MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Comparison of Machine Learning Techniques with Classical Statistical Models in Predicting Health Outcomes

Xiaowei Song[a], Arnold Mitnitski[b,c], Jafna Cox[b], Kenneth Rockwood[a,b]

[a] *Geriatric Medicine Research Unit, QEII Heath Sciences Centre, Canada*
[b] *Department of Medicine, Dalhousie University, Canada*
[c] *Faculty of Computer Science, Dalhousie University, Canada*

Abstract

Several machine learning techniques (multilayer and single layer perceptron, logistic regression, least square linear separation and support vector machines) are applied to calculate the risk of death from two biomedical data sets, one from patient care records, and another from a population survey. Each dataset contained multiple sources of information: history of related symptoms and other illnesses, physical examination findings, laboratory tests, medications (patient records dataset), health attitudes, and disabilities in activities of daily living (survey dataset). Each technique showed very good mortality prediction in the acute patients data sample (AUC up to 0.89) and fair prediction accuracy for six year mortality (AUC from 0.70 to 0.76) in individuals from epidemiological database surveys. The results suggest that the nature of data is of primary importance rather than the learning technique. However, the consistently superior performance of the artificial neural network (multi-layer perceptron) indicates that nonlinear relationships (which cannot be discerned by linear separation techniques) can provide additional improvement in correctly predicting health outcomes.

Keywords:

Health status, Machine learning, Classification, ROC curves.

Introduction

The prediction of health outcomes from available data is an important problem in health research and health management. It is usually assessed by calculating scores/indices for risk stratification [1]. Conventionally, such scores are based on statistical models, such as logistic regression and Cox proportional hazard model [1,2]. Most such applications are based on the belief (although not often explicit) that there exist a relatively small number of important variables (risk factors) and that careful selection of those variables is the key to successful performance of the models for outcome prediction. Unfortunately, however, risk factors typically interact with each other in a complicated and generally unknown way, and therefore often are eliminated from predictive models. More recently, new techniques based on machine learning have become available (*e.g.*, artificial neural networks, support vector machines). Such approaches are based on inductive inference rather than on classical statistics [3]. Machine learning algorithms are not widely available in statistical software packages and even when they are, their application demands skills, which are often outside the usual experience of biostatisticians. Recently, some reports have compared different learning techniques with "classical" statistical algorithms [4]. Such comparisons are generally few, and have explored only a small number of techniques in a limited number of data sets [2,5].

The aim of our report is to compare the performance of several well known machine learning techniques (including ones based on "classical" statistical models such as logistic regression) in two distinct data sets: (i) patients with the acute care hospital needs; and (ii) a population-based epidemiological survey that contains physician assessment data and self-reported data. We investigated the ability of the different techniques to learn from data (the training samples) and to correctly classify individuals at risk, and then compared the classification accuracy using the separate part of data (the testing samples).

Materials and Methods

Databases and variables

The data came from two different studies: administrative data from the Improving Cardiovascular Outcomes in Nova Scotia (ICONS) registry [6] and the Canadian Study of Health and Aging (CSHA) [7]. The ICONS database has a large sample (n~35,000 patients), with more than 500 variables per patient, and comprehensive follow-up. The database contains measures that have been traditionally used to assess disease severity in a cardiac population. The sample we used contains records of 4432 consecutive patients hospitalized in the Province of Nova Scotia with acute myocardial infarction (AMI) in 2000 and 2001. Among them 602 (13.7%) died within 30 days after admission to the hospital and 974 (22.2%) died within 1 year. Thirty-seven variables describing the histories of related illnesses (*e.g.* diabetes, congestive heart failure, laboratory test results (*e.g.*, creatinine, glucose, ejection fraction), and medications (*e.g.*, beta blocker, statine) were available [8].

The CSHA is a multi-year cohort study of cognitive impairment and other aspects of the health of older Canadians aged 65 years and older [7]. In the second wave of the study, 5586 community-dwelling people were interviewed and 1597 (28.6%) died at the

end of 72-month follow-up. Forty variables which recorded data about health attitudes, diseases (*e.g.*, cancer, heart conditions) and disabilities in activities of daily living were selected from the screening question pool for the present study. In total 2305 people either from institutions or from community had a clinical examination at CSHA. Among them, 1007 (43.7%) died by the end of 72-month follow-up. Seventy clinical measurements comprising disease were included in the CSHA clinical database for the present study.

Binary variables were represented as 0 (absent) and 1 (present). Multiple level variables were mapped into [0,1] interval using linear transformation. Continuous variables were dichotomized using conventional cut off values recommended by cardiologists (e.g., heart rat >100 was recoded as 1, otherwise 0. In addition we performed sensitivity analysis by varying the cut points and assessing the predictive accuracy of the model during validation. Age and gender were included as input variables. Survival status of the subjects over various periods in the various databases was used as output for the respective datasets. In ICONS, we investigated 30-day mortality prediction and separately 1 year mortality prediction. In CSHA, we considered separately the clinical data sample and the self-report data. Therefore, four datasets were used in the comparative analysis of the techniques further called: CSHA clinical; CHSA self-reported; ICONS 30-day; and ICONS 1- year.

Models and applications

Each technique was used to produce risk index score comprising the available variables in each data set. The techniques included: logistic regression (LR); single layer perceptron (SLP) with sigmoid activation function; multi layer perceptron (MLP) with back propagation algorithm; least squares linear separation (LSS); and least squares support vector machine (LS-SVM) with linear kernel [9].

Each dataset was divided into two non-overlapped samples: training (2/3 of cases, to calculate the parameters of the models and therefore to identify the formula for calculating the risk) and testing (1/3 of cases to check the accuracy of classification prediction). The accuracy of classification was assessed by the receiver operating characteristics (ROC) curves and the area under the ROC curve (AUC) was used for comparison of different learning techniques. All techniques were applied to each dataset. The samples were randomly selected and all the calculations were repeated 10 times, which allowed us to check the reproducibility of the risk assessment in different sub-samples, and to calculate the means and standard deviations in each setting. Comparisons of the mean values from each model were conducted using Analyses of Variance (ANOVA). The significance level was set at 0.05. Models, simulations, and statistical analyses were performed using Matlab software (version 6.5).

Results and Discussion

In Table 1, the accuracy of predictions, assessed by the area under the ROC curve (AUC), is summarized.

Each model shows similar values for the quality of separation in each data set. There is a substantial difference in the accuracy between ICONS and CSHA data sets. These differences reflect profoundly on the different nature of these databases. The ICONS database contains the information used in clinic, about the status of patients hospitalized with acute health conditions. There is no surprise that such information is of vital importance for the assessment of adverse health outcomes. All models better predict 1-year mortality than 30-day mortality because some measures crucial for short-term survival prediction (*e.g.*, troponin) contained too many missing cases to be readily applied at this point. Still the quality of prediction remains quite high (almost 0.90 in MLP and 0.85 in most other techniques).

Table 1: Performance results of different techniques in four data sets: areas under the Receive Operating Curves (AUC) (means and standard deviations).

Model	CSHA2 Clinical	CSHA2 Self-rprt	ICONS 1-year	ICONS 30-day
MLP	0.75 ± 0.01	0.76 ± 0.03	0.89 ± 0.02	0.87 ± 0.04
LSS	0.72 ± 0.03	0.74 ± 0.01	0.85 ± 0.02	0.83 ± 0.02
LS-SVM	0.72 ± 0.03	0.73 ± 0.03	0.85 ± 0.02	0.83 ± 0.01
LR	0.71 ± 0.02	0.72 ± 0.05	0.84 ± 0.01	0.82 ± 0.01
SLP	0.70 ± 0.02	0.70 ± 0.04	0.85 ± 0.01	0.82 ± 0.02

Note: Values in each cell (mean and standard deviation) represent the result of 10 simulations in the testing data set.

In the CSHA dataset, the accuracy of prediction is considerably lower (from 0.70 for the single layer to 0.76 for the multi layer perceptron). Relatively poor prediction accuracy compared to the ICONS dataset reflects both the longer period of follow-up and that the CSHA database was designed as an epidemiological study of elderly Canadians. Thus, it contains rather general information about individual conditions, and likely would not have any data on people with major acute problems, who would have been non-respondents [10,11].

The accuracy of prediction nevertheless seems to be rather high, given the nature of the variables in the models. Interestingly, there is no big difference in the prediction between data set containing the assessments provided by physicians and self reported assessment. The latter is conventionally considered as rather crude, but our results suggest that much more attention should be paid to self-report data. This may also imply that the quality of the questionnaires, rather than the degree of accuracy of answers, is more critical to the informativeness of the data.

Figure 1 shows an example of the ROC curves for the prediction of 1-year survival using the ICONS data. The sensitivity (true positive rate) increases quickly with increases in the false positive rate and reaches over 80% at a low cut-off point. It remained higher at the cut-off points of the risk score of above this level.

While the ROC curves are quite close to each other, the MLP consistently showed the best performance in all data sets (Table 1). Slightly below that is the performance of LSS and LS-SVM

(second and third lines of Table 1). Both are linear algorithms, as support vector machine is used here with linear kernels [9]. Logistic regression and SLP (which is structurally similar to LR) show inferior performance, although the difference is not statistically significant. The good performance of the algorithms based on the linear separation techniques (LSS and the linear kernel LS-SVM), may indicate that, in the first approximation, the relationships between the outcome (mortality) and the variables can be well represented by a linear function. In this case the interpretation of the parameters of the risk scores is straightforward: they indicate the relative importance of the variables in predicting mortality. The coefficients of such models were always highly correlated between different models.

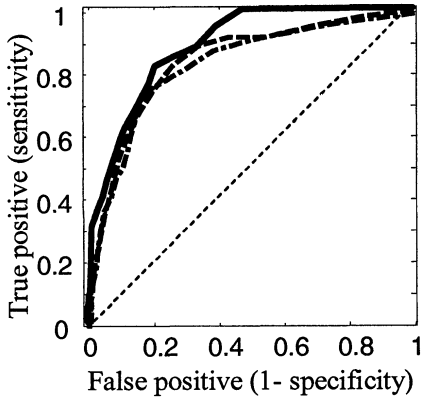

Figure 1 - The Receiver Operating Characteristic (ROC) curves for the risk scores obtained using several representative models. The x-axis represents the false positive rate, i.e., the proportion of subjects who survived but were classified as having died. The y-axis represents the true positive rate, i.e., the proportion of subjects who died and who were classified as to die. Solid line: Multi-Layer Perceptron (MLP); Dashed line: Least Square Separation Hyperplane (LSS); Dash-dot line: Logistic Regression (LR); the diagonal line shows the case of no useful information from the model. Below that, the instrument is more often wrong than right.

Small but consistent superiority of the MLP in all data sets is consistent with the importance of nonlinear relationships in biological data. This contributes to increasing accuracy of the model. Such nonlinear relationships are intrinsic in biological systems. Serious health conditions seldom appear as a single problem on their own, but usually present with a number of accompanying, less dramatic charges that, considered in isolation, might not appear to be very important. However, each of the variables may represent a small portion of the problem. Therefore any characteristic can contribute to the overall picture of an individual's health, which explains the productivity of approaches that express mortality in terms of as many variables as possible [12].

The statistical significance of the model can be addressed in simulations when the calculations are repeated with random subsamples of data (so-called "bootstrapping" [13]). The confidence interval measures of the model's performance and of the influence of variables included in the model can be addressed in such direct simulations.

Interestingly, the quality of the previously reported risk predictive models, expressed as AUCs, rarely extends 80% [1,2,14-16]. The question remains whether such accuracy can be exceeded using advanced machine learning techniques or whether there is a fundamental restriction of the particular data being evaluated, due to the "wrong" variables having been used, as is usually presumed.

Our results, however, suggest that properly integrating the various deficits (*i.e.*, using multiple dimensional variables) has great potential to successfully summarize individual health status [12]. Problems associated with an organism usually involve failures in multiple system functions and/or living abilities and are often displayed with multiple symptoms from many different aspects, as would be usual, for example, in summarizing the complex needs of frail elderly people in a clinical assessment [17-19].

Other nonlinear techniques need to be investigated in future studies. Among them, support vector machines (SVM) with polynomial kernels, radial basis networks (RBN) and nonlinear discriminant analysis. We made a preliminary assessment of the SVM with radial basis kernel in the ICONS dataset. The accuracy of the RBN model was not different from the linear models. If this is an intrinsic property of the kernel with the data or it can be improved by the appropriate selection of the tuning parameters remains to be seen.

It is interesting to observe that models of the same linear type that implemented different algorithms for case separation or classification did not show substantially different performances. This suggests that the choice of model amongst models of a similar type may be less critical in these cases.

The fact that the non-linear models such as MLP (often called an artificial neural network) further improved the accuracy of survival assessment suggests the existence of unknown interrelations between the variables. This would also be consistent with the observed inferiority of linear techniques in this inquiry. Attempts to impose, *a priori,* nonlinear relationships such as pair interactions may or may not be successful. Clearly, such relationships would be desirable to discover, because their interpretation can be very appealing. However, such efforts should be undertaken only after there is an indication that nonlinear relationships may be important in a particular dataset. Since such nonlinear relationships often can be established, more narrow searches of particular interactions can be investigated using more conventional models. Although the application of computational techniques is useful, the approach is entirely data driven. As shown in the results from the CSHA data, no matter which model was used and no matter how powerful it was, the model itself cannot create information that is lacking in the database in order to produce superior prediction performance. It is the variable set itself, *i.e.*, the amount of information contained, that limits the performance of the possibility of higher predictive accuracy. In the ICONS data, in contrast, the variables appear to be more descriptive, with more complete information that more validly represents the problems of the patients.

It is also of some practical interest to note that the speed of performance (efficiency) is quite different across the applied techniques. The most rapid (1-2 s) was seen with least squares linear separation (LSS). It was an order of magnitude faster than logistic regression and two orders faster than MLP or SVM.

We have suggested earlier that a simple but effective way of assessment of the health status in such databases by calculating frailty index as a simple proportion of deficits (binary variables) [12,20]. We have demonstrated that using such simple frailty index mortality can be predicted with the accuracy of AUC=0.66 [21] and even higher (0.70) in the CSHA data samples used in this study. Definitely, the assumption of equality of deficits is inadequate. Of note, each of the techniques considered here derives a risk index score which, in fact, can be regarded as a weighted frailty index [21]. Thus the importance of weighting needs to be incorporated in future uses of such index variables, although, unless the datasets are very large, the trade-off might be in generalizability.

In conclusion, our results suggest that integrating large numbers of diverse, clinically sensible measures, using various computational techniques is helpful for understanding determinants of health status. However, not only the method, but also, and probably most importantly, the levels of sensitivity as to how much information each dataset contains, determines the final performance of the model.

Acknowledgments

Dr. Mitnitski is supported by the Department of Medicine UMRF junior Research Grant fund, Dalhousie University. Dr. Cox receives salary support from a Canadian Institutes of Health Research/Regional Partnership Program Investigator Award and a Clinical Research Scholarship from the Faculty of Medicine, Dalhousie University. Dr. Rockwood is supported by an Investigator Award from the Canadian Institutes for Health Research and by the Dalhousie Medical Research Foundation as the Katheryn Allen Weldon Professor of Alzheimer Research.

References

[1] Singh M, Reeder GS, Jacobsen SJ, Weston S, Killian J, and Roger VL. Scores for post-myocardial infarction stratification in the community. *Circulation* 2002; 106: 2309-14.

[2] Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, and Jaulent AC. Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. *Proc AMIA Symp* 2000: 156-60.

[3] Vapnik VN. *The nature of statistical learning theory.* 2nd ed. Springer, New York, 2000.

[4] Knuiman MW, Vu HT, and Segal MR. An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. *J Cardiovasc Risk* 1997; 4: 127-34.

[5] Van Gestel T, Suykens JAK, Baesens B, Viaene S, Vanthienen J, Dedene G, De Moor B, and Vandewalle J. Benchmarking Least Squares Support Vector Machine Classifiers. *Neural Computation* 2002; 15: 1115-47.

[6] Cox JL. Optimizing disease management at a health care system level: the rationale and methods of the Improving Cardiovascular Outcomes in Nova Scotia (ICONS) *Study. Can J Cardiol* 1999; 15: 787-96.

[7] Canadian Study of Health and Aging Working Group. Canadian Study of Health and aging. *Int Psychogeriatr* 2001; Suppl.1: 1-237.

[8] Mitnitski A, Mogilner A, Song X., Cox J, and Rockwood. An index for prediction the risk of death following acute myocardial infarction. *J Am Coll Cardiol* (submitted)

[9] Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, and Vandewalle J. Support vector machines: least squares approaches and extensions. In: *Advances in Learning Theory: Methods, Models and Applications* (Eds.: J Suykens et al.) NATO Science Series, IOS press, 2003, pp.155-79.

[10]Helliewell B, Aylesworth R, McDowell I, Baumgarten M, and Syles E. Correlates of non-participation in the Canadian Study of Health and Aging. *Int Psychogeriatr* 2001; (1 Suppl): 49-56.

[11]Rockwood K, Stolee P, Robertson D, and Shillington ER. Response bias in a health status survey of elderly people, *Age Ageing* 1989;18: 177-82.

[12]Rockwood K, Mitnitski A, and MacKnight C. Some mathematical models of frailty and their clinical implications. *Rev Clin Gerontol* 2002; 12: 109-17.

[13]Efron B and Tibshirany R. *An introduction to the bootstrap.* NY London: Chapman and Hall, 1993.

[14]Fortescue EB, Kahn K, and Bates DW. Major adverse outcomes after percutaneous transluminal coronary angioplasty: a clinical prediction rule. *J Clin Epidemiol* 2003; 56: 17-27.

[15]Desai MM, Bogardus ST Jr, Williams CS, Vitagliano G, and Inouye SK. Development and validation of a risk-adjustment index for older patients: the high-risk diagnoses for the elderly scale. *J Am Geriatr Soc* 2002; 50:474-81.

[16]Van Ruiswyk J, Hartz A, Kuhn E, Krakauer H, Young M, and Rimm A. A measure of mortality risk for elderly patients with acute myocardial infarction. *Med Decis Making* 1993; 13: 152-60.

[17]Rockwood K, Silvius J, and Fox RA.Comprehensive geriatric assessment. *Postgrad Med* 1998; 103: 247-9, 254-8, 264.

[18]Rockwood K, Stadnyk K, MacKnight C, McDowell I, Hebert R, and Hogan DB. A brief clinical instrument to classify frailty in elderly people. *Lancet* 1999; 353: 205-6.

[19]Rockwood K, Stolee P, Robertson D, and Shillington ER. Response bias in a health status survey of elderly people, *Age Ageing* 1989; 18: 177-82.

[20]Mitnitski AB, Song X, and Rockwood K. The estimation of relative fitness and frailty in community dwelling older adults using self-report data. *J Geront Med Sci* 2003 (in press).

[21]Song X, Mitnitski AB, and Rockwood K. Assessment of Individual Risk of Death Using Self-report Data: an Artifi-

cial Neural Network Compared to a Frailty Index. *J Am Geriatr Soc* 2003 (in press).

Address for correspondence

Dr. Arnold Mitnitski,

Department of Medicine & Faculty of Computer Science,

Dalhousie University,

Rm. 229,5790 University Avenue, West Annex, 2nd Floor Halifax B3H 1V7, CANADA,

Email: Arnold.Mitnitski@Dal.Ca