Machine Learning Methods for Verbal Autopsy in Developing Countries

Sean T. Green, Abraham D. Flaxman

UW Institute for Health Metrics and Evaluation
2301 5th Avenue, Suite 600 Seattle, Washington 98121 USA
stgreen@uw.edu
abie@uw.edu

Abstract

Many resource-poor countries lack the capacity to accurately track vital registration data, such as cause of death, which are crucial inputs to health and development decision making. Verbal autopsy provides a means to ascertain cause of death in the poorest countries through the means of a standard questionnaire, but because doctors are scarce and their time is better spent treating the ill, methods of classifying deaths based on questionnaire input have become increasingly important. In this paper we present preliminary work on the use of machine learning algorithms to classify cause of death in developing countries.

Introduction

Recent health sector reforms and large-scale health and development efforts such as the establishment of the Millennium Development Goals (UN 2006) and the Grand Challenges for Global Health (Varmus et al. 2003) have helped to reinforce the need for evidence-based global health priorities. Accurate health metrics and improved statistics can provide crucial decision-making inputs that enable more efficient allocation of scarce financial resources towards the most pressing health needs (Murray and Frenk 2008). Mortality statistics are a widely-used resource for setting spending priorities, but out of 192 countries worldwide, only 23 have high-quality death registration data, and 75 have no cause-specific mortality fraction information at all (King and Lu 2008). Because most of the countries without complete vital registration systems are among the poorest, those countries that would most benefit from an accurate reporting of deaths are often among those with the least reliable data.

Verbal autopsies (VA) provide a way to diagnose cause of death in those countries without complete vital registration systems and those for which many deaths occur outside of the healthcare system (Anker et al. 1999), (WHO 2007). VA uses standard questionnaires and trained interviewers to try to elicit signs, symptoms, and other information relevant to the diagnosis of a disease from the primary caregiver or next-of-kin of a recently deceased person. The method is based upon the assumption that deaths are associated with features which can be recalled during an interview and later used to distinguish among unique causes of death. Traditionally, the interview responses are reviewed by a physician or team of physicians to ascertain the likely cause of death, but recent work has explored the use of expert algorithms and data-driven methods such as statistical models and machine learning algorithms (King and Lu 2008), (Murray et al. 2007) due to the scarcity of doctors in some countries and the premium on using doctors' available time for treatment rather than the assignment of causes of death for deceased persons. Although the various VA methods do not predict causes of deaths with vague symptoms as accurately as laboratory diagnostics can, verbal autopsy can predict causes of death with distinct symptoms with some degree of accuracy (WHO 2007). For some areas of the world verbal autopsies provide the only information about mortality currently available. Provided they can match or improve upon the accuracy of physician-coded VA and expert algorithms, data-driven methods should be used because they require less time from doctors or medical experts, and may provide valid reproducible results for the diagnosis of some causes.

Problem Formulation

Verbal autopsy diagnosis is a semi-supervised learning problem in which the responses to the questions in the questionnaire form binary, categorical, and continuous attributes and the disease classification is the categorical response. Disease classification categories typically number above 30 and may be as many as 150. The "true" labels for the response are the actual disease classification for a verbal autopsy case and come from select studies with "gold standard" diagnoses for which the evidence corroborating the diagnosis meets a compelling standard.

However, because the gold standard labels often come from studies in atypical locations within a country which possess monitoring and data collection capacity, or from hospital populations in a country, a question exists as to whether the sample of cases in a data set with gold standard diagnoses is representative of the larger population within a country.

The goal of verbal autopsy diagnosis, to build a classifier which is not overly attuned to the disease fractions observed in any study, is further complicated by the differences in disease prevalence that exist between countries. In order for a verbal autopsy classifier to be useful for classifying the death of an individual, it should be able to classify a death due to a disease with a sensitivity (true positive rate) near 90%; or in other words, it must have a generalization error (1-specificity) less than or equal to 10%. If the generalization error is higher, the classifier may be useful for predicting cause specific mortality fractions for age groups at the population level, but does not satisfy the ideal goal of being able to predict cause of death for individuals.

Preliminary Results

A sample dataset from a 2001 study in Bangladesh can be found at http://www.measureddhs.com; however, because the data do not contain gold standard diagnoses, the true causes of death are not known. The crude data for the 2001 Bangladesh study is typical of many VA studies. It has 928 rows, 1528 attributes (of which approximately 200 correspond to actual VA survey questions), and 140 different cause of death categories for the response. The attributes are a mixture of demographic data and survey response data with categorical survey responses consisting of either a “yes/ no,” or a response category selected from a list of 2-20 options. Continuous attributes typically correspond to some piece of demographic information such as household income, or the duration of disease symptoms in days or years.

Using an unpublished dataset with gold standard diagnoses we are experimenting with several classification algorithms including Support Vector Machines (Boser, Guyon, and Vapnik 1992), Boosting with CART using the Adaboost.M1 algorithm (Freund and Schapire 1996), Bagging with CART (Breiman 1996), and Random Forests (Breiman 2001). The experiments have been carried out using the R programming environment. Thus far no algorithm has been able to classify cause of death for all causes with an average generalization error below 60%, but some algorithms have been able to classify cause of death for easy-to-identify individual causes, such as HIV/AIDS and measles, with a generalization error below 10%. Attempts to improve classification of hard-to-recognize diseases, such as cancers and other neoplasms, usually result in increased generalization error for easy-to-recognize diseases.

We are also considering other learning algorithms as well as ways to combine the outputs of multiple models since some models appear to predict some causes of death better than others. The list of causes could also be adjusted, to reduce generalization error, by clustering causes which have similar signs and symptoms. Although individual-level predictions provide ultimate flexibility, an intermediate goal of predicting cause-specific mortality fractions for age- and sex-specific subpopulations would be sufficient for some applications.

References

UN., The Millennium Development Goals Report 2006: Statistical Annex, New York: United Nations, 2006.

H. Varmus, R. Klausner, E. Zerhouni, T. Acharya, A.S. Daar, and P.A. Singer, “PUBLIC HEALTH: Enhanced: Grand Challenges in Global Health,” Science, vol. 302, 2003, pp. 398-399.

C. Murray and J. Frenk, “Health metrics and evaluation: strengthening the science,” The Lancet, vol. 371, Apr. 2008, pp. 1191-1199.

G. King and Y. Lu, “Verbal Autopsy Methods with Multiple Causes of Death,” Statistical Science, vol. 23, Feb. 2008, pp. 78-91.

M. Anker, R. Black, C. Coldham, H. Kalter, M. Quigley, D. Ross, and R. Snow, A standard verbal autopsy method for investigating the cause of death in infants and children, Geneva, Switzerland: World Health Organization, 1999.

WHO, Verbal autopsy standards: ascertaining and attributing cause of death, Geneva, Switzerland: WHO Press, 2007.

C.J.L. Murray, A.D. Lopez, D.M. Feehan, S.T. Peter, and G. Yang, “Validation of the Symptom Pattern Method for Analyzing Verbal Autopsy Data,” PLoS Med, vol. 4, Nov. 2007, p. e327.

B.E. Boser, I.M. Guyon, and V.N. Vapnik, “A training algorithm for optimal margin classifiers,” Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, United States: ACM, 1992, pp. 144-152.

Y. Freund and R. Schapire, “Experiments with a new boosting algorithm,” Machine Learning: Proceedings of the 13th International Conference, Bari, Italy: Morgan Kaufmann Publishers Inc., 1996, pp. pp.148-156.

L. Breiman, “Bagging predictors,” Machine Learning, vol. 24, 1996, pp. 123 - 140.

L. Breiman, “Random Forests,” Machine Learning, vol. 45, Oct. 2001, pp. 5-32.