Using Support Vector Machines to Recognize Changes Characteristic to Obesity in Laboratory Results

T. Ferenci[1], L. Kovács[1], B. Benyó[1] and A. Kovács[2]

[1] Dept. of Control Engineering and Information Technology, Budapest University of Technology and Economics, Budapest, Hungary
[2] "Politehnica" University of Timisoara, Dept. of Mathematics, Timisoara, Romania

Abstract—**Obesity is an endemic in most part of the developed word. As the increased occurrence of many serious comorbodity (stroke, IHD, NIDDM) is casually linked to obesity, it represents a huge risk from the public health point of view. Focus should be placed on early identification of risked individuals, which calls for effective screening methods. This is largely hindered by the fact that current diagnostic methods are of either poor predictive value (anthropometric indices etc.) or unavailable for large-scale screenings (BIA, DXA etc.) Our idea to resolve this problem is based on the fact that obesity has a marked effect on many of the routinely used laboratory parameters. An obvious example is the serum level of various blood lipids: hyperlipidemia, hypercholesteremia are often observed in obese people. On the grounds of this fact, we hypothesized that obesity can be predicted using only routine laboratory parameters. Furthermore, we presumed that those are the best parameters to predict obesity (and obesity-associated risk), that best separate manifestly obese and healthy people. Our research aimed to investigate this possibility on adolescent population, as they are the most important from the public health point of view. To that end, we performed a cross-sectional clinical study that included the observation of n=148 male children (aged 12-16 year), consisting of healthy volunteers from four Hungarian secondary schools and obese patients treated with E66.9 "Obesity, unspecified" diagnosis. Observation included the recording of 27 laboratory parameter from a fasting blood sample. To investigate how well obese and healthy children can be separated based solely on their laboratory results, we employed a state-of-the-art classification tool, Support Vector Machines (SVM). Results were compared with the more classical approach of multivariate logistic regression. SVM's major drawback is its "black-box" nature; however, we concluded that its performance is excellent, superior to logistic regression.**

Keywords— **obesity, biostatistics, laboratory results, classification, Support Vector Machines (SVM).**

I. Introduction

A. About Obesity

Obesity [1] is considered an endemic in Hungary, just as in most part of the developed world. According to [2], 1.5 million Hungarian is definitely obese, whilst an additional 2.7 million can be considered overweight. Together, this is almost the half of the Hungarian population. To assess the risk of obesity we need to take into account the severity of the problems that obesity might cause, as risk is usually interpreted as the product of probability of occurrence and severity. In the last few decades, increased occurrence of several morbidities was casually linked to obesity, such as non-insulin dependent diabetes mellitus (NIDDM), stroke, ischemic heart disease (IHD), immunological and reproductive dysfunctions and certain neoplasms [3]. From these, stroke and IHD is amongst the first three in the list of mortality rates in Hungary [4].

These all confirm that obesity poses a significant risk, high enough to be interesting from the public health point of view as well. Early intervention is needed to prevent the onset of obesity and obesity-associated comorbidities, which intervention should be focused to be the most effective. Early, focused intervention presumes however an efficient screening method which can select those people who are prone to obesity even when they are healthy otherwise. This screening has many difficulties in practice, starting as early as how to measure obesity.

Measuring obesity and overweight is a complex task, hindered significantly by the varying definitions. However, as a bottom line, it is accepted that both state is characterized by elevated quantities of fat tissue, but we are still lacking exact, strict, widely used diagnostic criteria.

While indirect indicators, such as body mass index (BMI) or waist circumference (WC), are widely used [5], their predictive power is limited and they are especially unfit to predict obesity-related risk. Direct indicators, such as dual-emission X-ray absorptiometry (DXA) or bioelectric impedance analysis (BIA) can precisely characterize fat distribution on the other hand, giving a sound estimation of obesity-related risk. However, they are simply unfit for screening in wider populations, primarily either due to radiation exposure they cause or their expensiveness [6].

B. Our research project

A joint research project (started in 2007) of the Biomedical Engineering Laboratory of the Budapest University of Technology and Economics and the Heim Pál Children's

Á. Jobbágy (Ed.): 5th European IFMBE Conference, IFMBE Proceedings 37, pp. 215–218, 2011.
www.springerlink.com

Hospital (Budapest) aims to resolve this problem by investigating new ways in which obesity-related risks can be assessed. The combination of different techniques (anamnesis recording, anthropometric measurements, laboratory parameters, bioelectric impedance analysis) is the ultimate goal of the research. It is focused on children, especially on adolescent population (aged 14-18 years), as this is the optimal target population on which early intervention could and should be performed to best prevent the adverse health consequences.

This research project included a multicenter clinical observation of healthy and obese population focused on the abovementioned parameters. Such study has not yet been performed on Hungarian adolescent population from this aspect.

C. Problem statement

The study included the recording of routine blood test parameters of both obese and healthy subjects. This part of the database consists of 27 laboratory parameters.

The question arises whether these parameters alone are sufficient to recognize obesity. This is interesting because a positive answer would justify – in a multivariate view – that obesity causes systematical changes in laboratory results.

The rest of the paper is organized as follows. In Section II we will describe our clinical observation and review the methods used to solve the above problem: the support Vector Machines (SVMs). In Section III we will present the results and provide a short discussion of them. Finally, in Section IV we conclude the paper with a summary and suggest a few possibilities of further research.

II. Materials and methods

A. Our clinical study

We collected data from *n=393* subjects aged between 14 and 18 years, including both healthy and obese ones during our study.

The *healthy control group* consisted of volunteers from several Hungarian secondary schools. Each child participated with full written informed and the study was preauthorized by the Hungarian Regional Bioethical Commission. Participants were required to show up for fasting examination early in the morning. Examinations of healthy volunteers included anthropometrical measurements, body composition analysis (with InBody 3.0 multi-frequency bioelectric impedance analyzer), blood sample drawing for standard laboratory parameters and anamnestic data recording. Measurements were carried out by doctors and nurses of the Heim Pál Children's Hospital and results were manually recorded in electronic format [7].

The *obese group* consisted of children treated in the Heim Pál Children's Hospital, with their main diagnosis being E66.9 (according to ICD-10) “Obesity, unspecified”, with no comorbidity or significant other illness. Data (including laboratory parameters) of the obese children were extracted from the hospital's electronic records with a custom application developed by the authors [7].

In this paper, we use laboratory results extracted from the compiled database including 27 laboratory parameters. Their list, along with their basic descriptive indicators is shown in Table 1. (The abbreviations follow the international customs.)

For this analysis, only male subjects were used (as their number was far greater then female subjects), and we wanted to avoid mixing them, as sex has a profound impact on many laboratory parameter, which could have made the database heterogeneous [8].

Subject with missing value were excluded from the investigation. The remaining *n=148* subject included 64 healthy and 84 obese boys.

Table 1 Variables (laboratory results) of the database, and its basic descriptors in Mean ± SD (*n*) format

Variable	Healthy group	Obese group
AbsNeutrophilCount	3,441 ± 0,939 (64)	4,095 ± 1,232 (84)
AbsLymphocyteCount	2,267 ± 0,484 (64)	2,755 ± 0,603 (84)
AbsMonocyteCount	0,638 ± 0,152 (64)	0,712 ± 0,223 (84)
AbsEosinophilCount	0,185 ± 0,0959 (64)	0,264 ± 0,21 (84)
RBCCount	5,346 ± 0,281 (64)	5,178 ± 0,347 (84)
Haemoglobin	156,2 ± 9,225 (64)	142 ± 12,4 (84)
Haematocrit	0,463 ± 0,0218 (64)	0,423 ± 0,0317 (84)
MCV	86,76 ± 2,676 (64)	81,62 ± 3,848 (84)
MCH	29,23 ± 1,02 (64)	27,44 ± 1,648 (84)
MCHC	337 ± 8,522 (64)	336,2 ± 10,55 (84)
RDW	13,41 ± 0,619 (64)	13,88 ± 0,835 (84)
PlateletCount	238,6 ± 58,34 (64)	295,5 ± 57,69 (84)
MPV	10,86 ± 0,816 (64)	10,42 ± 0,827 (84)
CRP	1,007 ± 1,446 (64)	5,845 ± 9,545 (84)
seSodium	139 ± 1,442 (64)	138,3 ± 1,997 (84)
sePotassium	4,052 ± 0,341 (64)	4,336 ± 0,297 (84)
seChloride	102,1 ± 1,488 (64)	102,8 ± 2,133 (84)
seTotalProtein	78,12 ± 3,902 (64)	75,97 ± 4,402 (84)
seAlbumin	51,32 ± 2,756 (64)	47,78 ± 2,629 (84)
seUrea	4,767 ± 0,952 (64)	4,476 ± 1,048 (84)
seCreatinine	77,64 ± 9,879 (64)	61,1 ± 9,509 (84)
Triglycerides	0,921 ± 0,484 (64)	1,302 ± 0,656 (84)
TotalCholesterol	3,791 ± 0,747 (64)	4,216 ± 0,922 (84)
HDLCholesterol	1,287 ± 0,223 (64)	1,149 ± 0,226 (84)
GOT	21,44 ± 5,5 (64)	25,9 ± 8,318 (84)
GPT	20,38 ± 10,03 (64)	30,73 ± 19,46 (84)
GGT	23,28 ± 10,73 (64)	26,48 ± 12,75 (84)

Statistical analysis and visualization was performed with the R [9] statistical program package (version 2.12.0), including custom scripts (available from the corresponding author on request).

SVMs were also simulated within R, using the e1071 library that is based on the LibSVM [10] implementation.

B. Support Vector Machines

Support Vector Machine (SVM) [11], [12] represents a state-of-the-art classification (and regression) method that is essentially a linear classifier aiming to find a separating hyperplane that maximizes generalizability by maximizing distance from the training points. Using soft margin approach, SVM is applicable for cases were exact linear separation cannot be achieved. This maximization can be performed by the method of Lagrange multipliers, leading to a quadratic programming problem.

In addition, SVMs use a so called kernel trick [13], i.e. data is analyzed not in its feature space, but in a higher-dimensional (possibly infinite dimensional) kernel space. Hence, SVM is largely resistant to the "curse of dimensionality". We used Gaussian kernel in our investigation.

SVMs represent a very powerful, yet medically rarely used classification method. The reason why they are not employed frequently in medical literature is that their operation is largely "black-box" in nature: no matter how well they work (i.e. classify) their operation cannot be interpreted in any meaningful way.

Now we used SVMs despite these concerns in order to provide an "upper bound", i.e. a reference value for classification error rate to which the results of other (white-box) classification methods can be compared.

C. Classification with other classifiers

Examining the same database, we found [14] that multivariate logistic regression, a well known, completely white-box (and medically widely used) classical classification model has an overall true classification rate of 87,2% (if variable selection is employed) and 89,6% (if it is combined with Partial Least Squares (PLS) regression to provide dimension reduction, instead of the variable selection).

III. RESULTS AND DISCUSSION

SVM has two parameters that control its operation (and, hence, determines its performance): C that sets the trade-off between margin width and error penalty, and γ, which is the width parameter of the Gaussian kernel.

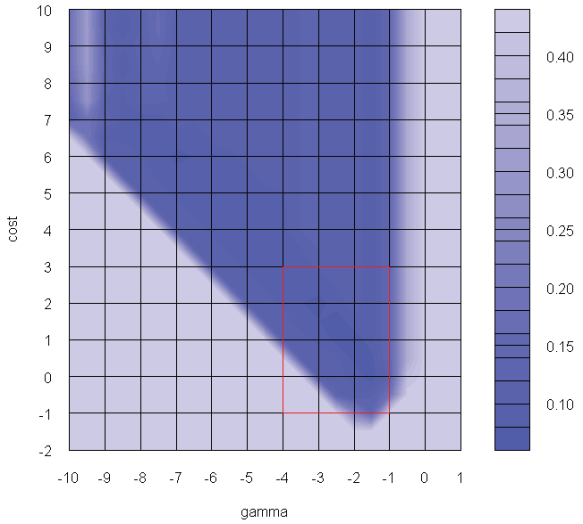

Fig. 1. Classification error with different parameters of SVM. (Parameters are shown on a logarithmic scale). Red border shows the best area.

The first problem is to find their optimal value of these parameters for our database. We employed a simple grid search strategy: we scanned the C-γ space and recorded classification errors in the points of our grid. This error was measured with 10-fold crossvalidation (to avoid overfitting). This was a bare minimum here, as hundreds of trainings were performed. Initial results are shown on Fig. 1.

Results were then refined by using a finer grid on promising areas (signed with red border on Fig. 1.). For the area indicated on Fig. 1, result is shown on Fig. 2.

Repeating this procedure, we concluded that best parameters for this database are: $C=7.9$ and $\gamma=0.005$.

Using these parameters to classify now the entire dataset, we obtained an overall true classification rate of 92.1% using 10-fold crossvalidation.

ROC-curve is shown on Fig. 3.; using a cutoff point of 0.5, the sensitivity was 96.5%, the specificity was 96.8% for the whole database.

IV. CONCLUSION

SVMs provide an excellent way to classify our database: using Gaussian kernel, and an optimized parameter set, SVM classified 93,2% of our database correctly into "healthy" and "obese" categories, using only routine laboratory parameters. Both the sensitivity and the specificity of the classifier were impressive.

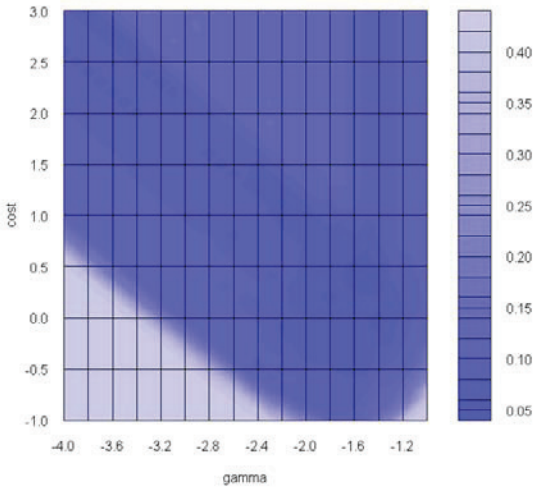

Fig. 2. Classification error with different parameters of SVM for a smaller parameter-space. (Parameters are shown on a logarithmic scale).

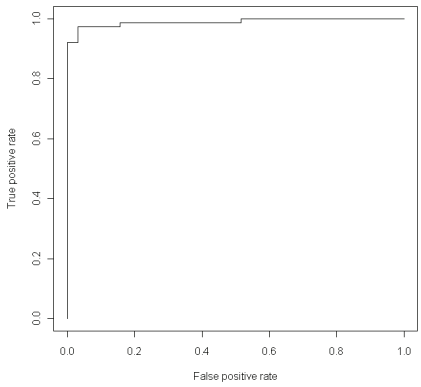

Fig. 3. ROC-curve of the final SVM classifier.

Although this result is better than logistic regression, they are not comparable in a sense that SVM is a "black-box" model and cannot be used for medical tasks. However, our result establishes a reference level, to which results of actually applicable (white-box) classification models can be compared to. (As they are supposed to reach it as closely as possible, so that their interpretation will not be misleading.)

Acknowledgment

The authors thank the Hungarian secondary schools involved in this project: Fazekas Mihály (Budapest), Leövey Klára (Budapest), Puskás Tivadar (Budapest) and Esze Tamás (Mátészalka) Primary and Secondary Grammar Schools. We say special thanks to Bétéri Csabáné for her cooperation in organizing the study at Mátészalka and Zuzsanna Almássy from Heim Pál Children's Hospital for her medical advices.

This work was supported in part by Hungarian National Scientific Research Foundation, Grants No. OTKA T69055, T82066. It is connected to the scientific program of the "Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" project, supported by the New Hungary Development Plan (Project ID: TÁMOP-4.2.1/B-09/1/KMR-2010-0002).

References

1. Andersen RE (2003) Obesity: etiology, assessment, treatment, and prevention. Champaign: Human Kinetics Publishers.
2. SRI for Health (2004) Hungary's healthcare and social system. Budapest: Strategic Research Institute for Health.
3. Avram MM, Avram AS, James WD (2005) Subcutaneous fat in normal and diseased states: 1. Introduction. J Am Acad Dermatol 53:663–670.
4. Hungarian Central Statistical Institute at Mortality by common death causes (1990-), http://portal.ksh.hu/pls/ksh/docs/hun/xstadat/xstadateves/i wnh001.html.
5. Mamtani M, Kulkarni H (2005) Predictive Performance of Anthropometric Indexes of Central Obesity for the Risk of Type 2 Diabetes. Archives Med Research 36:581–589.
6. Jebb SA, Elia M (1993) Techniques for the measurement of body composition: A practical guide. Int. J. Obes Related Metab Disorders 17:611–621.
7. Ferenci T (2009) Biostatistical analysis of obesity related parameters in Hungarian children (MSc Thesis, in Hungarian). Budapest University of Technology and Economics, Department of Informatics and Control Engineering, 2009.
8. Ferenci T, Kovács L, Almássy Zs et al. (2010) Differences in the Laboratory Parameters of Obese and Healthy Hungarian Children And Their Use in Automatic Classification. EMBC Proc., 32th An. Int. Conf. IEEE Eng. in Medicine and Biol. Society. Buenos Aires, Argentina, pp. 3883-3886.
9. R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
10. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines.
11. Vapnik V (1995) The Nature of Statistical Learning Theory. Springer-Verlag, Berlin.
12. Tan PN, Steinbach M, Kimar V (2006) Introduction to Data Mining. Addison-Wesley, New York.
13. Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, London.
14. Ferenci T, Kovács L, Almássy Zs et al. (2011) Impact of obesity on laboratory parameters: results of a Hungarian cross-sectional study and its implications in screening. Biomedical Engineering Online, Submitted

Author: Levente Kovács
Institute: Budapest University of Technology and Economics
Street: Magyar tudósok krt. 2.
City: Budapest
Country: Hungary
Email: lkovacs@iit.bme.hu