solving the following optimisation problem:

$$\min_{\alpha} : \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j a_i a_j K\left(x_i, x_j\right) - \sum_{j=1}^{l} a_i \tag{19}$$

$$\text{s.t.} \quad 0 \le \alpha_i \ \le C, i = 1, \cdots, l \tag{20}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{21}$$

where, y_i is the label of the *ith* sample x_i, α_i is the Lagrangian multiplier of x_i, C is the upper bound of α_i. $K(x_i, x_j)$ is the kernel, which can map the original data X into a high-dimensional Hilbert space, and can make the samples linear separable in the high-dimensional space. The samples with $\alpha_i > 0$ are called support vectors.

Accordingly, the decision function can be written as:

$$f(x) = sgn\left(\sum_{i=1}^{n_s} y_i \alpha_i^* K(x_i, x) + b^*\right) \tag{22}$$

where n_s is the number of support vectors.

SVMs have many distinctive advantages:

a. SVMs are function-based classifiers, which can be expressed in the standard form of quadratic optimization programming, which can be solved easily.
b. Kernel mapping techniques in SVM can cope with similarity in a high-dimensional Hilbert space, and many linear non-separable cases can be partitioned successfully by SVMs, which means that the performance of SVMs is very good.
c. SVMs conform to statistical theory. They minimize structural risk instead of empirical risk; the bound of the expected risk is given by maximizing the margin of the classification, so that when the optimal solution is obtained for training, the estimated prediction error is minimized.
d. SVMs can condense a large dataset into a comparably small dataset, which only includes the support vectors. The condensed database can save considerable memory, and simplify the testing procedure.

Different kernels define different SVMs; here we use the Radial Basis Function (RBF) kernel and the Linear kernel for prediction:

$$\text{Gaussian RBF} : K(x_1, x_2) = e^{-\|x_1 - x_2\|^2 / 2\sigma^2} \tag{23}$$

$$\text{Linear kernel} : K(x_1, x_2) = x_1.x_2 \tag{24}$$

where σ in Equation (23) is the parameter to be selected by the user.

Table 2 The attributes selected for training and their meaning

Attribute	Meaning of Attributes
Sex12	Sex = 1→male;sex = 2→female
sdsbwt	adjusted SDS Birth weight
sdslenv1	adjusted SDS length at the 1st visit (6 weeks)
Gestyrs	Time of gestation
sdsgainb2v1	adjusted SDS weight gain birth to 1st visit (6 weeks)
sdsgainwtv1v3	adjusted SDS weight gain between the 3rd visit (8 months) and the 1st visit (6 weeks)
bmiv3	Body mass index at the 3rd visit (8 months)
sdshtv4	adjusted SDS height at the 4th visit (2 years)
sdsbmiv4	adjusted SDS Body mass index at the 4th visit (2 years)
Overorobesev5	1→*overweight*; 0→*lean* at 3 years

Table 3 A comparison of SVM and Naïve Bayes for the data recorded at *6 weeks*

		Naïve Bayes			SVM		
		D	~D	Total	D	~D	Total
Classified	+	310	545	855	41	35	76
	-	2,465	13,203	15,668	2,734	13,713	16,447
	Total	2,775	13,748	16,523	2,775	13,748	16,523
Sensitivity Pr (+ | D)		11.2%			1.5%		
Specificity Pr (−| ~D)		96.0%			99.7%		
Positive predictive value Pr (D | +)		36.3%			53.8%		
Negative predictive value Pr (~D |−)		84.3%			83.4%		
Correctly classified		81.8%			83.2%		

3 Experiments and results

In the following, the algorithms described above are used to predict children who may be overweight or obese and the results compared with obtained by logistic regression (Buchan et al. 2007).

3.1 Data pre-processing

Data Cleaning The database contains 16,653 instances. There is a constraint that the height at visit *n*th cannot be lower than that recorded at visit (n−1) th; otherwise the instance is regarded as abnormal. Furthermore, the age of a child at visit *n*th cannot be less than that recorded at visit (n−1)th, otherwise the instance is regarded as abnormal. All such abnormal instances are discarded. After cleaning, 16,523 instances remain, of which 2775 (16.8%) samples are overweight, and 543 (3.29%) samples are obesity cases.

Discretization of continuous attributes Although many algorithms can handle continuous attributes while classifying a given sample (John and Langley 1995; Remco 2005), here we prefer to change continuous values into nominal values using the discretization method, which is supposed to be better than the normal Gaussian distributed method. The discretization method, which converts a continuous variable X into a discrete variable X' at different levels, was used here. In this paper, the minimum description length (MDL) method has been used to discretize values (Fayyad and Irani 1993).

3.2 Experiments to predict overweight children

Overweight cases are much more prevalent than obese cases in the database; the two are related in that being overweight will occur first and may lead to obesity. Consequently prediction of overweight at an early age is important.

In the Wirral database, there are six times as many overweight as obese children; approximately 20% of the database are either overweight or obese. In this section, several algorithms are evaluated for their accuracy in predicting overweight children.

Several attributes in the database are selected for training and prediction; the names of the attributes and their meaning are described in Table 2.

In the following the analysis is structured as follows (of the total):

A. Predict whether a child will be overweight after 6 weeks, using SVM and naïve Bayes.

Table 4 A comparison of SVM and Naïve Bayes for the data recorded by *8 months*

		Naïve Bayes			SVM		
		D	~D	Total	D	~D	Total
Classified	+	985	1,166	2,151	1,277	3,781	5,058
	-	1,790	12,582	14,372	1,498	9,967	11,465
	total	2,775	13,748	16,523	2,775	13,748	16,523
Sensitivity Pr (+ | D)		35.5%			46.0%		
Specificity Pr (−| ~D)		91.5%			72.5%		
Positive predictive value Pr (D | +)		45.8%			25.3%		
Negative predictive value Pr (~D |−)		87.5%			86.9%		
Correctly classified		82.1%			68.1%		

Table 5 A comparison of SVM and Naïve Bayes for the data recorded by "*2 years*"

		Naïve Bayes			SVM		
		D	~D	Total	D	~D	Total
Classified	+	1,518	949	2,467	1,665	2,805	4,470
	-	1,257	12,799	14,056	1,110	10,943	12,053
	total	2,775	13,748	16,523	2,775	13,748	16,523
Sensitivity Pr (+ | D)		54.7%			60.0%		
Specificity Pr (−| ~D)		93.1%			79.6%		
Positive predictive value Pr (D | +)		61.5%			37.3%		
Negative predictive value Pr (~D |−)		91.1%			90.8%		
Correctly classified		91.9%			76.3%		

B. Predict whether a child will be overweight at 8 months, using SVM and naïve Bayes.
C. Predict whether a child will be overweight after 2 years, using SVM and naïve Bayes.
D. Predict whether a child will be overweight after 2 years, using different algorithms.

For the first three of these tests, to aid result comparison, we have selected the corresponding set of attributes that had been selected for the logistic regression experiments (Cole et al. 1998). We note these selections had been made by medical domain experts. For the final test, the attributes are selected from the first three visits (to 8 months) and the first four visits (to 2 years old) respectively, to see the prediction rates at different stages.

3.2.1 Predict if a child will be overweight based on data recorded within 6 weeks of birth

This experiment predicts whether a child will be overweight at 3 years old, based on the data from the first visit, which is around 6 weeks after birth. In order to compare results, we selected the same attributes as for logistic regression: *sex12, sdsbwt, sdslenv1, sdsgainb2v1,* and *overorobesev5* (Cole et al. 1998). In this experiment, SVM with the RBF kernel and naïve Bayes are adopted for the prediction. The results are listed in Table 3, where "D" means the overweight case; "~D" means the lean case.

From the results, the sensitivity for naïve Bayes is 11.2%, which is reasonable for prediction at 6 weeks after birth; compared with naïve Bayes, SVM performs much worse, with only 1.5% of overweight cases predicted when the babies are 6 weeks old.

3.2.2 Predict whether a child will be overweight based on data recorded by 8 months

In this experiment, the data from the first three visits is used for prediction. In order to compare results, we selected the same attributes as for logistic regression: *Sex12, Sdsbwt, Sdslenv1, Sdsgainb2v1, Sdsgainwtv1v3, Bmiv3*, and *Overorobesev5* (Cole et al. 1998). The results are listed in Table 4.

Table 4 clearly shows that SVMs have much better sensitivity than naïve Bayes; however, the values of all the other parameters are lower than for those of naïve Bayes. For example, a positive prediction value of 25.3% is poorer than that obtained by naïve Bayes, This means that the SVM improves sensitivity by sacrificing the prediction accuracy of other parameters.

Table 6 The results of different algorithms

	Data recorded by 8 months			Data recorded by 2 years		
Algorithm	Sensitivity	Specificity	Overall accuracy	Sensitivity	Specificity	Overall accuracy
Decision tree	12.3%	97.8%	83.6%	29.2%	96.4%	85.2%
Association rules	17.8%	96.9%	83.7%	39.0%	95.4%	86.1%
Logistic regression	13.3%	**98.1%**	83.7%	29.1%	**97.1%**	83.2%
Neural network	14.0%	97.8%	**83.9%**	40.7%	93.5%	84.7%
Linear SVM	14.9%	95.9%	82.3%	59.6%	82.6%	78.7%
RBF SVM	**46.0%**	72.5%	68.1%	**60.0%**	79.6%	76.3%
Bayesian network	35.5%	91.5%	82.1%	54.7%	93.1%	**91.9%**
Naïve Bayesian	35.5%	91.5%	82.1%	54.7%	93.1%	**91.9%**

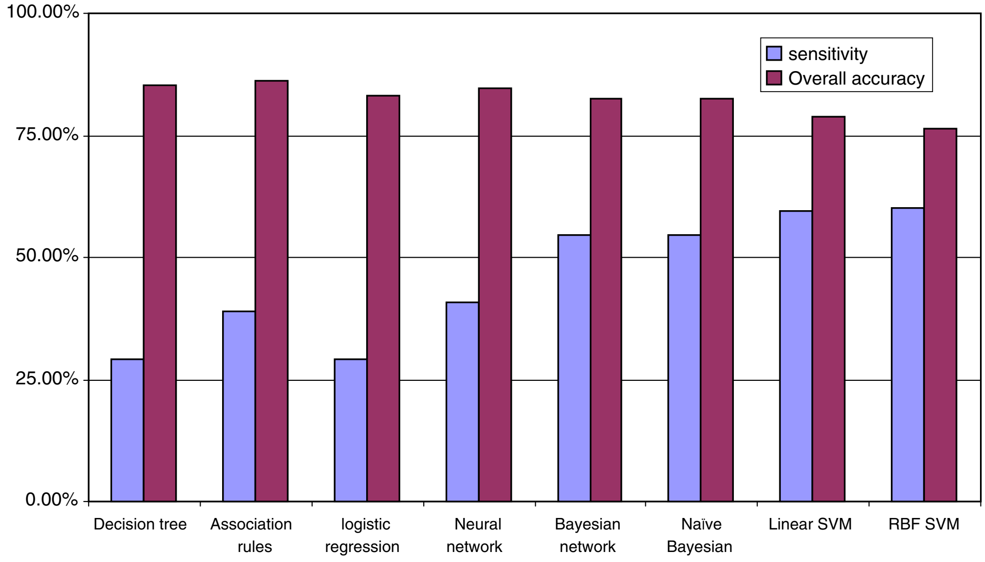

Fig. 2 Prediction rates with different algorithms, using the data recorded at first four visits

3.2.3 Predict whether a child will be overweight after 2 years

In the following, the overweight or obese cases are predicted using the first four visits; In order to compare results, we selected the same attributes as for logistic regression: *Sex12, Gestyrs, Sdsbwt, Sdslenv1, Sdsgainb2v1, Sdsgainwtv1v3, Bmiv3, Sdshtv4, Sdsbmiv4*, and *Overorobesev5* (Cole et al. 1998). The detailed results are listed in Table 5.

Table 5 illustrates that, by using the first four visits to predict overweight at 3 years of age, the RBF SVM achieves sensitivity of 60%, while naïve Bayes achieves sensitivity of 54.7%.

The other values obtained from SVM are lower than those obtained from naïve Bayes, especially for the specificity and positive predictive value. Here we hope to predict as many overweight cases as possible; however, we also wish to limit the number of lean cases predicted as overweight. From this point of view, SVM is not better than naïve Bayes as SVM has much poorer Pr (D | +) than naïve Bayes.

3.2.4 Prediction using more algorithms

In this section, different algorithms are used for overweight/obesity prediction, including decision trees (C4.5), association rules, neural networks, SVMs, logistic regression, naïve Bayes, and Bayesian networks.

The attributes are selected from the first three visits (to 8 months) and the first four visits (to 2 years old) respectively, to see the prediction rates at different stages. We used an attribute selection function before training, which requires two key objects to be set up: a feature evaluator and a search method (Liu and Yu 2005). The feature evaluator determines which method is used to assign a "worth" to each subset of features. The search method determines what style of search is performed. The *Cfs*- Subset Evaluator and *BestFirst* Search method were adopted for attribute selection:

- CfsSubsetEval—Evaluates the "worth" of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them; subsets of features that are highly correlated with the class while having low inter-correlation are preferred.
- BestFirst—Searches the space of feature subsets by greedy hill-climbing augmented with a backtracking facility.

The whole dataset was used for training and testing. The results are displayed in Table 6, and a separate bar chart is

	D	~D
+	0	187
-	0	5,431

Pr(+ | D) = 0.0 %
Pr(- | ~D) = 100.0%
Pr(D | +) = ~
Pr(~D | -) = 96.7%

Fig. 3 Indications at 2 years of obesity at 3 years, using decision trees (C4.5)

	D	~D
+	41	146
-	36	5,395

Pr(+ | D) = 21.9 %
Pr(- | ~D) = 99.3%
Pr(D | +) = 53.2%
Pr(~D | -) = 97.4%

Fig. 4 Indications at 2 years of obesity at 3 years using association rules (Apriori)

Table 7 Prediction rate of obesity from related values obtained via different algorithms using the data recorded before 2 years old

Algorithms	The values of related parameters (in %)	
	Sensitivity: $P_r(+\|D)$	Positive predictive value $P_r(D\|+)$
Decision tree	0	~
Logistic regression	11.2	**53.8**
Association rules	21.9	53.2
Neural network	24.6	29.3
Naïve Bayes	**62.0**	18.2
Bayesian network	**62.0**	18.2
RBF SVM	38.0	30.4

shown in Fig. 2 to demonstrate the sensitivity and overall prediction rate at the 4th visit.

Analysis:

- ❖ For prediction of overweight at 8 months old, it can be seen that all the results have comparable overall accuracy. SVMs gives a lower overall prediction accuracy, 68.1%, however, SVMs rate the highest for sensitivity, which is what we are primarily concerned with for overweight prediction. Naïve Bayes gives both high sensitivity and overall accuracy. Furthermore, notice that the Bayesian network does not appear to find any relationship among the attributes, for this case, it is equal to Naive Bayes. Considering that Bayesian network is more computationally intensive than Naïve Bayes, Naive Bayes is preferable.
- ❖ For prediction of overweight at 2 years, linear SVM, RBF SVM, naïve Bayes, and the Bayesian network have relatively better results, among which the sensitivities obtained from RBF SVM are highest (around 60%), and the sensitivity for the two Bayesian algorithms follows at around 55%. Bayesian classifiers have better prediction values.

3.3 Experiment to predict obesity

In this section, we will predict childhood obesity at 2 years, and use the whole database for training and testing. We used the same attribute selection function, as we did in the last experiment. The entire database is used with 67% of the samples randomly selected for training, and the other 33% for testing. Decision trees (C4.5), association rules (Apriori), Support Vector Machines (SVM), naïve Bayes, Bayesian network, and neural networks are used for prediction. The results are shown in Fig. 3, 4 and Table 7.

Analysis Seven algorithms have been used to predict obesity from the data recorded by 2 years old. All the results are relatively poor both for sensitivity and positive predictive value. Both naïve Bayes and the Bayesian network predict 62% of the obese cases when the children are around 2 years old, but at the same time, many lean cases are predicted to be obese, which make the positive predictive value lower than other algorithms. As was discussed earlier, for childhood obesity, sensitivity is most important. It should be noticed that when using a Bayesian network, no relationship appears to have been found among the attributes. As the Bayes network is much more computationally demanding, and as it has the same results as naive Bayes, naive Bayes is preferred.

The analysis suggests that obesity prediction at an early age is difficult. Part of the reason lies in the fact that the number of obesity cases is too small, being only 3.29% of the whole database. The samples between the two classes are seriously out of balance, this makes the analysis problematic; in addition, many non-obese samples are similar to obese samples at an early age.

4 Summary

In this paper, different 'data mining' algorithms have been applied to the prediction of overweight and obese children in their early years, based on the Wirral database. Generally speaking, prediction from early ages is difficult, partly because the reasons leading to overweight and obesity are complicated, involving not only physiological but also genetic, sociological and even psychological factors. The highest overweight prediction rate is 55–60% in this work. To get better prediction rates, more attributes may need to be recorded.

To compare the interpretation ability:

1) decision trees and association rules are popular for their ease of interpretation (Osei-Bryson and Giles 2006), but both fail to give appropriate rules here.
2) Bayesian networks identify relationships among the attributes; however, they do not work well here, as they

could not identify any rule, although their prediction rate is relatively high.

3) The interpretation ability of the algorithms appears too weak for the Wirral database, and they can only indicate some very simple rules.

To compare accuracy:

1) the prediction rates from logistic regression, decision tree and association rules are poor;
2) The neural network performs better than the above mentioned algorithms, but not as well as the Bayesian algorithms and SVMs.
3) SVMs have better sensitivity prediction rate than Bayesian algorithms, while Bayesian algorithms outperform SVMs in terms of overall prediction rate.

To summarise, SVM and Bayesian algorithms appear to be the best two algorithms for predicting overweight and obesity from the Wirral database. Generally, the results of this work have improved the prediction accuracy when compared to logistic regression and thus begun to show the value of incorporation of non-linear information in epidemiological prediction. In future work we plan to compare SVM and Bayesian algorithms with statistical models that are a more complete representation of the problem than logistic regression. Specifically, this will involve instrumental variable models using multi-level latent variable regression techniques to estimate unmeasured intermediating factors and handle the autocorrelation of growth measures over time.

References

Abu-Hanna, A., & de Keizer, N. (2003). Integrating classification trees with local logistic regression in intensive care prognosis. *Artificial Intelligence in Medicine*, *29*, 5–23.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *In Proc. 20th Int'l. Conf. Very Large Data Bases* (VLDB'94), (pp. 487–499).

BBC, Editor (2005). *TV watching link to child obesity*, Retrieved February 27, 2008, from BBCNews website: news.bbc.co.uk/2/hi/health/4562879.stm.

Bhargava, S. K., Sachdev, H. S., Fall, C., Osmond, C., Lakshmy, R., & Barker, D. J. P. (2004). Relation of serial changes in childhood body-mass index to impaired glucose tolerance in young adulthood. *New England Journal of Medicine*, *350*(9), 865–875.

Bouchard, C., Tremblay, A., Despres, J. P., Nadeau, A., Lupien, P. J., & Theriault, G. (1990). The response to long-term overfeeding in identical twins. *New England Journal of Medicine*, *322*(21), 1477–1482.

Buchan, I. (2005). Child obesity look harder approach via whole-population. School of Medicine, University of Manchester.

Buchan, I. E., Bundred, P. E., Kitchiner, D. J., & Cole, T. J. (2007). Body mass index has risen more steeply in tall than in short three year olds: serial cross-sectional surveys 1988–2003. *International Journal of Obesity*, *31*, 23–29.

Cole, T. J., Freeman, J. V., & Preece, M. A. (1998). British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. *Statistics in Medicine*, *17*, 407–429.

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *In Proc. 13th Int'l Joint Conf. on Uncertainty in AI*, (pp. 1022–1027). Morgan Kaufmann.

Gortmaker, S. L., Dietz Jr., W. H., Sobol, A. M., & Wehler, C. A. (1987). Increasing pediatric obesity in the United States. *American Journal of Diseases of Children*, *141*(5), 535–540.

Han, J. & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). Morgan Kaufmann.

John, G., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *In Proc. 11th Conf. Uncertainty in Artificial Intelligence* (pp. 338–345). Morgan Kaufmann.

Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, *7*, 317–337.

Liu, H., & Yu, L. (2005). Towards integrating feature selection algorithm for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, *17*(4), 491–502.

Osei-Bryson, K.-M., & Giles, K. (2006). Splitting methods for decision tree induction: an exploration of the relative performance of two entropy-based families. *Information Systems Frontiers*, *8*(3), 195–209.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

Remco, R. B. (2005). Naive Bayes classifiers that perform well with continuous variables. *In Proc. of the 17th Australian Conf. on AI* (AI 04), LNCS, *3339*, 1089–1094.

Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers—a survey. *IEEE Transactions on Systems, Man and Cybernetics*, *Part C*, *35*(4), 476–487.

Ross, J. G., & Pate, R. R. (1987). The national children and youth fitness study II: a summary of findings. *Journal of Physical Education, Recreation and Dance*, *58*(9), 51–56.

Tjortjis, C., Saraee, M., Theodoulidis, B., & Keane, J. A. (2007). Using T3, an improved decision tree classifier, for mining stroke related medical data. *Methods of Information in Medicine*, *46*(5), 523–529.

Vapnik, V. (1995). *The nature of statistical learning theory.* Springer-Verlag.

Vapnik, V. (1998). *Statistical learning theory.* Wiley.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and technique*s. 2nd Ed. Morgan Kaufmann.

Wolf, M. C., Cohen, K. R., & Rosenfeld, J. G. (1985). School-based interventions for obesity: current approaches and future prospects. *Psychology in the Schools*, *22*(2), 187–200.

Shaoyan Zhang received B.S. degree and PhD degree in Computational Engineering from Dalian University of Technology in China in 1994 and 2000, respectively. He received his 2nd PhD degree in Machine Learning from University of Manchester, UK in 2007. Dr Zhang is currently a mathematician at Unilever Discovery, Bedford, UK. His research interests include machine learning, signal processing, public health and Bioscience.

Christos Tjortjis is an adjunct Senior Lecturer at the Universities of Ioannina, Dept. of Computer Science, and of Western Macedonia, Dept. Engineering Informatics and Telecommunications, Greece, and an honorary lecturer at the University of Manchester, School of Computer Science, where previously he was a tenured Lecturer. He holds a DEng in Computer Engineering and Informatics from the University of Patras, and a BSc in Law from the Democritus University of Thrace, Greece. After gaining industrial experience as a consultant, he was awarded an MPhil in Computation from UMIST

and a PhD in Informatics from the University of Manchester, UK. His focal research area is data mining and aims to advance data mining in domains such as software engineering, biology and medicine. His work on data and has been published in over 30 international journal and conference papers. He has chaired or organised a number of international conferences and workshops and he has acted as guest editor for the Journal of Software Maintenance and Evolution.

Xiao-Jun Zeng received the B.Sc. degree in mathematics and the M.Sc. degree in control theory and operation research from Xiamen University, Xiamen, China, and the Ph.D. degree in computation from the University of Manchester, Manchester, U.K., in 1996. He has been with the University of Manchester since 2002, where he is currently a Senior Lecturer in the School of Computer Science. From 1996 to 2002, he was with Knowledge Support Systems, Ltd. (KSS), Manchester, where he was a Scientific Developer, Senior Scientific Developer, and Head of Research, developing intelligent decision support systems which won the European Information Society Technologies (IST) Award in 1999 and Microsoft European Retail Application Developer (RAD) Awards in 2001 and 2003. From 1985 to 1992, he was with the Department of Computer and Systems Sciences, Xiamen University, China, where he was a Lecturer and an Associate Professor. His current research interests include fuzzy systems and control, neural networks, machine learning, decision support systems, intelligent systems, and data mining. Dr. Zeng is an Associate Editor of the IEEE TRANSACTIONS ON FUZZY SYSTEMS, a member of the editorial board of the International Journal of Computational Intelligence Research, and a member of the peer review college of the U.K. Engineering and Physical Sciences Research Council.

Hong Qiao received the B.Eng. degree in hydraulics and control and the M.Eng. degree in robotics from Xi’an Jiaotong University, Xi’an, China, the M.Phil. degree in robotics control from the Industrial Control Center, University of Strathclyde, Strathclyde, U.K., and the Ph.D. degree in robotics and artificial intelligence from De Montfort University, Leicester, U.K., in 1995. She was a University Research Fellow with De Montfort University from 1995 to 1997. She was a Research Assistant Professor from 1997 to 2000 and an Assistant Professor from 2000 to 2002 with the Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong, China. Since January 2002, she has been a Lecturer with the School of Informatics, University of Manchester, Manchester, U.K. Currently, she is also a Professor with the Laboratory of Complex Systems and Intelligent Science, Institute of Automation, Chinese Academy of Sciences, Beijing, China (on leave from the University of Manchester). She first proposed the concept of “the attractive region in strategy investigation,” which has successfully been applied by herself in robot assembly, robot grasping, and part recognition. The work has been reported in Advanced Manufacturing Alert (Wiley, 1999). Her current research interests include information-based strategy investigation, robotics and intelligent agents, animation, machine learning (neural networks and support vector machine), and pattern recognition. Dr. Qiao is a member of the Program Committee of the IEEE International Conference on Robotics and Automation from 2001 to 2004. She is currently an Associate Editor of the IEEE TRANSACTION ON SYSTEMS, MAN, AND CYBERNETICS, PART B and the IEEE TRANSACTION ON AUTOMATION SCIENCE AND ENGINEERING.

Iain Buchan Professor of Public Health Informatics and Director of the Northwest Institute for Bio-Health Informatics, at the University of Manchester, and an honorary Consultant in Public Health in the English National Health Service (NHS). He has backgrounds in clinical medicine, public health and computational statistics, and runs a multi-disciplinary team bridging health sciences, computer science, social science, management science and mathematics. His work centres on the research and development of informatics methods for understanding and improving the public’s health, partly focusing on applications in obesity and metabolic health. He works closely with the English NHS to develop population-based e-infrastructure to enable both: large-scale, realistically-complex epidemiology; and future e-health interventions.

John Keane holds the MG Singh Chair of Computing Science at the University of Manchester and holds an Honorary Chair in Decision and Information Sciences at Manchester Business School. His primary research interest is high performance data mining. He is an Associate Editor of IEEE Transactions on Systems, Man and Cybernetics (Part C), and has been an expert advisor to the UN/World Bank, the Irish Research Council, the British Council and the Finnish Research Academy.