Naïve Bayes Classification of Public Health Data

With Greedy Feature Selection

Colors on Green by Robert Goodnough (1973)

A thesis submitted in partial fulfillment of the requirements

for the degree of Master of Science

Stephanie Hickey

Department of Computer Science, Iona College

May 2013

ABSTRACT

Public health issues feature prominently in popular awareness, political debate, and also 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 naïve 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 naïve Bayes classifier with greedy feature selection, applied to a robust pubic 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.

TABLE OF CONTENTS

Naïve Bayes Classification of Public Health Data With Greedy Feature Selection

Stephanie Hickey

May 2013

Introduction

Public health issues feature prominently in popular awareness, political debate, and also 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 naïve Bayes classifiers, aim to accomplish this with readily available data.

For this study, I applied a naïve Bayes classifier to a robust pubic 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 heath care spending, was a major issue in the 2012 elections. Yet despite the attention, apparent tradeoffs between the cost of health care, to both individuals and institutions, the quality of care received by most patients, and the efficiency of the system as a whole persist.

[1] Statistics obtained from Wikipedia.com entries for "World Health Organization ranking of health systems" and "World Health Report," December 8, 2012, and "Health care in the United States," December 3, 2012.

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.

Data Mining

Data mining is a relatively new discipline that builds on statistics, artificial intelligence, and pattern recognition and is focused on discovering useful patterns and relationships in large datasets. Useful patterns may be either descriptive - grouping data or attributes in a meaningful way - or predictive - enabling a generally unknown attribute to be surmised from available data. Both may be fruitful in the public health domain; this study focuses on predictive data mining.

Supervised Machine Learning

Machine learning refers to the subset of data mining approaches that are iterative, in which a model is programmatically "learned" in repeated processing of the data. Algorithms that use a training dataset to develop a model are considered supervised. Supervised approaches are generally appropriate when prediction is the objective; the training data contains correct or desired values of the target attribute and accurately predicting these drives the development of the model. By contrast, unsupervised machine learning is used to discover patterns in data that are not pre-determined, and aligns with descriptive approaches such as clustering - grouping data according to a similarity metric - and mining frequent patterns - finding recurring sequences or itemsets in data, and crafting association rules that describe them.

CLASSIFICATION

Predictive data mining on data with discrete attributes is classification: a discrete, target attribute is predicted for new data using a model honed on training data. Several classification approaches are explored in the literature reviewed below.

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks are designed to mimic the activity of neurons. A set of interconnected, multi-layer, inputs and outputs is created, with weights at each connection which are adjusted, iteratively, until accuracy predicting training data is optimized. Artificial neural networks can model any pattern or relationship in data, but the model itself is generally unintuitive; only the initial inputs (the values per attribute in the data) and outputs (the value for the target attribute) are interpretable, the weights and intermediate inputs/outputs in the interior layers are not.

DECISION TREE INDUCTION

Decision tree induction uses training data to construct a flowchart-like tree structure, where each branch represents the value of an attribute and each leaf represents a value of the target attribute. Decision trees require no domain knowledge and yet produce intuitive models for prediction and decision-making.

A refinement of this approach is random forest, in which multiple decision trees are constructed on subsets of the dataset - subsets of data or of attributes - and their predictions for new data are aggregated.

Naïve Bayes Classifiers

Naïve Bayes classifiers are statistical classifiers that use Bayes' theorem to convert the prevalence of attribute-value combinations, the prevalence of each value of the target attribute, and the prevalence of attribute-value combinations for each value of the target attribute, in the training data, into the most-likely value of the target attribute for a given attribute-value combination. Naïve refers to the approach's assumption that the impact of each attribute on the value of the target attribute is independent of the impact of each other attribute. Naïve Bayes classifiers are nicely transparent as a basis for prediction and decision-making.

A refinement of this approach are Bayesian belief networks, which allow for conditional dependence between attributes. A Bayesian belief network represents relationships between attributes, including the target attribute, as a directed acyclic graph; the probability distribution for each attribute (each node in the graph), given its related attributes (parent nodes), may be calculable from training data or may be based on domain knowledge. The relationships between attributes that determine the structure of the graph must be based on domain knowledge.

Support Vector Machines

Support vector machines are models that map training data, non-linearly, onto a higher-dimensional space in order to find a linear boundary between values of the target attribute. Support vectors refer to the edges of this boundary, and are separated by a margin which is maximized. Training a support vector machine may be a computationally-intensive process and the selection of an appropriate non-linear mapping may benefit from domain knowledge.

Instance-based Learners

Rather than building a model with training data and then considering new data, instance-based learners classify each new data instance according to the most-similar instances in the training data. These may be the *k*-nearest instances as mapped in Euclidean space or the instances with similar sets of attribute-value combinations - a matching subgraph where instances' attributes are represented as graphs. Instance-based learners offer no insight into patterns or relationships present in the data generally, but support incremental learning and are well-suited for complex data that may otherwise be difficult to model.

Whereas predictive data mining on data with discrete attributes is classification, regression is the appropriate method for data with continuous attributes.

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 (Bellizzi and Zupan, 2006), 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, 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.

Naïve Bayes classification has been demonstrated to be superior to several other classification methods when applied specifically to medical data (Al-Aidaroos, Bakar, and Othman, 2012). The authors evaluate naïve Bayes and five other methods on fifteen medical datasets from the University of California at Irvine (UCI) Machine Learning Repository. They favor naïve 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 naïve Bayes' independence assumption as the most pressing area for future work and suggest that hybrid methods might help.

An earlier study compared naïve 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 suggest that several methods be combined (Hassan and Verma, 2007); in their case, the output of three different classifiers was the input for an artificial neural network, all used to classify mammography data. While this may have scored well on their test data, I 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 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 (Popescu and Khalilia, 2011). 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. I 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: naïve Bayes, an IB1 instance-based learner, and C4.5 decision trees (Hwang, McCullagh, Black, and 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 naïve Bayes' classification accuracy on medical data (Abraham, Simha, and Iyengar, 2006, 2007, and 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. I 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, above: the importance of medical data as the basis for any conclusions applied to the public health domain, the promise of naïve Bayes classification for prediction in the public health domain, and the impact of feature selection on classification accuracy. I applied a naïve Bayes classifier to a robust pubic health dataset, with greedy feature selection, such that the n attributes which best predict a selected target attribute might be efficiently identified. My hypothesis was that this approach, detailed below, will assist prediction in the public health domain.

Methodology

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

DATA

I used the 2010 National Hospital Discharge Survey (NHDS) from the U.S. Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS).[2] The NHDS is an annual survey that includes demographic information, admission and discharge information, diagnoses, and procedures; the 2010 NHDS included 151,551 patients in 203 short-stay hospitals. Weights are included in each patients' record which allow for extrapolation of statistics to national or regional levels. For simplicity, I excluded children younger than one year old (leaving 135,418 patients).

Diagnoses and procedures[3] captured by the NHDS reflect The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), 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.

[2] For more information on the 2010 National Hospital Discharge Survey, see http://www.cdc.gov/nchs/nhds.htm. In particular, see Public Use Data File Documentation (NHDS_2010_Documentation.pdf) available under Public-Use Data Files / NHDS - Downloadable documentation via ftp. The dataset itself (NHDS10.PU.txt) may be downloaded from Public-Use Data Files / NHDS - Downloadable data files via ftp.

[3] For more information see http://www.cdc.gov/nchs/icd/icd9cm.htm.

Diagnosis Related Groups (DRGs)[4] are also captured; these are developed and used by the Centers for Medicare and Medicaid Services (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 used the CMS MS-DRG Grouper software Version 27.0 to assign the MS- DRG.

The following attributes were included as either predictive or target attributes:

Table 1: Attributes Considered

Attributes Considered	
Age (discretized to decades)	Weight
Sex	Number of diagnoses (calculated, up to 15)
Race	Primary diagnosis
Marital status	Number of procedures (calculated, up to 8)
Month of discharge	Primary procedure
Discharge status	Payment method
Length of hospital stay (discretized)	DRG
Region	Type of admission
Hospital size	Source of admission
Hospital ownership	Admitting diagnosis

Additional detail is provided as an Appendix.

[4] The complete list is available at http://www.cms.hhs.gov/AcuteInpatientPPS/downloads/FY_2010_FR_Table_5.zip.

Representation

I chose a naïve Bayes classifier for the purpose of predicting any one of the dataset's attributes, all of which are discrete or could be easily discretized. To pick the most-likely value for the target attribute for each instance, as the prediction, meant calculating the probability of each value of the target attribute for each instance's attribute-value combination, based on:

- the probability of that instance's attribute-value combination, based on their prevalence in the training data (number of instances with that combination divided by total number of instances),
- the probability of each value of the target attribute, based on their prevalence in the training data, and
- the probability of that instance's attribute-value combination given each value of the target attribute, based on their prevalence in the training data.

To predict length of hospital stay (LOS) for patient #400, for example:

prob (LOS = one day | patient #400’s age, sex, race, diagnoses, etc.) =

$$\frac{\text{prob (patient \#400's age, sex, race, diagnoses, etc. | LOS = one day) prob (LOS = one day)}}{\text{prob (patient \#400's age, sex, race, diagnoses, etc.)}}$$

however, prob (patient #400’s age, sex, race, diagnoses, etc.) =
prob (age = patient #400’s age) prob (sex = patient #400’s sex)
prob (race = patient #400’s race) prob (diagnoses = patient #400’s diagnoses) ...

similarly, prob (patient #400’s age, sex, race, diagnoses, etc. | LOS = one day) =
prob (age = patient #400’s age | LOS = one day)
prob (sex = patient #400’s sex | LOS = one day)
prob (race = patient #400’s race | LOS = one day)
prob (diagnoses = patient #400’s diagnoses | LOS = one day) ...

predicted LOS =
max [prob (LOS = one day | patient #400’s age, sex, race, diagnoses, etc.),
prob (LOS = two days | patient #400’s age, sex, race, diagnoses, etc.),
prob (LOS = one week | patient #400’s age, sex, race, diagnoses, etc.),
...
prob (LOS = long term | patient #400’s age, sex, race, diagnoses, etc.)]

Naïve Bayes’ assumption of attributes’ independence is frequently criticized, and was clearly a consideration here. 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. One team of theoreticians (Rish, Hellerstein, and Thathachar, 2001) consider naïve 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

I used classification accuracy as a simple, effective evaluation metric that is consistent with a naïve Bayes approach. Most of the dataset's attributes could be selected as the target attribute; I 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. I also considered number of diagnostic codes, as another potential proxy for cost and also severity of illness.

The dataset was divided into training and testing data using 10-fold cross-validation. Cross-validation is a method of model evaluation in which the dataset is randomly subdivided into roughly-equal subsets. For each of ten iterations, one subset is isolated for use as testing data and the remaining nine serve as training data. Thus the model is run ten times, with each subset serving as testing data once, and the results averaged. In this way, actual data values are available for both training and testing, but bias and variance are minimized.

OPTIMIZATION

I 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. Intuitively, the classification accuracy of the target attribute with each one other attribute might be calculated and ranked, based on the full dataset, and the n attributes with the highest individual classification accuracies selected for consideration as a set. It is possible that an attribute is effective in combination with another but not alone, however. In order to not preclude this possibility from surfacing, I used random selection of attributes with the above rankings as weights, such that highly ranked attributes had a greater probability of selection.

For the above example, predicting length of hospital stay (LOS):

Table 2: Probability of Feature Selection Targeting Length of Hospital Stay

Classification Accuracy	Attributes	Probability of Selection
45.7%	Age Sex Race Marital status Month of discharge Region Hospital size Hospital ownership Payment method Type of admission	1/48
45.8%	Source of admission	2/48
45.9%	Discharge status Number of diagnoses	3/48
46.0%	Number of procedures	4/48
53.2%	Admitting diagnosis	5/48
53.4%	Primary procedure	6/48
53.5%	DRG	7/48
55.3%	Primary diagnosis	8/48

The inspiration for this approach to greedy feature selection comes from simulated annealing, a local search algorithm itself inspired by a metallurgical process. In local search algorithms, a space is searched from a starting point by considering only the adjacent points, choosing one according to the criteria of the algorithm, and then considering the points adjacent to that one; this process is repeated until no further progress can be made - generally, until a local or global maxima is reached. In simulated annealing, an adjacent point is chosen at random. If it is an improvement, according to the criteria being used, then the selection is accepted. If it is not an improvement, however, it is still accepted with some probability that decreases with the "badness" of the selection and with the duration of the search as a whole. Thus, the search proceeds but with an occasional jump that allows it to escape local maxima. In an analogous way, this approach to greedy feature selection allows attributes to be considered based on their individual classification accuracy, but allows an attribute that may be effective in combination with another but not alone to be considered also.

Tools

Functional Programming in Clojure

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 I chose to use a functional programming language, namely Clojure[5], a Lisp that runs on the JVM and permits access to any Java libraries. Functional programming - that is, composing functions that take arguments and return val-

[5] More information is available at http://clojure.org

ues, and minimize side-effects, into programs which are akin to functions themselves - allows for concise, linear programs that mimic their mathematical underpinnings.

For example, for greedy feature selection, the probability of selection illustrated in Table 1 is calculated in the following function:

```
(defn prob-attrs
  "takes map of attribute index to one-to-one classification accuracy
   returns map of attribute index to probability that attribute should be selected"
  [scores]
  (let [distinct-scores (into (sorted-set) (vals scores))
        rankings (zipmap distinct-scores (range 1.0 (inc (count distinct-scores))))
        denom (reduce + (map rankings (vals scores)))]
   (reduce (fn [r [k v]] (assoc r k (/ (rankings v) denom))) {} scores)))
```

Prob-attrs is the function defined here. *Scores* is the argument and is the result of a prior function. *Distinct-scores*, *rankings*, and *denom* are named, intermediate results. The last line creates a new data structure from these intermediate results, using Clojure's *reduce* and *assoc* functions for manipulating collections, and returns it.

Weka

To improve performance, I accessed the naïve Bayes model in Weka 3.7.7 by using Weka's Java API from Clojure, as well as Weka's methods for 10-fold cross-validation. Weka is the Waikato Environment for Knowledge Analysis from The University of Waikato, New Zealand[6], a robust and widely-used application for data mining accessible as either a stand-alone

[6] More information is available at http://www.cs.waikato.ac.nz/ml/weka/documentation.html, http://weka.sourceforge.net/doc.dev/, and http://weka.wikispaces.com

application or a Java library. This improved calculation speed dramatically over my beginner's implementation of the same model.

Following are two functions which use Weka's naïve Bayes model:

```
(defn test-instance
  "takes model and instance, returns weight if model classifies instance correctly, 0.0 otherwise"
  [^NaiveBayes bayes ^Instance instance]
  (if (== (.classValue instance)
        (.classifyInstance bayes instance))
    (.weight instance)
    0.0))

(defn score
  "takes model and all instances, returns classification accuracy reflecting weights"
  [bayes instances]
  (let [correct-weight (reduce + (map (partial test-instance bayes) instances))
       total-weight (reduce + (map #(.weight ^Instance %) instances))]
   (/ correct-weight (double total-weight))))
```

classValue, *classifyInstance*, and *weight* are Weka methods. *^NaiveBayes* and *^Instance* are type hints.

Feasibility Study

At the outset, I implemented a naïve Bayes classifier in Clojure and applied it to the 2010 NHDS. I 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 naïve 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 the 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, I 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, I discretized the data into the following six categories: one day, two days, one week, one month, two months, and long term (the range w as 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 naïve 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 naïve 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. Representative results are included in Table 3.

Figure 1: Prevalence of Lengths of Hospital Stay

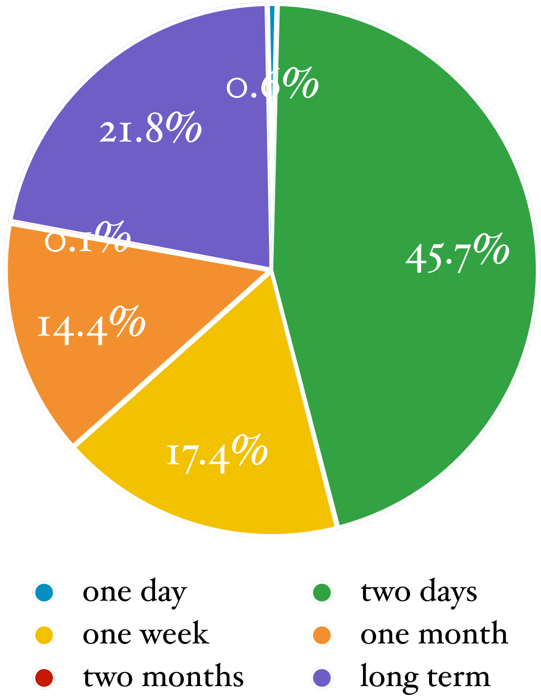


Table 3: Representative Results Targeting Length of Hospital Stay

Attributes Selected in 100 Iterations	Classification Accuracy
Primary diagnosis	61%
Primary diagnosis, primary procedure	67%

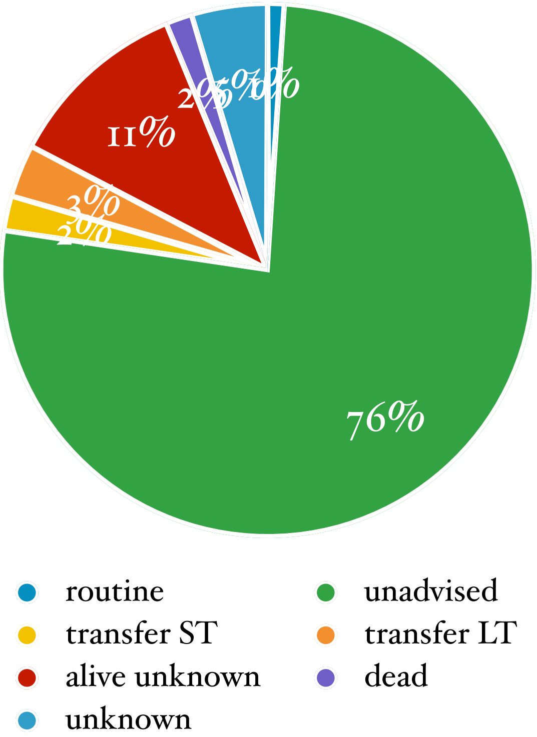
Attributes Selected in 100 Iterations	Classification Accuracy
Primary diagnosis, primary procedure, admitting diagnosis	69%
Primary diagnosis, primary procedure, admitting diagnosis, DRG	70%

I 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

I was optimistic that successfully predicting discharge status would have worthwhile implications, but the data was uncooperative. The 2010 NHDS 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 naïve 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.

Figure 2: Prevalence of Discharge Statuses



The same set of diagnosis code- and procedure code-related attributes were the better predictors. Representative results are included in Table 4.

Table 4: Representative Results Targeting Discharge Status

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%

Number of Diagnosis Codes

The 2010 NHDS 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 I 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 naïve 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. Representative results are included in Table 5.

Figure 3: Prevalence of Number of Diagnosis Codes

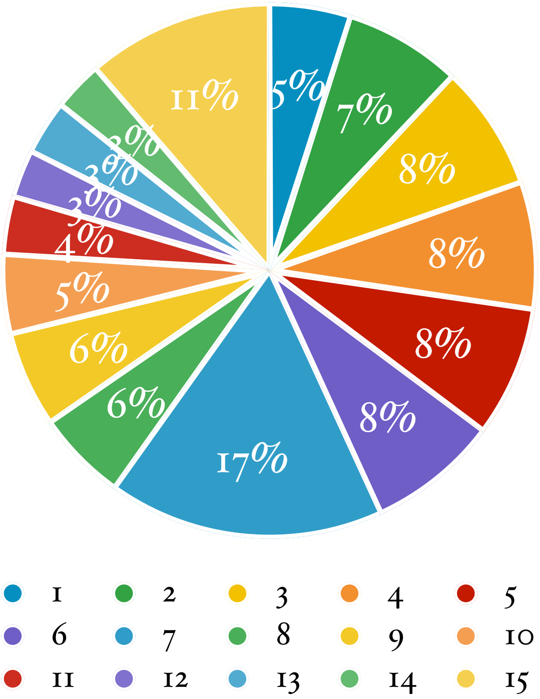


Table 5: Representative Results Targeting Number of Diagnostic Codes

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%

Concerns

The possibility that desired predictive relationships are not supported by the data must be considered. On the other hand, there is the possibility that predictive relationships inferred from the training data may be somehow unique to that data and have no applicability beyond it. Such over-fitting can result from spurious data, data that isn't fully representative, or luck; the latter is more likely when the number of attributes is high relative to the training data.

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 confidently assess 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 must hope that data is not uniformly distributed in input space, but sufficiently clumped to enable the calculation of useful classifiers.

Conclusions

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.

With respect to the second objective, 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 naïve 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 I think might be worthwhile.

Future Work

One aspect of the 2010 NHDS that I 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 using 3-digit ICD-9 codes, a la Stiglic (2011), or the hierarchy implicit in ICD-9 codes, a la Popescu and Khalilia (2011), might allay that concern.

The NHDS 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 data is 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

I am indebted to Dr. Smiljana Petrovic, Department of Computer Science, Iona College, for her tutelage, encouragement, advice, and patience over the past five years, and most particularly this last year.

I also thank Rich Hickey, for technical review.

BIBLIOGRAPHY

Abraham, R., Simha, J., & Iyengar, S. (2006). A comparative analysis of discretization methods for medical data-mining with Naïve Bayesian classifier. 9th International Conference on Information Technology (ICIT'06), pp. 235-236.

Abraham, R., Simha, J., & Iyengar, S. (2007). Medical datamining with a new algorithm for feature selection and Naïve Bayesian classifier. 10th International Conference on Information Technology (IEEE - ICIT 2007), pp. 44-49.

Abraham, R., Simha, J., & Iyengar, S. (2009). Effective discretization and hybrid feature selection using Naïve Bayesian classifier for medical datamining. International Journal of Computational Intelligence Research. ISSN 0974-1259 Vol.5, pp. 116-129, No.2.

Al-Aidaroos, K. M., Bakar, A. A. & Othman, Z. (2012). Medical data classification with Naive Bayes approach. Information Technology Journal, 11: 1166-1174.

Bellazzi, R., & Zupan, B. (2006). Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics, Volume 77, Issue 2 , Pages 81-97.

Cios, K., and Moore, G.W. (2002). Uniqueness of medical data mining. Artificial Intelligence in Medicine, Volume 26 Issue 1-2.

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.

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

Garrett, L. (1994). The Coming Plague: Newly emerging diseases in a world out of balance, New York: Farrar, Straus and Giroux.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts and Techniques, Third Edition, Waltham, MA: Morgan Kaufmann Publishers.

Hassan, S. Z. & Verma, B. (2007). A hybrid data mining approach for knowledge extraction and classification in medical databases. Seventh International Conference on Intelligent Systems Design and Applications.

Health care in the United States. (n.d.). In Wikipedia. Retrieved December 3, 2012 from http://en.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems

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, v.41 n.3, p.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):20-4.

McKenna, M. (2004). Beating Back the Devil: On the Front Lines with the Disease Detectives of the Epidemic Intelligence Service, New York: Free Press.

McKenna, M. (2010). Superbug: The Fatal Menace of MRSA, New York: Free Press.

O'Hagan, A., & Luce, BR. (2003). "A Primer on Bayesian Statistics in Health Economics and Outcomes Research," MEDTAP International, Incorporated.

Popescu, M. & Khalilia, M. (2011). Improving disease prediction using ICD-9 ontological features. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE).

Rish, I., Hellerstein, J., & Thathachar, J. (2001). An analysis of data characteristics that affect Naïve Bayes performance. Technical report, IBM T.J. Watson Research Center.

Rochester, E. (2013). Clojure Data Analysis Cookbook, Birmingham, UK: Packt Publishing Ltd.

Russell, S. & Norvig, P. (2010). Artificial Intelligence: A Modern Approach, Third Edition, Upper Saddle River, NJ: Prentice Hall.

Segaran, T., & Hammerbacher, J. (2009). Beautiful Data: The Stories Behind Elegant Data Solutions, Canada: O'Reilly Media, Inc..

Stiglic, G. (2011). Human disease network guided discovery of interesting itemsets in hospital discharge data. Proceedings of the 2011 workshop on Data mining for medicine and healthcare.

Tan, P., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining, Boston, MA: Addison-Wesley.

Vanderhart, L. & Sierra, S. (2010). Practical Clojure, New York: Springer-Verlag.

Witten, I.H., Frank, E., & Hall, M.A. (2011). Data Mining: Practical Tools and Techniques, Third Edition, Burlington, MA: Morgan Kaufmann Publishers.

Wolfe, N. (2011). The Viral Storm: The Dawn of a New Pandemic Age, New York: Times Books.

World Health Organization ranking of health systems. (2000). In Wikipedia. Retrieved December 8, 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

APPENDIX

The following statistics on the 2010 National Hospital Discharge Survey (NHDS) are provided by the U.S. Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS).

PATIENTS	VALUES	FREQUENCIES
All patients	Newborns	14,092
	0-1 year	2,041
	1 year & up	135,418
	Total	151,551

ATTRIBUTES	VALUES	FREQUENCIES
Age	Under 15	6,422
	15-44	38,160
	46-64	37,925
	65 & up	54,952
Sex	Male	56,295
	Female	81,164

Attributes	Values	Frequencies
Race	White	88,254
	Black	19,039
	Native-American	335
	Asian	1,691
	Pacific-Islander	81
	Other	6,654
	Multiple	94
	Unknown	21,311
Marital status	Married	26,024
	Single	17,702
	Widowed	9,020
	Divorced	4,936
	Separated	729
	Unknown	79,048

Attributes	Values	Frequencies
Month of discharge	January	11,736
	February	11,234
	March	12,388
	April	11,758
	May	12,027
	June	11,509
	July	11,840
	August	11,704
	September	11,501
	October	10,908
	November	10,185
	December	10,669
Discharge status	Routine	103,263
	Unadvised	1,544
	Transfer short-term	3,662
	Transfer long-term	15,651
	Alive otherwise unknown	6,949
	Dead	2,747
	Unknown	3,643
Length of hospital stay		Not Provided
Region	Northeast	26,282
	Midwest	40,784
	South	56,400
	West	13,993

Attributes	Values	Frequencies
Hospital size	Under 100 beds	22,752
	100-199	27,388
	200-299	21,498
	300-499	41,325
	500 beds & more	24,496
Hospital ownership	Proprietary	19,006
	Government	18,458
	Non-profit	99,995
Payment method	Workers Comp	456
	Medicare	60,672
	Medicaid	22,175
	Other government	2,322
	BCBS	11,280
	HMO or PPO	18,124
	Other private insurance	11,843
	Self pay	6,814
	No charge	401
	Other	1,266
	Unknown	2,106
Type of admission	Emergency	73,267
	Urgent	23,645
	Elective	34,700
	Unknown	5,847

ATTRIBUTES	VALUES	FREQUENCIES
Source of admission	Non-health care POA	62,565
	Clinic	9,143
	Hospital transfer	5,905
	SNF transfer	2,588
	Other transfer	1,127
	ER	48,050
	Court/law enforcement	302
	ASC transfer	147
	Hospice transfer	52
	Other	918
	Unknown	6,662