Predicting Healthcare Costs Using Classifiers

C.R. Stephens[1,2], H. Waelbroeck[1,3], S. Talley[1], R. Cruz[1], and A.S. Ash[4,5]

[1] Adaptive Technologies Inc., 6424 West Chisum Trail, Glendale, AZ 85310
[2] Instituto de Ciencias Nucleares, UNAM, A. Postal 70-543 México D.F. 04510
[3] eXa Inc., 51 East 42nd Street, Suite 602, New York, NY 10017
[4] Boston University School of Medicine
[5] DxCG Inc., Boston MA

1 Introduction

In the battle to control escalating health care costs, predictive models are increasingly employed to better allocate health care resources and to identify the "best" cases for preventive case management. In this investigation we predicted the top 0.5% most costly cases for year $N+1$, given a population in year N, with data for the period 1997-2001 taken from the MEDSTAT Marketscan Research Database for a cohort of privately insured individuals diagnosed with diabetes. We considered two performance metrics: i) classification accuracy, i.e. the proportion of correctly classified persons in the top 0.5% and ii) the total number of dollars associated with the predicted top 0.5% of most costly cases.

2 Methodology

The fundamental objects of interest are $P(\text{top } 0.5\%|\mathbf{X}_i)$ - the conditional probabilities to be in the top 0.5% cost category given a certain attribute vector $\mathbf{X}_i$ with components X_{ij}, where j runs over the values of the attribute i and $*$, where $*$ denotes a sum over all attribute values. We used 184 medical attributes associated with Diagnostic Cost Groups (DCGs) [2], which give a complete classification of medical conditions, and a set of quarterly cost data consisting of a further 30 variables. A GA was used to search for "fit" classifiers, fitness being measured by $\epsilon = N_{X_i}(P(\text{top } 0.5\%|\mathbf{X}_i) - P(\text{top } 0.5\%))/(N_{X_i}P(\text{top } 0.5\%)(1 - P(\text{top } 0.5\%)))^{1/2}$, where N_{X_i} is the number of individuals in the data with attribute vector $\mathbf{X}_i$. For the GA we determined the following optimal parameter values: population = 100, no. of generations = 50, mutation rate = 0.1 and crossover rate = 1. The fittest 100 classifiers over a set of runs were filtered out. This filtered list was then sorted according to a score function, S_1. Finally, a score, S_2, was assigned to an individual by an assignment algorithm between the classifiers and the individual. The highest scoring top 0.5% according to S_2 for year N was the prediction for year $N+1$ and was compared with taking the highest ranked cases based on each of two widely-used benchmarks in health care modeling: 1) year N cost [1]; and 2) "out-of-the box" year-N DCG prospective risk scores [2].

K. Deb et al. (Eds.): GECCO 2004, LNCS 3103, pp. 1330–1331, 2004.

3 Results

Best results were obtained using for S_1 - Winner-rerank-reevaluation. In this case the first, fittest, classifier is fixed. Individuals corresponding to this classifier are removed from the other classifiers in the list, the fitness recalculated and the list reranked. Passing to the subsequent classifiers this reevaluation procedure is iterated until one reaches the final classifier in the list. This procedure helps prevent poor classifiers from "hitchhiking" on the back of fitter ones. Finally, the final list is reranked using $P(\text{top } 0.5\%|\mathbf{X}_i)$ directly rather than ϵ. For S_2 a simple "winner-takes-all" assignment strategy was used, where an individual was assigned a score that was the final score of the winning classifier after the reevaluation and reranking procedure.

N		Benchmark 2	Benchmark 1	Score function
1997	# correct	29	31	52.5
	% correct	20	21.3	36.2
	$	11.7	11.4	15.9
1998	# correct	35	34	53.5
	% correct	18	17.5	27.6
	$	14.4	12.7	17.0
2000	# correct	82	93	132.1
	% correct	18.2	20.7	29.4
	$	35.6	35.3	46.6

We see above both in- and out-of-sample results. In-sample data were used to determine a set of optimal classifiers from predicting 1998 costs with $N = 1997$ data. These classifiers were then used for predicting the out-of-sample years 1999, using $N = 1998$ data, and 2001, using $N = 2000$ data. The results are averages over 10 runs. Both performance measures are shown - number/percentage of correctly identified individuals in the top 0.5% of year $N+1$ costs and the dollar amount ($ - millions) of costs associated with the predicted group.

The GA-discovered classifiers gave average out-of-sample improvements of 47% and 59% in predictive accuracy for individuals in the top 0.5% of next year costs over benchmarks 1 and 2 respectively - these being the most common in the industry. There were also significant improvements in the total dollar amount. The classifier system is also ideal for identifying important drivers of costs as that is precisely what the genetic search through the classifier space is determining.

References

1. W.F. Bluhm and S. Koppel, *Individual Health Insurance Premiums*, In *Individual Health Insurance*, ed. F.T. O'Grady, 59-61, Society of Actuaries, Schaumberg IL, (1988).
2. A. Ash, R.P. Ellis, G.C. Pope, J.Z. Ayanian, D.W. Bates, H. Burstin, L.I. Iezzoni, E. McKay and W. Yu, *Using Diagnoses to Describe Populations and Predict Costs*, *Health Care Financing Review* **10** (4), 17-29 (2000).