DATA MINING AND STATISTICS IN MEDICINE: AN APPLICATION IN PROSTATE CANCER DETECTION

Vida S. Tigrani, UCSF School of Medicine S245, San Francisco, CA 94143-0454
George H. John, Epiphany
vtigran@itsa.ucsf.edu gjohn@epiphany.com

Abstract:

Data mining is an umbrella term referring to the process of discovering patterns in data, typically with the aid of powerful algorithms to automate part of the search. These methods come from disciplines such as statistics, machine learning (artificial intelligence), pattern recognition, neural networks, and databases. Two data analysts with different heritages might approach a similar problem quite differently. In particular, this paper shows how the same problem, prostate cancer detection, is approached by an M.D. and a data mining analyst with a background in machine learning: hypothesis testing by the former, and bagged classification models by the latter. We then survey medical data analysis literature, describing the common statistical methods employed by physicians for clinical studies, and the advances in data mining and machine learning research that have been motivated by medical data analysis.

1. Introduction

Prostate cancer is the most common cancer among American men, affecting roughly one in five during the course of their lifetime. Over 330,000 new cases were diagnosed in 1997 (Reuters). In a sense, cancer is easy to detect — just kill the patient, perform a thorough autopsy, and you are likely to discover any existing cancer. Of course, this detection method defeats the purpose, but it serves to illustrate the fact that cancer detection is a spectrum from maximally invasive, expensive, and accurate methods to noninvasive, relatively inexpensive, and possibly less accurate methods. For prostate cancer, excision of the prostate and examination by a pathologist is the gold standard, with needle biopsy a close second, and blood tests of PSA level are the cheaper but less accurate alternative.

Prostate specific antigen (PSA) is produced exclusively in the body by prostate cells and is detected in low concentrations in the blood of normal healthy men. The concentration of PSA in the blood may be elevated above the normal range in patients who have cancer of the prostate. Today, the physician's guidelines recommend a biopsy for patients with PSA above a certain threshold. This method is known to have a high false positive rate — men without cancer can have elevated PSA levels — leading to wasted resources and significant anxiety and discomfort on the part of the patient.

We analyze different and more complex tests in this paper, first hypothesizing possibly superior tests and using statistical hypothesis tests to judge their merit, and then using data mining methods to explore a wide space of possible tests.

2. The Clinical Urology Data

We obtained a database of 682 urology patients. Table 1 describes the information we had on every patient. Race was included because previous work has shown that race may be an important factor in diagnosing prostate cancer. As mentioned above, PSA is known to be a useful predictor, but it is not per-

Table 1: Attributes in the prostate cancer database.

Attribute	Description
`Age`	Age of patient in years.
`Race`	Classification of patient into one of five groups: African-American, Asian, Filipino, Hispanic, or white.
`PSA`	Prostate Specific Antigen concentration in the blood.
`Size`	Prostate volume, as estimated from an ultrasound image.
`PSA Density`	PSA level divided by prostate volume.
`Abnormal DRE`	Whether digital rectal examination results were classified as abnormal, *i.e.*, a hard protuberance was found.
`Cancer`	Indicates whether the patient had cancer (based on biopsies).

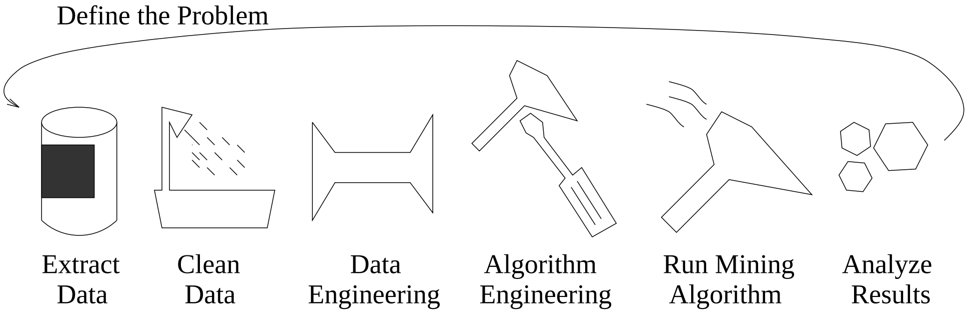

Figure 1: The data mining process.

fect. The test has a relatively low positive predictive value, meaning that when it predicts positive (cancer), it is often a false positive (see Table 3).

An elevated PSA is also seen in men who have other prostate diseases, most notably benign prostatic hyperplasia (BPH), an enlargement of the prostate gland which does not lead to cancer and is not life-threatening if treated. While BPH may in certain cases be the cause of serious illness to a patient, its treatment differs from that of prostate cancer. Prostate cancer may be treated by an open surgical removal of the prostate gland, possibly followed by radiation therapy or androgen ablation, but the treatment of BPH may be much less invasive.

Therefore, when confronted with an elevated PSA concentration in an otherwise healthy patient, the physician must somehow distinguish cancer from other causes. The definitive method to confirm a diagnosis of cancer is to examine a biopsy of the prostate tissue. Such a biopsy is taken using a spring-loaded needle through the rectum, under the guidance of a rectal ultrasound probe. This procedure incurs a financial cost and is quite uncomfortable for the patient. However, as we attempt to reduce the number of unnecessary biopsies, we also want to minimize the number of missed cancer diagnoses.

Attempts have been made to find a PSA-related measurement which would serve this purpose. PSA density, for example, is the concentration of PSA found in blood divided by the volume of the prostate gland. Because BPH causes a much greater enlargement of the prostate than does cancer, some investigators hypothesize that patients with a high PSA density are more likely to have cancer (Presti, Hovey, Carroll & Shinohara 1996).

Urologists also consider what they find on a digital rectal exam (DRE). The DRE is used to detect any hard protuberance on the part of the prostate that is adjacent to the rectum. Such a finding may be indicative of cancer of the prostate. Others have suggested that an elevated PSA in patients of different races may represent different risks of cancer. For example, African American men have a greater incidence of prostate cancer than white men with an earlier age of onset and higher mortality. We might hypothesize that physicians should be more aggressive in confirming a diagnosis of cancer in African Americans than in men of other races.

3. Analyzing the Data

The problem we investigate in this paper is how to support a physician's decision of whether a biopsy is warranted. Perhaps by analyzing existing or easily-measured data about a patient we can develop some means by which a physician caring for a patient can better decide when to biopsy. Figure 1 shows the general data mining process of defining the problem, collecting (or extracting) data, cleaning the data, data engineering, algorithm engineering, running the mining algorithm, and analyzing the results.

3.1 Data Cleaning and Exploration

Although the next two sections have the job of describing the statistical and data mining approaches to analyzing the data, the first step in each process is to check the quality of the data.

"Patient, English" – this was the name of patient 182 in a related database we looked at before the one analyzed in this paper. At some point, a secretary or medical student had a little joke to themselves and entered this patient's information in the database. It was unclear whether the data was real and the name was changed, or whether the data had been fabricated along with the name. The same database also contained roughly ten patients nearly 2000 years old – their year of birth had mistakenly been entered instead of their age.

Many patients in the database have ethnicity codes of `A`, `aa`, and `o`. Looking at surnames, it

Table 2: Comparison of PSA and PSA Density for cancer detection in all patients.

		PSA			PSA Density			Volume	
Race	Cancer	Median	25%ile	75%ile	Median	25%ile	75%ile	Median	Std Dev
White	Pos	8.6	5.4	15.3	0.21	0.13	0.46	37	20.0
White	Neg	5.8	3.6	8.2	0.11	0.07	0.16	46	33.3
Black	Pos	8.2	5.9	19.0	0.25	0.15	0.48	38	21.5
Black	Neg	6.2	4.4	10.1	0.12	0.07	0.18	48	37.7

Table 3: Common metrics for clinical tests.

Positive Predictive Value (PPV) Probability of positive outcome given positive prediction

Negative Predictive Value (NPV) Probability of negative outcome given negative prediction

Sensitivity Probability of positive prediction given positive outcome

True Positive Rate (same as above)

Specificity Probability of negative prediction given negative outcome

False Positive Rate Probability of positive prediction given negative outcome

seemed that `A` sometimes stood for *Asian*, sometimes *African-American*, and `aa` was similarly confused. Luckily, the hospital kept patient records online, and we manually scanned their patient notes to recover the correct ethnicity.

A few patients had a PSA value of 1000 ng/ml in the database. The normal range is less than 4 ng/ml. 1000 might have represented some special value, but it turned out to literally represent a thousand as the true "off the charts" value of PSA for those few patients.

A data analysis project in medicine is typically carried out in three steps: recording the raw data, compiling the raw data into a database, and analyzing the database. Recording of raw data is done for example by pathologists analyzing tissue samples, nurses and physicians recording patient histories, vital signs, and test results, and so forth. Often this data is recorded in ill- or semi-structured formats not amenable to analysis. A medical student or junior resident might be assigned the task of sifting through hundreds of patient files and filling out a standard form for each patient. A secretary might then take all of the forms and type the values into a spreadsheet. Only at this point do we have a database ready for analysis by the standard statistical packages.

Collecting data from physicians' handwritten notes is quite time-consuming. As a result, databases like the one used in this paper contain less information than is actually available. For example, the blood pressure of the patients in this study was available on their patient charts, but was not collected because it is not hypothesized that blood pressure is related to the presence of cancer.

4. The Statistical Approach

The data cleaning step just described is common in both the data mining and the statistical approaches. From that point on, they diverge. In the statistical analysis, we repeated the calculations of Presti et al. (1997) in the larger data set. There is no significant data engineering or algorithm engineering step, and for the most part, the step of running an algorithm on the data involves calculating statistics like the mean or median of a population, with confidence intervals. The emphasis is on running many statistics and in the analysis of the results.

The purpose of the study by Presti et al. (1997) was to determine whether racial variations in prostate cancer detection exist. In particular, they wanted to understand whether PSA and PSA density represented a different likelihood of cancer in African Americans and in whites. If this were the case, it might lead to different strategies in prostate cancer screening among men of different races. They evaluated patients who were referred to a urology clinic by another physician because the patient had either an elevated PSA level or an abnormal finding on digital rectal exam (DRE). Data on race was recorded, as were serum PSA. No patient was previously diagnosed with prostate cancer, and all patients underwent DRE, transrectal ultrasound with volume determination of the prostate, and prostate biopsies. Presti et al. (1997) studied 378 patients. Since that time, the database has expanded to include 682 patients. We replicated much of their anal-

Table 4: Properties of three tests for cancer detection.

	Sensitivity	Specificity	PPV	NPV
White men:				
PSA>4	87% (82-92)	28% (22-34)	47% (42-53)	64% (75-84)
psad>0.1	82% (76-88)	46% (39-53)	53% (47-59)	78% (71-85)
psad>0.15	67% (60-75)	71% (65-77)	63% (56-70)	75% (69-81)
Af.AM. men:				
PSA>4	99% (97-100)	21% (13-30)	55% (47-62)	95% (85-100)
psad>0.1	93% (87-98)	43% (32-53)	60% (52-69)	86% (76-97)
psad>0.15	74% (64-83)	62% (52-72)	65% (55-74)	71% (61-82)

ysis on this augmented data set.

We evaluated 682 patients who were referred for an elevated serum PSA or abnormal findings on digital rectal exam (DRE). Of the 682 patients, 174 were black and 380 were white. A normal DRE was reported in 280 patients. The mean age of the black patients was 67.6 years (SD=7.6 years), and the mean age of white patients was 68.8 years (SD=7.5 years). The median PSA, PSA density, and mean prostate size did not differ significantly between the 2 ethnic groups. (6.7 versus 7.1, 0.13 versus 0.17, and 50.1 versus 51.9 respectively. For the entire group, cancer was detected in 161 of 380 (42%) whites and 85 of 174 (49%) blacks. (chi-sqare p=0.15) Table 2 shows the median PSA, PSA density and prostate volume in positive and negative biopsy groups for white and black men. The sensitivity, specificity, and predictive values with 95% confidence intervals for cancer detection are shown in Table 3.

Physicians are often presented with new diagnostic tests, such as blood tests and radiology studies, and they must decide which of these provides new and better information than what they already use. For example, the PSA test has been available since 1987 for the detection of diseases of the prostate. Calculating PSA density requires an ultrasound image of the prostate to estimate volume. Is this additional procedure provide the clinician with better information about the patient's likelihood of having cancer?

Sensitivity, specificity, and predictive values are characteristics of a diagnostic test that physicians commonly use to evaluate its worth. For a certain cutoff value, there will be determined a number of patients who are correctly and incorrectly identified as having or not having a disease. Sensitivity is the probability that a person with a disease, in this case prostate cancer, will be correctly identified by the diagnostic test. It is defined as the number of true positives divided by the sum of true positives and false negatives. Specificity is the probability that a person without a disease is correctly identified. It is the number of true negatives divided by the sum of the true negatives and false positives. Positive predictive value is the probability that a person who has a positive test result actually has the disease. It is the number of true positives divided by the sum of true positives and false positives. Negative predictive value is the probability that a person with a negative test result does not have the disease. It is the number of true negatives divided by the sum of true negatives and false negatives.

These parameters with their 95% confidence intervals for cancer detection are shown in Table 4 when a positive test result is defined as either PSA > 4, PSA density > 0.1, and PSA density > 0.15. For example, when a PSA > 4 is used to define white patients at high risk for prostate cancer, we see that the sensitivity is relatively high, implying that this test will detect most patients who have prostate cancer. However, at this same cutoff value, the specificity is relatively low, indicating that many patients will unnecessarily undergo biopsies when they do not have cancer. In comparing predictive values, we can see that an African American with a PSA > 4 is more likely to actually have cancer than a white man.

Another standard statistical method to compare the groups of patients with and without cancer is to use the Student's t test. For example, patients without cancer had a mean PSA of 7.4 while those with had a mean PSA of 37.4, and this difference was highly significant (p<0.001). Likewise for PSA density, the mean for those without cancer was 0.145 while those with cancer was 0.853, and was highly significant (p<0.001) by the Student's t test. Thus race appears to be a significant factor. A "one size fits all" rule of thumb for deciding when to biopsy might well be replaced with a rule that takes race into account.

Patient #	Score	Outcome
...		
421	.9428	Cancer
013	.9190	Cancer
387	.8874	Benign
552	.8144	Cancer
...		

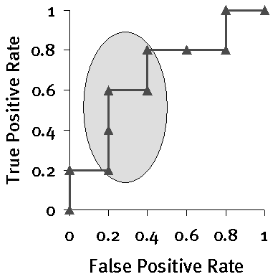

Figure 2: Example of building an ROC curve from scored data where the true outcome is known.

5. The Data Mining Approach

Returning to the data cleaning step where the data mining process and the statistical analysis process were the same, the next step for data mining is data engineering. For example, when analyzing the customers of a bank, a data miner might construct an "account balance acceleration" variable that is defined as some function of the change in balance over time. In our database, there are relatively few variables, and thus few opportunities to engineer new variables based on the existing ones. Interestingly, a concept related to balance acceleration is "PSA Velocity" – measuring the change in PSA in a single patient over time, the idea being that a marked increase indicates trouble, just as a marked decrease in a customers savings account indicates the customer might be moving assets into another account prior to closing their savings account. The use of PSA Density is one example of a constructed variable – the division of PSA by prostate volume. Another type of data engineering is choosing subsets of variables to use in the model. We will consider three options: using all variables, just PSA, and just PSA Density.

The next step is algorithm engineering. We will skip that step on the first pass through the process and see how "off the shelf" algorithms perform on the data, which leads us to running some mining algorihtms. The data we have is a standard classification problem: given a vector $\mathbf{x}$ predict y, where $\mathbf{x}$ is the vector describing a patient's PSA etc, and y is a 0-1 variable indicating whether the patient has cancer. We wish to build a model $\widehat{f} : \mathcal{X} \mapsto [0, 1]$, a function that maps from the observed variables to a number we may interpret as the probability of a patient having cancer (a "positive" result). Often in data mining this is called *scoring*.

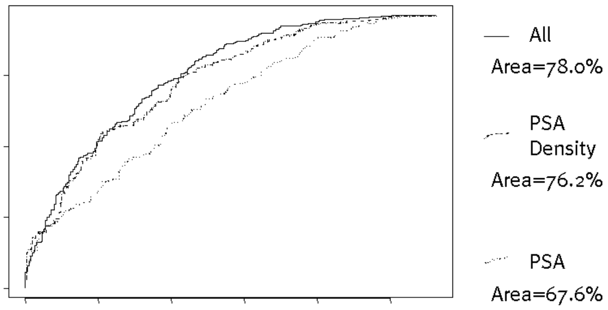

Figure 3: ROC curves for logistic regression (resubstitution).

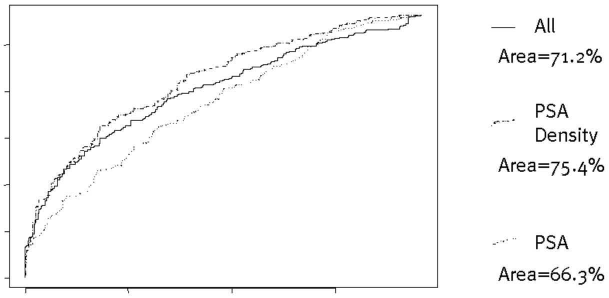

Figure 4: ROC curves for logistic regression (cross-validation).

To evaluate the quality of a model $\widehat{f}$ we will look at its ROC curve. If the ROC curve of $\widehat{f}$ passes through a point (x, y), this means that there exists some threshold δ such that if we classify all cases whose $\widehat{f}$ value is above δ as 1 and the rest as 0, then we expect to attain a true positive rate of y and a false positive rate of x. A perfect model's ROC curve passes through the point (0,1), meaning there exists a threshold value with a true positive rate of 1 (100%) and a false positive rate of 0. A completely random model's ROC curve is a straight line from (0,0) to (1,1). The area under the ROC curve is a statistic we will use to compare models – the higher the better. Figure 2 shows how a fragment of an ROC curve is constructed: we take a list of cases where the true outcome is known, sort the list in descending order by the score the model predicts for each case, and then draw the ROC curve with a pen, moving up each time we encounter a positive example and to the right each time we encounter a negative.

Figure 3 shows the ROC curve for three logistic regression models. It is a resubstitution estimate, meaning that the same data that was used to build the logistic regression model is also being used to estimate the ROC curve. The curves therefore ought

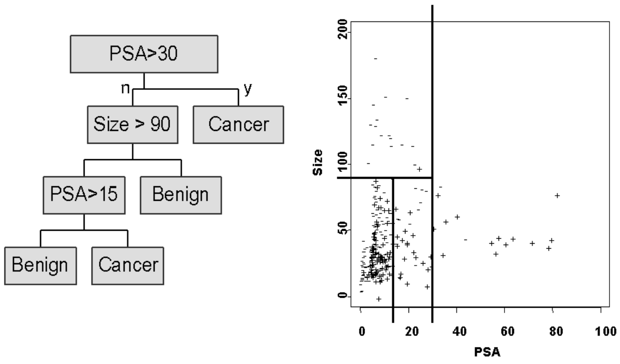

Figure 5: Classification tree example using PSA and Size.

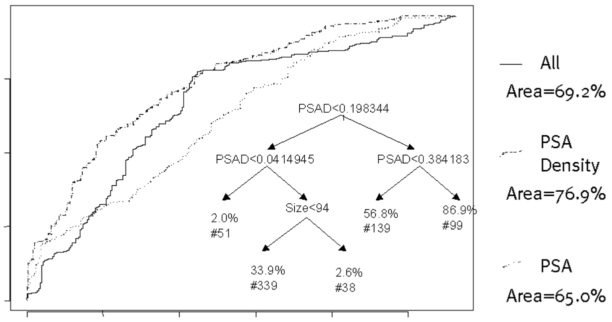

Figure 6: Pruned classification tree ROC curve (resubstitution).

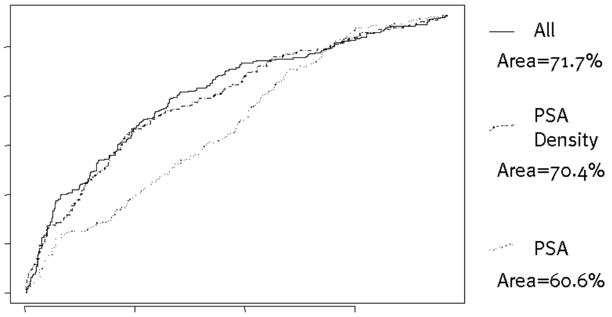

Figure 7: Pruned classification tree ROC curve (Cross Validation).

to be slightly optimistic. Figure 4 shows ROC curves built using cross-validation (Breiman, Friedman, Olshen & Stone 1984). These curves are unbiased. Note that the model built using all variables underperforms the model built using only PSA Density in the cross-validation curve but not the resubstition curve. This is expected – the more degrees of freedom in the model, the better it can fit the training (or in-sample) data, but the worse it might perform on test (out-of-sample) data. In both cases a model built using only PSA Density seems preferable to a model using only PSA.

A common scoring method used in data mining is the classification tree. Figure 5 shows an example classification tree. When a vector $\mathbf{x}$ is presented to the tree, it repeatedly tests the value of a component of $\mathbf{x}$ and then branches. For example, if a patient has PSA of 31 and prostate volume of 100ml, then the tree would classify them as benign. The graph on the right in Figure 5 shows another representation of the tree as axis-parallel cuts in the space of patients. The $(31, 100)$ patient would be in the upper left rectangle, where historically all patients have been free of cancer.

The example tree was hand-constructed, but we will use the CART algorithm (Breiman et al. 1984) to build trees that fit the training data well. Figure 6 shows the ROC curve for CART trees (pruned using cross-validation, and using only complete records in the training set, those with no missing values), again using all variables, PSA Density only, or PSA only. The graph shows the tree built using all variables. For example, the rightmost path in the tree says that if PSA Density is greater than .384183, then the patient is 86.9% likely to have cancer. Of the 628 patients in the database with no missing values, 99 of them fit this criterion (*i.e.*, had PSA Density $> .384$).

The trees we are looking at have been *pruned*, meaning that the algorithm has made an effort to ensure that the trees do not overfit the training data. Still, we expect to see worse results when we look at cross-validation ROC curves instead of resubstitution curves. Figure 7 shows the unbiased ROC curves. Note that the logistic regression model was actually doing better than the tree model – for the PSA-Density only model, it had an area of 75.4% whereas the best tree has an area of 71.7

This called for some algorithm engineering. We used *bagging* (Breiman 1996) in an attempt to get better results. Instead of building a single model $\widehat{f}(\mathbf{x})$, we build many models $\widehat{f}_i$ and combine them into a single model $\widehat{f}$ at the end where $\widehat{f}(\mathbf{x}) = \sum_i \widehat{f}_i(\mathbf{x})$. Each f_i is built from a slight different training set obtained by sampling with replacement N records from the original training set of N records. This procedure is designed to reduce the variance in the resulting model and there by reduce its error.

Figure 8 compares a bagged tree with the best logistic regression (using only PSA Density). Trees are generally ill-suited for problems where the output value varies smoothly with a continuous input value, but that is what seems to be happening in

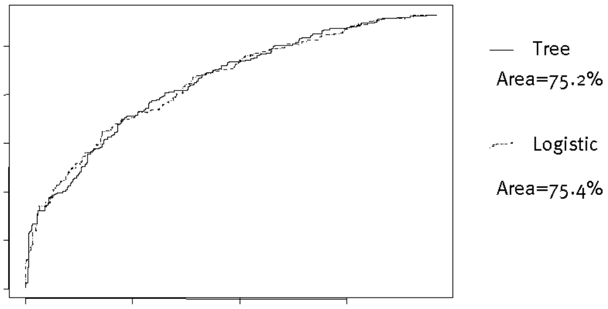

Figure 8: Bagged classification tree and logistic regression ROC (Cross Validation).

this database with PSA density. Bagging enables the tree to catch up with logistic regression, but not surpass it.

6. Literature Survey

A number of articles published in the medical literature report studies using CART and neural networks. Niederberger (1995) surveys recent work using neural networks in Urology, with many positive results. Snow, Smith & Catalona (1994) illustrates an important detail in evaluating a neural net – if the test set used for evaluation is used in constructing the model (as they used it to halt the training of the neural net), then the results will be optimistically biased. Many other papers examine data mining in urology (Decaestecker et al. 1996, Kattan, Cowen & Miles 1996, Montironi et al. 1995). Other papers in medicine examine a range of tasks including diagnosis (Barriga, Hamman, Hoag, Marshall & Shetterly 1996, Garbe et al. 1995, Selker et al. 1995, Soyer et al. 1995), mortality (Carmelli, Swan, Dame, McElroy, Gelb & Rosenman 1991), other postoperative risks (Kastrati et al. 1997). In an unusual medical application, Held et al. (1994) used CART to model travelers likely to not fully comply with malaria prophylaxis. Noncompliance can increase resistance to chloroquine and other malaria treatments, so ensuring compliance is important.

Moul et al. (1995) used neural networks to improve clinical staging ot nonseminomatous testicular cancer. They found that an experienced neural network researcher using custom software outperformed a less experienced researcher using commercial software in predicting which patients had occult disease beyond the testis (92

Machine learning and data mining researchers have often used medical domains for exploration or inspiration. Caruana, Baluja & Mitchell (1996) propose a novel type of neural net and shows its effectiveness is assigning standard or radical treatment to pneumonia patients. Pazzani, Mani & Shankle (1997) propose a novel method of building trees that are constrained so that their tests are consistent with physicians' knowledge and experience. John & Langley (1995) develop a method similar to logistic regression, and discover that its accuracy relative to a classification tree is comparatively high on medical databases.

7. Discussion and Future Work

The paper demonstrates the different styles of analysis common in medical statistics and data mining; the former commonly employs estimations of population means for various populations and hypothesis tests, while the latter tends to focus on building a model and analyzing its quality in comparison to other models.

In the comparison we found that a very simple statistical modeling algorithm performed as well as a sophisticated data mining algorithm. The mining algorithm's capacity to model a complex function may not have been exercised by the data, which seems to effectively have only a single very relevant observed variable. In commercial data mining, it often happens that a mining algorithm will discover relationships where none were thought to exist. In this database, few variables were collected on each patient (the collection process is manual and somewhat arduous), and they were variables that were strongly believed to be relevant. Future investigations of prostate cancer should collect a broader range of patient data.

Acknowledgements

We thank Dr. Joe Presti at UCSF for providing the data used in this paper.

References

Barriga, K. J., Hamman, R. F., Hoag, S., Marshall, J. A. & Shetterly, S. M. (1996), "Population screening for glucose intolerant subjects using decision tree analysis", *Diabetes Research and Clinical Practice* **34**, pp. S17–S29.

Breiman, L. (1996), "Bagging predictors", *Machine Learning* **24**(2), pp. 123–140.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Chapman & Hall, New York.

Carmelli, D., Swan, G. E., Dame, A., McElroy, M., Gelb, A. B. & Rosenman, R. H. (1991), “27-year mortality in the western collaborative group study: construction of risk groups by recursive partitioning”, *Journal of Clinical Epidemiology*, pp. 1341–1351.

Caruana, R., Baluja, S. & Mitchell, T. (1996), Using the future to "sort out" the present: Rankprop and multitask learning for medical risk evaluation, in *Advances in Neural Information Processing Systems 8*.

Decaestecker, C. et al. (1996), “Construction and application of hierarchical decision tree classification of ultrasonic prostate images”, *Medical and Biological Engineering and Computing* **34**, March, pp. 105–109.

Garbe, C. et al. (1995), “Primary cutaneous melanoma”, *Cancer* **75**, pp. 2484–2491.

Held, T. K. et al. (1994), “Malaria prophylaxis: identifying risk grops for non-compliance”, *Quarterly Journal of Medicine* **87**, pp. 17–22.

John, G. H. & Langley, P. (1995), Estimating continuous distributions in Bayesian classifiers, in P. Besnard & S. Hanks, eds, *Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco, pp. 338–345.

Kastrati, A. et al. (1997), “Predictive factors of restenosis after coronary stent placement”, *Journal of the American College of Cardiology* **30**, pp. 1428–1436.

Kattan, M. W., Cowen, M. E. & Miles, B. J. (1996), “Computer modeling in urology”, *Urology* **47**(1), pp. 14–21.

Montironi, R. et al. (1995), “Prostatic intraepithelial neoplasia (pin). performance of bayesian belief network for diagnosis and grading”, *Journal of Pathology* **177**, pp. 153–162.

Moul, J. W. et al. (1995), “Neural network analysis of quantitative histological factors to predict pathological stage in clinical stage i nonseminomatous testicular cancer”, *Journal of Urology* **153**, May, pp. 1674–1677.

Niederberger, C. S. (1995), “Commentary on the use of neural networks in clinical urology”, *Journal of Urology* **153**, May, p. 1362.

Pazzani, M., Mani, S. & Shankle, W. R. (1997), Comprehensible knowledge discovery in databases, in M. G. Shafto & P. Langley, eds, *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum, pp. 596–601.

Presti, J. C. et al. (1997), “Prospective evaluation of prostate specific antigen and prostate specific antigen density in the detection of carcinoma of the prostate: ethnic variations”, *Journal of Urology* **157**, March, pp. 907–912.

Presti, J. C., Hovey, R., Carroll, P. R. & Shinohara, K. (1996), “Prospective evaluation of prostate specific antigen and prostate specific antigen density in the detection of nonpalpable and state "t1c" carcinoma of the prostate”, *Journal of Urology* **156**, November, pp. 1685–1690.

Selker, H. P. et al. (1995), “A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ichemia among emergency department patients”, *Journal of Investigative Medicine* **43**(5), October, pp. 468–476.

Snow, P. B., Smith, D. S. & Catalona, W. J. (1994), “Artificial neural networks in the diagnois and prognosis of prostate cancer: a pilot study”, *Journal of Urology* **152**, November, pp. 1923–1926.

Soyer, H. P. et al. (1995), “Diagnostic reliability of dermoscopic criteria for detecting malignant melanoma”, *Dermatology* **190**, pp. 25–30.