



Data mining a large health insurance database

A. F. Gualtierotti

*IDHEAP and University of Lausanne,
Chavannes-près-Renens, Switzerland.*

Abstract

There are many data mining problems which do not fit easily into the text-book mold of data mining exposition and software tools. In fact, most real life problems have a wider, less precisely defined scope than that of typical data mining discourse. One such example, for which a data mining project is under way, is presented and commented. The attendant difficulties as well as some potentially available tools (those that are presently tried, or eventually will be) are described and their (lack of) merit discussed.

1 Introduction

Health insurance is ruled, in Switzerland, by a law (Loi fédérale du 18 mars 1994 sur l'assurance-maladie - LAMal) which requires from every resident of Switzerland that he or she be insured against illness (Art. 3). Each resident is free to choose the insurance company that will provide the required insurance (Art. 4). Insurance companies, though private enterprises, cannot abstain from providing a given person with the required insurance (Art. 4).

Insurance costs have been rising steadily over the years and modifications to the law have only succeeded in slowing the rate of increase of costs over these years.

Insurance companies reimburse part of the costs of the health care provided by the various health care professionals on the basis of bills that patients or third parties forward to their respective insurances. The law (Art. 28) also requires from insurance companies that they provide the central office for social insurance (Office fédéral des assurances sociales - OFAS) with data that pertains to these bills, and the office in question has, among its duties (Art. 32), that of using the data thus gathered for all pertinent



administrative questions, such as surveillance of the performance of the insurance companies, monitoring of the health system, forecasting of health costs, and so forth. The data base of the title is that that gathers this data. As there are between six and seven millions residents in Switzerland, the data base has a potential rate of yearly increase which is of the order of ten to the power seven.

The number of insurance companies is quite large (in the hundreds), the companies involved quite heterogeneous (very small, "manually" managed, to very large, computerized firms), and health insurance is geographically determined as premia vary from state to state (there are twenty six states, also quite heterogeneous).

The "health insurance problem" is a daily fixture of the news and all kinds of claims are proffered which can never be checked for lack of access to the data and possibly insufficient data information content.

2 Objectives for the analyses of the data base

As is often the case of similar areas of human endeavour, what is expected from an analysis of the data is very generally obvious, but operationally obscure. First of all one would like to know what "the data has to say" irrespectively of any *a priori* queries. Secondly one would like to delimit the areas where the costs are (the usual "eighty percent of the costs are due to twenty percent of the insured" - in fact it seems that about half the costs are due to imaginary or best left alone illnesses) and what costs one may expect. Finally, and perhaps most importantly, one would like to know if premia increases requested yearly by the insurance companies are justified by the health care provided rather than by unwarranted company extravaganza. Usually it is very difficult to obtain, from the user, any requests for knowledge that are closer to an operational query than the general requests just outlined.

One can thus only expect to narrow the scope of the queries by trial and error, showing some attempts at data analysis and being quite often told that they are inadequate. Given the mass of data, only data mining procedures can be expected to provide adequate service, all the more so that there is often the further request that all the available data be used, rather than a sample.

A second objective assigned to the analyses is that of finding some data support for a procedure, based on health delivery knowledge, which purports to assign to each insured, for a given period of time, typically one year, a "principal illness" and a "principal provider," which are means of achieving the broad goals already outlined. Indeed, if one wants to know, for instance, how much cardiology weighs in the insurance burden, one will sum all the costs for which the "principal illness" is "cardiology." Such an approach may induce biases as there is always the danger of aggregating similarly sized expenditures in quite different areas of health care.



3 Data

The original data is contained in several administrative files which must be merged and which provide informations of three basic types:

1. data on health providers, mostly the medical specialty they practice;
2. insurance profile of the insured person (for example, one may find out that the insurance regime is HMO, what the premium is and for what duration);
3. records of treatments received by the insured person (for example, one may find out that the insured person was treated, on a given date, by a cardiologist, on recommendation of an ophtalmologist, and incurred some pharmacy expenses, with reimbursment in a given amount).

As each insurance company has its own administrative system, the data the companies globally, and often reluctantly, provide to OFAS is of mixed quality and requires quite a bit of cleansing. The result is of rather poor quality, the meaning of one entry of one variable often depending on the value and meaning of that of another. Furthermore, whenever one needs a statistical file, that is a file with statistical variables registered as columns and statistical units and records registered as rows, extensive programming (in FORTRAN) must be performed, often to find some data incompatibility which is and must be fixed in an *ad hoc* manner. As concepts are not "stabilized," two successive files pertaining to the same object of study are practically never identical.

In such a context, what constitutes a data warehouse is ill-defined, the data warehouse of Charybdis being the messy flat files of Scylla.

4 Data characteristics

The data present a number of features, beyond the obvious problems of data quality, that are unusual for data mining in that standard data mining techniques do not accomodate such features. These are:

1. the heterogeneity of the data types

Part of the data is nominal, part is ordinal, and part is interval/ratio. Some of the "status" values of this data change rarely (sex), and some infrequently (type of insurance - which is most often settled yearly), whereas the benefits data are registered at a much higher rate, at unpredictable times.

2. the unstructured temporal dimension

Each person insured receives benefits from health insurance which can be pictured as a set of stochastic processes paths: there is a counting process, which counts how many acts have been performed, there



is a cost (to insurance companies) process which sums the expenditures caused by the acts performed, there is the process of gaps which records the time elapsed between medical bills (which one would like to interpret as medical acts), and finally there is the process of the successive health providers specialty.

Typically, one process path contains too few events for it to be studied as a single unit, and too many for the events to be "standardized" into a (small) finite number of regular time categories.

3. the mismatch between the nature of the data and the information sought

The mismatch occurs at two levels:

- (a) first of all, the only somewhat accurately available time points are those which induce the initial recording of a bill (thus, for a doctor that sends a bill every three months, illnesses may appear as seasonal);
 - (b) secondly, one is interested in illnesses and not in bills, and thus one has the problem of translating billing periods into illness periods.
4. the large number of missing values

There are many missing values, and, though the data base is large, the high number of possible modalities may deplete the base excessively, if one discards statistical units with missing values.

5 Required data mining techniques and possible solutions

Basically two approaches to such data are available. One leads to "transforming" time into categories so that the usual classification and other data mining techniques can apply. The categories can be of two types. The first is that resulting from expanding time functions into known bases: the categories are obtained using the expansion coefficients. The other approach consists in segmenting the time series into illnesses bouts and gaps between illnesses. The bouts replace then, as statistical units, the benefits provided to the insured by the health insurance system.

The second approach one may pursue makes one look at each insured person's "health trajectory" as an independent realisation of a stochastic process.

Both approaches have their pluses and minuses. For instance basis representation of functions requires "time padding;" one must insert the actual times of billing

$$T_1 < T_2 < \dots T_n$$

into a much more numerous, regular time sequence

$$\left\{ t_i^{(n)}, 1 \leq i \leq p_n \right\},$$

and extend the definition of time functions to these new values. Some imprecision is then necessarily introduced. Indeed, when

$$t_i^{(n)} < T_j < t_{i+1}^{(n)},$$

as one must set

$$T_j \approx t_k^{(n)}, \text{ and } \varphi(T_j) \approx \varphi(t_k^{(n)}), k \in \{i, i+1\},$$

φ being the time function, there results an uncontrollable, albeit small, time shift.

Segmentation into successive periods of health and illness based on gap length will crucially depend on the size of the gap (Are gaps longer than three weeks indicative of different illnesses? What about chronic diseases?), and the ability of making sense of the segments obtained (What should one do with an occasional, but expensive visit to an ophthalmologist, while the insured visits a cardiologist on a regular basis?).

Finally, it is rather hard to find a criterion that would help in deciding that “health trajectories,” corresponding to different persons, are independent paths of the same stochastic process!

The “natural” stochastic process types that may help describe “health trajectories” are semi-Markov processes [1] and hidden Markov processes.[2] Indeed, the semi-Markovian idea “sees” the insured in successive “states” of health and illness, states in which he or she spends time whose length depends on the state. For the hidden Markov model approach, the hidden states are those of health and illness. However, the sojourn in those states can only be “observed” indirectly through the billing process.

5.1 One technique to classify time step functions: the Haar wavelet

5.1.1 Reasons to classify step functions

As mentioned, the best conceptual description of an illness or, here, a billing process, is that of an increasing path of a stochastic process (the path of cumulated costs: the choice to resort to cumulated costs rather than to raw costs is dictated by the use of the Haar wavelet technique [3] described below for which decreases associated to the raw costs enter as “noise”). Indeed, as stated above, one can then try to model the process as, for instance, a semi-Markov or hidden Markov model, which are techniques within which, for instance, forecasting is not only natural, but also a well supported operation. However, to use that path as a vehicle for statistical description, one would need more “events.” One way to increase this number is to obtain independent realisations of the process, and one can hope that similar



paths will yield these independent realisations. The Haar wavelet technique should then provide an efficient classification tool.

5.1.2 The Haar wavelet

The Haar wavelet is in fact an orthonormal basis (in the mathematical sense) for the space of square integrable functions over the unit interval, with respect to Lebesgue measure. Its definition is:

$$h_{0,0}(x) = 1, \quad 0 \leq x \leq 1,$$

and, for $n \in \mathbb{N}$, and $k \in [1, 2^n]$,

$$h_{n,k}(x) = \begin{cases} 2^{\frac{n}{2}} & \text{if } \frac{k-1}{2^n} \leq x < \frac{k-\frac{1}{2}}{2^n} \\ -2^{\frac{n}{2}} & \text{if } \frac{k-\frac{1}{2}}{2^n} \leq x < \frac{k}{2^n} \\ 0 & \text{if otherwise} \end{cases}.$$

Any step function will only require a finite number of $h_{n,k}$ functions to be represented and it is the coefficients of that representation which serve as classifiers, for instance in a regression tree.[4] One first advantage of the Haar wavelet, with regard to data mining (this is a feature common to all other wavelets, hence their recent success), is that the representations it provides are rather economical (very few coefficients are required to produce a very good approximation to the original function). A second advantage, again with regard to data mining, is that very efficient algorithms exist and that they are very easy to program.

5.1.3 Example of use of the Haar wavelet

Two features of the billing process which are of interest relate to the “density” of events: this density may serve as an identifier of acute illnesses as well as a separator for illnesses bouts. The Haar wavelet representation can be used to that end as follows. The Haar wavelet algorithm starts indeed by computing, at adjacent dyadic “time points,” the differences of the values of the function to be represented. If this function is the counting function, the difference in adjacent values will be different from zero only when an event occurs. It suffices thus to localize and examine these differences to produce easily a density process. The practical difficulty is the segmentation of time: when is there a gap in the density process? To arrive at an empirical solution one must simultaneously use the features of the different functions representing billing, and the Haar representation may provide the “universal tool.”

5.1.4 Problems with use of the Haar wavelet

The Haar wavelet is one possible cure to the lack of structure of the temporal dimension of the billing process. One replaces the continuous time span $[0, T]$ with 2^n dyadic time intervals: for instance, for a year of 365 days, one



will have to use 2^9 dyadic points at least, as

$$2^8 = 256 < 365 < 2^9 = 512.$$

As a bill can contain several medical acts for a single session at the doctor's, registering these acts will require the inclusion of intermediate dyadic points, which can bring the required power to 2^{11} . The data required is thus large and redundant, and adds one problem to those that must be solved: should one store the redundant data, or compute it whenever necessary?

Another factor driving up the number of dyadic points is the following. Whenever an event coincides temporally with a dyadic division, there is the risk that the corresponding value "disappear" from the picture, as it does not enter the computation of a difference. If the corresponding cost, for instance, is that of a very expensive intervention, there is the risk of seriously underestimating that cost. The only protection against such a problem seems to lie in an increase in the number of dyadic points, and thus in computation time and storage space.

5.2 One technique to investigate gaps between illnesses: semi-Markovian ideas, the EM algorithm, and related techniques

Typically, a gap between illnesses can be looked at as an interarrival (random) time. As seen, this leads to semi-Markovian ideas: the "sojourn time at an illness" will depend on the last (several) state(s). One can thus try to fit a mixture of exponential distributions [5] to the gap data and then use this mixture as a segmenting mechanism. This could be compared to a direct search using a "frequent sequences" algorithm.[6] The problem here will be to reduce as much as possible the noise created, for instance, by missing values (sequences of the type $\# \rightarrow$ GENERAL PRACTITIONER, where $\#$ is a missing value).

Techniques to fit a mixture of exponentials typically require use of the EM algorithm,[7] a procedure not found in typical data mining software. The ability of the data mining software to interface with efficient (for large data sets), but more traditional statistical packages is in that respect crucial. One such combination is provided by the SAS statistical package which can be combined with SAS Enterprise Miner,[8] though one still may be required to do some programming, all statistical routines not being supported by any statistical software!

6 Conclusion

Most "decision support" techniques have their origin in quite specific problems, and data mining is no exception in that it seems that the engines for development have been the huge data bases arising from retail chains and merchandising.[9] The needs of quite different organisations and their problems are however often superficially analogous and, given the propensity



of software merchants and proselites to take liberties with the meaning of words, it quickly seems that the same tools can be mustered to solve these rather different, and often more complex problems. A short analysis often shows however that the new problems can at best be marginally lessened by existing technologies. The example considered in this paper is typical, and only tenacious practice will bring an answer to the question of whether or not the particular data base under consideration can be data mined!

References

- [1] Haccou, P. & Meelis, E., *Statistical Analysis of Behavioural Data*, Oxford University Press, Oxford, pp. 256-308, 1992.
- [2] Petrie, T., Probabilistic functions of finite state Markov chains, *Annals of Mathematical Statistics*, **40**, pp. 97-115, 1969.
- [3] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J., *Classification and Regression Trees*, Wadsworth, Belmont CA, 1984.
- [4] Nievergelt, Y., *Wavelets Made Easy*, Birkhäuser, Boston, pp. 3-35, 1999.
- [5] Titterton, D.M., Smith, A.F.M. & Makov, U.E., *Statistical Analysis of Finite Mixture Distributions*, Wiley, Chichester, pp. 137-146, 1985.
- [6] Srikant, R. & Agrawal, R., Mining sequential patterns: generalizations and performance improvements, *Proceedings of Extended Data Base Technology (EDBT'96)*, Avignon, France, 1996.
- [7] Dempster, A.P., Laird, N.M. & Rubin, D.B., Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society B*, **39**, pp. 1-38, 1977.
- [8] SAS System, Cary, North Carolina: SAS Institute Inc.
- [9] Agrawal, R., Imielinski, T. & Swami, A., Mining associations between sets of items in massive databases, *Proceedings of the International Conference on Management of Data (ACM SIGMOD'93)*, Washington D.C., 1993.