Joint Prediction of Chronic Conditions Onset: Comparing Multivariate Probits with Multiclass Support Vector Machines

Shima Ghassem Pour[1,2(✉)] and Federico Girosi[1,2]

[1] Centre for Health Research, Western Sydney University, Sydney, Australia
S.GhassemPour@WesternSydney.edu.au
[2] Capital Markets CRC, Sydney, Australia

Abstract. We consider the problem of building accurate models that can predict, in the short term (2–3 years), the onset of one or more chronic conditions at individual level. Five chronic conditions are considered: heart disease, stroke, diabetes, hypertension and cancer. Covariates for the models include standard demographic/socio-economic variables, risk factors and the presence of the chronic conditions at baseline. We compare two predictive models. The first model is the multivariate probit (MVP), chosen because it allows to model correlated outcome variables. The second model is the Multiclass Support Vector Machine (MSVM), a leading predictive method in machine learning. We use Australian data from the Social, Economic, and Environmental Factory (SEEF) study, a follow up to the 45 and Up Study survey, that contains two repeated observations of 60,000 individuals in NSW, over age 45. We find that MSVMs predictions have specificity rates similar to those of MVPs, but sensitivity rates that are on average 12 % points larger than those of MVPs, translating in a large average improvement in sensitivity of 30 %.

1 Introduction

While infectious disease continue to pose a threat to world health, in the words of the World Health Organization "it is the looming epidemics of heart disease, stroke, cancer and other chronic diseases that for the foreseeable future will take the greatest toll in deaths and disability"[1]. In fact, already 10 years ago the total number of people dying from chronic diseases was double that of all infectious diseases, maternal/perinatal conditions, and nutritional deficiencies combined [1]. The rise of these conditions can be traced to a complex web of interactions of common factors, such as genes, nutrition and life-style, with socio-economic status.

Since chronic conditions can be very costly but are also preventable there is great interest in building models that allow to simulate the costs and benefits of health interventions in this area, and that can be used for planning and policy purposes by government agencies and other interested stakeholders [2–4].

The risk of developing a chronic condition is highly dependent on factors such as obesity or smoking and on individual characteristics such as income

A. Gammerman et al. (Eds.): COPA 2016, LNAI 9653, pp. 185–195, 2016.
DOI: 10.1007/978-3-319-33395-3_13

and education. These factors vary greatly within the population, and therefore it is particular important to develop models that predict the onset of chronic conditions at individual level, that can then be used as components of simulation models to be applied to an entire population [2].

Since chronic conditions are quite correlated (for example diabetes and heart disease often go together) it is imperative to use models that make joint predictions, rather than modeling each condition separately. In the biostatistics literature this is usually done using multivariate probit models (MVP) [5,6]. While MVP are very attractive because they are easily interpretable, they rely on a very simple and rather restrictive specification and they were designed more for the purpose of understanding the determinants of the outcomes, rather than for predicting the future.

From a machine learning viewpoint it is somewhat surprising that there have not attempts to use more sophisticated, and appropriate, type of models, such as Support Vector Machines (SVMs) or Deep Learning (DL) methods. We start to fill this gap by presenting, in this paper, a comparison between the predictive ability of MVPs and SVMs. We have chosen SVMs to start with mostly because the biostatistics community is very comfortable with `R` and at the moment there is somewhat more support in `R` for SVMs than for DL.

It is important to underscore that the ability to improve the accuracy of MVP predictions is not an academic exercise. What is of interest to policy makers is long term predictions (20 to 30 years), that can only be made by repeatedly applying shorter term predictions (from one to three years, depending on the availability of longitudinal data). Therefore even a small improvement in the accuracy of short-term predictions can result in large reduction in the uncertainty of the long-term estimates, having a large impact on the policy outcomes.

The rest of this article is organized as follows. Section 2 describes the data used in our experiment. Section 3 briefly describes the MVP and SVM models. Section 4 discusses the experimental results and Sect. 5 concludes the paper.

2 Data

In order to build a predictive model of chronic disease it is necessary to have longitudinal data, in which the same individual has been observed at least twice. Since we are interested in predicting several chronic conditions at once, and since the joint prevalence of certain conditions is not very high, the data sets needs to be quite large in order to capture some of those combinations. There is a dearth of longitudinal data that can be used for this purpose, and one of the largest is the Australian Social, Economic, and Environmental Factory (SEEF) study, a follow up to the 45 and Up Study survey [7]. The approval for this study is provided by the NSW Population & Health Services Research Ethics Committee (AU RED reference:HREC/15/CIPHS/4).

The 45 and Up Study survey (www.saxinstitute.org.au), which was carried out between 2006 and 2009, contains information regarding the health and social wellbeing of 267,153 individuals aged 45 years and older living in New South

Wales (NSW), Australia. Eligible individuals, sampled from the Medicare population of NSW, were mailed the questionnaire, an information sheet and a consent form and provided with a reply paid envelope. The survey over-sampled individuals aged 80 years and over and residents of rural areas by a factor of two. In addition, all residents aged 45 years and older in remote areas were sampled. The overall response rate of the 45 and Up Study is 18 %, accounting for approximately 10 % of all individuals of age 45 years or older living in NSW. While the response rate is low and participants tended to be of more favorable socioeconomic circumstances than average for the age group, previous work has shown that analytical findings based on internal comparisons, such as odd-ratios, are generalizable and comparable to those derived from smaller but more representative population health surveillance [8].

Data captured in the 45 and Up Study baseline include a number of self-reported chronic conditions such as (ever diagnosed) heart disease, high blood pressure, diabetes, stroke, asthma, depression and different types of cancer.

Questionnaire data also include information on key potential confounder and mediating factors, including age, sex, household income, level of education, smoking history, alcohol use, physical activity, height and weight, functional status, psychological distress, medical and surgical history and dietary habits. A full description of all the variables available in the 45 and Up Study together with basic summary statistics can be found elsewhere [7].

The SEEF study data, that include all the original variable in the 45 and Up Study plus a host of additional variables, were collected in 2010 from a random sub sample of the baseline 45 and Up Study cohort. One hundred thousands 45 and Up Study participants were mailed an invitation and the SEEF questionnaire. About 60,000 individuals joined the SEEF study by completing the consent form and the questionnaire and mailing them to the study coordinating center.

Our dependent variables are 5 binary variables denoting the presence or absence of the following chronic conditions at follow-up: heart disease, hypertension, diabetes, stroke, and cancer. These health conditions were self-reported and based on the responses to survey questions formulated as follows: “Has a doctor ever told you that you have [name of condition]?”.

Since individuals can develop any of those five conditions we consider the multi-class problem of predicting in which of the $2^5 = 32$ combinations of conditions individuals will fall at follow-up. We report in Table 1 the size of each of the 32 classes in the SEEF data. Since some of the classes are very small and neither of the two methods out-performed the other in those cases, we have eliminated from our data the classes with fewer than 100 cases (outlined in bold in Table 1).

The two main risk factors that we used as covariates were obesity and smoking status. Possible values of smoking status are “Not Smoking”, “Smoker” and “quit smoking”, which are derived from the combined answers to the following two questions “Have you ever been a regular smoker?” and “Are you a regular smoker now?”.

Table 1. Class size (bold font shows the classes which we removed)

Condition	Size
No condition	16421
Cancer	11896
Cancer-hypertension	7614
Diabetes	737
Diabetes-Cancer	523
Diabetes-Cancer-Hypertension	1195
Diabetes-Hypertension	1389
Diabetes-Stroke	**19**
Diabetes-Stroke-Cancer	**20**
Diabetes-Stroke-Cancer-Hypertension	137
Diabetes-Stroke-Hypertension	**95**
Heart	1319
Heart-Cancer	1925
Heart-Cancer-Hypertension	2847
Heart-Diabetes	169
Heart-Diabetes-Cancer	251
Heart-Diabetes-Cancer-Hypertension	749
Heart-Diabetes-Hypertension	506
Heart-Diabetes-Stroke	**19**
Heart-Diabetes-Stroke-Cancer	**32**
Heart-Diabetes-Stroke-Cancer-Hypertension	343
Heart-Diabetes-Stroke-Hypertension	111
Heart-Hypertension	1787
Heart-Stroke	**89**
Heart-Stroke-Cancer	154
Heart-Stroke-Cancer-Hypertension	360
Heart-Stroke-Hypertension	197
Hypertension	8475
Stroke	166
Stroke-Cancer	199
Stroke-Cancer-Hypertension	376
Stroke-Hypertension	284

Obesity status was based on the values of the body mass index (BMI), which is the body weight in kilograms divided by the square of the body height in meters. We used the standard World Health Organization classification system to categories individuals as Underweight ($\mathrm{BMI} < 18.5$), Normal ($18.5 \leq \mathrm{BMI} < 25$), Overweight($25 \leq \mathrm{BMI} < 30$) and Obese ($\mathrm{BMI} \geq 30$).

Additional covariates used in the analysis are the five chronic conditions at baseline, age category, gender, income, work status, private health insurance status, Body Mass Index (BMI) and smoking status.

The SEEF study includes many more variables (such as education, dietary habits or family history) that could be used in the analysis but we have restricted ourselves to this set because we found that adding more variables did not significantly improve the predictions.

Since individuals were recruited in the 45 and Up Study over a period of few years the interval between interviews is not always the same, resulting in follow-up data being collected between 2 and 4 years after baseline. Therefore we also included as a covariate the time to follow up, which on average was 2 and half years. The summary statistics for the covariates used in the model are shown in Table 2.

Table 2. Summary statistics of the SEEF Study. All quantities measured at baseline except when reported otherwise. Quantities in parenthesis are proportions.

Age	[45,50]	(50,55]	(55,60]	(60,65]	(65,70]	(70,75]	(75,80]	(80,85]	(85,100]
	8902	10302	10196	9115	8007	5197	4132	3062	926
	(0.15)	(0.17)	(0.17)	(0.15)	(0.13)	(0.09)	(0.08)	(0.05)	(0.01)
Income	>20K	20K-30K	30K-40K	40K-50K	50K-70K	70K+			
	13954	7874	6510	5769	8208	17524			
	(0.23)	(0.13)	(0.10)	(0.10)	(0.14)	(0.30)			
Health Insurance	No Private	Veteran	Health CareCard	Private Extras	Private No Extras				
	9236	1045	9186	30970	9402				
	(0.15)	(0.02)	(0.15)	(0.52)	(0.16)				
BMI	Normal	Obese	Overweight	Underweight					
	22780	12344	23950	765					
	(0.38)	(0.21)	(0.40)	(0.01)					
Smoking	Not smoker	Quit smoking	Smoker						
	34910	21485	3444						
	(0.59)	(0.35)	(0.06)						
Work status	Full	Not working	Part time						
	18002	29608	12229						
	(0.30)	(0.50)	(0.20)						
Gender	Female	Male							
	32128	27711							
	(0.54)	(0.46)							
Conditions	Heart	Diabetes	Stroke	Cancer	Hypertension				
	7059	4441	1256	22283	20598				
	(0.12)	(0.07)	(0.02)	(0.37)	(0.34)				
Conditions at follow-up	Heart	Diabetes	Stroke	Cancer	Hypertension				
	10564	5973	2036	28278	26233				
	(0.18)	(0.10)	(0.03)	(0.47)	(0.44)				

3 Methodology

3.1 Multivariate Probit

Let us denote by $Y_{i\alpha}^{(1)}$ the binary variable indicating the presence at follow-up of chronic condition α for individual i, where $i = 1, 2, \ldots, N$ with $N = 60,000$, and $\alpha = \{\text{heart disease, diabetes, hypertension, stroke, cancer}\}$. Let us also denote by

$Y_{i\alpha}^{(0)}$ the corresponding variable measured at baseline, and by $\mathbf{Z}_i \in R^d$ a vector of other covariates measured at baseline. To simplify the notation we denote by $\mathbf{Y}_i^{(1)}$ ($\mathbf{Y}_i^{(0)}$) the vectors whose components are $Y_{i\alpha}^{(1)}$ ($Y_{i\alpha}^{(0)}$).

The MVP model is a latent variable model with the following specification:

$$\hat{\mathbf{Y}}_i^{(1)} = \Gamma \mathbf{Y}_i^{(0)} + \Theta \mathbf{Z}_i + \boldsymbol{\epsilon}_i, \qquad Y_{i\alpha}^{(1)} = 1 \text{ if } \hat{Y}_{i\alpha}^{(1)} > 0, \text{ 0 otherwise} \qquad (1)$$
$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \Sigma)$$

where Γ and Θ are matrices of coefficients, of dimensions 5×5 and $5 \times d$ respectively, that need to be estimated. The key to the MVP model of Eq. 1 is the presence of the 5×5 (unknown) covariance matrix Σ. The off-diagonal elements of its inverse capture the correlations across chronic conditions and the fact that developing, say, heart disease and diabetes are not independent events. Prediction of the MVP model are performed probabilistically, by feeding samples of the multivariate normal distribution $\mathcal{N}(0, \Sigma)$, one for each individual, in Eq. 1.

The estimation of the full MVP model is notoriously computationally intensive, although recent advances in computational methods [6] make it much more approachable. For the purpose of our experiments we have developed an approximation of the traditional method in which we use observed correlation among chronic conditions to approximate the matrix Σ, which makes the estimation of the model much simpler. Since we have not observed deterioration in performance by using the approximate method, all the experiments performed for the production of this paper have been performed using the approximation rather than the full implementation.

3.2 Support Vector Machines

Support Vector Machines (SVMs) have been around the machine learning community for more than 20 years now [9], and for the sake of brevity we simply refer the reader to standard textbooks and references [10,11]. SVMs have many attractive features, but one that should be emphasized in the context of this paper is that, unlike MVP, they do not rely on distributional assumption regarding the process that generates the data. Instead, SVMs relies on two key modeling choices: one is

1. the parameter (usually denoted by C) that controls the penalty associated with the misclassification of a data point;
2. the kernel, that is associated with the choice of the (possibly infinite dimensional) feature space onto which the input variables are projected [12]. For the purpose of this paper we have mainly experimented with polynomial kernels of the form $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i \cdot \mathbf{x}_j)^p$, that are uniquely parametrized by the degree p.

SVMs were originally designed for binary classification problems, but several extensions exist that allow to deal with multi-class problems.

In the `R` package we use for the SVM implementation, `Kernlab`, there are several options for dealing with multi-class problems [13,14]. We found that for

this problem best results were obtained by using the "one vs one" approach, in which one trains $K(K-1)/2$ binary classifiers (with $K = 32$ in our case). Each of the classifier separates one class from another class, and in order to classify a new sample, all classifiers are applied and the class that gets the highest number of votes is selected. While it is not fully clear why the "one vs one" approach worked better than the alternatives (such as the "one vs all" approach [15]), the fact that in this particular application many of the events we are trying to predict are quite rare seems to play a role, since it can lead to very imbalanced data sets.

4 Experimental Results

4.1 Performance Evaluation Metrics

We used a 10-fold cross-validation approach to estimate the performance of the MVP and SVM methods. The full data sets was first randomly partitioned in 10 subsets of equal size (approximately 6,000 data points each). For each of the 10 replication trials we withhold one of the 10 partitions and use it for testing, while the remaining 9 partitions are used for training. For each of the 10 trials we compute 4 performance measures, and we report the average of the performance measures over the 10 replications.

As performance measures we report sensitivity and specificity, since they are the ones most commonly used in health studies, as well as accuracy and the F1 score. We report the definitions below, where TP, TN, FP and FN refer to the total number of true positives, true negatives, false positives and false negatives respectively.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

Sensitivity (or true positive rate, or recall) is important because it measures the ability to identify who is going to develop the disease.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

Specificity (or true negative rate) is important because it measures the ability to identify who is *not* going to develop the disease.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

Accuracy indicates how many samples are correctly classified overall. Accuracy can be misleading when the dataset is imbalanced. Therefore an alternative performance measure is the F1 Score, defined as:

$$\text{F1 Score} = 2 \times \frac{pr}{p + r} \tag{5}$$

where p is the precision and r is the recall (or sensitivity). Here precision is defined as the ratio of true positives (TP) to all predicted positives ($TP + FP$). Since the F1 score is the harmonic mean of precision and recall a high score is obtained when precision and recall are both high.

4.2 Results

The average of the performance measures over the 10 replication sets for Multiclass SVMs (MSVMs) and MVP are shown in Figs. 1 and 2. In Fig. 1 we report specificity and sensitivity for both methods. The key message of this figure is that while the specificity of the two methods are comparable, the sensitivity of MSVM is, on average about 12 % points better than the one of multivariate probit. Since sensitivities are in general not very high, this translates in a large relative improvement, of approximately 30 %.

A similar pattern is seen on accuracy and F1 scores. With very few exceptions SVMs are more accurate than MVPs, although by not too much. That the difference is not great relates to the fact that in most cases the classification

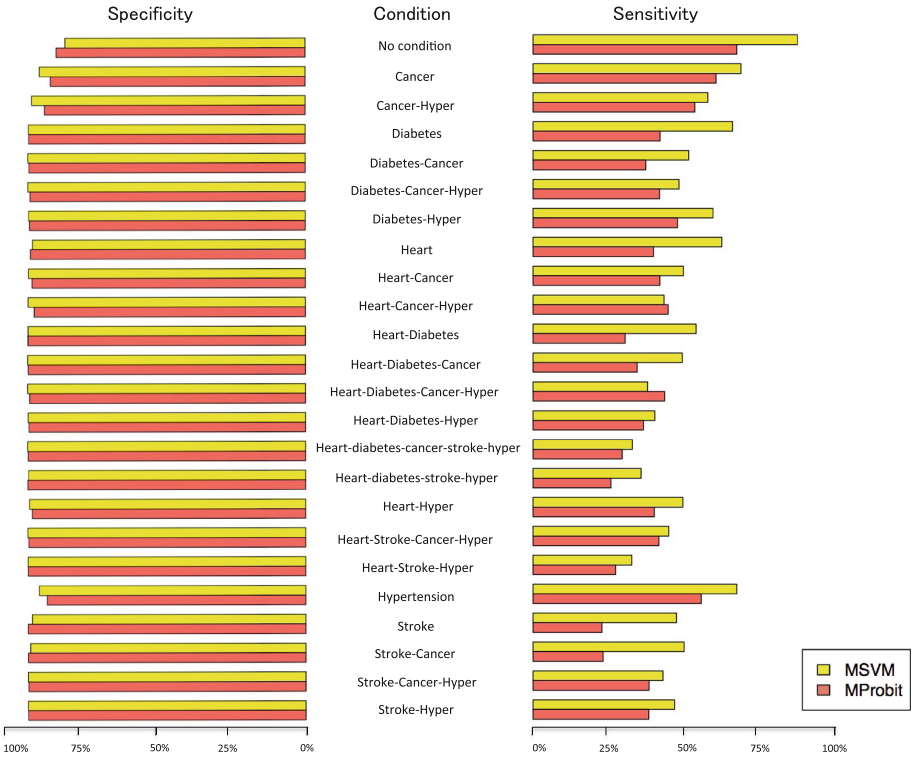

Fig. 1. Comparison between MSVM and MVP using 10-fold cross-validation: sensitivity and specificity.

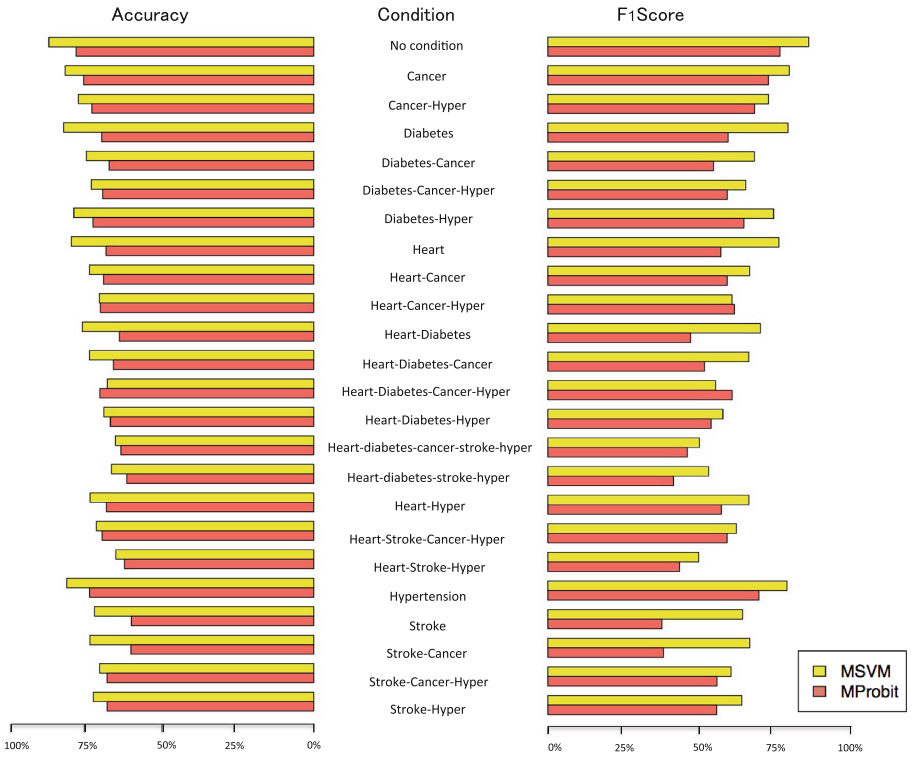

Fig. 2. Comparison between MSVM and MVP using 10-fold cross-validation: accuracy and F1 score.

problem is quite imbalanced, for which accuracy is not a good performance measure. The F1 score shows show larger differences between SVMs and MVPs, which is not surprising since a component of the F1 score is the sensitivity of the method, that is greatly improved using MSVMs.

5 Lessons Learned

Few lessons have emerged from this study. First of all, independently of which method we use, predicting who is going to develop some combination of chronic conditions in the near future, based on a handful of individual characteristics and the current chronic conditions, is quite hard. While maintaining specificity rates above 90 %, most of the sensitivity rates, obtained using MSVMs, fell within 50 % and 75 %.

In our experience including additional risk factors, such as diet or family history, will only lead to marginal improvements. What is likely to have a major impact on the predictive ability of any method is a more accurate measurement of people's health status, such as actual results of pathology and imaging tests.

Unfortunately it seems unlikely that data sets of this type, that in principle exist, can be made available to researchers any time soon.

This implies that it is crucial to make the best possible use of the current data, and that is why the choice of predictive model is highly relevant. Given that short-term predictions are of particular value in the process of making long-term predictions, which carry enormous policy implications, even a small improvement in accuracy could have serious policy implications. Put in this context, an average improvement in sensitivity of 12 % points, which translates into a 30 % relative improvement, is enormous.

We do not claim to have produced the best possible classifier, and it is likely that better methods can be devised, especially if they start taking advantage of prior information we have on the development of chronic conditions. However the main lesson learned is that the choice of predictive model can make a big difference. This seems particular important because in the area of health analytics we have not seen a high rate of adoption of methods such as SVMs or Deep Learning, which have proved to be extremely successful in a wide range of applications. Therefore we hope that this study will be a first step toward a broader use of methods that carry the potential of leading to large improvement over the status quo.

Acknowledgment. This research was completed using data collected through the 45 and Up Study (http://www.saxinstitute.org.au). The 45 and Up Study is managed by the Sax Institute in collaboration with major partner Cancer Council NSW, and partners the National Heart Foundation of Australia (NSW Division), NSW Ministry of Health, beyondblue, NSW Government Family & Community Services Carers, Ageing and Disability Inclusion, and the Australian Red Cross Blood Service. We thank the many thousands of people participating in the 45 and Up Study. We also thank Capital Markets CRC,that has sponsored this research.

References

1. World Health Organization: Preventing Chronic Diseases. A Vital Investment, World Health Organization, Geneva (2005)
2. Goldman, D., Zheng, Y., Girosi, F., Michaud, P.-C., Olshansky, J., Cutler, D., Rowe, J.: The benefits of risk factor prevention in americans aged 51 years and older. Am. J. Public Health **99**(11), 2096–2101 (2009)
3. Lymer, S., Brown, L., Duncan, A.: Modelling the health system in an ageing Australia, using a dynamic microsimulation model. Working Paper 11/09, NATSEM at the University of Canberra, June 2011
4. Boyle, J.P., Thompson, T.J., Gregg, E.W., Barker, L.E., Williamson, D.F.: Projection of the year 2050 burden of diabetes in the US adult population: dynamic modeling of incidence, mortality, and prediabetes prevalence. Popul. Health Metr. **8**(1), 29 (2010)
5. Greene, W.: Econometric Analysis. Pearson Education, New Jersey (2003)
6. Cappellari, L., Jenkins, S.: Calculation of multivariate normal probabilities by simulation, with applications to maximum simulated likelihood estimation. Stata J. **6**(2), 156–189 (2006)

7. Collaborators, U.S.: Cohort profile: the 45 and up study. Int. J. Epidemiol. **37**(5), 941 (2008)
8. Mealing, N., Banks, E., Jorm, L., Steel, D., Clements, M., Rogers, K.: Investigation of relative risk estimates from studies of the same population with contrasting response rates and designs. BMC Med. Res. Methodol. **10**(1), 26 (2010)
9. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
10. Scholkopf, B., Smola, A.J.: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
11. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge (2000)
12. Smits, G.F., Jordaan, E.M.: Improved SVM regression using mixtures of kernels. In: Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN 2002, vol. 3, pp. 2785–2790. IEEE (2002)
13. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. J. Mach. Learn. Res. **2**, 265–292 (2002)
14. Weston, J., Watkins, C.: Support vector machines for multi-class pattern recognition. ESANN **99**, 219–224 (1999)
15. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. J. Mach. Learn. Res. **5**, 101–141 (2004)