OPERATIONS RESEARCH
DOI 10.1287/opre.1080.0619ec
pp. ec29–ec41

e-companion

ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Algorithmic Prediction of Health Care Costs"

by Dimitris Bertsimas, Margrét V. Bjarnadóttir, Michael A. Kane, J. Christian Kryder, Rudra Pandey, Santosh Vempala, and Grant Wang,

Operations Research, DOI 10.1287/opre.1080.0619.

A Appendix - Detailed List of Variables and Examples of Coding Groups

Table 1 on page 8 summarizes the variables used in our study, what follows is a more detailed list.

1. Variables 1 thought 218 are the number of claims for a member, that have coding belonging to different diagnosis groups. A diagnosis groups is a collection of ICD-9 codes that have been put together to form a group. As an example, Table 15 in Appendix D shows all ICD-9 codes that fall into the diabetic group.

2. Variables 219 through 398 are the number of claims for a member, that have coding belonging to different procedure groups. These groups were created in the same fashion as the diagnosis groups. Table 16 in Appendix D contains an example of all ICD-9 codes that fall into our MRI-scan group.

3. Variables 399 through 734 are the number of claims for a member, that have coding belonging to different drug groups. The grouping is bases on the NDC codes, and drugs from the same drug class are grouped together. Table 17 in Appendix D contains an example of all drugs in the Insulin group.

4. Variables 735 through 1485 are indicators of additional specified risk factors. We can divide the risk factors into 4 main categories:

   (a) Interaction between illnesses, an example is: diabetes and obesity.

   (b) Interaction between diagnosis an age, an example is: CAD and age above 65.

   (c) Noncompliance to treatment an example is: a pattern of ER care without office visits.

   (d) Illness severity, an example is: diabetes with foot ulcers.

5. Variables 1486 through 1489 are counts of the number of different diagnosis, procedures, drugs and risk factor a member has during the observation period.

6. Variables 1490 through 1511 are cost variables. Those are:

   (a) Monthly costs of the last twelve months of the observation period (12 variables).

   (b) Overall pharmacy costs (1 variable).

   (c) Overall medical costs (1 variable).

   (d) Overall cost (the sum of medical and pharmacy costs) (1 variable).

   (e) Overall cost in the last 6 months of the observation period (1 variable).

   (f) Overall cost in the last 3 months of the observation period (1 variable).

   (g) Positive and negative trends, found by fitting a line through the last monthly costs of the observation period (2 variables).

(h) Acute indicator, a indicator variable found by comparing the highest month with the average monthly cost. If these are significantly different, the indicator takes on the value 1. (1 variable)

(i) Number of months above average. This variable is an indicator of the shape of the cost profile. If the cost is relatively constant over the period, this variable takes on a value around six, which is an indicator for a chronic cost profile (1 variable).

(j) The cost of the highest month in the observation period.

7. Variable 1512 is an indicator variable for the gender being female.

8. Variable 1513 is the age of the member at the beginning of the observation period.

B Appendix - Classification Trees

Classification trees recursively partition the independent variable space into a set of subspaces and assign a separate classification rule to each subspace. This partitioning can be represented as a tree. We start with the whole sample space at a root node, and then partition the data set into two subsets according to a splitting rule designed to minimize a node impurity measure that has been defined. This first split is shown in Figure 4. The process then continues dividing up the subspaces, until a defined stopping criteria is satisfied.

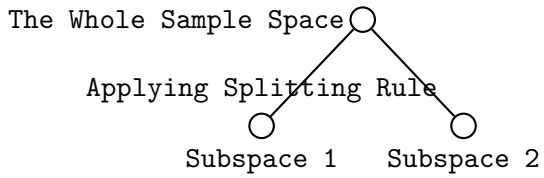

Figure 4: The first step in recursive partitioning, creating the first two sub nodes.

We used the software CRUISE [34] to create our classification trees. What follow is a discussion on some of the specifics of the algorithm, we refer the reader to [34] and [35] for further details.

Notation

Let X be a vector of independent variables and Y be a categorical dependent variable that takes k

different values. Let t be any node in a classification tree and let p_{it} be the fraction of the observations for which $y = i$ at node t. Let N_t be the collection of observations at node t.

Node impurity measure

The most common node impurity measure, and the one that we use, is the Gini-index, which is defined as

$$1 - \sum_{i=1}^{k} p_{it}^2,$$

which can be rewritten as

$$\sum_{i \neq j} p_{it} \cdot p_{jt}.$$

In the case of uneven misclassification costs, as in the case of the penalty matrix error measure, the Gini-index is adjusted to

$$\sum_{i \neq j} W(i,j) p_{it} \cdot p_{jt}$$

where $W(i,j)$ is the cost of misclassifying as i a class j case. Note that if all the observations belong to a single class, then the Gini-index is zero.

Splitting rule

The split is chosen to give the largest reduction in the defined node impurity measure. The split can either rely on a single variable or multiple variables. To preserve interpretability, we choose to use single variable splits. There are two main categories of univariate splitting methods: exhaustive search and methods that do statistical hypothesis tests at each node to assess the significance of a split.

Loah and Shih [37] show that the key to avoiding selection bias is the separation of variable selection from split point selection. This separation differs from the exhaustive search approach of simultaneously finding the variable to split on and the splitting value. It has also been noted that exhaustive search methods are biased toward categorical variables over numerical variables as well as toward continuous variables over discrete variables, because the continuous variables afford more splits. An algorithm that overcomes this bias is 1D as described in [35], where at each node an analysis of variance (ANOVA) F-statistic is calculated for each variable. The variable with the largest F-statistic is selected, and linear

discriminant analysis is applied to it to find the split value. This is the algorithm used in this paper.

Classification rule at end nodes

At an end node, the class that minimizes a error measure is assigned to the node. For example, for our penalty error we assign class i to an end node t if i minimizes

$$\sum_{j \in N_t} P(i, y_j),$$

where P is the penalty matrix defined in Section 2.4 and y_j is the observed class of member j. In the case when the costs for all misclassifications are equal, the assignment rule simplifies to assign the most frequently observed class at an end node as the classification rule. For the average absolute dollar error we assign the median cost of the learning sample at each node.

Pruning and selecting the "best" tree

At some point, the recursive partitioning needs to be stopped. This stopping point can be predefined, limited by the number of levels, minimum number of observations at a node, or when improvement in the node impurity measure is negligible. A methodology introduced by Brieman *et al.* [29], which is used here, is to overgrow the tree and then prune it back using the pruning sample. After overgrowing the tree, the classification rule is applied to the pruning sample and the misclassification cost is calculated at each node and at each parent node. We then cut off the nodes that result in the smallest increase (or the biggest decrease) in the overall misclassification cost. The result is a sequence of trees, each associated with a certain misclassification cost. The tree with the smallest misclassification cost is called 0-SE [29]. The smallest tree within one standard error of the minimum is called 1-SE and is the tree we choose to use.

C Appendix - Clustering

This Appendix explains how the Eigencluster algorithm was adapted for medical data mining. For details on the original version of Eigencluster we refer the reader to [38] and [33].

C.1 Notation and Outline

Given a learning set L and a validation set V of patients for whom we know the result period's cost, we predict the cost for the test set T according to the below procedure. We view each member as a vector of features, which belong in one of two categories – cost and medical. Let L_i, V_i, T_i be the subset of members in the learning, validation and test set respectivly that belong to bucket i in the observation period. The procedure is as follows, and applies to any measure (the hit ratio, the penalty error, the average absolute prediction error, etc.).

1. Define feature weights
2. Apply feature weights to $L_i \cup V_i \cup T_i$.
3. Use EigenCluster to cluster $L_i \cup V_i \cup T_i$ based only on cost features. Let R_i be the resulting hierarchical clustering tree.
4. Using L_i and V_i, compute the frontier F_i of R_i for which clustering based on medical information is at least as good as clustering based on cost information.
5. For each node C in F_i, let x be the single prediction that optimizes the sum of the measure on $C \cap (L_i \cup V_i)$. Use x as a prediction for each test member in $C \cap T_i$.

Below we briefly discuss each of the outlined steps.

C.2 Define feature weights

We define two sets of weights: cost weights and medical weights. The cost weights apply to the cost features, whereas the medical weights apply to the medical features. As the last months of the observation period have stronger correlation with the result period, the last months are given more weight than the first. Equal weight is given to each of the medical features.

C.3 Applying weights

For every member u (vector of features) in $L_i \cup V_i \cup T_i$, we apply the weights $\vec{w}$ by setting $u_i \leftarrow \sqrt{w_i} u_i$. Thus, the inner product between members is now the weighted similarity.

C.4 Using EigenCluster

The goal in applying EigenCluster is to put members who have similar cost patterns together in a "cluster". The hypothesis is that members with similar cost patterns in the observation period will have similar future cost patterns. In each "cluster", there will be members of the learning, validation and test sets. Thus, we will make a prediction for each of the members of the test set based on the result period behavior of the learning set and validation sets.

Technical details

We apply EigenCluster to the set of members $L_i \cup V_i \cup T_i$, where each member is only described by cost features. The result is a hierarchical clustering tree R_i. Each node is a subset of the members and the root node is the entire set $(L_i \cup V_i \cup T_i)$. Each interior node has two child nodes, whose subsets comprise the subset at the parent node. Each leaf node (a node with no child nodes) represents a subset of size at least 50.

C.5 Compute the frontier

We would like to make predictions that are based on medical information, as well as based on cost information. It appears that cost information can distinguish members with different result period costs at a coarse level, but medical information cannot. On the other hand, medical information can distinguish members with different result period costs at a more fine level, whereas cost patterns cannot. This is the motivation behind the frontier – the "coarsest" level at which medical information can distinguish members.

The frontier consists of nodes in R_i for which we can improve the clustering using medical information, that is the resulting prediction is at least as good as if we had clustered those nodes further using cost information. We describe next how to compute this frontier.

Technical details

We walk up the tree R_i and apply EigenCluster to the member subset at each interior node, but only using medical features. Suppose we are at some interior node, and c_1 and c_2 are the best error measures

for our child nodes C_1 and C_2 as determined by the learning sample ($L_i \cap C_1$ and $L_i \cap C_2$) and applied to the validation sample ($V_i \cap C_1$ and $V_i \cap C_2$). We apply EigenCluster to the subset at our current node to obtain a hierarchical clustering tree $\hat{R}$. For every leaf node $\hat{C}$ in $\hat{R}$, we compute the single answer $\hat{x}$ that optimizes the sum of the error measure on $\hat{C} \cap L_i$ and apply it to the validation sample $\hat{C} \cap V_i$. Let $\hat{c}$ be the cost incurred by $\hat{x}$. If $\hat{c}$, summed over all leaf nodes is more optimal than the sum of c_1 and c_2, we designate this interior node to lie on the frontier F_i, replace its subtree with $\hat{R}$, set its cost to be the sum of the $\hat{c}$'s and continue up R_i. After we have walked up the whole tree we have replaced parts of the tree R_i, which was built using cost information only, with number of new subtrees $\hat{R}_i$'s that use medical information and improve prediction.

C.6 Prediction

Each leaf node contains a subset of patients. Roughly two thirds are in the learning and validation sets, and the a third are in the test set. The idea is that every member of this node is similar – otherwise they would not be put in the same node. Therefore, it is natural to think that the result period behavior of the patients in the learning and validation sets is similar to the result period behavior of the patients in the test set. This motivates our prediction technique, described below.

Technical details

We now have at our disposal leaf nodes $C_1, \ldots, C_n$. Each C consists of members of the learning, validation and the test sets. We compute the answers x_j that optimizes the sum of the measure on $C \cap L_i \cap V_i$, and use this as a prediction for each member in $C \cap T_i$.

D Examples of Group Coding

Table 15

ICD-9 Code	**Description**
250	Diabetes Mellitus
2500	Diabetes Mellitus without complications
2500x	Diabetes Mellitus without complications

Continued on next page...

Table 15 – Continued

ICD-9 Code	Description
2501	Diabetes with Ketoacidosis
2501x	Diabetes with Ketoacidosis
2502	Diabetes with Hyperosmolarity
2502x	Diabetes with Hyperosmolarity
2503	Diabetes with Coma
2503x	Diabetes with Coma
2504	Diabetes with Renal Manifestations
2504x	Diabetes with Renal Manifestations
2505	Diabetes with Ophthalmic Manifestations
2505x	Diabetes with Ophthalmic Manifestations
2506	Diabetes with Neurological Manifestations
2506x	Diabetes with Neurological Manifestations
2507	Diabetes with Peripheral Circulatory Disorders
2507x	Diabetes with Peripheral Circulatory Disorders
2508	Diabetes with Manifestations
2508x	Diabetic Hypoglycemia
2509	Diabetes with Complication
2509x	Diabetes with Complication
3572	Polyneuropathy in Diabetes
3620	Diabetic Retinopathy
36201	Background Diabetic Retinopathy
36202	Proliferative Diabetic Retinopathy
36203	Nonproliferative Diabetic Retinopathy
36204	Mild Nonproliferative Diabetic Retinopathy
36205	Moderate Nonproliferative Diabetic Retinopathy
36206	Severe Nonproliferative Diabetic Retinopathy

Continued on next page. . .

Table 15 – Continued

ICD-9 Code	Description
36207	Diabetic Macular Edema
36641	Diabetic Cataract
6480	Diabetes Mellitus - Complications of Delivery
6480x	Diabetes Mellitus - Complications of Delivery
V4585	Insulin Pump Status
V5391	Fitting/Adjust Insulin Pump

Table 15 An example of ICD-9 diagnosis codes in a diabetes diagnosis group ("x" at the end of a code stands for any number.

Table 16

Code	Description	Code Origin
0159T	Computer-aided detection, including computer algorithm analysis of MRI	CPT4
0160T	Therapeutic repetitive transcranial magnetic stimulation treatment pla	CPT4
70336	Magnetic Image, Jaw Joint	CPT4
7054x	MRI of Face, Neck and Head	CPT4
7055x	MRI of the Brain	CPT4
7155x	MRI Chest	CPT4
7214x	MRI Neck, Lumbar or Chest Spine	CPT4
72150	Magnetic Resonance (proton)	CPT4
72156	MRI (proton) of Chest, Lumbar or Angio Spine W/O&w Dye	CPT4
7219x	MRI Pelvis	CPT4
73218/9	MRI Upper Extremity	CPT4
7322x	MRI Uppr Extremity	CPT4
73718/9	MRI Lower Extremity	CPT4
7372x	MRI Lower Extremity	CPT4

Continued on next page...

Table 16 – Continued

Code	Description	Code Origin
7418x	MRI Abdomen	CPT4
7555x	Heart/Cardiac MRI	CPT4
76093	Magnetic Image, Breast	CPT4
76094	Magnetic Image, Both Breasts	CPT4
76394	MRI for Tissue Ablation	CPT4
76400	Magnetic Image, Bone Marrow	CPT4
76498	MRI Procedure	CPT4
7702x	Magnetic resonance guidance	CPT4
77084	Magnetic resonance (eg, proton) imaging, bone marrow blood supply	CPT4
C8903-8	MRI , Breast	HCPCS
C9723	Dynamic Infrared Blood Perfusion Imaging	HCPCS
Q0070	Magnetic Image, Spine	HCPCS
I8891	MRI of Brain & Brainstem	ICD9 Procedures
I8892	MRI Chest & Heart	ICD9 Procedures
I8893	MRI Spinal Canal	ICD9 Procedures
I8894	MRI Musculoskeletal	ICD9 Procedures
I8895	MRI Pelvis,prostate,bladder	ICD9 Procedures
I8896	Other Intraoperative Magnetic Resonance Imaging	ICD9 Procedures
I8897	Magnetic Resonance Image Unspecified	ICD9 Procedures
I8899	Unspecified MRI	ICD9 Procedures
R483	MRI	Rev Code
R61x	MRI	Rev Code

Table 16 An example of procedure codes in a procedure group. The table displays all codes within the MRI-scan group ("x" at the end of a code stands for any number). In general the codes in a procedure group come from various sources: ICD-9, DRG, Rev Coding, CPT4 and HCPCS.

Table 17

NDC Code	NDC Description	Rx10	Rx10 Description
00003352115	Insulin	6820080000	Insulins
00069006119	Exubera chamber	6820080000	Insulins
00069009741	Exubera release unit	6820080000	Insulins
00002811201	Iletin ii pzi beef	6820080010	Insulin- Beef
00002821201	Iletin ii reg. beef	6820080010	Insulin- Beef
00002831201	Iletin ii nph beef	6820080010	Insulin- Beef
00002841201	Iletin ii lente beef	6820080010	Insulin- Beef
00003244510	Insulin, purified ultralente B	6820080010	Insulin- Beef
00169352215	Insulin standard nph	6820080010	Insulin- Beef
00169352815	Insulin standard lente	6820080010	Insulin- Beef
00169355215	Insulin standard semilente	6820080010	Insulin- Beef
00169357215	Insulin standard ultralente	6820080010	Insulin- Beef
00002811101	Iletin ii pzi pork	6820080020	Insulin- Pork
00002821101	Iletin ii regular pork	6820080020	Insulin- Pork
00002831101	Iletin ii nph pork	6820080020	Insulin- Pork
00002841101	Iletin ii lente pork	6820080020	Insulin- Pork
00002850001	Iletin ii regular pork	6820080020	Insulin- Pork
00003244110	Insulin, purified semilente po	6820080020	Insulin- Pork
00169010001	Insulin purified	6820080020	Insulin- Pork
00169020001	Insulin purified	6820080020	Insulin- Pork
00169030001	Insulin purified	6820080020	Insulin- Pork
00169244010	Insulin purified regular pork	6820080020	Insulin- Pork
00169244210	Insulin purified lente pork	6820080020	Insulin- Pork
00169244710	Insulin purified nph pork	6820080020	Insulin- Pork
00169351215	Insulin standard regular	6820080020	Insulin- Pork
54569165200	Iletin ii reg. pork	6820080020	Insulin- Pork

Continued on next page...

Table 17 – Continued

NDC Code	NDC Description	Rx10	Rx10 Description
54569165202	Iletin ii reg. pork	6820080020	Insulin- Pork
54569281600	Insulin purified lente pork	6820080020	Insulin- Pork
54569281700	Insulin purified regular pork	6820080020	Insulin- Pork
54569289100	Iletin pork nph	6820080020	Insulin- Pork
54569289101	Iletin pork nph	6820080020	Insulin- Pork
00002811001	Iletin pzi	6820080030	Insulin- Beef & Pork
00002824001	Iletin regular i	6820080030	Insulin- Beef & Pork
00002831001	Iletin nph i	6820080030	Insulin- Beef & Pork
00002844001	Iletin lente i	6820080030	Insulin- Beef & Pork
00002851001	Iletin semilente	6820080030	Insulin- Beef & Pork
00002864001	Iletin ultralente	6820080030	Insulin- Beef & Pork
54569165101	Iletin nph i	6820080030	Insulin- Beef & Pork
54569165102	Iletin nph i	6820080030	Insulin- Beef & Pork
54569295100	Iletin regular i	6820080030	Insulin- Beef & Pork
54569295101	Iletin regular i	6820080030	Insulin- Beef & Pork
54868142801	Iletin nph i	6820080030	Insulin- Beef & Pork
54868208901	Iletin regular i	6820080030	Insulin- Beef & Pork
00002751001	Humalog	6820080050	Insulin- Human
00002751101	Humalog mix 75/25	6820080050	Insulin- Human
00002751559	Humalog	6820080050	Insulin- Human
00002751659	Humalog	6820080050	Insulin- Human
00002821501	Humulin r	6820080050	Insulin- Human
00002821601	Humulin br	6820080050	Insulin- Human
68115083910	Lantus	6820080050	Insulin- Human

Table 17 An example of drugs in a drug group. The table contains examples of drugs that belong to the Insulin group, as well as their Rx10 and NDC codes.