Large Scale Diagnostic Code Classification for Medical Patient Records

Lucian Vlad Lita and **Shipeng Yu** and **Stefan Niculescu** and **Jinbo Bi**
Siemens Medical Solutions
firstname.lastname@siemens.com

Abstract

A critical, yet not very well studied problem in medical applications is the issue of accurately labeling patient records according to diagnoses and procedures that patients have undergone. This labeling problem, known as coding, consists of assigning standard medical codes (ICD9 and CPT) to patient records. Each patient record can have several corresponding labels/codes, many of which are correlated to specific diseases. The current, most frequent coding approach involves manual labeling, which requires considerable human effort and is cumbersome for large patient databases. In this paper we view medical coding as a multi-label classification problem, where we treat each code as a label for patient records. Due to government regulations concerning patient medical data, previous studies in automatic coding have been quite limited. In this paper, we compare two efficient algorithms for diagnosis coding on a large patient dataset.

1 Introduction

In order to be reimbursed for services provided to patients, hospitals need to provide proof of the procedures that they performed. Currently, this is achieved by assigning a set of CPT (Current Procedural Terminology) codes to each patient visit to the hospital. Providing these codes is not enough for receiving reimbursement: in addition, hospitals need to justify why the corresponding procedures have been performed. In order to do that, each patient visit needs to be coded with the appropriate diagnosis that require the above procedures. There are several standardized systems for patient diagnosis coding, with ICD9 (International Classification of Diseases, (Organization, 1997)) being the official version. Usually a CPT code is represented by a five digit integer whereas an ICD9 code is a real number consisting of a 2-3 digit disease category followed by 1-2 decimal subcategory. For example, a CPT code of 93307 is used for an Echo Exam. An ICD9 code of 428 represents Heart Failure (HF) with subcategories 428.0 (Congestive HF, Unspecified), 428.1 (Left HF), 428.2 (Systolic HF), 428.3 (Diastolic HF), 428.4(Combined HF) and 428.9 (HF, Unspecified).

The coding approach currently used in hospitals relies heavily on manual labeling performed by skilled and/or not so skilled personnel. This is a very time consuming process, where the person involved reads the patient chart and assigns the appropriate codes. Moreover, this approach is very error prone given the huge number of CPT and ICD9 codes. A recent study (Benesch et al., 1997) suggests that only 60%-80% of the assigned ICD9 codes reflect the exact patient medical diagnosis. This can be partly explained by the fact that coding is done by medical abstractors who often lack the medical expertise to properly reach a diagnosis. Two situations are prevalent: "over-coding" (assigning a code for a more serious condition than it is justified) and "under-coding" (missing codes for existing procedures/diagnoses). Both situations translate into significant financial loses: for insurance companies in the first case and for hospitals in the second case. Additionally, accurate coding is extremely important because ICD9 codes are widely used in determining patient eligibility for clinical trials as well as in quantifying hospital compliance with quality initiatives.

Another recent study (Sonel et al., 2006) stresses the importance of developing automated methods for patient record information extraction by demonstrating how an automated system performed with 8% better accuracy than a human abstractor on a task of identifying guideline compliance for unstable angina patients. In the study, differences between the automated system and the human abstractor were adjudi-

cated by an expert based on the evidence provided.

In this paper we compare several data mining techniques for automated ICD9 diagnosis coding. Our methods are able to predict ICD9 codes by modeling this task as a classification problem in the natural language processing framework. We demonstrate our algorithms in section 4 on a task of ICD9 coding of a large population of patients seen at a cardiac hospital.

2 Related Work

Classification under supervised learning setting has been a standard problem in machine learning or data mining area, which learns to construct inference models from data with known assignments, and then the models can be generalized to unseen data for code prediction. However, it has been rarely employed in the domain for automatic assignment of medical codes such as ICD9 codes to medical records. Part of the reason is that the data and labels are difficult to obtain. Hospitals are usually reluctant to share their patient data with research communities, and sensitive information (e.g. patient name, date of birth, home address, social security number) has to by anonymized to meet HIPAA (Health Insurance Portability and Accountability Act) (hip,) standards. Another reason is that the code classification task is itself very challenging. The patient records contain a lot of noise (misspellings, abbreviations, etc), and understanding the records correctly is very important to make correct code predictions.

Most of the ICD9 code assignment systems work with a rule-based engine as, for instance, the one available online from the site *http://www.icd9coding.com/*, or the one described in (reb,), which displays different ICD9 codes for a trained medical abstractor to look at and manually assign proper codes to patient records.

A health care organization can significantly improve its performance by implementing an automated system that integrates patients documents, tests with standard medical coding system and billing systems. Such a system offers large health care organizations a means to eliminate costly and inefficient manual processing of code assignments, thereby improving productivity and accuracy. Early efforts dedicated to automatic or semi-automatic assignments of ICD9 codes (Larkey and Croft, 1995; Lovis et al., 1995) demonstrate that simple machine learning approaches such as k-nearest neighbor, relevance feedback, or Bayesian independence classifiers can be used to acquire knowledge from already-coded training documents. The identified knowledge is then employed to optimize the means of selecting and ranking candidate codes for the test document. Often a combination of different classifiers produce better results than any single type of classifier. Occasionally, human interaction is still needed to enhance the code assignment accuracy (Lovis et al., 1995).

Similar work was performed to automatically categorize patients documents according to meaningful groups and not necessarily in terms of medical codes (de Lima et al., 1998; Ruch, 2003; Freitas-Junior et al., 2006; Ribeiro-Neto et al., 2001). For instance, in (de Lima et al., 1998), classifiers were designed and evaluated using a hierarchical learning approach. Recent works (Halasz et al., 2006) also utilize NGram techniques to automatically create Chief Complaints classifiers based on ICD9 groupings.

In (Rao et al.,), the authors present a small scale approach to assigning ICD9 codes of Diabetes and Acute Myocardial Infarction (AMI) on a small population of patients. Their approach is semi-automatic, consisting of association rules implemented by an expert, which are further combined in a probabilistic fashion. However, given the high degree of human interaction involved, their method will not be scalable to a large number of medical conditions. Moreover, the authors do not further classify the subtypes within Diabetes or AMI.

Very recently, the Computation Medicine Center was sponsoring an international challenge task on this type of text classification problem.[1] About $2,216$ documents are carefully extracted (including training and testing), and 45 ICD9 labels (with 94 distinct combinations) are used for these documents. More than 40 groups submitted their results, and the best macro and micro F1 measures are 0.89 and 0.77, respectively. The competition is a worthy effort in the sense that it provides a test bed to compare different algorithms. Unfortunately, public datasets are to date much smaller than the patient records in even a small hospital. Moreover, many of the documents are very simple (one or two sentences). It is difficult to train

[1] http://www.computationalmedicine.org/challenge/index.php

good classifiers based on such a small data set (even the most common label 786.2 (for "Cough") has only 155 reports to train on), and the generalizability of the obtained classifiers is also problematic.

3 Approach

This section describes the two data mining algorithms used in section 4 for assigning ICD9 codes to patient visits as well as the real world dataset used in our experiments.

3.1 Data: ICD-9 Codes & Patient Records

We built a 1.3GB corpus using medical patient records extracted from a real single-institution patient database. This is important since most published previous work was performed on very small datasets. Due to privacy concerns, since the database contains identified patient information, it cannot be made publicly available. Each document consists of a full hospital visit record for a particular patient. Each patient may have several hospital visits, some of which may not be documented if they choose to visit multiple hospitals[2]. Our dataset consists of 96557 patient visits, each of them being labeled with a one or more ICD9 codes. We have encountered 2618 distinct ICD9 codes associated with these visits, with the top five most frequent summarized in table 1. Given sufficient patient records supporting a code, this paper investigates the performance of statistical classification techniques. This paper focuses on correct classification of high-frequency diagnosis codes.

Automatic prediction of the ICD9 codes is a challenging problem. During each hospital visit, a patient might be subjected to several tests, have different lab results and undergo various treatments. For the majority of these events, physicians and nurses generate free text data either by typing the information themselves or by using a local or remote speech-to-text engine. The input method also affects text quality and therefore could impact the performance of classifiers based on this data. In addition to these obstacles for the ICD9 classification task, patient records often include medical history (i.e. past medical conditions, medications etc) and family history (i.e. parents' chronic diseases). By embedding unstructured medical information that does not directly describe a patient's state, the data becomes noisier.

A significant difference between medical patient record classification and general text classification (e.g. news domain) is word distribution. Depending on the type of institution, department profile, and patient cohort, phrases such as "discharge summary", "chest pain", and "ECG" may be ubiquitous in corpus and thus not carry a great deal of information for a classification task. Consider the phrase "chest pain": intuitively, it should correlate well with the ICD-9 code 786.50, which corresponds to the condition chest pain. However, through the nature of the corpus, this phrase appears in well over half of the documents, many of which do not belong to the 786.50 *category*.

3.2 Support Vector Machines

The first classification method consists of support vector machines (SVM), proven to perform well on textual data (Rogati and Yang, 2002). The experiments presented use the SVM Light toolkit (Joachims, 2002) with a linear kernel and a target positive-to-negative example ratio defined by the training data. We experiment with a cost function that assigns equal value to all classes, as well as with a target class cost equal to the ratio of negative to positive examples. The results shown in this paper correspond to SVM classifiers trained using the latter cost function. Note that better results may be obtained by tuning such parameters on a validation set.

3.3 Bayesian Ridge Regression

The second method we have tried on this problem is a probabilistic approach based on Gaussian processes. A Gaussian process (GP) is a stochastic process that defines a nonparametric prior over functions in Bayesian statistics (Rasmussen and Williams, 2006). In the linear case, i.e. the function has linear form $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, the GP prior on f is equivalent to a Gaussian prior on $\mathbf{w}$, which takes the form $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w})$, with mean $\boldsymbol{\mu}_\mathbf{w}$ and covariance $\boldsymbol{\Sigma}_\mathbf{w}$. Then the likelihood of labels $\mathbf{y} = [y_1, \ldots, y_n]^\top$ is

$$P(\mathbf{y}) = \int \prod_{i=1}^{n} P(y_i|\mathbf{w}^\top \mathbf{x}_i) P(\mathbf{w}|\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w})\, d\mathbf{w} \quad (1)$$

with $P(y_i|\mathbf{w}^\top \mathbf{x}_i)$ the probability that document $\mathbf{x}_i$ takes label y_i.

[2]Currently, there is a movement to more portable electronic health records

ICD9	Freq	Coverage	Description
786.50	59957	0.621	Chest pain, unspecified
401.9	28232	0.292	Essential hypertension, unspecified
414.00	27872	0.289	Unspecified type of vessel, native or graft
427.31	15269	0.158	Atrial fibrillation
414.01	13107	0.136	Coronary atherosclerosis of native coronary artery

Table 1: Statistics of the top five ICD-9 codes most frequent in the patient record database. Frequency of ICD-9 code in corpus and the corresponding coverage (i.e. fraction of documents in the corpus that were coded with the particular ICD-9 code).

In general we fix $\boldsymbol{\mu}_{\mathbf{w}} = \mathbf{0}$, and $\boldsymbol{\Sigma}_{\mathbf{w}} = \mathbf{I}$ with $\mathbf{I}$ the identity matrix. Based on past experience we simply choose $P(y_i|\mathbf{w}^\top \mathbf{x}_i)$ to be a Gaussian, $y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$, with σ^2 a model parameter. Since everything is Gaussian here, the *a posteriori* distribution of $\mathbf{w}$ conditioned on the observed labels, $P(\mathbf{w}|\mathbf{y}, \sigma^2)$, is also a Gaussian, with mean

$$\hat{\boldsymbol{\mu}}_{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}, \qquad (2)$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ is a $n \times d$ matrix. The only model parameter σ^2 can also be optimized by maximizing the likelihood (1) with respect to σ^2. Finally for a test document $\mathbf{x}_*$, we predict its label as $\hat{\boldsymbol{\mu}}_{\mathbf{w}}^\top \mathbf{x}_*$ with the optimal σ^2. We can also estimate the variance of this prediction, but describing this is beyond the scope of this paper.

This model is sometimes called the *Bayesian ridge regression*, since the log-likelihood (i.e., the logarithm of (1)) is the negation of the ridge regression cost up to a constant factor (see, e.g., (Tikhonov and Arsenin, 1977; Bishop, 1995)):

$$\ell(\mathbf{y}, \mathbf{w}, \mathbf{X}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|^2,$$

with $\lambda = \sigma^2$. One advantage of Bayesian ridge regression is that there is a systematic way of optimizing λ from the data. Feature selection is done prior to calculation (2) to ensure the matrix inverse is feasible. Cholesky factorization is used to speed up calculation. Though the task here is classification, we treat the classification labels as regression labels and normalize them before learning (i.e., subtract the mean such that $\sum_i y_i = 0$).

4 Experiments

In this section we describe our experimental setup and results using the previously mentioned dataset and approaches. Each document in the patient database represents an event in the patient's hospital stay: e.g. radiology note, personal physician note, lab tests etc. These documents are combined to create a hospital visit profile and are subsequently preprocessed for the classification task. No stemming is performed for the experiments in this paper.

We limit our experiments on hospital visits with less than 200 doctor's notes. As a first pre-processing step, we eliminate redundancy at a paragraph level and we perform tokenization and sentence splitting. In addition, tokens go through a number and pronoun classing smoothing process, in which all numbers are replaced with the same token, and all person pronouns are replaced with a similar token. Further classing could be performed: e.g. dates, entity classing etc, but were not considered in these experiments. As a shared pre-processing for all classifiers, viable features are considered all unigrams with a frequency of occurrence greater or equal to 10 that do not appear in a standard lists of function words.

After removing consolidating patient visits from multiple documents, our corpus consists of near $100,000$ data points. We then randomly split the visits into training, validation, and test sets which contain 70%, 15%, and 15% of the corpus respectively. The classifiers were tested on an 15% unseen test set. Thus, the training set consists of approximately $57,000$ data points (patient visits), which is a more realistic dataset compared to the previously used datasets – e.g. the medical text dataset used in the Computation Medicine Center competition.

This paper presents experiments with the five most frequent ICD9 codes. This allows for more in-depth experiments with only a few labels and also ensures sufficient training and testing data for our experiments. From a machine learning perspective, most of the ICD9 codes are unbalanced: i.e. much less than half of the documents in the corpus actually have a given label. From a text processing perspective, this

	Average F1 Measure	
	Micro	**Macro**
SVM	0.683	0.652
BRR	0.688	0.667

Table 3: F1 measure for the ICD-9 classification experiments

is a normal multi-class classification setting.

Prior to training the classifiers on our dataset, we performed feature selection using χ^2. The top $1,500$ features with the highest χ^2 values were selected to make up the feature vector. The previous step in which the vocabulary was drastically reduced was necessary, since the χ^2 measure is unstable (Yang and Pedersen, 1997) when infrequent features are used. To generate the feature vectors, the χ^2 values were normalized into the ϕ coefficient and then each vector was normalized to an Euclidean norm of 1.

In these experiments, we have employed two classification approaches: support vector machine (SVM) and Bayesian ridge regression (BRR), for each of the ICD9 codes. We used the validation set to tune the specific parameters parameters for these approaches – all the final results are reported using the unseen test set. For the Bayesian ridge regression, the validation set is used to determine the λ parameter as well as the best cutting point for positive versus negative predictions in order to optimize the $F1$ measure. Training is very fast for both methods when $1,500$ features are selected using χ^2.

We evaluate our models using Precision, Recall, AUC (Area under the Curve) and F1 measure. The results on the top five codes for both classification approaches are shown in Table 2. For the same experiments, the receiver operating characteristic (ROC) curves of prediction are shown in Figure 1 and in Figure 2. The support vector machine and Bayesian ridge regression methods obtain comparable results on these independent ICD9 classification problems. The Bayesian ridge regression method obtains a slightly better performance.

It is important to note that the results presented in this section may considerably underestimate the true performance of our classifiers. Our classifiers are tested on ICD9 codes labeled by the medical abstractors, who, according to (Benesch et al., 1997), only have a 60%-80% accuracy. A better performance estimation can be obtained by adjudicating the differences using a medical expert (as the small scale approach presented in (Sonel et al., 2006)), but we did not have access to such a resource.

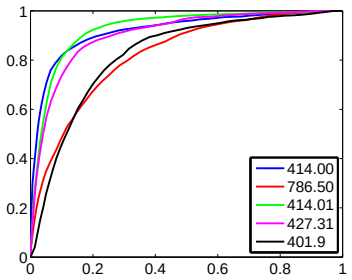

Figure 1: ROC curve for the SVM ICD9 classification

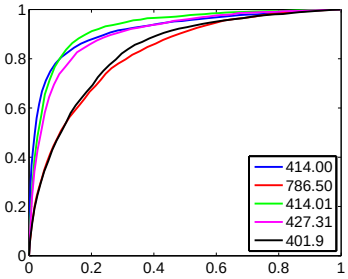

Figure 2: ROC curve for the BRR ICD9 classification

5 Conclusions & Future Work

Code classification for medical patient records is becoming a critical task in the healthcare industry. This paper presents two automatic code classification approaches and applies them on a *r*eal, *l*arge hospital dataset. We view this problem as a multi-label classification problem and seek automatic solutions, specifically targeting ICD9 code classification. We have tested two state-of-the-art classification algorithms: support vector machines and Bayesian ridge regression) with promising performance.

The data set in our study contains more than 90,000 patient visits, and is by far the largest corpus for research purpose to the best of our knowledge. The features extracted from patient visits were selected for individual ICD9 codes based on χ^2 score. Low and high-frequency features were filtered out. Several other feature selection methods were considered (including information gain), yielding comparatively moderate performance levels.

ICD9	Support Vector Machine				Bayesian Ridge Regression			
	Prec	Rec	F1	AUC	Prec	Rec	F1	AUC
786.50	0.620	0.885	0.729	0.925	0.657	0.832	0.734	0.921
401.9	0.447	0.885	0.594	0.910	0.512	0.752	0.609	0.908
414.00	0.749	0.814	0.784	0.826	0.784	0.763	0.772	0.827
427.31	0.444	0.852	0.584	0.936	0.620	0.625	0.623	0.931
414.01	0.414	0.906	0.568	0.829	0.575	0.742	0.648	0.836

Table 2: Top five ICD-9 codes most frequent in the patient record database showing the performance of support vector machine-based method (SVM) and of bayesian ridge regression-based method (BRR).

Both Support Vector Machines and Bayesian ridge regression methods are fast to train and achieve comparable results. The F1 measure performance on the unseen test data is between 0.6 to 0.75 for the tested ICD9 codes, and the AUC scores are between 0.8 to 0.95. These results support the conclusion that automatic code classification is a promising research direction and offers the potential to change clinical coding dramatically.

Current approaches are still an incipient step towards more complex, flexible and robust coding models for classifying medical patient records. In current and future work we plan to employ more powerful models, extract more complex features, and explore inter-code correlations.

Patient record data exhibits strong correlations among certain ICD9 codes. For instance the code for fever 780.6 is very likely to co-occur with the code for cough 786.2. Currently we do not consider inter-code correlations and train separate classifier for individual codes. We are currently exploring methods that can take advantage of inter-code correlations and obtain a better, joint model for all ICD9 codes.

References

C. Benesch, D.M. Witter Jr, A.L. Wilder, P.W. Duncan, G.P. Samsa, and D.B. Matchar. 1997. Inaccuracy of the international classification of diseases (icd-9-cm) in identifying the diagnosis of ischemic cerebrovascular disease. *Neurology*.

C. M. Bishop. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.

Luciano R. S. de Lima, Alberto H. F. Laender, and Berthier A. Ribeiro-Neto. 1998. A hierarchical approach to the automatic categorization of medical documents. In *CIKM*.

H. R. Freitas-Junior, B. A. Ribeiro-Neto, R. De Freitas-Vale, A. H. F. Laender, and L. R. S. De Lima. 2006. Categorization-driven cross-language retrieval of medical information. *JASIST*.

S. Halasz, P. Brown, C. Goodall, D. G. Cochrane, and J. R. Allegra. 2006. The NGram cc classifier: A novel method of automatically creating cc classifiers based on ICD9 groupings. *Advances in Disease Surveillance*, 1(30).

Health insurance portability and accountability act. 2003. *http://www.hhs.gov/ocr/hipaa*.

T. Joachims. 2002. *Learning to Classify Text Using Support Vector Machines. Dissertation.* Kluwer.

L. Larkey and W. Croft. 1995. Automatic assignment of icd9 codes to discharge summaries.

Christian Lovis, P. A. Michel, Robert H. Baud, and Jean-Raoul Scherrer. 1995. Use of a conceptual semi-automatic icd-9 encoding system in a hospital environment. In *AIME*, pages 331–339.

World Health Organization. 1997. Manual of the international statistical classification or diseases, injuries, and causes of death. *World Health Organization, Geneva*.

R.B. Rao, S. Sandilya, R.S. Niculescu, C. Germond, and H. Rao. Clinical and financial outcomes analysis with existing hospital patient records. *SIGKDD*.

C. E. Rasmussen and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

PhyCor of Corsicana. In *Book Chapter of Information Technology for the Practicing Physician*. Springer London.

B. A. Ribeiro-Neto, A. H. F. Laender, and L. R. S. De-Lima. 2001. An experimental study in automatically categorizing medical documents. *JASIST*.

Monica Rogati and Yiming Yang. 2002. High-performing feature selection for text classification. *CIKM*.

P. Ruch. 2003. *Applying natural language processing to information retrieval in clinical records and biomedical texts*. Ph.D. thesis, Department of Informatics, Universite De Genéve.

A.F. Sonel, C.B. Good, H. Rao, A. Macioce, L.J. Wall, R.S. Niculescu, S. Sandilya, P. Giang, S. Krishnan, P. Aloni, and R.B. Rao. 2006. Use of remind artificial intelligence software for rapid assessment of adherence to disease specific management guidelines in acute coronary syndromes. *AHRQ*.

A. N. Tikhonov and V. Y. Arsenin. 1977. *Solutions of Ill-Posed Problems*. Wiley, New York.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. *ICML*.