MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220161

Accuracies of Training Labels and Machine Learning Models: Experiments on Delirium and Simulated Data

Yan Cheng[a,b], Yijun Shao[a,b], James Rudolph[c], Charlene R. Weir[d,e], Beth Sahlmann[f], Qing Zeng-Treitler[a,b]

[a] George Washington University, Washington, DC, USA
[b] Washington DC VA Medical Center, Washington, DC, USA
[c] Providence VA Medical Center, Providence, RI, USA
[d] Salt Lake City VA Medical Center, Salt Lake City, UT, USA
[e] University of Utah, Salt Lake City, UT, USA
[f] Office of Analytics and Performance Integration, Veterans Health Administration, Fort Myers, FL, USA

Abstract

Supervised predictive models require labeled data for training purposes. Complete and accurate labeled data is not always available, and imperfectly labeled data may need to serve as an alternative. An important question is if the accuracy of the labeled data creates a performance ceiling for the trained model. In this study, we trained several models to recognize the presence of delirium in clinical documents using data with annotations that are not completely accurate. In the external evaluation, the support vector machine model with a linear kernel performed best, achieving an area under the curve of 89.3% and accuracy of 88%, surpassing the 80% accuracy of the training sample. We then generated a set of simulated data and carried out a series of experiments which demonstrated that models trained on imperfect data can (but do not always) outperform the accuracy of the training data.

Keywords:

weak supervised learning; support vector machine; delirium

Introduction

Data consistency and reliability are both challenges that arise in the context of complex health systems.[14] There is considerable variability in clinical data, such as diagnostic codes, assessments, labs, and medical condition descriptions.[1; 14] Clinical data may be labeled or unlabeled to indicate an associated outcome or classification. Labeled data could be in the form of structured data or text; however, both may be prone to low reliability. For example, ICD codes often have errors, with an error rate varying from 17.1% to 76.9%.[3; 12] Even when trained reviewers examine a clinical text sample, the inter-rater agreement cannot reach 100%, causing variability in the label category.[2; 10] However, supervised machine learning usually assumes the labels in the reference standard to be completely correct. While this assumption may be appropriate for particular outcomes (e.g., death), the variability in other documented clinical outcomes challenges the 'completely correct' assumption. [13]

Many studies have explored methods to handle labelled data with errors/noise, such as label noise-robust methods, data cleansing methods, and noise tolerant methods.[6; 11; 17] Among these methods, weak supervised machine learning (i.e. learning from imperfect data), such as support vector machine (SVM), has been evidenced to accommodate incomplete, inexact, or inaccurate labels to some degree by reducing their dependence on a gold standard assumption.[17; 19] However, there is a lack of studies about the relationship between label accuracy and learning performance explored in both real-world and simulated datasets, especially in the clinical care setting. It still remains unclear if the accuracy of the imperfect data creates a performance ceiling for the trained models, and how it changes with varied sample size, error type and rate, and underlying relationship between features and outcome.

The objectives of this study are 1) to explore the relationship between label accuracy and learning performance in the context of a real clinical use case of delirium identification, which is commonly under-reported and is documented irregularly.[7]; and 2) to further explore the relationship between label accuracy and learning performance using simulated data.

Methods

Delirium data

The sample for this analysis was drawn from the Veteran's Affairs (VA) External Peer Review Program (EPRP) and from across 118 VA medical centers with inpatient facilities.[9] Medical records of inpatients who were deemed to be at high risk for delirium were randomly selected for electronic medical record (EMR) review by trained nurses. Inter-rater reliability, performed routinely within the EPRP program, found 92% agreement between reviewers for prevalent delirium.

A total of 22,851 cases from the 2015-2017 fiscal year were reviewed by EPRP for delirium documentation. Since EPRP reviewers were instructed to review H&P (history and physical), Admission, and Emergency notes at the time of admission, we only selected patients with at least one of these 3 types of notes within 2 days before or after the admission date on record. This left us with 21,458 patients with 23,230 H&P notes, 10,430 admission notes, and 42,849 emergency notes.

EPRP reviewers answered 5 delirium-related questions associated with what was documented: 1) a problem of delirium, 2) a change in mental status, 3) presence of confusion, 4) presence of disorientation for each patient case, and 5) notation that the risk of delirium was assessed. They were also instructed to focus on the "Assessment and Plan" sections in the notes. Upon review of the notes, we observed that reviewers did not consistently differentiate between the first 4 questions, nor did they consistently limit their review to the "Assessment and Plan" section. Nevertheless, in most cases, EPRP reviewers appeared to have captured the presence of delirium by answering one of the first 4 questions as a yes. Thus, we combined the first 4 questions into one question that would receive a "yes" or "no" answer: Is delirium present? If a reviewer answered yes to one or more of the first 4 questions, the answer was yes, delirium was considered present. We also utilized the entire text of documents for the analysis to be consistent with the human reviewers. We want to note here that EPRP is contracted by VA to a commercial company. Multiple reviewers participated in the reviews varying from month to month due to regular turn-over. Inter-rater reliability (IRR) assessments were performed on a sample of records previously reviewed by the EPRP reviewers. The IRRs were performed by the Regional Managers (RM), who were blinded to the original abstractor's responses. The IRR process involved the comparison of the answers from the RM review to the answers from the abstractor's review for every question available for the same records. An agreement rate was calculated based on the question level comparisons. IRR rates were calculated for the delirium study questions in 2015 and 2017. The IRR agreement rates for the five delirium questions were 98.1% and 93.6% for question 1; 92.6% and 83.3% for question 2; 95.4% and 82.1% for question 3; 97.2% and 83.3% for question 4; and 98.1% and 98.7% for question 5, respectively.

Three models were trained in our study: a logistic regression model, an SVM model with a linear kernel, and an SVM with a polynomial kernel. One- and two-gram word features were extracted from the dataset. A customized feature selection method based on discriminating power was used.[15] We randomly sampled 80% of the entire text corpus for training and 10% for testing. The testing dataset was used to empirically fine-tune the hyper-parameters for better performance. The remaining 10% of the EPRP data were used to validate the final model selected based on testing, which we refer to as internal validation. The accuracy was used to evaluate model performance.

For external validation, we randomly sampled another 100 documents from the EPRP corpus for expert review. This was a slow process because delirium is rarely documented explicitly and most of the documents were lengthy (e.g. some exceeded 5 pages) with numerous clinical details. In addition, since the EPRP sample is from high-risk patients, many patients had underlying dementia and/or other comorbid conditions which complicated the determination of delirium. For example, "anger" and "yelling" may or may not be caused by delirium. The informaticians and clinicians on the team held a series of meetings to create an annotation guideline. Two informaticians then pre-reviewed the notes together. This was followed by further discussions with the clinicians about each case that required judgment beyond what was specified in the guideline until consensus was reached by the clinicians. The difficulty of creating a reliable annotation guideline for these delirium questions highlights the benefits of this use case for this study.

Simulated data

We generated datasets for different scenarios based on four factors: a) data type defined as an underlying relationship between features and outcome (i.e. non-linear vs linear function), b) error type (systematic vs random error), c) error rate, and d) sample size.

Table 1 – Variables and parameters for data simulation set-up

Variables/Parameters	Value/Distribution of Value	Meaning
B_i	*$P(B_i=1)=0.3$*	Binomial variables
C_j	*Uniform(0, 1)*	Continuous variables
S_k	*Squared terms of randomly selected C_j*	Squared terms
I_m	*Interaction terms between two randomly selected variables from B_i and C_j*	Interaction terms
β_0	-1	Intercept
β_i, γ_j, δ_k, θ_m	*Uniform(-2, 2)*	Coefficients
ε	*Gaussian (0, 2)*	Noise

We created 100 random variables (Table 1) as the features, among which 90 were binary variables (B_i) with the probability of 30% for the value of 1, and 10 were continuous variables (C_j) with uniform distribution on the interval of 0 and 1. We set up features in this fashion to mimic data in real clinical cases. The features used in the delirium use case are binary. In clinical predictive modeling we often have a mixture of binary (e.g. diagnosis) and continuous variables (e.g. lab result). In the delirium dataset, for example, the average proportion of "true/yes" value for each feature was around 30% in general. To represent the linear relationship between features and outcome, the true value of binary outcome variable (Y) was determined by the value of a variable (V) that was derived according to the linear equation as follows:

$$V = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \cdots + \beta_{90} B_{90} + \gamma_1 C_1 + r_2 C_2 + \cdots + \qquad \text{Eq. (1)}$$

In the Eq. (1), β_0 is a constant, β_i and γ_j were coefficients for B_i and C_j, respectively, and ε was the noise representing the effect from unknown or unmeasured variables. All β_i and γ_j were randomly set with the uniform distribution on the interval of -2 and 2, and ε was set with a normal distribution with a mean of 0 and standard deviation of 2. The Y was determined by the value of V: if V > 0 then Y = 1; otherwise, Y = 0.

To set up data with a non-linear relationship between features and outcome, we added 5 squared terms (S_k) of randomly selected continuous variables and 20 interaction terms (I_m) between randomly selected variables into the equation. Adding squared and interaction terms made V no longer linearly correlated to B_i and C_j. The non-linear equation was as follows:

$$V = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \cdots + \beta_{90} B_{90} + \gamma_1 C_1 + r_2 C_2 + \cdots + r_{10} C_{10} + \delta_1 S_1 + \delta_2 S_2 + \cdots + \delta_5 S_5 + \theta_1 I_1 + \theta_2 I_2 + \cdots + \theta_{20} I_{20} + \varepsilon$$

Eq. (2)

In the Eq. (2), δ_k and θ_m were coefficients for S_k and I_m, respectively. All δ_k and θ_m were randomly set with the uniform distribution on the interval of -2 and 2.

In this simulation, we tested two types of errors occurring on the outcome variables: random errors and systematic errors. Random errors were defined as errors occurring randomly and not dependent on any condition. To generate random errors, we randomly flipped a proportion of the true value of Y according to the error rate. Systematic errors were defined as errors occurring depending on the value of certain features. To generate systematic errors, we randomly selected 10 features that were combined to determine where the true value of Y should be flipped. For each error type, we generated training data with varying sample size of 1000, 5000, 10000, 20000, and 40000 and with errors at 5 levels (0%, 5%, 10%, 20%, and 40%). We did not test on outcome classifiers with error rates more than 40%, because error rates close to 50% are almost no different from a random guess, and error rates higher than 50% will lead the classification in the wrong direction.

We trained a logistic regression model, an SVM model with a linear kernel, and an SVM model with a polynomial kernel, respectively, using a set of simulated data as well as an unseen, gold standard dataset (N=1000). Each model was trained on the 90 binary and 10 continuous feature variables. Given the randomness of error assignment, we created 10 training datasets, and then trained a model on the 10 training datasets. The performance of the model is estimated as the average of accuracies on the test datasets. The model performance was estimated for each scenario (i.e., each combination of the 4 factors described above).

Results

Delirium data

On the internal validation data, the accuracies for the logistic regression model, the SVM with a linear kernel, and the SVM with a polynomial kernel were 86.7%, 87.3%, and 86.8%, respectively, indicating good discrimination. Among these, the SVM with a linear kernel performed best. When we maximized the accuracy of the SVM with a linear kernel at 86.3%, the sensitivity reached 56.3%, specificity 93.9%, positive predicted value (PPV) 70.0% and negative predicted value (NPV) 89.4%.

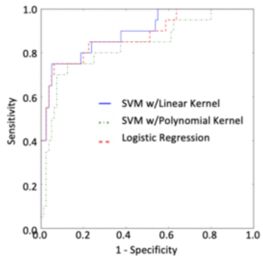

Figure 1 – ROC curves of three models against gold standard labels

Using the reference standard of 100 documents, we assessed all three models and the EPRP human annotation. Since the three models provided a prediction score, we were able to create ROCs (Figure 1). EPRP provides one label per case (there were multiple reviewers on the EPRP team, but one reviewer per case), we thus calculated accuracy, sensitivity, specificity, PPV, and NPV without ROC.

Table 2– machine learning models and human annotation performance against gold standard

	Logistic Regression	SVM with Linear Kernel	SVM with Polynomial Kernel	EPRP Annotation
Accuracy	88.0%	88.0%	87.0%	80.0%
Sensitivity	55.0%	55.0%	65.0%	70.0%
Specificity	96.3%	96.3%	92.5%	82.5%
PPV	78.6%	78.6%	68.4%	50.0%
NPV	89.5%	89.5%	91.4%	91.7%

Tested on the external validation data, all three models were more accurate than the EPRP annotation, with the SVM with a linear kernel achieving the best accuracy (Table 2). This is because large datasets annotated by a group of reviewers can yield collective wisdom or knowledge. At the same time, since each document is only reviewed by an individual EPRP reviewer, the labels are not completely accurate. It is also worth noting that the Accuracies of the logistic regression model and the SVM model with a linear kernel in the external validation (87.9% and 89.3%, respectively) are better than their Accuracies in the internal validation (86.7% and 87.3%, respectively).

Simulated data

The accuracies of three models were displayed in Figure 2. When an error rate was at 0%, 5% or 10%, accuracies were in the range of 0.8-0.9, except for scenarios with the smallest sample size of n=1000. In Figure 2, we added a reference line for the error rates of 20% and 40%. If the accuracy of a model was higher than the reference line, then it indicated that the model outperformed the accuracy of datasets (where accuracy > 1 - error rate). We found all three models outperformed the accuracy of the datasets in many scenarios where error rates are 20% and 40%.

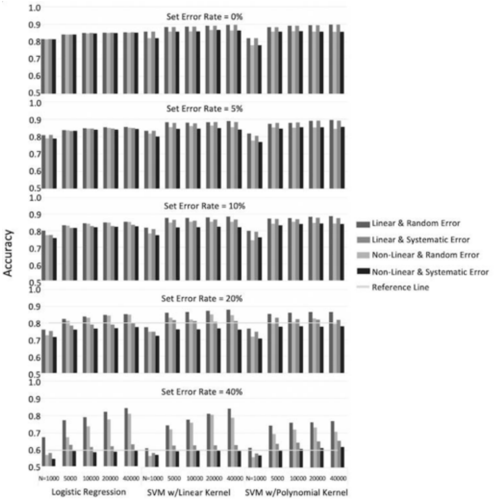

Figure 2 – Three model performance comparison in different scenarios

In general, all three models performed better on data with imputed random error than with imputed systematic error, and better on data generated using the linear function than using the non-linear function. The SVM models consistently performed better than logistic regression when the error rate is no higher than 10%, especially for data with a large sample size. For an error rate of 20%, the SVM with a linear kernel worked best on data generated using linear function and random error, and the SVM with a polynomial kernel worked best on data generated using a linear function and random error, while logistic regression worked best on data generated using the non-linear function and random error.

Discussion

In clinical research, large and perfectly labeled data is not easy to obtain. Many studies have applied different learning methods (such as semi-supervised learning, positive-only learning, and learning from consensus-based labels) in processing data with assuming that labeled data are perfect or close-to-perfect. Semi-supervised learning usually starts with some labeled data but also makes use of unlabeled data. [18] The correct labels of unlabeled data are inferred from an original set of labeled data. The inference is not expected to be completely correct, nevertheless, various studies have been able to leverage the information from unlabeled data. Positive-only learning is applied when only positive labels are available. [5; 16; 20] The remaining sample does not have any labels. Two approaches have been taken to utilize unlabeled data in the context of only positive labels: 1) inferring negative instances from those that are unknown, or 2) treating unknowns as negatives while taking into consideration that there is noise (misclassification) in the negatives.[5; 16; 20] Learning from consensus-based labels is another approach that is commonly used.[8] Reference standards are often produced through chart review and consensus building. This means labels from individual reviewers are not viewed as accurate enough for use, but rather the aggregated consensus is preferred. The machine learning models trained using this kind of reference standard have been shown to outperform each individual reviewer, but not the reference standard itself, because the consensus label is still deemed to be 100% accurate. In consideration of this, these learning methods assume the originally available labeled data are without errors. Weak supervised learning has advantages over these methods, because it assumes that an unknown proportion (≤50%) of the data is mistakenly labeled.[19] Please note a distinction between imperfect and wrong or inaccurate data. Imperfect data has >50% correctly labeled instances while wrong data has ≥ 50% instances labeled incorrectly.

Although a number of studies have utilized imperfect data, [4; 5; 8; 16; 18-20] a key question remains: Does the accuracy/error rate of the training data set a performance ceiling for the trained model? In addition, factors that affect the weak supervised learning performance also needed further study.

In both delirium and simulated experiments, weak supervised learning performed well on imperfect data (Table 2 and Figure 2). The weak supervised learning on delirium was successful because: a) we had a very large sample (about 76,000) of documents with labels; b) the labels are mostly correct (80% accuracy according to gold standard); and c) the errors are more random rather than systematic given that it was annotated by a group of annotators composed of clinicians and informaticians.

Based on our knowledge, a majority of the published informatics studies involve one or two specific datasets. Because of this, one may argue that each dataset is unique and could question the generalizability of the findings. This is the motivation behind our creation of the simulated datasets, which allowed us to generate data with varying amounts and types of error, and test the learning performance on a larger number of datasets. In addition, in each simulated dataset, we included a noise variable with a random value to introduce more variability.

On the simulated data, within each modeling algorithm (logistic regression, SVM with linear kernel, and SVM with polynomial kernel), the models were more sensitive to systematic errors and nonlinearity. In fact, with systematic errors, the model performance started to drop slightly at the highest sample size we tested, indicating a pattern of overfitting to the errors. The SVM with a linear kernel outperformed logistic regression in general and the SVM with a polynomial kernel performed best on data with systematic error especially when the error rate was at a high level. A key implication of our findings is that we can and should leverage big, imperfect datasets, but we also need to carefully assess the nature of errors, linearity of data, and modeling algorithm.

External validation is critical to weak supervised learning. In supervised learning, cross-validation or hold out data are commonly used for validation. These are, of course, not appropriate when the labels have errors. It is important to establish a small dataset with “ground truth” for validation and for selecting a threshold when needed. In the delirium use case, for instance, we need to select a cutoff with high accuracy to operationalize the automated quality measure, which is only possible if we have a validation set.

One may question that if we can create “ground truth” for a small dataset, why not do it for a larger sample. The issue is cost-effectiveness. Our team of clinicians and informaticians had multiple rounds of discussions over several weeks to reach a consensus and create a gold standard for 100 cases because clinical experts had limited availability each week, delirium state is usually not explicitly stated, and clinical judgments are involved. We could have annotated many more documents if we were not striving for 100% consensus and accuracy. Nevertheless, we could not have annotated more than 1000 documents, with confidence that we will have an error rate of <10%. As our experiments on simulated data show, having a very large but imperfect dataset can sometimes lead to a higher accuracy than having a perfect but small dataset in terms of training.

Conclusions

Our study demonstrated that machine learning models can achieve accuracy that is higher than that of training data, using both a real clinical use case and simulated data. The results of the study support the usefulness of imperfect data in clinical research via weak supervised learning.

Acknowledgements

This study is supported by the Veterans Health Administration Health Services Research & Development: # CRE 12-321; CREATE: A VHA NLP Software Ecosystem for Collaborative Development and Integration #CRE 12–315; and Center of Innovation in Long Term Services and Supports (CIN 13-465), and National Institutes of Health (NIH) Clinical and Translational Science Award (CTSA) program, grant UL1TR001876.

References

[1] A. Belle, R. Thiagarajan, S.M. Soroushmehr, F. Navidi, D.A. Beard, and K. Najarian, Big Data Analytics in Healthcare, *Biomed Res Int* **2015** (2015), 370194.

[2] M. Boguslav and K.B. Cohen, Inter-Annotator Agreement and the Upper Limit on Machine Performance: Evidence from Biomedical Natural Language Processing, in: *MEDINFO 2017: Precision Healthcare through Informatics*, 2017, pp. 298-302.

[3] J.P. Burnham, J.H. Kwon, H.M. Babcock, M.A. Olsen, and M.H. Kollef, ICD-9-CM Coding for Multidrug Resistant Infection Correlates Poorly With Microbiologically Confirmed Multidrug Resistant Infection, *Infect Control Hosp Epidemiol* **38** (2017), 1381-1383.

[4] A. Cocos, T. Qian, C. Callison-Burch, and A.J. Masino, Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation, *J Biomed Inform* **69** (2017), 86-92.

[5] C. Elkan and K. Noto, Learning Classifier from Only Positive and Unlabeled Data, in: *KDD'08*, Las Vegas, Nevada, USA, 2008, pp. 213-220.

[6] B. Frénay and M. Verleysen, Classification in the Presence of Label Noise: a Survey *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS* **25** (2014), 845-869.

[7] C. Hope, N. Estrada, C. Weir, C.C. Teng, K. Damal, and B.C. Sauer, Documentation of delirium in the VA electronic health record, *BMC Res Notes* **7** (2014), 208.

[8] G. Hripcsak, C. Friedman, P.O. Alderson, W. DuMouchel, S.B. Johnson, and P.D. Clayton, Unlocking clinical data from narrative reports: a study of natural language processing, *Ann Intern Med* **122** (1995), 681-688.

[9] S.J. Hysong, C.R. Teal, M.J. Khan, and P. Haidet, Improving quality of care through improved audit and feedback, *Implement Sci* **7** (2012), 45.

[10] O. Inel, K. Khamkham, T. Cristea, A. Dumitrache, A. Rutjes, J. van der Ploeg, L. Romaszko, L. Aroyo, and R.J. Sips, CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data, *Semantic Web - Iswc 2014, Pt Ii* **8797** (2014), 486-504.

[11] G. Lugosi, Learning with an Unreliable Teacher, *Pattern Recognition* **25** (1992), 79-87.

[12] K.J. O'Malley, K.F. Cook, M.D. Price, K.R. Wildes, J.F. Hurdle, and C.M. Ashton, Measuring diagnoses: ICD code accuracy, *Health Serv Res* **40** (2005), 1620-1639.

[13] Z. Obermeyer and E.J. Emanuel, Predicting the Future - Big Data, Machine Learning, and Clinical Medicine, *N Engl J Med* **375** (2016), 1216-1219.

[14] W. Raghupathi and V. Raghupathi, Big data analytics in healthcare: promise and potential, *Health Inf Sci Syst* **2** (2014), 3.

[15] J.A. Walsh, Y. Shao, J. Leng, T. He, C.C. Teng, D. Redd, Q. Treitler Zeng, Z. Burningham, D.O. Clegg, and B.C. Sauer, Identifying Axial Spondyloarthritis in Electronic Medical Records of US Veterans, *Arthritis Care Res (Hoboken)* **69** (2017), 1414-1420.

[16] M. Yousef, S. Jung, L.C. Showe, and M.K. Showe, Learning from positive examples when the negative class is undetermined--microRNA gene identification, *Algorithms Mol Biol* **3** (2008), 2.

[17] J. Zhang and Y. Yang, Robustness of Regularized Linear Classification Methods in Text Categorization in: *SIGIR'03*, Toronto, Canada, 2003.

[18] X. Zhang, J. Yin, and X. Zhang, A Semi-Supervised Learning Algorithm for Predicting Four Types MiRNA-Disease Associations by Mutual Information in a Heterogeneous Network, *Genes (Basel)* **9** (2018).

[19] Z. Zhou, A brief introduction to weakly supervised learning, *National Science Review* **5** (2018), 44-53.

[20] M.A. Zuluaga, D. Hush, E.J. Delgado Leyton, M. Hernandez Hoyos, and M. Orkisz, Learning from only positive and unlabeled data to detect lesions in vascular CT images, *Med Image Comput Comput Assist Interv* **14** (2011), 9-16.

Address for correspondence

Corresponding Author: Qing Zeng-Treitler

Email: zengq@email.gwu.edu

Present Address: George Washington University Biomedical
Informatics Center, 800 22nd St NW, Washington, DC 20052