Communications of the IIMA

Volume 13 | Issue 2 Article 7

2013

Naive Bayes Classification of Public Health Data with Greedy Feature Selection

Stephanie J. Hickey
Iona College

Naive Bayes Classification of Public Health Data with Greedy Feature Selection

Stephanie J. Hickey
Iona College, USA
stephhickey@gmail.com

ABSTRACT

Public health issues feature prominently in popular awareness, political debate, and in data mining literature. Data mining has the potential to influence public health in a myriad of ways, from personalized, genetic medicine to studies of environmental health and epidemiology, and many applications in between. Authors have asserted the importance of medical data as the basis for any conclusions applied to the public health domain, the promise of naive Bayes classification for prediction in the public health domain, and the impact of feature selection on classification accuracy. In keeping with this perspective, this study explored the combination of a naive Bayes classifier with greedy feature selection, applied to a robust public health dataset, with the goal of efficiently identifying the one or several attributes, which best predict a selected target attribute. This approach did consistently identify the most-predictive attributes for a given target attribute and produced modest increases in classification accuracy. For each choice of target attribute, the most predictive attributes were those relating to diagnosis or procedure codes, a result, which points to several opportunities for future work.

Keywords: Data mining, feature selection, health care, naive Bayes

INTRODUCTION

Public health issues feature prominently in popular awareness, political debate, and in data mining literature. Data mining, the work of discovering patterns in data, has the potential to influence public health in a myriad of ways, from personalized, genetic medicine to studies of environmental health and epidemiology, and many applications in between. Classification of new data based on patterns previously observed holds promise for applying specific advances to public health more generally. Classification algorithms that take advantage of Bayes' Theorem and prevalence statistics, dubbed naive Bayes classifiers, aim to accomplish this with readily available data.

For this study, we applied a naive Bayes classifier to a robust public health dataset, with greedy feature selection, with the objective of efficiently identifying that the n attributes which best predict a selected target attribute, without searching the input space exhaustively. For example, is length of hospital stay impacted by insurance type, by region, by type of hospital, or by something else? Do diagnoses and procedures drive outcomes (discharge status) or does something else?

This study may contribute toward applying data mining approaches to public health data, specifically, to predicting attributes that represent a measure of treatment outcome or a proxy for cost, for patients receiving health care services in U.S. hospitals, based on readily accessible patient data.

PUBLIC HEALTH CARE IN THE U. S.

The U.S. health care system has had no shortage of attention recently. According to the World Health Organization, health care spending amounted to $7,146 per capita and 15.2% of the gross domestic product in 2008, the highest of any nation. In its World Health Report 2000, its most recent survey of population health and health systems financing, however, the U.S. ranked 38th. As recently as 2010, 49.9 million residents had neither public nor private insurance to help allay the cost of health care[1]. The debate surrounding the Patient Protection and Affordable Care Act and the Health Care and Education Reconciliation Act of 2010, designed to extend insurance options to more residents and curtail further increases in health-care spending, was a major issue in the 2012 elections. Yet, despite the attention, apparent tradeoffs between the costs of health care, both to individuals and institutions, the quality of care received by most patients, and the efficiency of the system as a whole persist.

The recent explosion in data available for analysis is as evident in health care as anywhere else. Private and public insurers, health care providers, particularly hospitals, physician groups and laboratories, and government agencies are able to generate far more digital information than ever before. This data presents an opportunity; clues to the varied challenges faced by the health care system may lie in this data. The insights gained from effectively mining public health data have implications for several types of stakeholders in the current health care system: planning implications for hospital administrators, treatment protocol implications for physician groups, public health implications for legislators, government agencies, and think tanks.

LITERATURE REVIEW

Not surprisingly, a great deal of data mining analysis is being done in the public health domain, particularly predictive data mining in clinical medicine (Bellazzi & Zupan, 2008), and the potential influence of such work is broad and compelling (Kulikowski, 2002). Further, data mining in the public health domain presents unique challenges (Cios & Moore, 2002): heterogeneity of medical data, ethical, legal, and social constraints on use of that data, statistical approaches that address heterogeneity and these constraints, and the special status of medicine as a revered and scrutinized field responsible for life-and-death decisions that may affect all of us.

Naive Bayes classification has been demonstrated to be superior to several other classification methods when applied specifically to medical data (Al-Aidaroos, Bakar, & Othman, 2012). The authors evaluate naive Bayes and five other methods on fifteen medical datasets from the UCI

[1] Statistics were obtained from multiple sources (“Health Care,” 2012; “World Health Organization,” 2000; & “World Health Report,” 2010).

machine-learning repository. They favor naive Bayes both for its predictive performance and for its transparency and interpretability; both would be key for any results to be embraced by health care practitioners. They identify naive Bayes' independence assumption as the most pressing area for future work and suggest that hybrid methods might help.

An earlier study compared naive Bayes and seven other methods, plus four hybrid methods, in predicting pneumonia mortality (Cooper et al., 1997). The eight methods had absolute error rates within 1% or each other; the four hybrid methods were deemed promising but not statistically reliable.

Hassan and Verma (2007) suggest that several methods be combined; in their case, the output of three different classifiers was the input for a neural network, all used to classify mammography data. While this may have scored well on their test data, we would be concerned that such a model was unintuitive for real-world decision-making.

The importance of preprocessing is frequently cited in studies based on medical data, as well as more generally. Popescu and Khalilia (2011) use the hierarchy implicit in ICD-9 codes as a measure of similarity among patients; including so-designated similar patients during classification improved the predictive performance of random forest and support vector machine methods on three prevalent diseases. See more on ICD-9 codes below. The authors' future work involves extending their work to more diseases, still using the Nationwide Inpatient Sample data from the Agency for Healthcare Research and Quality's Healthcare Cost and Utilization Project. We would like to see their approach paired with other classification methods, as well as to health care data more broadly.

Several studies focus on feature selection as a key preprocessing step. Feature selection has been used to identify the most salient attributes in a dataset and thereby improve the accuracy and efficiency of classification under several methods: naive Bayes, IB1, and C4.5 (Huang, McCullagh, Black, Harper, 2007). In this study, features were selected based on their alignment with the target attribute. The authors focused on predicting diabetes control status for patients with type 2 diabetes; however, feature selection based on likely strength of prediction could be considered for health care data more broadly.

One series of articles describes the combination of several preprocessing steps to boost naive Bayes' classification accuracy on medical data (Abraham, Simha, & Iyengar, 2006, 2007, & 2009): entropy-based discretization and feature selection that involves filtering features with low chi-squared statistics and greedy selection among the remaining features. The authors based their work on medical datasets from the UCI machine-learning repository. Greedy feature selection using other measures of likely strength of prediction would be worth investigating.

Association rules, specifically itemset discovery using supervised beam search, have been used to investigate co-morbidity among diagnoses reflected in the 2005-2009 National Hospital Discharge Survey (Stiglic, 2011); see more on this survey below. The author identifies visualization as suitable for future work. We find two other aspects of this study interesting, however: using 3-digit ICD-9 codes to consolidate the input space, and using similarity, in their case three separate co-morbidity measures, in feature selection.

HYPOTHESIS

Several threads emerge from the literature review: the importance of medical data as the basis for any conclusions applied to the public health domain, the promise of naive Bayes classification for prediction in the public health domain, and the impact of feature selection on classification accuracy. We applied a naive Bayes classifier to a robust public health dataset, with greedy feature selection, such that the n attributes which best predict a selected target attribute might be efficiently identified. Our hypothesis was that this approach, detailed below, would assist prediction in the public health domain.

METHODOLOGY

The methodology we used follows and is divided into the components identified by Domingos (2012).

Data

We used data[2] published in the *Hospital Discharge Survey* (National Center for Health Statistics [NCHS], 2010b). This annual survey includes demographic information, admission and discharge information, diagnoses, and procedures; the 2010 study included 151,551 patients in 203 short-stay hospitals. Weights are included in each patient's record, which allow for extrapolation of statistics to national or regional levels. For simplicity, we excluded children younger than one year of age (leaving 135,418 patients).

Diagnoses and procedures captured by the *NHDS* reflect *The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM),* published by the NCHS (2010a), i.e., the World Health Organization's Ninth Revision, International Classification of Diseases (ICD-9). *ICD-9-CM* is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States.

DRGs—diagnosis related groups—are also captured (Centers for Medicare and Medicaid Services [CMS], n.d.); these are developed and used by the CMS to determine payment for inpatient hospital care of Medicare patients, and represent types of hospital cases that are expected to be similar in terms of resource use. For the 2010 *NHDS*, (NCHS, 2010b) used the CMS MS-DRG Grouper software Version 27.0 to assign the MS- DRG.

Representation

We chose a naive Bayes classifier for predicting any one of the dataset's attributes, all of which are discrete or could be easily discretized.

[2] Data may be downloaded from ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHDS/nhds10/

Naive Bayes' assumption of attributes' independence is frequently criticized. While independence almost certainly isn't the case between attributes in health care data, numerous studies have shown this approach to be accurate and efficient nonetheless. Theoreticians Rish, Hellerstein, and Thathachar (2001) consider naive Bayes classification to be most accurate when attributes are either independent or most-closely correlated. The latter seems likely for many attributes concerning health care data.

Evaluation

We used classification accuracy as a simple, effective evaluation metric that is consistent with a naive Bayes approach. Most of the dataset's attributes could be selected as the target attribute; we focused on length of hospital stay, as a potential proxy for both health care quality and cost, and discharge status, for its obvious humanitarian implications. We also considered the number of diagnostic codes, as another potential proxy for cost and severity of illness. We divided the dataset into training and testing data, using 10-fold cross-validation. Cross-validation is a method by which the data is randomly subdivided into subsets, in this case, ten. One subset is isolated for use as testing data and the remaining serve as training data. In successive iterations, the model is run with each subset serving as testing data, and the results averaged. In this way, actual data values are available to calculate classification accuracy but bias is minimized.

Optimization

We implemented greedy feature selection, such that the n attributes which best predict a selected target attribute might be efficiently identified, without searching the input space exhaustively. The classification accuracy of the target attribute with each one other attribute was calculated and ranked. It is possible that an attribute is effective in combination with another but not alone, however. So as not to preclude this possibility from surfacing, we used random selection of attributes with the above rankings as weights, such that highly ranked attributes had a greater probability of selection, reminiscent of a simulated annealing approach.

Tools

Predictive data mining is primarily a mathematical exercise, the application of algorithms to data and the resulting computation, as opposed to one in which the behavior of objects in the real world is modeled. For this reason, we chose to use a functional programming language, namely Clojure, a Lisp that runs on the JVM and permits access to any Java libraries.

To improve performance, we accessed the naive Bayes model in Weka 3.7.7 by using Weka's Java API from Clojure, as well as methods for 10-fold cross-validation. This improved calculation speed dramatically over our beginner's implementation of the same model.

FEASIBILITY STUDY

At the outset, we implemented a naive Bayes classifier in Clojure and applied it to the 2010 NHDS. We focused on patients' age, discretized into completed decades, Diagnosis Related

Group, which has 730 distinct values, and length of hospital stay, discretized into six categories. The prevalence of each length of stay category in the full dataset, alone and for each combination of age and diagnosis, was used to predict the length of stay for each patient. For a sample of 50,000 patients drawn from regular intervals of the full (unordered) dataset, 53% were predicted correctly. These results were duplicated using Weka 3.6.6 Explorer.

RESULTS AND DISCUSSION

This combination of a naive Bayes classifier with greedy feature selection did identify the most-predictive n attributes for a given target attribute. Both the level of classification accuracy and the increase in classification accuracy achieved varied by target attribute, however. In some cases that increase was modest; in others, it alluded to the (disappointing) reality that such predictive relationships may not be hiding in the data after all.

Because our greedy feature selection algorithm has an element of randomness to it, the attributes selected as most-predictive after 100 iterations (or some number of iterations smaller than the input space) are generally consistent, but not exactly the same, from run to run. Classification accuracy varied little from run to run, however.

To explore this approach, we chose as the target attribute length of hospital stay, discharge status, and number of diagnostic codes, in turn.

Length of Hospital Stay

For length of hospital stay, we discretized the data into the following six categories: one day, two days, one week, one month, two months, and long term (the range was 0 to 497 days). A worst-case, blind guess, therefore, might be 1/6, or 17%, classification accuracy. An estimate based on the data and the starting point for a naive Bayes approach, would be to predict whichever is the most prevalent value (here, two days); in this case that would achieve 46% classification accuracy. Using naive Bayes model, the best single attribute achieved 61%; the best two, three, and four attributes selected in 100 iterations achieved 67%-70% classification accuracy. In all iterations, the most successful attributes, alone or in combination, were those relating to diagnosis or procedure codes. More details are included in Table 1.

Attributes Selected in 100 Iterations	Classification Accuracy
Primary diagnosis	61%
Primary diagnosis, primary procedure	67%
Primary diagnosis, primary procedure, admitting diagnosis	69%
Primary diagnosis, primary procedure, admitting diagnosis, DRG	70%

Table 1: Length of Hospital Stay.

We did explore other categories besides these six, as well as investigated whether misclassifications were "close," meaning one category removed from the correct one. Neither proved to be significant.

Discharge Status

We were optimistic that successfully predicting discharge status would have worthwhile implications, but the data was uncooperative. The 2010 *NHDS* (NCHS, 2010b) uses seven discharge statuses: routine, unadvised, transfer to short-term facility, transfer to long-term facility, alive otherwise unknown, dead, and status unknown. The most prevalent value was present in 76% of the patients (thankfully, routine discharge), making that a potent predictor. Using naive Bayes model, the best single, two, three, and four attributes selected in 100 iterations were only able to improve that to 80-83% classification accuracy. The same set of diagnosis code- and procedure code-related attributes were the better predictors. More details are included in Table 2.

Attributes Selected in 100 Iterations	Classification Accuracy
Primary diagnosis	80%
Primary diagnosis plus either primary procedure or admitting diagnosis	82%
Primary diagnosis, primary procedure, and either admitting diagnosis or admission source	82%
Primary diagnosis, primary procedure, admitting diagnosis, admission source	83%

Table 2: Discharge Status.

Number of Diagnostic Codes

The 2010 *NHDS* (NCHS, 2010b) captures up to fifteen diagnosis codes and up to eight procedure codes for each patient; the first of each is designated to be primary, but following that, they are generally unranked. Two of the attributes we included, throughout, were calculations of the number of each present in the data. Possible values for the number of diagnosis codes were 1-15. The distribution was fairly flat, with the most prevalent value present in 17% of the patients. Using naive Bayes model, the best single attribute improved classification accuracy to 40%. The best two, three, and four attributes selected in 100 iterations improved that to 49%-56%. The same set of diagnosis code- and procedure code-related attributes were the better predictors. More details are included in Table 3.

Attributes Selected in 100 Iterations	Classification Accuracy
Primary diagnosis	40%
Primary diagnosis, admitting diagnosis	49%
Primary diagnosis, primary procedure, admitting diagnosis	54%
Primary diagnosis, primary procedure, admitting diagnosis, DRG	56%

Table 3: Number of Diagnostic Codes.

Concerns

The possibility that predictive relationships are not supported by the data or, on the other hand, that predictive relationships inferred from the data are somehow unique to that data and have no applicability beyond it (over-fitting), must be considered.

The curse of dimensionality may be a factor, that is, the possible combinations of all the values of all the attributes may create an input space large enough that available data is insufficient to assess confidently any predictive relationships therein. Even a dataset of 150,000 patients covers a small subset of the input space created by this dataset's attributes; we hope that data is not uniformly distributed in input space, but sufficiently clumped to enable the calculation of useful classifiers.

CONCLUSION AND FUTURE WORK

This study had two objectives, to aid stakeholders in the current health care system and to aid researchers applying data mining techniques to the public health domain.

The combinations of attributes selected as the best predictors of the selected target attributes, which themselves represent a measure of treatment outcome or a proxy for cost, might have utility for stakeholders in the current health care system. Several types of stakeholders make decisions based on this type of information. The best combinations of attributes might have planning implications for hospital administrators, treatment protocol implications for physician groups, and public health implications for legislators, government agencies, and think tanks.

In addition, the impact of feature selection on classification accuracy has been reiterated throughout the literature. The best combinations of attributes identified here might prove useful to researchers looking for dimensionality reduction for their own studies. These combinations of attributes might be the input to more refined models, tailored to the specific attributes.

The combination of a naive Bayes classifier with greedy feature selection did consistently identify the most-predictive n attributes for a given target attribute, although greater than modest increases in classification accuracy would have been gratifying. For each choice of target attribute, the most-predictive attributes, alone or in combination, were those relating to diagnosis or procedure codes. This result points to some refinements or expansions that we think might be worthwhile.

One aspect of the 2010 NHDS that we have not fully explored is the specific diagnosis and procedure codes. Because they are generally unranked after the first of each, their positions in each patient's record are not informative. They might be represented as an array of binary attributes, each representing an individual code and its inclusion, or not, in each patient's record. It might be interesting to consider whether certain codes are stronger predictors than either the primary code or the set of codes is; this effort might be hampered, however, by the input space it implies. Using 3-digit ICD-9 codes, in the style of Stiglic (2011), or the hierarchy implicit in ICD-9 codes, in the style of Popescu and Khalilia (2011), might moderate that concern.

The *NHDS* (NCHS, 2010b) has been conducted annually since 1965 and captures a fairly consistent set of data elements; changes in scope and methodology have been well-documented. Expanding this effort to include more years' data might impart greater confidence that the input space was well-covered.

The *NHDS* (NCHS, 2010b) data are sampled at both the hospital level and the patient level, and is meant to be representative of hospital utilization nationally. In this dawning era of Big Data, however, were there another, direct source of de-identified patient data that might permit more nuanced analysis. It would be exciting to participate in the discovery of any predictive relationships that may have been masked by sampling that was necessary until now.

ACKNOWLEDGEMENT

The author thanks Dr. Smiljana Petrovic, Department of Computer Science, Iona College, for guidance and technical review, and Rich Hickey, for technical review.

This research satisfies the thesis requirements for an M.S. in Computer Science.

REFERENCES

Abraham, R., Simha, J. B., & Iyengar, S. S. (2006). A comparative analysis of discretization methods for medical data-mining with naive Bayesian classifier. In S. P. Mohanty & A. Sahoo (Eds.) *Proceedings of the 9th International Conference on Information Technology* (pp. 235-236). Bhubaneswar, India: IEEE Computer Society. doi: 10.1109/ ICIT.2006.5

Abraham, R., Simha, J. B., & Iyengar, S. S. (2007). Medical datamining with a new algorithm for feature selection and naive Bayesian classifier. In S. K. Rath & P. Mohapatra (Chairs) *10th International Conference on Information Technology* (pp. 44-49). Rourkela, India: IEEE. doi: 10.1109/ICIT.2007.41

Abraham, R., Simha, J. B., & Iyengar, S. S. (2009). Effective discretization and hybrid feature selection using naive Bayesian classifier for medical datamining. *International Journal of Computational Intelligence Research, 5*(2), 116-129. doi: 10.5019/j.ijcir.2009.175

Al-Aidaroos, K. M., Bakar, A. A., & Othman, Z. (2012). Medical data classification with naive Bayes approach. *Information Technology Journal, 11*, 1166-1174. doi: 10.3923/itj.2012.1166.1174

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics, 77*(2), 81-97.

Centers for Medicare & Medicaid Services. (n.d.). *Details for Title: Table 5—List of MS-DRGs, Relative weighting factors and geometric and arithmetic mean length of stay*. Retrieved from http://www.cms.hhs.gov/AcuteInpatientPPS/downloads/FY_2010_FR_Table_5.zip

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine, 26*(1-2), 1-24.

Cooper, G. F., Aliferis, C. F., Ambrosino, R., Aronis, J., Buchanan, B. G., Caruana, R., . . . Spirtes, P. (1997). An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine, 9*(2), 107-138.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87. doi: 10.1145/2347736.2347755

Hassan, S. Z., & Verma, B. (2007). A hybrid data mining approach for knowledge extraction and classification in medical databases. In L. M. Mourelle, N. Nedjah, J. Kacprzyk, & A. Abraham (Eds.) *Seventh International Conference on Intelligent Systems Design and Applications* (pp. 503-510). Rio de Janerio, Brazil. IEEE Computer Society.

Health care in the United States. (2012). In *Wikipedia*. Retrieved December 2, 2012, from http://en.wikipedia.org/wiki/Health_care_in_the_United_States

Huang, Y., McCullagh, P., Black, N., & Harper, R. (2007). Feature selection and classification model construction on type 2 diabetic patients' data. *Artificial Intelligence in Medicine, 41*(3), 251-262.

Kulikowski, C. A. (2002). The micro-macro spectrum of medical informatics challenges: From molecular medicine to transforming health care in a globalizing society. *Methods of Information in Medicine; 41*(1), 20-24.

National Center for Health Statistics. (2010a). *International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)*. Retrieved from http://www.cdc.gov/nchs/icd/icd9cm.htm

National Center for Health Statistics. (2010b). *National hospital discharge survey: Public use data file documentation*. .Retrieved from ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHDS/NHDS_2010_Documentation.pdf

Popescu, M., & Khalilia, M. (2011). *Improving disease prediction using ICD-9 ontological features*. In proceedings of the 2011 IEEE International Conference on Fuzzy Systems (FUZZ). Taipei, Taiwan.

Rish, I., Hellerstein, J., & Thathachar, J. (2001). *An analysis of data characteristics that affect naive Bayes performance* (Technical Report). Hawthorne, NY: IBM T. J. Watson Research Center. Retrieved from http://www.cs.iastate.edu/~honavar/rish-bayes.pdf

Stiglic, G. (2011). Human disease network guided discovery of interesting itemsets in hospital discharge data. In *Proceedings of the 2011 workshop on Data Mining for Medicine and Healthcare* (pp. 76-79). San Diego, CA: Association for Computing Machinery (ACM). doi: 10.1145/2023582.2023597

World Health Organization ranking of health systems. (2000). In *Wikipedia*. Retrieved December 3, 2012, from http://en.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems

World health report. (2010). In *Wikipedia*. Retrieved December 8, 2012, from http://en.wikipedia.org/wiki/World_Health_Report

This Page Was Left Blank Intentionally.