

ALTERNATIVE APPROACH TO DETERMINE BLOOD LIPIDS WITH INTELLIGENT SYSTEMS

Petra POVALEJ, Peter KOKOL
University of Maribor – FERI
Laboratory of System Design
Smetanova 17, SI-2000 Maribor
Slovenia

{petra.povalej, kokol}@uni-mb.si

Abstract

Invasive medical examinations (like blood examination) are often rather unpleasant especially for children. Therefore an attempt was made to find a non-invasive method for determining the level of blood lipids in children. With that goal in mind two different machine learning methods were used on the database of five-year old children. The results gained with both methods were compared and will be presented in this paper.

1. Introduction

The role of cholesterol in infants and the measurement and management of blood cholesterol in children is a subject of controversy among physicians. Several scientific evidences are supporting relationship between elevated blood cholesterol in children and high cholesterol in adults and development of adult arteriosclerotic diseases such as cardiovascular and cerebrovascular disease.

For most of the children every medical examination causes anxiety. Therefore finding some painless, timesaving methods as a substitute for expensive and time-consuming medical examinations would be a great success.

In this paper we will present current results of our efforts to find a non-invasive method for determining the level of lipids in children. This research was performed in cooperation with the Adolf Drolc Health Centre in Maribor, where 729 five-year-old children were examined. For each child the following data was gathered: gender, height, weight, head circumference, chest circumference, upper arm circumference, skinfold thickness, pulse, systolic and diastolic blood pressure, blood cholesterol level, LDL and HDL cholesterol level and triglycerides. Our aim was to find some relations among these attributes and the blood cholesterol level in order to predict (with a considerate degree of accuracy) whether the child has elevated blood cholesterol without blood testing.

2. Blood lipids

Lipids are fats in the bloodstream and in all of the body's cells. Among the components of the lipids in blood are triglycerides, which come from the fats eaten or being made by our body from other things eaten, including carbohydrates. If the calories consumed are not used immediately for energy, they are stored as triglycerides in the fat cells. They are released when the body needs energy, such as between meals. Having too many triglycerides has been linked to coronary artery disease.

A certain amount of cholesterol is important to the healthy function of our body. It is an oily substance that is used to build cell walls and form some hormones and tissues. High level of cholesterol in the blood, known as hypercholesterolemia, is a major risk factor for heart disease and can lead to a heart attack.

We accumulate cholesterol in two ways. The liver produces about 1,000 milligrams of cholesterol a day. Another 150 to 250 milligrams comes from the foods we eat.

Cholesterol and triglycerides are carried in the bloodstream by lipoproteins. Two kinds - low-density lipoproteins (LDL) and high-density lipoproteins (HDL) - are the most important.

Low-density lipoproteins, sometimes called "bad" cholesterol, are the primary cholesterol carrier. If there is too much LDL in the bloodstream, it can build up on the walls of the arteries that lead to the heart and the brain. This buildup forms plaque, a thick, hard substance that can block arteries. If a blood clot forms and gets jammed in a clogged artery leading to the heart or the brain, you could have a heart attack or a stroke.

The remainder of body's cholesterol, about one-third to one-fourth of it, is moved through the blood by high-density lipoproteins, or HDL. These are sometimes known as "good" cholesterol because they carry the cholesterol away from the arteries and back to the liver, where it is passed from the body.

High levels of LDL cholesterol (the bad cholesterol) are related to a risk for heart disease and stroke, whereas high levels of HDL cholesterol (the good cholesterol) can protect against these. Conversely, low levels of LDL cholesterol are good and low levels of HDL cholesterol are bad.

High levels of triglycerides and cholesterol are a major risk factor for coronary artery disease, also known as atherosclerosis. An adult's heart attack or stroke has its origins in the development of atherosclerosis, which begins in earnest during the late teen years. Paying attention to cholesterol levels in children can lead to proper diet and medical treatment throughout life, slowing the progress of atherosclerosis, and either delaying or preventing heart attacks and stroke.

Childhood cholesterol levels were not tracked until recently, and some experts think that high cholesterol in kids is a major underreported public health problem. The health risks associated with high cholesterol - heart disease and stroke, for example - generally don't show up for years, even decades, so making the connection between kids and cholesterol is difficult for many people. Children with elevated cholesterol levels may be a precursor, some doctors believe, to a generation of teenagers with cardiovascular disease.

3. Methods

Machine learning is all about learning rules from the data with a purpose to do predictions as well as analysis. As described in [1], machine learning is about learning from the past experiences with respect to some performance measure.

Various techniques have been developed on machine learning. These include neural networks, genetic algorithms, decision trees, hybrid systems, etc.

In the next section two machine learning methods used in our research will be shortly described. Each of the techniques is essentially about learning from prior knowledge from several training objects and is based on decision trees.

Decision trees are used extensively for classification. In practice they are often used for decision support in various fields like medicine, economy, etc [7]. A decision tree has one big advantage in compare with other machine learning techniques: very simple and clear representation of the path to acquired decision. It is mostly for that reason the decision trees are very often used for decision support in medicine.

A decision tree is inducted on a training set, which consists of training objects. Every training object is completely described by a set of attributes (object properties) and class (decision, outcome). Attributes can be numeric or discrete, but numeric attributes are not suitable for learning a tree. Therefore they must be mapped into a discrete space.

There are two types of nodes in a decision tree: internal and external nodes. Each internal node (non-terminal node) contains a test of a specific attribute value. External nodes (terminal nodes, decision nodes, leaves) are labeled with a class, which represents a decision. Nodes are connected with edges (links). Edges are labeled with different outcomes of a test performed on an attribute in a source node.

For testing a decision tree a testing set is used. Testing set consists of testing objects described with the same attributes as training objects except that testing objects are not included in training set.

The results are described with specificity, sensitivity and total accuracy. Specificity is defined as the number of correctly classified children with normal cholesterol level divided by the number of all children with normal cholesterol level. Sensitivity is the number of correctly classified children with abnormal cholesterol level divided by the number of all children with abnormal cholesterol level. The overall quality of a decision tree is described with total accuracy.

3.1. Classic decision tree induction

Classic decision tree induction uses legendary Quinlan's ID3 algorithm that was presented in 1986. The basic idea behind the ID3 is that in the decision tree at each node should be associated a non-categorical attribute which is most informative among the attributes not yet considered in the path from the root. Entropy is used to measure how informative is a node.

The tool we used is called MtDecit 3.0 [3]. It basically follows the same principles as many other decision tree building tools, but additionally implements different ways of numeric attribute's discretization [4].

3.2. Genetically induced decision trees

Disadvantages of classic decision tree induction such as sensitivity to noise (missing or corrupted data) [9], encouraged us to try another method of machine learning that combines two methods: decision tree induction and genetic algorithms. This hybrid method merges all advantages of both methods and therefore usually gives better results.

Genetic algorithms are based on the evolutionary ideas of natural selection and genetic processes of biological organisms [5,9]. They are often capable of finding optimal solutions even in most complex search spaces or at least they offer significant benefits over other search and optimization techniques.

The first phase of genetic process is generation of initial population. Enough individuals have to be constructed to fulfill the whole population. Every individual in this method is represented as a decision tree.

Second phase of genetic process is evolution of population with the use of three genetic operators. First genetic operator is selection when individuals are evaluated on the basis of fitness function and the best ones (parents) are chosen for creating new individuals (children) with a second genetic operator - crossover. This way a new population is created.

After the new individual is constructed by crossover, a genetic operator of mutation is applied with certain (low) probability. Mutation serves as a random change of individuals with intention to find an optimal solution to the given problem faster and more reliably. The tool we used in research is Vedec [6].

4. Data collection

The database used in the study included 729 subjects described with 14 properties (gender, height, weight, head circumference, chest circumference, upper arm circumference, skinfold thickness, pulse, systolic and diastolic blood pressure blood cholesterol level, LDL cholesterol level, HDL cholesterol level and triglycerides). The last four properties were defined as outcomes (class attributes). All other properties were used as attributes.

Because of the nature of decision trees, numeric class attribute had to be mapped into two discrete values (normal/abnormal). For discretization standard recommended values presented in table 1 were used.

Table 1: Normal levels for blood lipids in children

<u>Lipids</u>	<u>Normal level</u>
Blood cholesterol level	< 5 mmol/l
HDL cholesterol level	> 1 mmol/l
LDL cholesterol level	< 3 mmol/l
Triglycerides	< 1,4 mmol/l

All objects in the database with more than 90% of missing parameters and all objects with unknown class attribute were deleted, so the new database has remaining 712 objects.

4.1. Training and test sets

For the training purposes a training set had to be build. The percentage of objects with abnormal levels of blood cholesterol, HDL and LDL cholesterol and triglycerides were substantially lower than 50% (see table 2). Such distribution could influence decision trees in such a way that they would learn more about normal blood lipids than abnormal.

Table 2: The number of objects with normal / abnormal levels of blood lipids

<u>Class</u>	<u>Triglycerides</u>	<u>Blood cholesterol</u>	<u>HDL cholesterol</u>	<u>LDL cholesterol</u>
NORMAL	680	539	565	436
ABNORMAL	32	173	147	276

Therefore the training set was carefully built in such a way that all classes of an outcome attribute were represented equally.

For a purpose of assessing decision trees constructed from the training set a test set was used. The test set included all objects from initial database that were not used for training purposes.

5. Results

Many different experiments have been conducted in order to obtain some useful results and to compare efficiency of used methods.

The most interesting experiments will be shortly described in the next section.

Experiment No.1

From the table 2 it can be seen that the number of abnormal cases of triglycerides was too small for experimenting therefore this attribute was deleted from the database. Consequently in our first experiment the goal was to predict the level of blood cholesterol, HDL cholesterol and LDL cholesterol. Therefore all three attributes were defined as an outcome with 8 different classes. The training and test sets were built and applied to both tools.

Different types of discretization were used for classic decision tree induction, but the dynamic discretization on the whole data set gave best results. Both methods resulted in poor accuracy. The best decision tree in fact was genetically induced and it classified testing objects with the total accuracy of 46%. The method of classic tree induction produced even worse results.

Experiment No.2

Based on the poor results of the first experiment some of the classes were combined following the evaluation of the physicians. For that reason an output attribute “degree” was defined. Different combinations of blood cholesterol level, LDL and HDL levels were evaluated with a degree from 1 to 4, where 1 represents the worst possible combination of cholesterol levels and 4 represents the best (see table 3) from the medical point of view

Table 3: Evaluation of comparison among cholesterol levels with a degree from 1 to 4, where 1 represents the worst state of child’s blood lipids an 4 represents the best

<u>Blood cholesterol level</u>	<u>HDL cholesterol level</u>	<u>LDL cholesterol level</u>	<u>Degree</u>	<u>Number of objects</u>
abnormal	abnormal	abnormal	1	129
abnormal	abnormal	normal	2	0
abnormal	normal	abnormal	3	485
abnormal	normal	normal	3	
normal	normal	abnormal	3	
normal	abnormal	normal	3	
normal	abnormal	abnormal	3	
normal	normal	normal	4	98

Once again a training set including 60 randomly chosen objects from each of the outcome classes was built. All other objects were used in test set. The decision trees induced with both methods had highest total accuracies 50%.

Experiment No. 3

In the third experiment the number of classes in outcome attribute was reduced even more. All three cholesterol levels (blood cholesterol level, HDL and LDL cholesterol level) were combined in one outcome – “lipids” with two possible values:

- lipids = “normal” if an object has normal all three cholesterol levels or
- lipids = “abnormal” otherwise.

In the database was approximately the same proportion of object with normal (349) and abnormal lipids (363).

Training set included 178 objects with normal lipids and the same number of objects with abnormal lipids. All other objects were included in a testing set. The best results are presented in table 4. All decision trees had very low accuracies.

Table 4: Results of the comparison of two methods applied on training and test sets for determining the level of lipids in blood

Method	Specificity	Sensitivity to abnormal lipids.	Total accuracy
Classic decision tree induction	75%/ 53,8%	73,5% / 53%	73%/ 53,4%
Genetically induced dec. trees	73,5%/ 50,8%	70,7%/ 53,5%	72,2%/ 52,3%

Experiment No. 4

Poor results stimulated us to add a new attribute to objects in the data set – Roher index (ROI), which is similar to body mass index (BMI), but often used with children because it correlates less with height than BMI but equally well with skinfold thickness. It is calculated as follows:

$$ROI = \frac{weight}{height^3} .$$

Children with $ROI > 1,5$ are obese and those with $ROI < 1,1$ are lean. On this bases three values for the new attribute ROI (obese, normal and lean) were defined. Then both decision tree induction methods were used on new data sets (training and test set).

Nonetheless the accuracy of induced decision trees was not much higher than without ROI (see table 5).

Table 5: Results of the comparison of two methods applied on training and test sets with added attribute ROI for determining the level of lipids in blood

Method	Specificity	Sensitivity to abnormal lipids	Total accuracy
Classic decision tree induction	73,3%/ 55%	70,4% / 53%	72%/ 54%
Genetically induced dec. trees	71,3%/ 58,4%	70,2%/ 51,8%	70,8%/ 55%

Our next attempt to improve results was distribution of the initial database in three data sets according to Roher index. First data set included objects with ROI=obese (175 objects), second data set included objects with ROI=normal (530 objects) and third with ROI=lean (7 objects). The third data set was too small for experimenting, so only the first two were used. Both methods of decision tree induction were used separately for objects with ROI=obese and for objects with normal ROI.

The training set for objects with ROI=obese included 100 objects with the same proportion of object with normal and abnormal lipids. All other objects were included in a test set (normal lipids: 21 objects, abnormal lipids: 54 objects). Once again the results were bad and the classic decision tree induction was just a little bit better than genetic algorithms (see table 6).

Table 6: Results of the comparison of two methods applied on data sets where only objects with ROI=obese are included

Method	Specificity	Sensitivity to abnormal lipids	Total accuracy
Classic decision tree induction	78% / 57,1%	75% / 53,7%	76% / 54,7%
Genetically induced dec. trees	88% / 52,3%	78% / 51,8%	83% / 52%

Objects with ROI=normal were also divided into training and test data sets. 200 objects were included in the training dataset with the same proportion of normal and abnormal level of lipids. The test set included all other objects. In the table 7 it can be seen that there was no significant difference between the accuracies for objects with ROI=obese and objects with ROI=normal.

Table 7: Results of the comparison of two methods applied on data sets where only objects with ROI=normal are included

Method	Specificity	Sensitivity to abnormal lipids	Total accuracy
Classic decision tree induction	78% / 57,1%	70% / 53,7%	75% / 54,7%
Genetically inducted dec. trees	69% / 52,5%	80% / 52,5%	74,5% / 52,4%

At the end of the fourth experiment it was concluded that combining all cholesterol levels in one output leads to bad results. Both methods of decision tree induction gave similar accuracies.

Experiment No.5

In our last experiment our research was limited to the outcome on blood cholesterol level only. LDL and HDL cholesterol were eliminated from our database. In order to improve the power of classifiers (see table 2) the number of objects with normal blood cholesterol level was reduced in the training set so that both values were represented in approximately the same proportion (149 classified as ‘normal’ and 108 classified as ‘abnormal’). All other objects were used for testing purposes.

The best decision tree classified test objects with 72,6% total accuracy, but the sensitivity to abnormal cholesterol level was only 37,5%. That shows that the decision tree specialized for classifying objects with normal cholesterol level (specificity 74,3%) and therefore it is not applicable for practical usage.

Consequently a hybrid method of genetically induced decision trees was used in order to achieve better results. The problem of classifying objects with abnormal cholesterol level was not solved either.

As in previous experiment Roher index was added as a new attribute in object’s description. The most interesting results are shown in the table 8.

Table 8: Results of the comparison of two methods applied on data sets with new attribute (Roher index) added to the description of objects. Accuracies of induced decision trees are represented for training set / testing set

Method	Specificity	Sensitivity to abnormal chol.	Total accuracy
Classic decision tree induction	74,5% / 60,3%	41,7% / 63,2%	60,7% / 60,5%
Genetically inducted dec. trees	77,1% / 47,3%	75,9% / 52,6%	76,7% / 47,6%

The results were (as expected) better than before. Sensitivity to abnormal cases of blood cholesterol level was much higher in both methods, but genetic induction of decision trees was less accurate on the testing set than classic decision tree induction.

Since height and weight were already included in the calculation of Roher index, it was presumed that they themselves might not be regarded as an influencing factor for blood cholesterol level determination. Therefore they were excluded from object’s description. The best results of inducing decision trees using classic and genetic method are presented in table 9.

Table 9: Results of the comparison of two methods applied on data sets with added attribute (Roher index) and attributes height and weight excluded from an object's description

Method	Specificity	Sensitivity to abnormal chol.	Total accuracy
Classic decision tree induction	91,9% / 62,8%	74,1% / 52,6%	84,5% / 62,3%
Genetically induced dec. trees	87,2% / 65,3%	43,5% / 42,1%	68,9% / 63,8%

In comparison with previous results it can be seen that new decision trees were less accurate. Therefore it can be established that height and weight have some influence on blood cholesterol level.

Further the database was divided in two data sets on the basis of ROI. After examining both data sets it was established that within obese children 75,5% had normal cholesterol level and 24,5% had abnormal cholesterol level. Among children with normal Roher index 79,6% had normal cholesterol level and 20,4% had abnormal.

Each data set was first divided into training and test sets considering the ratio of objects with normal cholesterol level and object with abnormal cholesterol level in the training set (see table 10).

Table 10: The Number of objects in training and testing sets for two data sets: first with obese children and second with normal children according to Roher index

	ROI = obese		ROI = normal	
	Normal chol.	Abnormal chol.	Normal chol.	Abnormal chol.
Training set	55	40	130	110
Testing set	70	10	274	16

First the training set for obese children was used with both tools. The results are described in table 11 where it can be seen that genetically induced decision tree has higher accuracy than classic decision tree. The decision tree induced with genetic algorithms classified test objects with accuracy over 70%, which is high enough to be used for medical purposes.

Table 11: Results of the comparison of two methods applied on data sets where only objects with ROI=obese are included

Method	Specificity	Sensitivity to abnormal chol.	Total accuracy
Classic decision tree induction	84,6% / 61,6%	85,7% / 71,4%	85,1% / 62,5%
Genetically induced dec. trees	86,8% / 78%	65,5% / 71,4%	77,6% / 77,5%

After encouraging results we were inquisitive if the results would change for the better in case of excluding weight and height from the list of attributes in object's description (see table 12). As it was established before the results are better when the height and weight are included in the induction of decision trees.

Table 12: Results of the comparison of two methods applied on data sets where only objects with ROI=obese are included and attributes weight and height are excluded

Method	Specificity	Sensitivity to abnormal chol.	Total accuracy
Classic decision tree induction	100% / 69,9%	82,1% / 71,4%	92,5% / 70%
Genetically induced dec. trees	86,8% / 72,6%	68,9% / 71,4%	79,1% / 72,6%

The same experiments that were conducted with the obese children data set were then repeated with the data set that included only children with normal Roher index. The results are presented in tables 13 and 14.

Table 13: Results of the comparison of two methods applied on data sets where only objects with ROI=normal are included

Method	Specificity	Sensitivity to abnormal chol.	Total accuracy
Classic decision tree induction	94,5% / 61,6%	70,5% / 71,4%	84,6% / 62,5%
Genetically inducted dec. trees	88,1% / 49,1%	73% / 72,7%	81,9% / 50,2%

Table 14: Results of the comparison of two methods applied on data sets where only objects with ROI=normal are included and attributes weight and height are excluded

Method	Specificity	Sensitivity to abnormal chol.	Total accuracy
Classic decision tree induction	91,8% / 62,9%	73,1% / 54,5%	84% / 62,5%
Genetically inducted dec. trees	88,1% / 52,9%	60,2% / 54,5%	76,6% / 53%

Surprisingly the decision trees for children with normal Roher index were less accurate than decision trees for obese children. Accuracies on training sets were much higher than accuracies on testing set therefore those decision trees cannot be useful in practice.

The results of this experiment show that Roher index has an influence on determining blood cholesterol level. The decision tree induction on the dataset with obese children by Roher index was more successful than with children classified as normal according to Roher index. From the comparison of the methods used for decision tree induction can be seen that genetic algorithms usually gave decision trees with higher accuracy than classic decision tree induction.

6. Discussion

After examining the database it was expected that better results will be gained with the use of genetic algorithms because of their insensibility to missing and corrupted data.

First we tried to classify objects according to the level of blood cholesterol, HDL and LDL cholesterol, but there were too many classes for outcome attribute and therefore the results were poor. Reducing the number of classes in outcome attribute lead us to a little bit better results.

Higher results were gained when we reduced our classification on blood cholesterol level only. The initial database was divided into three data sets according to Roher index: first for obese, second for normal and third for lean children. The accuracy that was achieved during our experiments on obese children was substantially higher than the accuracy achieved on children with normal Roher index. The only reason for that which arises at the moment is that with obese children the influencing factors for determining blood cholesterol level are clearer. From the results gained with all four experiments it can be concluded that the attributes in our database are not significant enough for determining cholesterol levels in children using presented methods of machine learning.

Experiments for determining cholesterol level with the use of machine learning have not yet been performed on the children so young. That is why this research is very interesting from medical point of view.

7. Conclusion

In this paper two methods of machine learning (classic decision tree induction and decision tree induction with genetic algorithms) were compared on the problem of determining blood lipids in children. Presented results show that with the use of genetics the induction of decision trees was in most cases more successful than classic decision tree induction. Many different experiments were conducted. The best results were obtained in our last experiment where we tried to determine blood cholesterol level on the data set with only obese children included (according to Roher index). These results show that the obesity has an influence on the level of blood cholesterol.

To summarize we have reached the conclusion that attributes used in the database are not enough for determining cholesterol level in children. For that purpose we should obtain some other attributes that are more significantly linked with the cholesterol level.

In the future we will try some other new methods of machine learning such as rough sets on the present database and we are expecting some interesting results.

References

- [1] Mitchell T.: Machine Learning, Addison Wesley, MA, 1997.
- [2] J. R. Quinlan: C4.5: Programs for machine learning. Morgan Kaufmann publishers, San Mateo, CA, 1993.
- [3] Zorman Milan, Hleb Špela, Šprogar Matej: Advanced tool for building decision trees MtDecit 2.0. In: Kokol Peter (ed.), Welzer-Družovec Tatjana (ed.), Arabnia Hamid R (ed.). International conference on artificial intelligence, June 28 – July 1, 1999, Las Vegas, Nevada, USA. Las Vegas: CSREA, (1999), book. 1: 315-318.
- [4] Zorman Milan, Kokol Peter: Dynamic discretization of continuous attributes for building decision trees. In: Fyfe C. (ed.). Proceedings of the second ICSC symposium on engineering of intelligent systems, June 27-30, 2000, University of Paisley, Scotland, U.K.: EIS 2000. Wetaskiwin; Zürich: ICSC Academic Press, (2000): 252-257.
- [5] Baeck T.: Evolutionary Algorithms in Theory and Practice, Oxford University Press, Inc., 1996.
- [6] Sprogar Matej, Kokol Peter, Zorman Milan, Podgorelec Vili, Yamamoto Ryuichi, Masuda Gou, Sakamoto Norihiro: Supporting Medical Decisions with Vector Decision Trees In: V: PATEL, V. L. (ur.), ROGERS, R. (ur.), HAUX, R. (ur.). 10th World Congress on Medical Informatics MEDINFO, London, 2001. MEDINFO 2001: proceedings of the 10th World congress on medical informatics, London, UK, 2-5 September, 2001, Amsterdam: IOS Press: Ohmsha, (2001): 5.
- [7] Kokol Peter, Završnik Jernej, Zorman Milan, Malčić Ivan, Kancler Kurt: Participative Design, Decision Trees, Automatic Learning and Medical Decision Making. In: Brender J. (ed.). Medical Informatics Europe '96, (Studies In Health Technology And Informatics, Vol.34). Amsterdam [Etc.]: Ios Press; Tokyo: Ohmsha, (1996): 501-505.
- [8] Stewart, Kerry J., Carol S. Brown, Colleen M. Hickey, Linda D. McFarland, John J. Weinhofer, and Sheldon H. Gottlieb. Physical fitness, physical activity, and fatness in relation to blood pressure and lipids in pre-adolescent children: Results from the FRESH Study. *Journal of Cardiopulmonary Rehabilitation*, (1995): 122-129.
- [9] Podgorelec, Vili, Kokol, Peter. Induction of medical decision trees with genetic algorithms. In: Proceedings of the international ICSC congress on Computational intelligence methods and applications, June 22-25, 1999, Rochester, N.Y. USA. [s. l.]: ICSC - International Computer Science Conventions, (1999): 7.