

# Predicting Healthcare Costs using GAs

C. R. Stephens\*  
Instituto de Ciencias  
Nucleares, UNAM  
A. Postal 70-543  
México D.F. 04510

stephens@nucleares.unam.mx

S. Talley  
Adaptive Technologies Inc.  
6424 West Chisum Trail  
Glendale, AZ 85310  
susantalley@at-inc.biz

## ABSTRACT

Predicting prospective healthcare costs is of increasing importance. Genetic search is used to discover attribute sets and associated posterior probability classifiers that predict the top 0.5% most costly individuals in year  $N + 1$  based on previous medical conditions and costs in year  $N$ . The predictive performance of single-variable classifiers (cost-drivers), found using statistical measures familiar from datamining, as well as Naive Bayesian analysis, are compared and contrasted with that of classifiers found using genetic search. Comparison is also made with two well known benchmarks from the healthcare literature.

## 1. INTRODUCTION

In the battle to control escalating health care costs predictive models of varying degrees of sophistication are increasingly being employed to try and predict from a given population who is most likely to be a high cost case. This is clearly of great importance for many different types of player in the healthcare sector, both public and private: ranging from private sector insurance companies, to large government agencies, such as Medicare and Medicaid, and healthcare service providers, such as hospitals. The estimated costs of an individual or group of individuals play a crucial role in setting insurance rate premiums and estimating risk, estimating public healthcare costs, pricing healthcare services and identify the “best” cases for preventive case management, among others. In general, better prediction permits a more optimal allocation of healthcare resources.

Healthcare cost distributions are known to be highly skewed, with a few people generating the majority of costs [1, 2]. For instance, a 1996 government survey [2] found that the most costly 1 percent of the population consumed 27 percent of the resources, while the top 5 percent consumed 55 percent. This suggests the importance of prospectively iden-

\*Part of this work was done while on sabbatical leave at the Department of Computer Science, University of Essex, Wivenhoe, CO4 5SQ.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*Genetic and Evolutionary Computation Conference (GECCO) '05*, 25th-29th June, 2005, Washington D.C.

Copyright 2005 ACM 1-59593-097-3/05/0006 ...\$5.00.

tifying those individuals most likely to incur excessively high healthcare costs. Various methods have been used to predict high cost individuals. A common actuarial technique [3] uses total year  $N$  cost, age and gender to predict year  $N + 1$  costs. Such methods, however, do not focus on medical conditions as the primary drivers of cost, and therefore their conclusions fail to provide actionable information to medical practitioners. Diagnostic Cost Groups (DCGs) [4, 5], on the other hand, use age, gender and the range of medical problems encountered in year  $N$  to predict year  $N + 1$  costs. Specifically, cost weights for different classes of medical problems are determined via a linear regression analysis of year  $N + 1$  costs versus year  $N$  medical problems, age and gender. More sophisticated techniques, such as Bayesian Markov Chain Monte Carlo methods [6], in the context of predicting mean annual costs, and neural networks [7], in the context of predicting length of hospital stay, have been used, but not in the context of identifying high cost individuals, i.e. those individuals most likely to lead to very high medical costs and who might benefit from medical intervention.

The goal of this research was to predict the top 0.5% most costly individuals for year  $N + 1$  given a population in year  $N$ . Two performance metrics were considered: i) classification accuracy, i.e. the proportion of correctly classified individuals in the top 0.5% and ii) the total number of dollars associated with the predicted top 0.5% of most costly individuals. Predictions are produced by using a Genetic Algorithm (GA) to search for sets of attribute values,  $\mathbf{X}$ , that are predictive of the class,  $Y$ , of most costly individuals. Posterior probability classifiers of the form  $P(Y|\mathbf{X})$  then form the “raw material” from which predictions are made. Of course, classification is an enormous field, with a large number of associated tools, both from standard statistical analysis and artificial intelligence (see [8] or [9] for good overviews).

## 2. THE DATA SET AND BENCHMARKS

Data for the period 1997-2001 was taken from the MED-STAT Marketscan Research Database for a cohort of privately insured individuals diagnosed with diabetes.<sup>1</sup> The years  $N = 1997, 1998$  and 2000 were considered. For each year  $N$ , an individual was associated with a set of year  $N$  data that forms an attribute array  $X_{ijk}^N$ , where  $k$  refers to the individual,  $N$  the year,  $i$  to the attribute and  $j$  to the

<sup>1</sup>Diabetes is an expensive, progressive disease in which prospective identification and case management can be very useful.

corresponding attribute value.<sup>2</sup> Also, associated to each individual is their year  $N + 1$  total costs.<sup>3</sup> For example, the data for 1997 individuals were 1997 data - DCGs, 1997 costs etc. - and 1998 total costs, while the data for 1998 individuals contained 1998 DCGs etc. and 1999 total costs. The size of the cohort varied significantly from year to year with 29,062 1997 cases, 38,879 1998 cases and 90,104 2000 cases.

Predicted next year costs were determined principally from current year medical conditions and current year medical costs. Current costs were given quarterly in three classes - inpatient, outpatient and pharmacy. Medical conditions were classified according to DCGs (Hierarchical Condition Categories - HCCs) in which all medical conditions are classified into one of 184 HCCs. The HCC variables are binary in that an individual either did or did not manifest the condition in year  $N$ . Additional variables were also utilized, such as an individual's age and gender.

To evaluate performance two benchmarks that are widely used in the industry were used for comparison. Benchmark 1 [3] used year  $N$  cost as a predictor of year  $N + 1$  costs, the intuition being that people who are expensive in one year are likely to be expensive in the next. Of course, this is less true for someone with an acute condition. Benchmark 2 [4] used "out-of-the box" year- $N$  DCG prospective risk scores from DxCG Inc.'s proprietary software. Note that Benchmark 2 purposely avoids using year  $N$  cost information, because it seeks to measure medical need, independent of practise style variations, as much as possible. The goal in this paper is different: to predict people likely to be expensive, whether or not that expectation is driven by pure need. Finally, we compared our results to those found using a Naive Bayesian classifier, the latter being widely regarded as an outstanding model for classification [10].

### 3. THE BASIC METHODOLOGY

The basic methodology is to search for "fit" classifiers that identify those year  $N$  attributes, and/or attribute values, that are most predictive of individuals in the top 0.5% for next-year ( $N + 1$ ) costs. These classifiers then serve as the "raw material" by which, using an assignment rule from classifiers to individuals, individuals are scored and ranked according to their probability of being in the top 0.5% of next-year costs.

For data corresponding to an individual,  $k$ , in a given year,  $N$ , the HCCs, cost and other data form an attribute array with components  $X_{ijk}^N$ . The fundamental objects of interest are  $P(\text{top } 0.5\% | \mathbf{X}^N)$  - the conditional probabilities to be in the top 0.5% cost category in year  $N + 1$  given a certain attribute set  $\mathbf{X}^N$  with components  $X_{ij}^N$ , the probabilities having been summed over the individuals  $k$  that correspond to  $i$  and  $j$ . For example, for three binary attributes  $P(\text{top } 0.5\% | \mathbf{X}^N = 1 * 0)$  represents the probability to be in the top 0.5% when attributes 1 and 3 take value 1 and 0, while attributes 2 can be either 0 or 1.<sup>4</sup>

In the search for fit classifiers three approaches are compared and contrasted: i) statistical measures familiar from

<sup>2</sup>We also add to the attribute values \*, which denotes any attribute value. For instance,  $1 * 6$  would signify that attribute 1 has value 1, attribute 2 has any value and attribute 3 has value 6.

<sup>3</sup>All costs were normalized to year 2000 dollars.

<sup>4</sup>This probability is just a marginal of the underlying probability distribution.

datamining to identify the key *single* drivers of high costs; and ii) a GA to identify fit attribute sets/classifiers; and iii) a Naive Bayesian analysis. For i) a score,  $S_d$ , that depends on the most important cost drivers was used to rank individuals and determine those most likely to be in the top 0.5%. In contrast, for ii) the fittest 100 classifiers over a single run, or 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, a classifier only being a candidate for contributing to  $S_2$  for an individual if the individual matches or "activates" the classifier  $\mathbf{X}^N$ . For instance, using the three attribute example above: an individual with attribute values 100 activates the classifier  $1 * *$  but not  $* 1 *$ . The highest scoring top 0.5% according to  $S_2$  for year  $N$  is then the prediction for year  $N + 1$ .

### 4. KEY SINGLE-VARIABLE DRIVERS

To identify the most important drivers of high next-year costs the data was first analyzed using some basic "signal-to-noise" ratios. One of these is  $\epsilon(X_{ij})$ , which measures the utility of a given feature value  $X_{ij}$  for classifying data into a class  $Y$ . Explicitly,

$$\epsilon = N_{X_{ij}}(P(Y|X_{ij}) - P(Y)) / (N_{X_{ij}}P(Y)(1 - P(Y)))^{1/2} \quad (1)$$

where  $N_{X_{ij}}$  is the number of individuals in the data with value  $j$  for attribute  $i$ .

Another useful measure, that concentrates on the relevance of features, rather than feature values, is  $\epsilon'$  defined as

$$\epsilon' = (\langle X_i \rangle_Y - \langle X_i \rangle_{\bar{Y}}) / \left( \frac{\sigma_Y^2}{N_Y} + \frac{\sigma_{\bar{Y}}^2}{N_{\bar{Y}}} \right)^{1/2} \quad (2)$$

where  $\langle X_i \rangle_Y$  is the mean of the feature  $X_i$  for class  $Y$  and  $\langle X_i \rangle_{\bar{Y}}$  is the mean for another class,  $\bar{Y}$ , such as the complement of  $Y$ .  $\sigma_Y^2$  and  $\sigma_{\bar{Y}}^2$  are the corresponding variances, and  $N_Y$  and  $N_{\bar{Y}}$  the number of observations in  $Y$  and  $\bar{Y}$ .

To identify key drivers of next year's costs the average value of an attribute in three different classes of year  $N + 1$  costs was examined: top 0.5%, next top 0.5% and next 4%, and compared to an overall population average. The drivers were selected according to how well they discriminated among the different cost classes and the overall average using  $\epsilon'$ . All year  $N$  costs, including overall costs, inpatient and outpatient, as well as their quarterly components, gave good discrimination. Also of interest was a strong "seasonal" trend, apparent in the quarterly figures, with approximately equal costs in Q1 and Q2, a strong decrease and a strong increase of 40% in Q4. High cost weight HCCs are more discriminating than low cost weight ones.

To enhance performance relative to Benchmark 2, features which best discriminated between false positives in the top 0.5% and false negatives from the next top 0.5% of Benchmark 2 were sought. Year  $N$  cost was useful in this regard, the implication being that some of those cases from the top 0.5% of Benchmark 2, but with low year  $N$  cost, should be replaced by cases from the next top 0.5% of Benchmark 2 that have much higher year  $N$  costs. Among the annual cost variables, inpatient costs showed relatively poor discrimination while greater outpatient costs were found to be most prevalent in true positives and false negatives, with increases of 95% for true positives over false positives and

150% for false negatives over true negatives. Hence, this feature is very relevant. It was also noted that Q4 total costs gave ample discrimination. High Q4 costs, above and beyond the previously identified seasonal trend, indicate a deterioration, as opposed to an improvement, in health.

With these important drivers in hand a score function,  $S_d$ , was created that depended on them. Individuals were then ranked with respect to  $S_d$ . In Table 1 below, the performance comparison between Benchmark 1, Benchmark 2 and the Score function is shown. \$ figures are in millions of dollars. The two benchmarks have comparable performance for both performance measures.  $S_d$  was constructed using 1997 data only, leaving the larger 1998 and 2000 datasets for out-of-sample validation. The improvement in classification accuracy over the average from the benchmarks is 33% in 1997 (in sample) and 28% on average in the out of sample dataset. This demonstrates the existence of significant predictors of next year costs other than DCGs (HCCs) and year  $N$  total costs.

N		Benchmark 2	Benchmark 1	Score function $S_d$
1997	# correct	29	31	40
	% correct	20	21.3	27.6
	\$	11.7	11.4	13.2
1998	# correct	35	34	40
	% correct	18	17.5	20.6
	\$	14.4	12.7	14.1
2000	# correct	82	93	116
	% correct	18.2	20.7	25.8
	\$	35.6	35.3	41.4

## 5. GENETIC SEARCH IN THE SPACE OF CLASSIFIERS

Although interesting and useful classifiers may be constructed by the analysis of section 4, the curse of dimensionality demands an intelligent search mechanism if one wishes to move beyond single-variable classifiers. For example, for the medically based HCC attributes alone, there are  $2^{186}$  potential attribute combinations! In this study a GA was used to search the space of attribute combinations/classifiers, a classifier being coded as an  $L$ -bit schema. The initial population was determined at random, but using a high probability, usually 0.98, for putting an \* at any given bit position, as higher order schemata correspond to so few cases in the data they would be statistically unreliable or non-existent.

Experiments were carried out to establish the performance of the GA as a function of: number of generations, population size, mutation rate and crossover rate. Simple one-point crossover and proportional selection were used throughout. Elitism with memory was also used, in which a list of fixed size, usually 100, of the best classifiers found over a particular run or set of runs was kept. After a fixed number of generations, in the case of a single run, or a fixed number of runs in the case of multi-runs, a final, filtered, ranked list of classifiers was used to determine the individuals in the top 0.5% class.

Two fitness functions were considered:  $f_1$  - the value of  $P(Y = \text{top } 0.5\% | \mathbf{X}^N)$  for a given classifier  $\mathbf{X}^N$ ; and  $f_2$ , given by  $\epsilon$  in equation (1) above, once again with  $Y = \text{top } 0.5\%$ . Both have advantages and disadvantages.  $f_1$  has the disadvantage that it does not take into account statistical reliabil-

ity, as the maximum fitness value,  $f_1 = 1$ , may be associated with a very small sample. For example, one may find that there are only two individuals with a certain set of attribute values and that both are high cost individuals.  $f_2$ , on the other hand, is a better measure of signal to noise. However, given that the objective is to optimize precision, not statistical reliability, this fitness function may overemphasize less predictive features associated with large sample sizes.

A potential defect of a classifier-based approach, as outlined thus far, is the mutual dependence among classifiers. This can most simply be illustrated in the language of conditioning information as schemata. Imagine three binary features -  $X_1$ ,  $X_2$  and  $X_3$ . There are  $3^3$  associated schemata, e.g. 111, 11\*, \*\*0 etc. Consider the schemata: 111, 11\* and 10\*, with fitnesses  $f_{111}$ ,  $f_{11*}$ , and  $f_{10*}$  respectively, where  $f_{111} > f_{11*} > f_{10*}$ , i.e. the classifiers are ranked in order: 111, 11\*, 10\*. The first two schemata are clearly not linearly independent. However, 111 and 11\* - 111, where - means set difference, are independent. If  $f_{(11*-111)} < f_{10*}$  then the ranking 111, 11\*, 10\* is erroneous in the sense that  $f_{11*} > f_{10*}$  only because of the fit individuals inherited from 111. However, these have already been counted. Removing this redundancy and forming the classifiers: 111, 11\* - 111 (= 110) and 10\*; as  $f_{110} < f_{10*}$  the new classifiers should be reranked to order: 111, 10\*, 110.

Having determined a fixed number of fit classifiers, using fitness as a filter, one then ranks this filtered set using a score function,  $S_1$  (which may be just the already assigned fitness) and specifies a rule that assigns to an individual a score,  $S_2$ , measuring the probability that this individual is in the top 0.5% cost category for year  $N + 1$  and derived from the classifiers associated to this individual. Several ranking and scoring algorithms were considered:

- (i) **Winner:**  $S_1$  is one of the fitness functions,  $f_1$  or  $f_2$ .  $S_2$  for an individual is given by the value of  $S_1$  associated with the highest ranked classifier from the final list activated by that individual. If there is a tie between individuals, priority is given to the individual that activated the most classifiers.
- (ii) **Winner-rerank:** The same as i) except that the final list of classifiers is reranked before individuals are assigned. Thus, if fitness function  $f_1$  was used in the GA the final list of classifiers is reranked using  $f_2$  and vice versa. Ties are resolved as in i). The point of this reranking is to try and ameliorate the fitness function defects discussed above.
- (iii) **Average:**  $S_1$  is as in i), however,  $S_2$  is calculated by considering all the classifiers from the final list activated by the individual and taking the average of their  $S_1$  values. Ties are resolved as in i).
- (iv) **Match:**  $S_1$  is as in i).  $S_2$  for an individual however, is given by the number of classifiers activated by that individual. Ties are resolved by giving priority to the individual that activated the classifier of highest fitness.
- (v) **Winner-reevaluation:**  $S_1$  is obtained by iteratively removing redundancy using the algorithm discussed above. Thus: the first classifier is fixed. Individuals associated with this classifier are removed from the other classifiers in the list, fitness recalculated and the list reranked. The second classifier in the new list is

now fixed, the first classifier being already fixed, and any individuals associated with this classifier are removed from the rest, fitness is recalculated and the list reranked. This procedure is iterated until one reaches the final classifier in the list. Individuals are now assigned as in i) but using the new final classifier list.

- (vi) **Winner-rerank-reevaluation:** This works identically to v) except after the last reranking a further reranking is done according to ii) above. In other words this assignment and ranking algorithm both removes the redundancy problem and ameliorates the defects of the fitness functions. Individuals are now assigned as in i) but using the new final classifier list.

## 6. RESULTS

The most useful parameters for the GA were determined first. For population size:  $P = 10, 50, 100, 500, 1000$  were tested; for number of generations:  $G = 5, 10, 50, 100, 500$ ; for mutation rate:  $\mu = 0.001, 0.01, 0.05, 0.1, 0.2, 0.3$ ; and for crossover rate  $p_c = 0, 0.5, 1$ . Each parameter was evaluated by considering single runs over all values of the other parameters and taking the average performance in terms of percentage of correctly identified individuals in the top 0.5% class for next year costs. Hence, 900 runs were performed in total. Averages were also taken over all 6 scoring algorithms of section (5). The parameter values  $P = 100, G = 50, \mu = 0.1$  and  $p_c = 1$  were found to be the most appropriate.

The performance of the prediction classifiers found by genetic search was compared against Benchmarks 1 and 2, the Score function,  $S_d$ , of section 4 and the results of a Naive Bayesian analysis. The reported results are averages over 10 independent runs. Both performance measures - number/percentage of correctly identified individuals in the top 0.5% of year  $N+1$  costs and the dollar amount of costs associated with the predicted group - were considered. Results for the different ranking and assignment functions from section 5 are also given, where in the tables A\_SCORE = Average, M\_SCORE = Match, W\_SCORE = Winner, R\_SCORE = Winner-rerank, WR\_SCORE = Winner-reevaluation and RR\_SCORE = Winner-rerank-reevaluation.

Figure 1 shows the results associated with taking as training set classifiers for predicting 1998 data from 1997 data, with "1997 training data" representing the performance on this training data, while "1998 test data" shows the performance of these classifiers on test data consisting of predictions for 1999 data from 1998 data. Similarly, "2000 test data" shows the results of the 1997 trained classifiers on test data consisting of predictions for 2001 given 2000 data. Analogous results were obtained using as training set classifiers found by predicting 1999 data from 1998 data and 2001 data from 2000 data respectively.

The performances of the different assignment and ranking functions are compared first. Most notable is the poor performance of the Average method. Naive intuition might lead one to think that such a "consensus"-type function should lead to a robust performance. One can envision different explanations as to why this does not occur here. Primary among them is that averaging includes classifiers that are relatively unfit compared to their more predictive counterparts. Certainly the performance of the Winner algorithm is superior. However, note that the performance of the Match

1997						
29062 individuals (145 individuals in the top 0.5%)						
AVERAGES OF NUMBER OF CORRECTS IN THE TOP 0.5%						
	A_SCORE	M_SCORE	W_SCORE	R_SCORE	WR_SCORE	RR_SCORE
1997 training data	36.5	51.8	48.8	50.4	50.4	52.5
1998 test data	12.8	49	46.4	48	50.1	53.5
2000 test data	96.3	120.6	129.9	126.7	130	132.1
AVERAGES OF CORRECT % IN THE TOP 0.5%						
	A_SCORE	M_SCORE	W_SCORE	R_SCORE	WR_SCORE	RR_SCORE
1997 training data	0.25175	0.35722	0.33652	0.34756	0.34756	0.36205
1998 test data	0.06598	0.25257	0.23918	0.24743	0.25823	0.27578
2000 test data	0.214	0.26801	0.28867	0.28155	0.28888	0.29355
AVERAGES OF \$ (DOLLARS) IN THE TOP 0.5%						
	A_SCORE	M_SCORE	W_SCORE	R_SCORE	WR_SCORE	RR_SCORE
1997 training data	12696266.9	16197381.1	15426708.1	15513532.2	15660352.6	15945161.4
1998 test data	6104883.62	17016515.1	16562169.1	16631128.8	16724903.2	16976880
2000 test data	36132774.2	44103103.4	46542682.3	45547032.8	46335143.6	46589881

Figure 1: Performance table for training data from  $N = 1997$

algorithm, which is also based on consensus, is better than Average and comparable with Winner. A possible explanation for this is the following: Average suffers from the phenomenon of classifier redundancy. Hence, classifiers are weighted incorrectly in the averaging procedure. For instance, if five agents are effectively detecting the same illness from slightly different viewpoints, this cost component would be overestimated relative to an equally severe illness that only activates a single classifier. The effect of removing redundancy seems to be significant and worthy of further study.

Considering performance relative to the Benchmarks - apart from the Average algorithm, which showed relatively small improvements over the benchmarks in the main - all assignment and ranking functions showed large improvements over both Benchmarks and with respect to both performance measures. In the case of the most predictive assignment algorithm - Winner-rerank-reevaluation - the average out-of-sample performance increases for predictive accuracy for individuals in the top 0.5% of next-year costs over Benchmarks 1 and 2 were 47% and 59% respectively. Comparing with the performance of  $S_d$ , an average out-of-sample improvement of 25% is seen, thus proving the value of an intelligent genetic search as opposed to a "hand made" approach. It is also interesting to note that there is very little decline in performance when passing from training data to test data, thus showing that the system is generalizing well and not suffering from significant overfitting.

We have seen that genetic search was capable of discovering predictive classifiers that led to superior performance when compared to a pair of industry benchmarks and some simple data mining models. However, there are a large number of other potential competitor techniques that could have been used. Obviously, one cannot compare with them all. Here we compare with a well known technique that has been found to be better, or at least competitive, on large classes of different problems - Naive Bayesian classifiers [10] that yield posterior class probabilities of the form  $P(Y|\mathbf{X}) = \prod_i P(\mathbf{X}_i|Y)P(Y)/P(\mathbf{X})$ . In the Table below we see the out of sample results for 2000-2001 for Naive Bayesian classifiers formed from likelihood functions  $P(\mathbf{X}|Y)$  determined on the (in sample) 1997 data. In order to further investigate the relative predictability of the HCCs and the cost based variables Naive Bayesian classifiers were calculated separately for all variables, just the HCCs and just the non-HCC variables.

N		ALL	No HCC	HCC only
2000	# correct	102	106	68
	% correct	22.7	23.6	15.1
	\$	39.5	41.1	31.3

Comparing with the results for Benchmarks 1 and 2 and the score function  $S_d$  the Naive classification leads to an improvement with respect to the two industry benchmarks for both all variables and only non-HCC variables. Thus we see that cost based variables are much more predictive than HCCs in the context of a Naive Bayesian analysis, both in terms of predicting high cost individuals as well as \$ spent. Noticably, however, even the best Naive predictions are slightly inferior,  $\approx 10\%$ , to the score function  $S_d$  and, comparing with Figure 1, up to 25% worse than the best GA based results. One of the most likely reasons for the superior performance of the GA-discovered classifiers is that there are substantial correlations associated with certain attribute sets that the Naive classifiers neglect.

## 7. CONCLUSIONS

In this paper an important problem in the healthcare industry has been addressed - prediction of those individuals - the top 0.5% - most likely to lead to high medical costs and who, for example, may benefit from medical intervention. Using both cost and medical data, for the first time in this problem area, a classifier-based approach was used. Predictive classifiers were determined using: i) statistical measures familiar from datamining; and ii) a GA. With the former, only "single-variable" classifiers that correspond to key cost drivers were determined, thus precluding any accounting of non-linear interactions between variables.

Predictive performance was compared to that of two standard industry benchmarks. It was shown that a classifier-based approach led to significant performance gains relative to these benchmarks. However, the efficacy of the genetic search was significantly better than that of the more traditional data mining approach or of a more sophisticated Naive Bayesian analysis. Average out-of-sample improvement for correctly classifying the top 0.5% most costly individuals was 28% when using the "standard" datamining approach and 53% when using the genetic search. These results show the utility of a classifier-based approach in general, as it allows for a more democratic analysis (both cost and medical data together) than either of the benchmarks, and, more particularly, the advantage of using an intelligent genetic search in the very high-dimensional space of potential classifiers, where non-linear interactions between different attributes may be accounted for. Several subtle issues that can greatly impact the predictive capacity of the classifiers found in such a search were also highlighted. Among these were the type of fitness function used, the ranking algorithm for comparing classifiers, the mutual statistical dependency of related classifiers and the score assignment algorithm between individuals and classifiers.

## Acknowledgements

CRS wishes to thank the EPSRC for support (grant number GR/T2461/01) while on sabbatical leave at the University of Essex, as well as DGAPA, UNAM and Conacyt project 30422-E.

## 8. ADDITIONAL AUTHORS

Additional authors: R. Cruz (Adaptive Technologies Inc.) and A. Ash (Boston University School of Medicine and DxCG Inc., Boston MA).

## 9. REFERENCES

- [1] G.F. Anderson and J. Knickman, *Patterns of Expenditures Among High Utilizers of Medical Care Services*, *Medical Care* **22** (2), 143-149 (1984).
- [2] M.L. Berk and A.C. Monheit, *The Concentration of Health Care Expenditures Revisited*, *Health Affairs* **20** (2), 9-18 (2001).
- [3] W.F. Bluhm and S. Koppel, *Individual Health Insurance Premiums*, In *Individual Health Insurance*, 59-61, Society of Actuaries, Schaumburg IL, (1988).
- [4] 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).
- [5] Y. Zhao, A.S. Ash, J. Haughton, B. McMillan, *Identifying future high-cost cases through predictive modeling*, *Disease Management and Health Outcomes* **11** (6), 389-397 (2003).
- [6] N.J. Cooper, P.C. Lambert, K.R. Abrams and A.J. Sutton, *Predicting the cost of illness over time using Bayesian Markov chain Monte Carlo methods: An Application to Early Inflammatory Polyarthritis* Dept. of Health Science, Univ. of Leicester Technical Report 03-04 (2003).
- [7] M.R. Kraft, K.C. Desouza and I. Androwich, *Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population*, In *Proceedings of the 36th Hawaii International Conference on System Sciences*, IEEE (2002).
- [8] D. Hand, H. Mannila and P. Smyth, *Principles of Data Mining*, MIT Press, MA (2001).
- [9] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification*, Wiley-Interscience (2000).
- [10] P. Domingos and M. Pazzani, *Beyond independence: Conditions for the optimality of the simple Bayesian classifier*, in *Proceedings of the 13th International Conference on Machine Learning*, 105-112 (Morgan Kaufmann 1996).