Data Mining Application to Healthcare Fraud Detection: Two-Step Unsupervised Clustering Method for Outlier Detection with Administrative Databases

Data Minining Applicato al Riconoscimento Frodi in Sanita': Algoritmo a Due Step per l'Identificazione di Outliers con Database Amministrativi

Massi Michela C., Ieva Francesca, Lettieri Emanuele

Abstract This study aims at exploiting Administrative Databases to identify potentially fraudulent providers. It focuses on the DRG upcoding practice, i.e. the tendency of coding within Hospital Discharge Charts (HDC), codes for provided services and inpatients health status so to make the hospitalization fall within a more remunerative DRG class. The model here proposed is constituted by two steps: one first step entails the clustering of providers, in order to spot outliers within groups of similar peers; in the second step, a cross-validation is performed, to verify the suspiciousness of the identified outliers. The proposed model was tested on a database relative to HDC collected by Regione Lombardia (Italy) in a time period of three years (2013-2015), focusing on the treatment of heart failure.

Massi Michela Carlotta
MOX - Modellistica e Calcolo Scientifico
Dipartimento di Matematica
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy

Center for Analysis Decisions and Society
Human Technopole
Palazzo Italia, Via Cristina Belgioioso, 28, 20157 Milano, Italy
e-mail: michelacarlotta.massi@polimi.it

Ieva Francesca
MOX - Modellistica e Calcolo Scientifico
Dipartimento di Matematica
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
e-mail: francesca.ieva@polimi.it

Lettieri Emanuele
Department of Management, Economics and Industrial Engineering
Politecnico di Milano
Lambruschini Street 4/c, 20100 Milano, Italy
Tel.: +39-02-23994077
e-mail: emanuele.lettieri@polimi.it

Abstract *Questo studio ha lo scopo di sfruttare i Database Amministrativi per identificare operatori sanitari potenzialmente fraudolenti. Si focalizza sulla pratica dell'upcoding, i.e. la tendenza a registrare sulle Schede di Dimissione Ordinaria, codici relativi ai servizi prestati e allo stato di salute del paziente, in modo da far ricadere il ricovero in una classe DRG pi remunerativa. Il modello proposto costituito da due step: il primo riguarda il clustering degli ospedali, per identificare outliers tra gruppi di operatori simili; il secondo punta a validare la sospettosit degli outlier identificati. Questo modello stato testato su un database amministrativo relativo al trattamento di Scompenso Cardiaco, collezionato da Regione Lombardia in un periodo di tre anni (2013-2015).*

Key words: Data Mining, Healthcare Fraud, DRG Upcoding, Administrative Database

1 Introduction

Fraud in the context of Healthcare has different perpetrators (e.g. hospitals, medical figures, private facilities) and different dynamics.
The *upcoding* practice consists in classifying a patient in a DRG, or registering treatment codes (in case of Fee for Service[1] payment systems) that produce higher reimbursements [Simborg (1981)]. This practice by public hospitals is more likely to be due to unintentional errors by coders, or misunderstandings with doctors; while, when talking about private hospitals or medical practices, it could actually be due to profit maximising purposes [Silverman (2004), O'Malley (2005)].
Given that Healthcare is the target of large public and private investments (on average, in OECD countries, 15% of the governament budget is allocated for this purpose), independently from the geographical and political setting, this sector is a rather appetible one for frauds. In this ever-growing healthcare industry, using manual countermeasures to fight frauds is not enough. Given the ever-growing availability of digital data, the adoption of data mining techniques might help to reach better results and more efficient processes, in terms of both time and costs. Developing tailored algorithms would allow for the restriction of the pool of investigated providers, including those acting cautiously and perpetrating fraud within the limits of those indicators monitored according to general policies [Musal (2010)].
In this paper, we deal with the development of a novel systematic and quantitative approach to fraud detection, with a focus on the upcoding practice, because of the extremely relevant economic impact this kind of fraud has on the system [Steinbusch (2007)]. The objective is to propose a novel tool to support human decisions in the preliminary phases of screening providers to spot suspects eligible for a more in depth investigation, including those with a more cautious approach to fraud. Because of the large availability of Administrative Databases, we decided to exploit this type of data as our source of information. All the analysis described in

[1] A method in which doctors and other health care providers are paid for each service performed. Examples of services include tests and office visits. [Healthcare.gov]

the following Sections were performed using R [R Core].
The proof of concept of our method, which is generally adaptive to any kind of system, was developed in the Italian context, Lombardy Region in particular, studying the *behavior* of hospitals by exploiting regional administrative data. This region represents a relevant benchmark in the Italian healthcare landscape and as such it was deemed an interesting starting point for our testing, but the flexibility of the developed method allows for its generalizability to other contexts.

2 Literature Review

Data mining approaches to healthcare fraud detection are still at their infancy, but recently gaining momentum. Even though limited, some contributions to the literature exist. We can group the existing methodologies of fraud detection into three main groups: supervised, unsupervised, or hybrids of the two [Aral (2012)]. Supervised techniques [Pearson (2006), Kumar (2010), Francis (2011), Kirlidog (2012)] despite their undeniable potential and predictive power, exhibit the risk of focusing on old patterns and losing predictive capability as new records are evaluated over time [Joudaki (2015)]. For this reason, unsupervised techniques are the most adopted [Yang (2006), Shan (2008), Shan (2009), Luo (2010), Musal (2010), Tang (2011), Konijin (2012), Shin (2012), Liu (2013), Konijin (2013), Konijin (2015), Bauder (2016), Capelleveen (2016)], tackling the intrinsic characteristic of fraud of changing in time, according to the arising regulations or control systems. Hybrid techniques or on-line processing systems take the best of both approaches [Aral (2012), Ngufor (2013), Kose (2015)], creating a combination of supervised and unsupervised methods.
As previously mentioned, this paper focuses on the upcoding practice, that can be declined in several ways and has drawn the attention of most data mining researches because of the extremely high impact it has on healthcare expenditures. The majority of the available literature attempts to spot providers with very high claiming episodes, which distinguish themselves as *evident* outliers - one clear example in [Capelleveen (2016)]. However, the evolving nature of modern fraud, has pushed some researchers to change direction and try and identify providers that have a more cautious approach to fraud, which would be neglected in the aforementioned high-claiming groups. The *behavioral* models suggested by Musal *et al.* [Musal (2010)] and Shin *et al.* [Shin (2012)] try indeed to respond to this objective. However, those techniques have still very limited application, even though their usefulness in reducing the overall level of fraud in the system is undeniable. This lack of interest can be justified by the risk of these models to exhibit higher false positive rates, and the lower amount of recovery from each correctly spotted fraudster [Musal (2010)]. However, as reported in [Shin (2012)], types of fraud are growing increasingly sophisticated, and patterns detected from fraudulent and non-fraudulent behaviors become rapidly obsolete because of rapid changes in behavior, and fraudulent providers are becoming smarter in finding more cautious approaches

which usually prevent them from being investigated [Bolton (2002)].
In conclusion, from the study of previous contributions emerges the need for more data mining methods capable to adapt to changing behaviors of fraudsters, and identifying more cautious approaches to fraud. For this reason, the algorithm proposed in this study belongs to this more complex and less tackled pool of unsupervised *behavioral* methodologies.

3 Methods

The proposed model was tested on an Administrative Database collected by Lombardy Region, composed by Hospital Discharge Charts (HDCs) from 2013 to 2015. Patients' data were at first anonymized, and hospital confidentiality was preserved. To improve the model's performance, one specific disease was selected to search for anomalous behaviors by hospitals. Therefore, an extraction process was performed as first step, focusing on Heart Failure cases.
Since this research was performed in an unsupervised setting (ie. no indication about fraudulent records was available at the beginning of the process), and no interaction with experts was planned, all choices were driven by insights gathered from the Literature Review. At first, three different datasets were built: the Hospital Discharge Charts (HDCs) Dataset, with one HDC per row, the Hospital Dataset, aggregating data with hospitals as statistical unit, and the Patients Dataset, where each observation is a patient treated at least once by the providers. From literature, a list of variables deemed interesting to evaluate fraudulent behaviors was collected (e.g. [Berta (2010), Ekin (2017)], or the Upcoding Index in [Silverman (2004)]). Such variables have subsequently been extracted or reproduced to be added to the available datasets.
The algorithm proposed in this research is constituted by two fundamental steps:

One *First Step*, in which the Hospitals Dataset is exploited. This dataset was composed by descriptive variables (i.e. number of patients treated by the hospital, average cost, average length of stay, degree of specialization for treatment of HF, etc.), and integrated with the values of r_{ij} [Ekin (2017)], which represents the ratio between the probability that the HF-related DRG i is registered by provider j and the probability that DRG i is registered in all the population of N hospitals, calculated for every HF-related DRG registered by the provider. The vector of r_{ij} represents, for each provider, a description of the '*behavior*' of that provider w.r.t. the treatment of HF.
Gouping hospitals on the basis of their characteristics and their behavior in the treatment of HF, makes it possible to distinguish homogeneous groups of providers, allowing us to spot the providers *behaving* differently from their similar peers. For this reason, the algorithm here proposed entails a k-mean clustering on the aforementioned hospitals' information, with the aim of identifying clusters of similar providers.
Once the clusters have been identified, the Euclidean distance of each element of the

Table 1 Variables adopted for cross-validation with hospitals' characteristics and casemix

Variabile	Dataset	Note
Specialization	Hospital	Number of HF specific cases
Perc. DRGs with CC	Hospital	Complexity of the casemix
Upcoding Index	Hospital	Berta *et al.*, 2010[Berta (2010)]
Average Cost	Hospital	Expensiveness of the casemix
Age	Patients	Patient complexity indicator
Lenght of Stay	Patients, HDC	Patient complexity indicator
Comorbidity	Patients, HDC	Patient complexity indicator, etimated as in [Gagne (2011)]
Total Costs	Patients, HDC	Cost of care per Patient
Cost / Lenght of Stay	HDC	Cost of care against quantity of care needed
Cost / Comorbidity	Hospital	Cost of care against quantity of care needed

cluster by the corresponding centroid is computed. Figure 2 shows such distribution for all the hospitals of our dataset. Given such distribution, it is possible to define a threshold (say the 95th quantile), where hospitals having a distance greater than this threshold are labelled as 'providers needing further investigation'.

A *Second Step* is then performed by the algorithm. Since *behavioral methods* demonstrate to have higher false positive rates - even though stronger in identifying cautious fraudsters [Musal (2010)] - a further validation of results is recommended. This phase is indeed useful to controllers to create a visual dashboard to support an informed skimming of the suspects identified in the first phase. In this way, their attention and in-depth investigations will be focused on a restricted number of cases, with a higher probability of detecting actual frauds. This passage entails the use of variables describing Hospitals deemed interesting in literature to identify fraudulent providers, together with indexes estimating the complexity of the casemix of patients they faced. Both groups of variables are useful to try and verify whether such outliers could be considered suspects w.r.t. upcoding fraud or, on the opposite, justified in their particular behavior because of the complex patients' population they treated. In Table 1 are listed the variables adopted for the cross-validation.

4 Results

From the application of the clustering algorithm on the data at hand - after the estimation of the optimal nunmber of clusters with the WSS method - resulted the six clusters of hospitals represented in Figure 1.

The number of clusters was chosen on the basis of a robustness analysis where the different values of Silhouette obtained for different k values were compared to identify the optimal clusters' configuration. In addition, evaluating clusters characteristics, clear differences among clusters were recognized w.r.t. some dimensions of interest: for this reason, the result was deemed satisfactory.

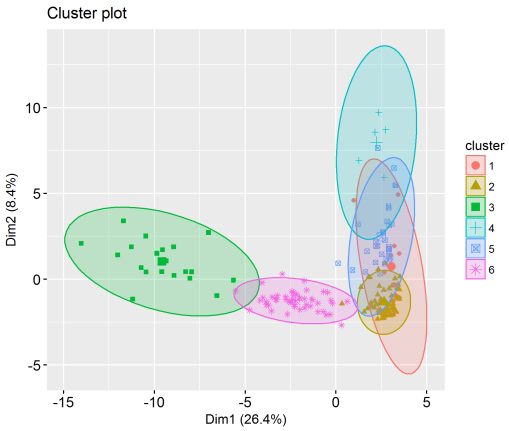

Fig. 1 Clusters of hospitals resulting from *k-mean* clustering. The clusters are plotted against the first two Principal Components resulting from a PCA.

Once the clusters were generated and all distances calculated, 10 hospitals resulted as outliers adopting the 95^{th} percentile threshold.

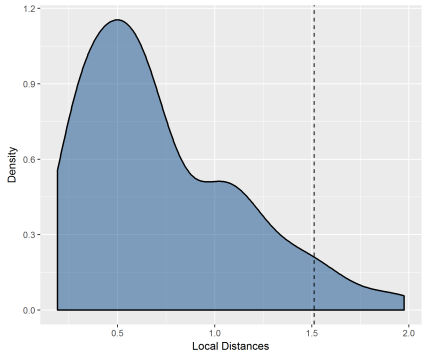

Fig. 2 Local distances distribution of hospitals from the center of their cluster. The 95th percentile threshold is highlighted by the vertical dotted line.

Among these, 8 were private providers, and 2 public.
The complete analysis can be found in [Massi (2018)], but one illustrative case could be that of a private provider with very low level of specialization, and a very high Upcoding Index, that when cross-validated with its casemix of patients confirmed its suspect behavior (e.g. both Comorbidity Score and average LOS were low, but the reimbursements requested per hospitalization were higher than the average of providers).
This is just an example of how this procedure could constitute an important method to support performances' monitoring in the Healthcare domain. Indeed, combining the first and the second step of this method we get a twofold goal: first, we provide a flexible tool for screening a huge amount of data which are currently available (e.g. Administrative Databases), exploiting their use for performance assessment; moreover we limit the number of false positive to be deeply investigated by controllers.

References

[Aral (2012)] Aral K.D., Güvenir H.A., Sabuncuoğlu I, Akar A.R., (2012) A Prescription Fraud Detection Model, *Comput. Methods Prog. Biomed.*, **106 (1)**, pp. 37–46.

[Bauder (2016)] Bauder R., Khoshgoftaar T.M., Seliya N., (2017) A survey on the state of healthcare upcoding fraud analysis and detection, *Health Services and Outcomes Research Methodology*, **17 (1)**, pp. 31–55.

[Berta (2010)] Berta P., G. Callea, G. Martini, G. Vittadini, (2010) The effects of upcoding, cream skimming and readmissions on the Italian hospitals efficiency: A population-based investigation, *Economic Modelling*, **27**, pp. 812–821.

[Bolton (2002)] Bolton R.J., Hand D.J., (2002) Statistical Fraud Detection: A Review, *Statistical Science*, **17 (3)**, pp. 235–255.

[Capelleveen (2016)] Van Capelleveen G., Mannes P., Roland M., Dallas T., Van Hillegersberg J., (2016) Outlier detection in healthcare fraud: A case study in the Medicaid dental domain, *International Journal of Accounting Information Systems*, **21**, pp. 1831.

[Ekin (2017)] Ekin T., Ieva F., Ruggeri F., Soyer R., (2017) On the Use of the Concentration Function in Medical Fraud Assessment, *The American Statistician*, **71 (3)**, pp. 236–241.

[Francis (2011)] Francis C., Pepper N., and Strong H., (2011), Using support vector machines to detect medical fraud and abuse *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 8291–8294.

[Gagne (2011)] Gagne J.J., Glynn R.J., Avorn J., Levin R., Schneeweiss S., (2011) A combined comorbidity score predicted mortality in elderly patients better than existing scores, *Journal of clinical epidemiology*, **64 (7)**, pp. 749–759.

[Healthcare.gov] Healthcare.gov, (2018), [online] https://www.healthcare.gov/glossary/fee-for-service/

[Joudaki (2015)] Joudaki H., Rashidian A. *et al*., (2015) Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature, *Global Journal of Health Science*, **101**, **7**, p. 378.

[Kirlidog (2012)] Kirlidog M., Cuneyt A., (2012) A Fraud Detection Approach with Data Mining in Health Insurance, *Procedia - Social and Behavioral Sciences*, **62**, pp. 989–994.

[Konijin (2012)] Konijin R.M., Kowalczyk W., (2012) Hunting for Fraudsters in Random Forests, *Hybrid Artificial Intelligent Systems: 7th International Conference - Proceedings*, **Part 1**, pp. 174–185.

[Konijin (2013)] Konijin R.M., Duivesteijn W., Kowalczyk W., Knobbe A., (2013) Discovering Local Subgroups, with an Application to Fraud Detection, *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference - Proceedings*, **Part 1**, pp. 1–12.

[Konijin (2015)] Konijin R.M., Duivesteijn W., Meeng M., Knobbe A., (2015) Cost-based quality measures in subgroup discovery, *Journal of Intelligent Information Systems*, **45 (3)**, pp. 337–355.

[Kose (2015)] Kose I., Gokturk M., Kilic K., (2015) An Interactive Machine-learning-based Electronic Fraud and Abuse Detection System in Healthcare Insurance, *Appl. Soft Comput.*, **36 (C)**, pp. 283–299.

[Kumar (2010)] Kumar M., Ghani R., Mei ZS, (2010) Data Mining to Predict and Prevent Errors in Health Insurance Claims Processing, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 65–74.

[Liu (2013)] Liu Q., Vasarhelyi M., (2013) Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information, *29th World Continuous Auditing and Reporting Symposium*, **Part 1**.

[Luo (2010)] Luo W., Gallagher M., (2010) Unsupervised DRG Upcoding Detection in Healthcare Databases, *2010 IEEE International Conference on Data Mining Workshops*, pp. 600–605.

[Massi (2018)] Massi M.C., Ieva F., Lettieri E., (2018) Data Mining Application to Healthcare Fraud Detection: A Two-Step Unsupervised Clustering Model for Outlier Detection with Administrative Databases, *Mox Report*, 49/2018

[Musal (2010)] Musal R.M., (2010) Two models to investigate Medicare fraud within unsupervised databases, *Expert Systems with Applications*, **37 (12)**, pp. 8628–8633.

[Ngufor (2013)] Ngufor C., Woytusiak J., (2013) Unsupervised labeling of data for supervised learning and its application to medical claims prediction, *Computer Science*, **14 (2)**.

[O'Malley (2005)] O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM., (2005) Measuring diagnoses: ICD code accuracy., *Journal of Health Economics*, **40 (5)**, pp. 1620–1639.

[Pearson (2006)] Pearson R.A., Murray W., Mettenmeyer T., (2006) Finding anomalies in Medicare, *Electronic Journal of Health Informatics*, **1 (1)**, p. 2.

[R Core] R Development Core Team (2009), R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. [Online] http://www.R-project.org

[Silverman (2004)] Silverman E., Skinner J, Medicare upcoding and hospital ownership, *Journal of Health Economics*, **23 (2)**, pp. 369–389 (2004)

[Simborg (1981)] Simborg DW, (1981) DRG creep: a new hospital-acquired disease, *New Engl J Med*, **304(26)**, pp. 16021604.

[Shan (2008)] Shan Y, Jeacocke D., Murray D.W., Sutinen A., (2008) Mining Medical Specialist Billing Patterns for Health Service Management, *Proceedings of the 7th Australasian Data Mining Conference*, **87**, pp. 105–110.

[Shan (2009)] Shan Y., Murray D.W., Sutinen A., (2009) Discovering Inappropriate Billings with Local Density Based Outlier Detection Method, *Proceedings of the Eighth Australasian Data Mining Conference*, **101**, pp. 93–98.

[Shin (2012)] Shin H., Park H., Lee J., Jhee W.C., (2012) A Scoring Model to Detect Abusive Billing Patterns in Health Insurance Claims, *Expert Systems with Applications*, **39 (8)**, pp. 7441–7450.

[Steinbusch (2007)] Steinbusch P., Oostenbrink B., Zuurbier J., Schaepkens F., (2007) The Risk of Upcoding in Casemix Systems: A Comparative Study, *Health Policy*, **81**, pp. 298–299.

[Tang (2011)] Tang M., Mendis B., Sumudu U., Murray D.W., Hu Y., Sutinen A., (2011) Unsupervised Fraud Detection in Medicare Australia, *Proceedings of the Ninth Australasian Data Mining Conference*, **121**, pp. 103–110.

[Yang (2006)] Yang WS, Hwang SY, (2006) A process-mining framework for the detection of healthcare fraud and abuse, *Expert Systems with Applications*, **31 (1)**, pp. 56–68.