Misplacing the Code: An Examination of Data Quality Issues in Bayesian Text Classification for Automated Coding of Medical Diagnoses

Eitel J. M. Lauria, Marist College, Poughkeepsie, NY, USA; E-mail: Eitel.Lauria@marist.edu
Alan D. March, Universidad del Salvador, Buenos Aires, Argentina; E-mail: amarch@conceptum.com.ar

ABSTRACT

In this article we discuss the effect of dirty data on text mining for automated coding of medical diagnoses. Using two Bayesian machine learning algorithms (naive Bayes and shrinkage) we build ICD9-CM classification models trained from free-text diagnoses. We investigate the effect of training the classifiers using both clean and (simulated) dirty data. The research focuses on the impact that erroneous labeling of training data sets has on the classifiers' predictive accuracy.

Keywords: Text classification, Bayesian machine learning, health care coding, ICD9-CM

INTRODUCTION

Most of the data in health care settings are recorded as free text in narrative form, and are therefore prone to typographical errors and misinterpretations of ambiguous terms and phrases. To try to solve this issue, researchers and practitioners have recurred to the manual coding of information contained in clinical documents, using different coding schemes. One of the most widely used coding systems is the International Classification of Diseases (ICD), published by the World Health Organization, and in particular the Clinical Modification of its 9th edition, known as ICD-9-CM. ICD-9-CM has a hierarchical structure through which diagnose codes may be aggregated into blocks of decreasing level of detail.

The problem with ICD-9-CM is that manual coding is a costly, non-trivial task, requiring well-trained human resources. ICD-9-CM is not a mere list of codes: it is a complex ruled-based system devised to assign codes to free text based diagnoses and medical procedures. The extant literature is replete with examples depicting the relationship between coding errors and the level of expertise of health care coders. The vast amount of data generated by health care production environments imposes a restriction on the feasibility of coding all the information in a cost-efficient and timely manner. For these reasons several authors have explored the possibility of automating the coding process. Different techniques have been considered to fulfill this task, including rule-based approaches that rely on grammar-based rules (Friedman et al, 2004), and statistical text classifiers based on machine learning algorithms (March et al, 2004)

When dealing with statistical classification for automated coding, the quality of the input data used for training purposes becomes an item of concern. The effective use of statistical machine learning algorithms requires that the input data attain a certain degree of quality. There is a tradeoff between the cost of guaranteeing input data quality and the cost of misclassification given by inadequate predictive accuracy of the models developed with the input data at hand.

Two types of input data errors can be analyzed: (a) free text diagnoses containing misspellings or semantic ambiguities; (b) erroneous assignation of ICD-9-CM codes. In previous work we have focused on text errors in diagnoses (Lauría & March, 2006), disregarding potential erroneous coding. In this paper we center on coding errors: we address the issue of building text classification models based on statistical machine learning algorithms using training data in which the quality of ICD9-CM codes is questionable. Our research deals with Bayesian classifiers, specifically naive Bayes and shrinkage-based naive Bayes (McCallum et al, 1998).

ICD9-CM codes are assigned by human experts who manually review cases. There are multiple factors that can give way to errors of judgment, including the amount of time dedicated to review each case, the resources at hand, the training and expertise of the coders and the complexity of the coding process. The training data set could therefore contain clean free text diagnoses but "dirty" codes.

BAYESIAN TEXT CLASSIFIERS

Text classification can be seen as the task of estimating the unknown target function $\mathsf{f} : \mathsf{D} \rightarrow \mathsf{C}$ that assigns each document $d_j \in \mathsf{D}$ to a given category value $c_i \in \mathsf{C}$, where C is a predefined set of categories, and D is a domain of free text documents. Through supervised learning from a set of documents $D \subseteq \mathsf{D}$, a model $\hat{\mathsf{f}} : \mathsf{D} \rightarrow \mathsf{C}$ can be built to approximate the target function f. Text classification is a well studied problem, with numerous machine learning techniques that have been proposed in the literature, including probabilistic (Bayesian) methods, regression methods, decision trees, neural networks, support vector machines, maximum entropy algorithms, and classifier committees.

Naive Bayes learners have proven to be quite successful when applied to text classification, as reported by Joachims (1997). In the naive Bayes learning framework, a document *d* is classified by computing the posterior probability of each class $\mathrm{P}(c_i \mid d) \propto \mathrm{P}(d \mid c_i) \cdot \mathrm{P}(c_i)$, and assigning the most probable class given the document's words. Naive Bayes makes the simplifying assumptions that a) the probability of each word in a document is independent of its surrounding words given the class; b) the probability of each word in a document is independent of its position in the document. The naive Bayes classification criterion results in:

$$c_{NB} = \arg\max_{c_i \in \mathsf{C}} P(c_i \mid d) = \arg\max_{c_i \in \mathsf{C}} P(c_i) \prod_{k=1}^{|d|} P(w_{d_k} \mid c_i) \qquad (1)$$

where w_{d_k} identifies the word in position *k* of document *d*. The subscript d_k indicates an index into the vocabulary V of training data set *D*. Priors $P(c_i)$ are calculated by computing frequency counts on training data set *D*. Each conditional probability $P(w_{d_k} \mid c_i)$ is calculated as:

$$P(w_{d_k} \mid c_i) = \frac{N_{ik} + 1(\ )}{\sum_k N_{ik} + |\mathcal{V}|} \qquad (2)$$

We define N_{ik} to be the count of the number of times that word w_{d_k} is present in the concatenation of all sample documents that belong to category c_i. Note that the relative frequencies are supplemented by standard Laplace smoothing to

avoid probability estimates equal to zero. For a detailed analysis of naive Bayes text classification see Mitchell (1997).

HIERARCHICAL NAIVE BAYES CLASSIFICATION

For text classification problems with a large number of categories, the training data for each category are sparse, rendering less reliable conditional probability estimates, which in turn affect the performance of naïve Bayes learners as effective classifiers. But if the set of categories has a hierarchical structure, as in the case of ICD9-CM, the accuracy of a naïve Bayes classifier can be significantly improved by taking advantage of the class hierarchy. McCallum et al (1998) have used a well known statistical technique, known as *shrinkage*, that smoothes the conditional probability estimates of data-sparse leaf nodes in the class hierarchy with those of their ancestors. Intuitively, it is easy to see that the probability estimates at the leaf level are more specific but less reliable since they are calculated using less training data. The probability estimates at higher levels are calculated using more data, and are therefore more reliable; but are less specific than their corresponding children levels. For each node (class value) in a class hierarchy of *r* levels, the algorithm computes maximum likelihood (ML) estimates

$$\hat{p}_{ik}^{(h)} = N_{ik}^{(h)} / \sum_{k} N_{ik}^{(h)}, \quad h = 1 \ldots r$$

(as in equation 1, but without Laplace regularization), using all documents in the training data set labeled with that class value. Each node's training data records are filtered to eliminate its child's data before computing the ML estimate, in order to ensure that the probability estimates along a given path remain independent. A uniform probability estimate $\hat{p}_{ik}^{(0)} = 1/|\mathcal{V}|$ is added beyond the root level to deal with unreliable (e.g. zero frequency) estimates caused by rare words. An improved estimate of each leaf node $\hat{p}_{ik}$ is then calculated by "shrinking" (i.e. interpolating) its ML estimate towards the ML estimates of its *(r+1)* ancestors in the tree path

$$\hat{p}_{ik} = \lambda_i^{(0)} \cdot \hat{p}_{ik}^{(0)} + \lambda_i^{(1)} \cdot \hat{p}_{ik}^{(1)} + \ldots + \lambda_i^{(r)} \cdot \hat{p}_{ik}^{(r)} \qquad (3)$$

where $\lambda_i^{(0)}, \lambda_i^{(1)}, \ldots, \lambda_i^{(r)}$ (interpolation weights among the ancestors of class c_i) add to 1.

McCallum et al use an iterative approach (resembling Dempster's EM algorithm) to calculate optimal values of the interpolation weights. For details of the algorithm see McCallum et al (1998).

EXPERIMENTAL SETUP

Training data were gathered and cleaned from 11776 free-text outcome diagnoses occurring in 7380 hospitalizations, which were previously coded by domain experts using the 1999 Spanish Edition of ICD-9-CM. Codes were aggregated at level 3 and level 4 of the hierarchy, corresponding to the Section and 3-digit code levels of ICD-9-CM. Level 3 contained a total of 408 leaf codes, of which 2687 were part of the data set; level 4 included 2687 leaf codes, of which 651 were used. We assessed the representativeness of the test data set to the training data, both in terms of vocabulary and ICD9-CM codes (class labels).

The experiments followed these guidelines:

i. Generate multiple dirty data sets with incremental perturbations of the set of training cases (5%, 10%, 15%,20%,25% 30%,35% 40%, 45%, 50%, 55%, 60%, 65% and 70% of the cases)
ii. For each of these data sets, randomly select 10% of the sample (1178 documents) to be used as hold-out data sets for testing purposes. Use the remaining 90% (10598 documents to train the text classifiers (Note: a test sample of 10% was selected to maximize the amount of training data)
iii. To simulate an error of judgment in the assignation of ICD codes, replace a correct code with another one picked from the ICD-9-CM catalog. The replacement code is selected using combined criteria that include the numeric proximity to the correct code, the semantic similarity of the corresponding diagnoses and the frequency of occurrence of the replacement code in the training sample.
iv. Train the statistical text classifiers using both clean and dirty data. Classifiers are built for every combination of machine learning algorithm (naive Bayes and shrinkage), class hierarchy (level 3 and level 4) and training data set (1 clean , 14 dirty), 2 x 2 x 15 = 60 models all in all
v. Evaluate the classifiers' performance by measuring their predictive accuracy (mean value, standard error, 95% confidence interval)

RESULTS

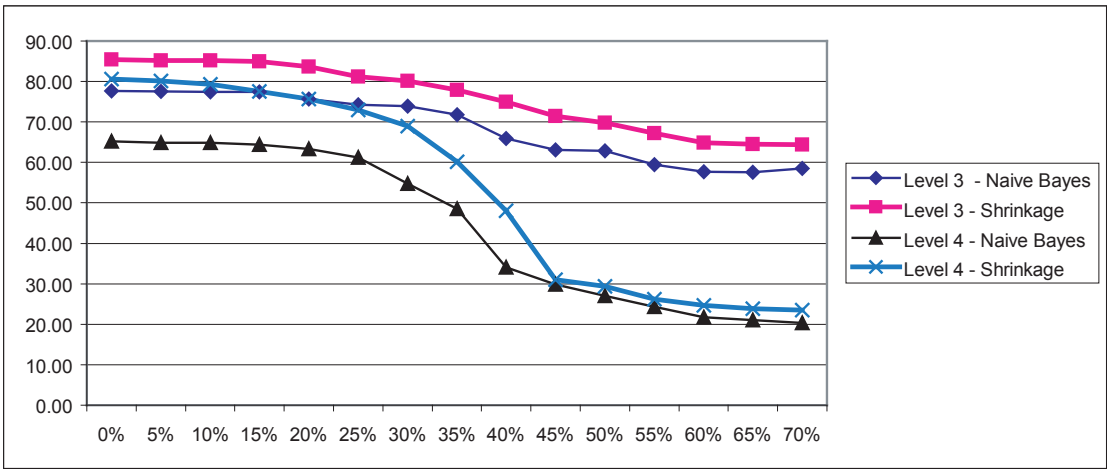
Table 1 shows the assessment of predictive accuracy performance of both text classifiers. Figure 1 displays the mean accuracy of the classifiers as a function of the percentage of label errors. The shrinkage algorithm

Table 1. Predictive accuracy of Bayesian text classifiers

	Level 3 hierarchy								Level 4 hierarchy							
	Naive Bayes				Shrinkage				Naive Bayes				Shrinkage			
%errors in data	Mean	SE	Lo (*)	Hi (*)	Mean	SE	Lo (*)	Hi (*)	Mean	SE	Lo (*)	Hi (*)	Mean	SE	Lo (*)	Hi (*)
0%	77.67	1.21	75.20	79.96	85.40	1.03	83.27	87.30	65.20	1.39	62.43	67.87	80.65	1.15	78.30	82.80
5%	77.50	1.22	75.03	79.79	85.23	1.03	83.09	87.14	64.86	1.39	62.09	67.53	80.14	1.16	77.77	82.32
10%	77.41	1.22	74.94	79.71	85.14	1.04	82.99	87.06	64.80	1.39	62.03	67.48	79.29	1.18	76.88	81.51
15%	77.39	1.22	74.91	79.69	84.97	1.04	82.82	86.90	64.35	1.40	61.57	67.03	77.59	1.21	75.12	79.88
20%	75.64	1.25	73.11	78.01	83.62	1.08	81.40	85.62	63.33	1.40	60.54	66.03	75.72	1.25	73.19	78.08
25%	74.20	1.27	71.63	76.62	81.15	1.14	78.82	83.28	61.21	1.42	58.40	63.95	72.92	1.29	70.31	75.38
30%	73.85	1.28	71.27	76.28	80.14	1.16	77.77	82.32	54.75	1.45	51.90	57.57	68.93	1.35	66.23	71.51
35%	71.82	1.31	69.18	74.31	77.84	1.21	75.38	80.12	48.47	1.46	45.63	51.32	60.19	1.43	57.37	62.95
40%	65.96	1.38	63.21	68.61	74.96	1.26	72.41	77.35	34.13	1.38	31.48	36.89	48.05	1.46	45.21	50.90
45%	63.07	1.41	60.28	65.78	71.48	1.32	68.84	73.99	29.80	1.33	27.26	32.47	30.98	1.35	28.40	33.68
50%	62.90	1.41	60.10	65.61	69.78	1.34	67.10	72.33	27.08	1.29	24.62	29.69	29.37	1.33	26.84	32.03
55%	59.50	1.43	56.67	62.27	67.20	1.37	64.47	69.82	24.33	1.25	21.97	26.86	26.20	1.28	23.77	28.79
60%	57.64	1.44	54.80	60.43	64.86	1.39	62.09	67.53	21.73	1.20	19.47	24.17	24.70	1.26	22.32	27.24
65%	57.55	1.44	54.71	60.34	64.50	1.39	61.72	67.18	21.02	1.19	18.79	23.44	23.90	1.24	21.55	26.42
70%	58.57	1.44	55.73	61.35	64.35	1.40	61.57	67.03	20.37	1.17	18.17	22.76	23.51	1.24	21.18	26.02

(*) confidence: 95%, test sample size: 1178 (10%)

Figure 1. Predictive accuracy as a function of % of errors in training data



surpasses naive Bayes in all cases (all combinations of level 3 / level 4 hierarchies and clean and dirty training data). Both text classifiers s are quite robust when subjected to training data with incremental perturbations in the labels. In particular, shrinkage evidences a rather high level of predictive accuracy with 25% of label errors (81% for level 3 and 73% for level 4). Beyond 30% the algorithm experiences an abrupt decline in performance.

CONCLUSION

Bayesian text classifiers are robust, useful tools for automated ICD9-CM coding. Preliminary results show that Bayesian text classifiers (shrinkage in particular) perform at an acceptable level, even with training data containing partially dirty labels (ICD9-CM codes). This may have a direct impact on the cost incurred in producing training data sets: predictive accuracy ca be maximized with minimum data quality enhancement cost. This kind of research could help derive policy associated with data quality procedures that precede automated coding. Investing in text classification tools should help enhance automated ICD9-CM coding while maintaining low operational costs.

REFERENCES

1. Friedman C, Shagina L, Lussier Y, et al (2004), Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 11(5):392-402
2. Lauría, E., March, A., "Effect of Dirty Data on Free Text Discharge Diagnoses used for Automated ICD-9-CM Coding", Proceedings of AMCIS 2006, the 12th Americas Conference on Information Systems, Acapulco, Mexico, August 4-6, 2006
3. March A, Lauría E., Lantos J. (2004),"Automated ICD9-CM coding employing Bayesian machine learning: a preliminary exploration", Proceedings of SIS2004 (Informatics & Health Symposium, SADIO), 33rd International Conference on Computer Science & Operational Research (JAIIO), Buenos Aires, Argentina
4. McCallum A, Rosenfeld R, Mitchell T, Ng AY (1998), Improving Text Classification by Shrinkage in a Hierarchy of Classes. In: Proceedings of the Fifteenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, pp 359-367.
5. Mitchell T (1997), *Machine Learning*, McGraw-Hill
6. Thorsten Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of the 1997 International Conference on Machine Learning (ICML '97)*, 1997.