ICD9 to ICD10 update: effects in SVM fitting & prediction

Elies Ramon[1a], *Víctor Moreno*[2a,b]

[1]egurrea@idibell.cat, [a]Unit of Biomarkers and Susceptibility (UBS), Oncology Data Analytics Program (ODAP), Catalan Institute of Oncology (ICO). ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL). Centro de Investigación Biomédica en Red de Epidemiologia y Salud Pública (CIBERESP).

[2]v.moreno@iconcologia.net, [b]Department of Clinical Sciences, Faculty of Medicine and Health Sciences and Universitat de Barcelona Institute of Complex Systems (UBICS), University of Barcelona

Introduction: The International Classification of Diseases (ICD) is a worldwide healthcare classification system that codifies diagnoses and procedures encountered in medical settings. The 10th revision greatly expanded the number of codes of the 9th version, including 5 times as many diagnosis codes and 19 times as many procedure codes. Structural changes include slightly different arrangements of chapters and subchapters and a higher level of granularity and detail (3-5 characters for diagnosis codes in ICD9 vs 3-7 in ICD10). ICD9→10 translation is done with the General Equivalence Mapping (GEM) but, in most cases, there is not a one-to-one match between codes. Here, we analyse if the translation noise impacts the fitting and performance of machine learning models. Support Vector Machines (SVM) are especially suited for this task, as their dual form allows us to assess perturbations on the dataset as changes in the relationship between samples via the kernel matrices.

Materials & methods: The training set consisted of 9103 colorectal cancer cases and 9103 matched controls from the Catalan Institute of Healthcare (PADRIS database). For all patients, personal story of previous hospital admissions was coded in ICD9. We translated this data to ICD10 using GEM; if one-to-one match was not possible, one of the candidate ICD10 codes was chosen at random. Then we collapsed both datasets at the chapter, subchapter and category (3-code digit) level. For each level of granularity, we trained a linear SVM model using the original (SVM-9) and translated (SVM-10) data; we denote their respective kernel matrices as K and K'. The optimal value for the C hyperparameter was chosen with 5-Cross-Validation. Once we had the definitive model, the noise of the ICD9→10 translation was compared in five different ways: 1) Cosine similarity and relative spectral distance between K and K', 2) Stability of the support vectors in SVM-9 and 10, 3) Comparison of the feature importances recovered from the SVM-10 model and the SVM-9 model, 4) Accuracy of SVM-9 model fitting, and "refit" accuracy when switching the original K matrix for K' and 5) Prediction performance of both models in an independent test set (1602 cases and 7718 controls) coming from the same PADRIS database.

Results: Our original ICD9 data was grouped in 17 chapters, 129 subchapters and 845 categories, expanded to 20 chapters, 232 subchapters and 1221 categories in the ICD10 translation. Similarity between K and K' ranged between 75-92% and decreased with higher granularity. We observed the same pattern with support vectors (90-97% of SVM-9 also present in SVM-10). Prediction performance was modest (AUC≈0.59, weighted accuracy≈0.56) without significant differences between both models. Top most important features were also similar and included a history of previous neoplasms, anaemia and neurodegenerative or mental disorders. In summary, although translation slightly distorted the original kernel matrix, it did not have a strong effect in the fitting and prediction of the SVM.

Keywords: ICD, SVM, kernel matrix