Technical Brief ■

Automatic Classification of Foot Examination Findings Using Clinical Notes and Machine Learning

SERGUEI V.S. PAKHOMOV, PHD, PENNY L. HANSON, SUSAN S. BJORNSEN, STEVEN A. SMITH, MD

Abstract We examine the feasibility of a machine learning approach to identification of foot examination (FE) findings from the unstructured text of clinical reports. A Support Vector Machine (SVM) based system was constructed to process the text of physical examination sections of in- and out-patient clinical notes to identify if the findings of structural, neurological, and vascular components of a FE revealed normal or abnormal findings or were not assessed. The system was tested on 145 randomly selected patients for each FE component using 10-fold cross validation. The accuracy was 80%, 87% and 88% for structural, neurological, and vascular component classifiers, respectively. Our results indicate that using machine learning to identify FE findings from clinical reports is a viable alternative to manual review and warrants further investigation. This application may improve quality and safety by providing inexpensive and scalable methodology for quality and risk factor assessments at the point of care.

■ **J Am Med Inform Assoc.** 2008;15:198–202. DOI 10.1197/jamia.M2585.

Introduction

Regular foot examinations are part of a program to reduce the risk of serious diabetes complications[1,2] and the assessment of compliance with this guideline is central to quality assurance reporting in diabetes management.[3–5] Assessing the evidence of foot examinations is among five process quality measures for diabetes management used by the Agency for Healthcare Research and Quality (AHRQ) to report on national health care effectiveness.[6] Such quality assurance (QA) measures are important for the assessment of health care provider performance and for improving patient safety.[7] Clinical performance measures based on administrative structured data in the electronic medical record (EMR) are currently available but may be suboptimal for certain measures as compared to the information reported in the unstructured, free-text part of the EMR.[8,9] Manual audit of the medical record that relies on the free text of the EMR for many quality indicators including foot examinations is subject to lack of internal consistency and poor inter-rater reliability.[10–13] Furthermore, it is expensive, time consuming[11] and therefore is not feasible in large ambulatory patient populations. Thus QA measurements are performed on small samples of patient populations. While this sampling strategy is informative at the aggregate level of a health care organization, it does not provide actionable data that can be used to improve the health and safety of individual patients or assess the performance of individual physicians. The increasing adoption of the EMR[14] and advances in Natural Language Processing (NLP) and machine learning make it possible to analyze the unstructured text of clinical reports automatically or semi-automatically in ways which were historically impossible or cost prohibitive using paper based records.[6,15,16]

Affiliations of the authors: Department of Pharmaceutical Care and Health Systems, University of Minnesota (SVSP), Twin Cities, MN; Department of Health Care Policy and Research (PLH, SSB, SAS), Department of Endocrinology, Mayo Clinic (SAS), Rochester, MN.

Dr. Pakhomov's efforts were supported by the NIH Roadmap Multidisciplinary Clinical Research Career Development Award Grant (K12/NICHD-HD49078).

Correspondence: Serguei V.S. Pakhomov, PhD, 7-125F Weaver-Densford Hall, 308 Harvard Street S.E., Minneapolis, MN 55455; e-mail: <pakh0002@umn.edu>.

Received for review: 08/07/07; accepted for publication: 12/10/07.

Background

This study uses clinical documentation of in- and out-patient visits to the Mayo Clinic, Rochester, MN. Figure 1 shows an example of the text contained in a physical examination section of a clinical note that documents a foot examination. According to the clinical care guidelines for diabetes management, a complete foot examination should document the following three components: structural, neurological, and vascular.[17] The evidence of documentation is determined by examining the medical record by a licensed health care provider and must include at least two of the three components.[17] More recently, manual administrative coding systems tailored to National Committee for Quality Assurance (NCQA) requirements are being used as well (e.g., Kaiser Permanente).

Several systems using NLP and machine learning have been developed to process unstructured text of clinical reports.[6,16,18–22] The system described in this article relies on Support Vector Machines (SVMs) which represent a set of automatic classification algorithms widely used in medical text categorization.[23–27]

Methods

We trained three SVMs (one for each foot examination component) to associate a set of predictive covariates with the correct classification (normal, abnormal, not assessed) of the text in physical examination sections of clinical notes. The overall architecture of the system is illustrated in Figure 2. In Phase I, the text of the physical examination section is

Physical Examination*
General: On examination, patient is alert, no acute distress.
Heart: Within normal limits.
Lungs: Within normal limits.
Abdomen: Within normal limits.
Extremities: His foot examination shows good dorsalis pedis pulses bilaterally with good hair growth. He does have onychomycosis of several toenails in both feet with tinea pedis in between his toes, particularly on the left foot. There is also sensation to light touch which is also intact.

* Mayo Clinic notes are created according to the HL7 Clinical Document Architecture where each note section is represented as a coded element. The terms associated with that element (e.g., "PE", "phys exam", "exam", "physical exam", etc) may vary by service, provider and due to historical changes in the EMR documentation system.

Figure 1. An example of the text of a physical examination report.

searched electronically to determine if it contains evidence of a foot examination.[28] If such evidence is found, then the text of the note is converted to a vector of predictive covariates in Phase 2 and presented to three SVM classifiers.

Participants

We randomly selected 145 eligible patients who provided research authorization from 6000 Mayo patients who were part of the primary care diabetes registry at the Mayo Clinic and seen in the primary care clinic between July and September 2004. Compliance with diabetes care guidelines was assessed by manually examining medical records for these patients for the 12 months prior to the index visit.

Data

A total of 492 physical examination sections from 430 clinical notes representing 145 patients were used in this study. The text of each section was manually examined by two Mayo Clinic staff (PH and SS) independently from each other.

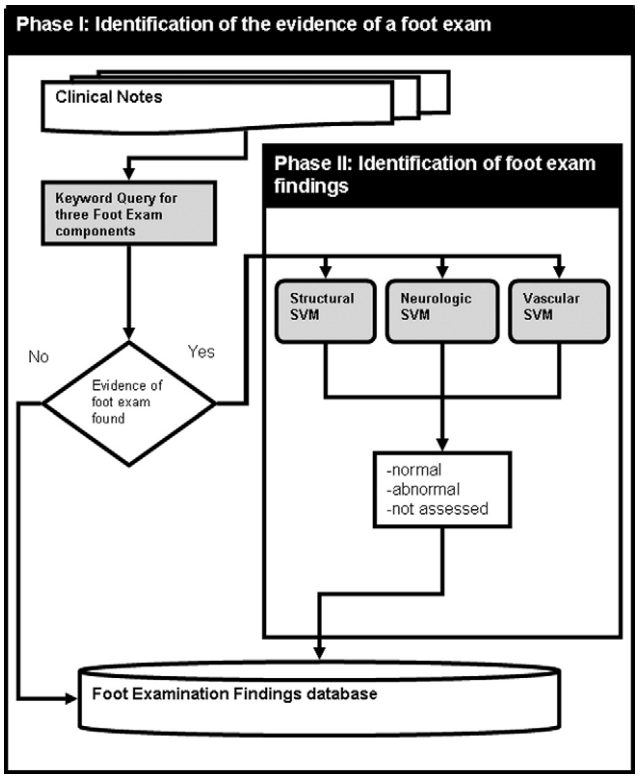

Figure 2. Architecture and process-flow of the foot exam identification system.

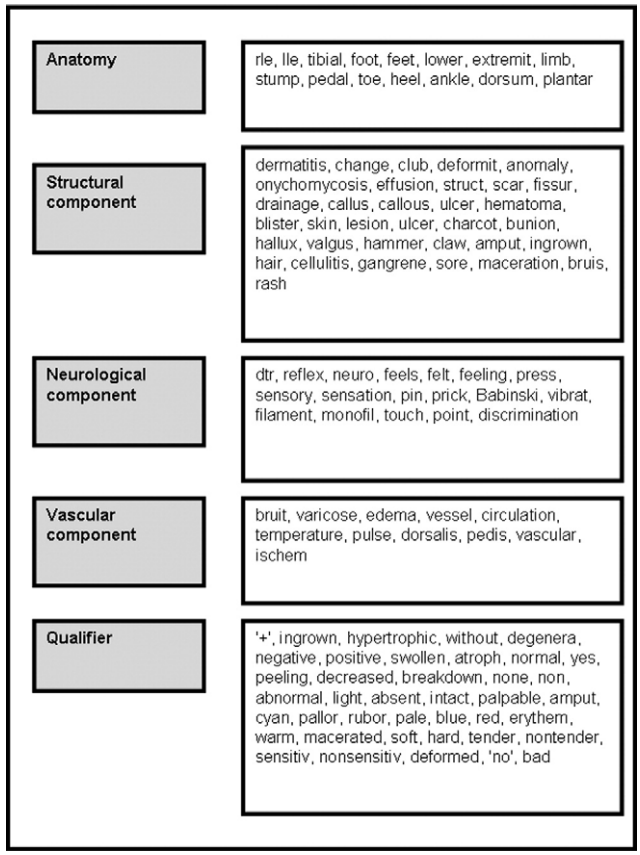

Figure 3. Keywords used as predictive covariates.

Disagreements were resolved after agreement statistics were calculated but prior to validating the approach presented in this manuscript.

Machine Learning

To train and validate the SVM classifiers, we represented each of the 492 physical examination sections in terms of predictive covariates. The covariates for the primary analysis consisted of a vocabulary illustrated in Figure 3 corresponding to five categories of indicators: anatomy, neurological findings, structural findings and qualifiers. The list of keywords was generated by using a combination of expert opinion and statistical methods. Keywords based on expert opinion were obtained first by manually examining the Measurement Manual followed by a consultation with an endocrinologist specializing in diabetes (SS). The manual contains measurement criteria and definitions based on the American Diabetes Association/National committee for Quality Assurance (ADA/NCQA)[29] and is available from the first author upon request. For the statistical approach, we split the text of the physical examination sections of all clinical notes into single words and extracted the vocabulary of covariates. Each covariate had a binary value: "1" if it was present in the note and "0" otherwise. The text of each physical examination section was converted using Perl regular expressions to a "bag-of-words"[30] vector consisting of these covariates. The covariates under the "Qualifier" category in Figure 3 we used only if these keywords were found within the scope of the Structural, Neurological or Vascular component covariates. This notion of contextual scope is exemplified in Figure 4 where each word (w) represents a

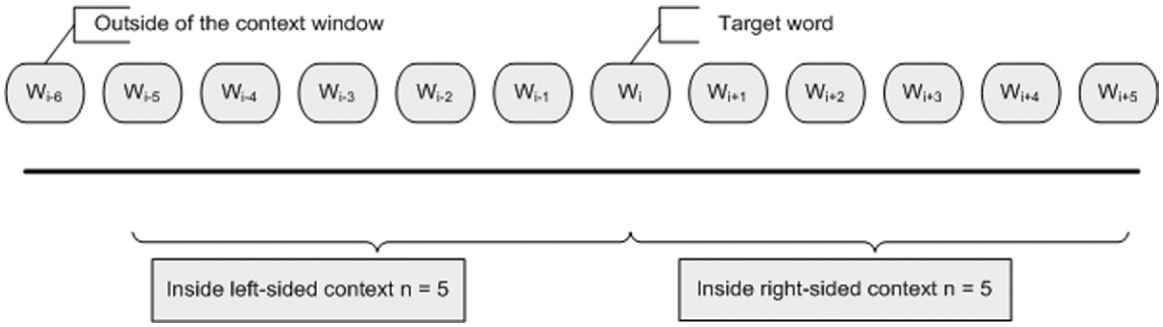

Figure 4. Schematic representation of the contextual scope for qualifiers.

predictive covariate. For example, the sentence in (1) is represented as a vector of covariates corresponding to the structural component of a foot exam shown in (2).

(1) . . . General: She is a AGE woman in no apparent distress. . . . Vessels: No abdominal aortic, femoral, renal, or carotid bruits. Dorsalis pedis is 0. Heart: . . . . Lungs: . . . . Abdomen: . . . Extremities: Venostasis and skin discoloration. Bilateral lower extremity +1 edema. No ulcers. No skin breakdown. Bilateral changes of Onychomycosis. . . .

(2) {skin 1,lower 1,extremit 1,change 1,no 1,ulcer 1,onychomycosis 1, fe_outcome abnormal}

In the example (1), the qualifier covariate "no" is set to "1" because it occurs in the context of the word "ulcers" which represents a structural component covariate. The contextual scope here is defined as the span of text within five words to the left or right of the target. The optimal size of five words was determined empirically by iterating through several rounds of validation ranging from three to ten words.

Regular expressions for covariates were defined in terms of the stemmed as well as unstemmed variants, and since the stemmed form is typically shorter it would match before the unstemmed form. Thus, while the example in (1) contains the plural form "ulcers", it is represented as stemmed "ulcer" in the covariate vector in (2). The words "dorsalis," "pedis," and "edema" are not represented in the vector in (2) because they are defined as part of the vascular component, while the example in (2) represents the covariate extraction for the structural component only.

We used WEKA implementation of the sequential minimal optimization algorithm (SMO) classifier[31] with the default parameter settings—attribute normalization, polynomial kernel with exponent of 1.0, and complexity of 1.0. No parameter optimization was performed for this study.

Statistical Analysis

Testing of each SVM component was performed with a 10-fold cross-validation strategy.[30,32] We report the overall accuracy of each of the components in addition to true and false positive rates for each of the three categories: abnormal, normal, and not assessed. Overall accuracy is computed as the ratio of all true positives and true negatives to the total number of samples. True positive rate is calculated as the ratio of true positives to the total number of positive examples for a given category in the reference standard, while the false positive rate is the ratio of false positives to the total number of negative examples in the reference standard. Kappa statistic was used to measure the reliability of the reference standard with 95% confidence intervals based on binomial error distribution.

Observations

Manual Classification of Physical Examination Text

Manual assessment of 492 physical examination sections resulted in the distribution of samples across classification categories shown in Table 1. Kappa statistic for the inter-annotator agreement was 0.93 (95% CI: 0.89–0.97) for the structural component, 0.96 (95% CI: 0.93–0.99) for the neurological component, and 0.95 (95% CI: 0.92–0.98) for the vascular component. The majority of discordances (22 out of 27; 82%) were attributable to disagreement on whether a component of the foot exam was assessed. The remaining 18% (5 out of 27) were attributable to disagreements on whether the outcome of the foot examination was normal or abnormal.

Evaluation of SVM Classifiers

The results of the primary analysis of automatic classification with SVM classifiers trained with covariates in Figure 3 are presented in Table 2. The overall accuracy of the three classifiers varied between 80% for the structural component classifier to 87% for the neurological, and 88% for the vascular components with the average accuracy of 85%. The product of accuracies across all three components is 62% indicating the worst case scenario where the final outcome is determined based on the information from all three components.

Discussion

The proposed methodology is scalable to any number of patients or providers and thus offers a low-cost, timely and

Table 1 ■ Distribution of Samples in the Manually Examined Data Set of Physical Examinations across Foot Examination Categories (Structural, Neurological and Vascular)

N = 492	Neurological n (%) (95% CI)	Vascular n (%) (95% CI)	Structural n (%) (95% CI)
Abnormal	155 (32) (27–35)	116 (24) (20–27)	144 (29) (25–33)
Normal	155 (32) (27–35)	135 (27) (23–31)	154 (31) (27–35)
Not Assessed	182 (37) (33–41)	241 (49) (45–53)	194 (39) (35–44)
Total*	100	100	100

*The percentages in columns do not add up to 100% due to rounding.

Table 2 ■ Cross-validation Results for Automatic Classification of Foot Examinations with SVM

Classifier	Neurological		Vascular		Structural	
Overall Accuracy	87%		88%		81%	
Individual Category	TP Rate (%)	FP Rate (%)	TP Rate (%)	FP Rate (%)	TP Rate (%)	FP Rate (%)
Abnormal	81	6	81	7	77	11
Normal	85	9	82	5	77	6
Not assessed	93	5	95	4	85	13

reliable strategy for investigating diabetes risk factors in large populations as well as longitudinally for individual patients. This methodology will also enable more accurate quality assurance measurements and may be used at the point of care to provide actionable information to physicians about patient profiles at high risk for diabetes complications.

Lessons Learned

Table 3 summarizes the distribution of misclassification errors. These fall into six classes depending on which of the three categories (i.e. normal, abnormal, not assessed) were predicted by the SVM for the erroneous sample and which category was manually assigned to that sample. The majority of misclassifications involved the confusion between "abnormal" and "normal" categories. We also find a large proportion of errors where the predicted category is "not assessed" while the actual category is "normal." The latter is probably due to the similarity between these two categories and is not as important in practical terms as the former type of misclassification because "not assessed" category may be treated as "normal" by default.

Informal observation of the erroneous samples also indicates that some of the misclassification errors resulted because our approach only loosely associates qualifiers with the appropriated term in the text of the report. In some instances this strategy may have resulted in associating the wrong qualifier with a given term or failing to associate a qualifier due to fixed scope boundaries. A more sophisticated strategy based on syntactic phrase boundaries identified by an automatic parser with a rule-base capable of processing semi-structured clinical reports may improve the results.

While the keywords may be extracted from available corpora automatically, domain expertise is still required to classify the clinical records for training data and to suggest additional keywords that may not have been captured by the statistical methods. The latter may be accomplished with a general purpose NLP system. One advantage of the techniques proposed in this paper over general-purpose NLP applications is that they are simple and may be focused on a specific type of a physical exam at the system design stage allowing for greater flexibility. Also, physical examination sections of clinical reports tend to contain telegraphic and sometimes semi-structured speech that varies by clinic and provider. While a general-purpose NLP system is more versatile than the methods described in this paper, to be most effective, such system's rule-base may need to be modified to account for the idiosyncrasies of the language used in physical examination sections. It may be beneficial to use a general purpose NLP system, however, to aid in extracting optimal covariates for machine learning.

Table 3 ■ Distribution of Misclassification Errors

Actual	Predicted	Structural N (% Total)	Neurological N (% Total)	Vascular N (% Total)
abnormal	normal	30 (25.9)	15 (20.5)	23 (18.9)
abnormal	not assessed	13 (11.2)	13 (17.8)	29 (23.8)
normal	abnormal	29 (25.0)	24 (32.9)	23 (18.9)
normal	not assessed	10 (8.6)	8 (10.9)	19 (15.6)
not assessed	abnormal	24 (20.7)	7 (9.6)	17 (13.9)
not assessed	normal	10 (9.0)	6 (8.3)	11 (9.0)
Total classification errors		116	73	122

Conclusion

This study demonstrates an application of machine learning for identification of foot examination findings from clinical reports. This application may improve quality and safety of patient care by providing inexpensive and scalable methodology for conducting quality assessments as well as for assessing potential risk factors at the point of care.

References ■

1. Birke JA, Horswell R, Patout CA, Jr., Chen SL. The impact of a staged management approach to diabetes foot care in the Louisiana public hospital system. J La State Med Soc 2003; 155(1):37–42.
2. Bruckner M, Mangan M, Godin S, Pogach L. Project LEAP of New Jersey: lower extremity amputation prevention in persons with type 2 diabetes. Am J Manag Care 1999;5(5):609–16.
3. Mangione CM, Gerzoff RB, Williamson DF, et al. The association between quality of care and the intensity of diabetes disease management programs. Ann Intern Med 2006;145(2):107–16.
4. Kuo S, Fleming BB, Gittings NS, Han LF, Geiss LS, Engelgau MM, Roman SH. Trends in care practices and outcomes among Medicare beneficiaries with diabetes. Am J Prev Med 2005;29(5): 396–403.
5. Saaddine JB, Cadwell B, Gregg EW, Engelgau MM, Vinicor F, Imperatore G, Narayan KM. Improvements in diabetes processes of care and intermediate outcomes: United States, 1988–2002. Ann Intern Med 2006;144(7):465–74.
6. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005;12(4):448–57.
7. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: Institute of Medicine; 2001.
8. Tang P, Ralston M, Fernandez A, Qureshi L, Graham J. Comparison of Methodologies for Calculating Quality Measures Based on Administrative Data versus Clinical Data from an Electronic Health Record System: Implications for Performance Measures. J Am Med Inform Assoc 2007;14:10–5.
9. Persell SD, Wright JM, Thompson JA, Kmetik KS, Baker DW. Assessing the validity of national quality measures for coronary artery disease using an electronic health record. Arch Intern Med 2006;166(20):2272–7.
10. Burton C, Pennington L, Roddam H, Russell I, Russell D, Krawczyk K, Smith HA. Assessing adherence to the evidence

base in the management of poststroke dysphagia. Clin Rehabil 2006;20(1):46–51.
11. Cassidy LD, Marsh GM, Holleran MK, Ruhl LS. Methodology to improve data quality from chart review in the managed care setting. Am J Manag Care 2002;8(9):787–93.
12. Gompertz PH, Irwin P, Morris R, Lowe D, Rutledge Z, Rudd AG, Pearson MG. Reliability and validity of the Intercollegiate Stroke Audit Package. J Eval Clin Prac 2001;7(1):1–11.
13. Jutley RS, McKinley A, Hobeldin M, Mohamed A, Youngson GG. Use of clinical audit for revalidation: is it sufficiently accurate? J Qual Clin Pract 2001;21(3):71–3.
14. Ford E, Menacheni N, Phillips T. Predicting the Adoption of Electronic Health Records by Physicians: When Will Health Care be Paperless? J Am Med Inform Assoc 2006;13(1):106–12.
15. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1994;1(2):142–60.
16. Friedman C. A broad-coverage natural language processing system. AMIA Annu Symp Proc 2000:270–4.
17. Diabetes Physician Recognition Program Measures for Adult Patients. November 18, 2006. Available at: http://www.ncqa.org/tabid/139/Default.aspx. Accessed Jan. 2008.
18. Hazlehurst B, Sittig DF, Stevens VJ, Smith KS, Hollis JF, Vogt TM, et al. Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. Am J Prev Med 2005;29(5):434–9.
19. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7(6):593–604.
20. Fiszman M, Chapman WW, Evans SR, Haug PJ. Automatic identification of pneumonia related concepts on chest x-ray reports. Proceedings AMIA Annu Symp 1999:67–71.
21. Pakhomov S, Weston S, Jacobsen S, Chute C, Meverden R, Roger VL. Electronic Medical Records for Clinical Research: Application to the Identification of Heart Failure. Am J Manag Care 2007;13(6 Part 1):281–8.
22. Pakhomov SS, Hemingway H, Weston SA, Jacobsen SJ, Rodeheffer R, Roger VL. Epidemiology of angina pectoris: role of natural language processing of the medical record. Am Heart J 2007;153(4):666–73.
23. Cohen AM. An effective general purpose approach for automated biomedical document classification. AMIA Annu Symp Proc 2006:161–5.
24. Xu H, Markatou M, Dimova R, Liu H, Friedman C. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. BMC Bioinform 2006;7:334.
25. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis C. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc 2005;12(2):207–16.
26. Hiissa M, Pahikkala T, Suominen H, Lehtikunnas T, Back B, Karsten H, et al. Towards automated classification of intensive care nursing narratives. Stud Health Technol Inform 2006;124:789–94.
27. Joshi M, Pakhomov SV, Pedersen T, Chute CG. A comparative study of supervised learning as applied to acronym expansion in clinical reports. AMIA Annu Symp Proc 2006:399–403.
28. Pakhomov S, Bjornsen S, Hanson P, Smith S. Quality Performance Measurement Using the Text of Electronic Medical Records. Medical Decisions Making 2007;(in press).
29. Bjornsen S, Murphy M, Bryant S, Holm D, Dinneen S, Smith S. Reliability of Data Collection for the Assessment of Diabetes Performance Indicators. Diabetes 1998;47(Suppl 1):A184.
30. Manning C, Shutze H. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press; 1999.
31. Platt J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Schoelkopf B, Burges C, Smola A, eds. Advances in Kernel Methods—Support Vector Learning. Boston, MA: MIT Press; 1998.
32. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. San Francisco: Elsevier; 2005.