Chest Pain in the Emergency Department: Use of Asymmetric Penalties in Sequential Minimal Optimization with Feature Selection to Improve Clinical Decision Making Accuracy

Acknowledgements: I would like to extend my gratitude to Dr. Judd Hollander for providing the data set compiled at the Hospital of the University of Pennsylvania.

Abstract

Background: Disposition and assessment of patients presenting with acute chest pain is one of the most difficult decisions made in the Emergency Department. Computational tools have been employed with limited and variable success in this decision space.
Objective: Apply an Asymmetric Cost Support Vector Machine to a large data set of patients seen in an urban academic Emergency Department to assess the separability of the data and develop an accurate classifier over the data set.
Methods: An Asymmetric Cost Support Vector Machine and Principal Components Analysis were implemented in Matlab with appropriate variations on John Platt's Sequential Minimal Optimization algorithm. Non-compound linear, radial basis, and polynomial kernels were examined. Using Hold Out Cross Validation, optimal parameters for a radial basis and polynomial kernels, the optimal non-compound kernel, the optimal penalty values for false positives/negatives, and optimal dimensional reduction "factor" were assessed.
Results: The data is highly non-separable, but by selecting optimal parameters, we achieved a sensitivity of 94.5% and a specificity of 43.5%. The sensitivity and specificity that would have resulted from selecting based upon best error alone were 87.5% and 76.7%. Test set and Training Set error rates were comparable and compare favorably to previous classification methods applied in the same decision space.
Conclusions: Support Vector Machines show promise in providing an accurate classifier with good performance for patients presenting with chest pain to an urban academic Emergency Department. Even robust methods have difficulty accurately classifying the data. Asymmetric penalties mitigate the problems encountered with imperfect separation. Further optimizations are possible both to the asymmetric cost support vector machine itself and to accurately assess the impact of feature selection.

Introduction

Background: Chest pain represents one of the most common presenting complaints of patients to Emergency Departments (EDs) nationally[1-3]. The range of diseases that map to the complaint of chest pain vary widely in both severity and organ system involved (from psychosomatic, to "heart burn" on up to potentially lethal cardiovascular pathology) complicating the clinical assessment[4, 5]. National data suggests that only 11% of patients who present to Emergency Departments with the complaint of chest pain are subsequently found to have acute cardiac ischemia or another diagnosis requiring admission[2]. Furthermore, of admitted patients, only 15%-20% are thought to have chest pain related to acute ischemic cardiovascular pathology[2, 4-9]. This incurs tremendous costs to the system without identifiable benefit[10]. In 1997, over three million patients were admitted to US hospitals with chest pain and the costs to the system for those not found to have an ischemic etiology for the pain is well over $3 billion dollars by the most conservative estimates[11]. However, injudicious discharge of these patients home can result in major patient morbidity and mortality[9, 12]. In fact, untreated myocardial infarctions have at least a 25% 6 month mortality[12]. Several studies estimate the rate of discharge of chest pain with

ischemic heart disease to be roughly 4 to 4.4%[13, 14]. Different computational techniques have been used to differentiate patients as far back as 1982[15]. An algorithms attributable to Goldman et al. suggest that no one (not even the healthy male in his mid 20's) is "safe" for discharge home based upon an ad hoc threshold for posterior probability of disease (usually <1%)[15]. The "1%" rubric, in and of itself, has not been established with any sort of methodological rigor beyond expert opnion[5, 16].

Despite tremendous clinical advances, the rate of missed myocardial ischemia has remained around 4% since 1996 [15, 16]. Neural networks, Bayesian methods, and computer designed decision rules have had variable efficacy in improving on the sensitivity of experienced clinicians[16-25]. Generally these studies establish improvements in specificity to levels anywhere from ~30% on up to ~88%, but often at considerable cost in sensitivity down to 80% - 88%. There is also the challenge of practitioner acceptance. Hollander et al found that out of 432 patients enrolled in a study, feedback to physicians with a neural network affected decisions in only two cases[16].

Expert clinical opinion has major limitations as well. Ting et al found that each year of post-graduate clinical experience resulted in a 1.4 increased odds of admitting a patient with suspected ischemic chest pain without an increase in the frequency of detecting legitimate cases resulting in a marked increase in the rate of unnecessary admission[26]. Dreiseitl et al analyzed four standard statistical computing techniques to identify which features of a data set were most predictive of ischemic causes for chest pain where $|X| = 43$. There was significant inter-method variability in which features were selected, but overall eleven features were selected as important by most algorithms. Interestingly, a consulting cardiologist identified only three of the eleven selected features. Surprisingly, five of the nine features identified by the cardiologist as important were not selected by ANY of the learning algorithms[23].

Objective: Application of a support vector machine (SVM) to the classification of patients presenting to the ED with a complaint potentially referable to ischemic heart disease. SVMs provide a simple and intuitive method by which to differentially handle false positive and false negative classifications as the consequences of each are not symmetric: the "costs" of a missed myocardial infarction are clearly not comparable to the costs of an unnecessary admission both in monetary and health measures("Quality Adjusted Life Years": QALYs)[27, 28]. A review of the published literature suggests that there is no published work on the use of SVMs in this clinical space[29]. Feature selection with PCA was examined. Dimensional reduction has several potential advantages: i) reduced computational complexity, ii) lower dimensional data can be more effectively compiled in a clinical scenario, iii) mapping data that is not separable in a higher feature space to a lower feature space where a chosen kernel can more accurately separate the data. The goal is to draw a distinction between those that require admission and those that do not even if the final diagnosis is not attributable to a cardiac etiology. Clinical evidence suggests that in an expansive enough data set, perfect classification is not possible. Success should be measured by a reduction in false positives without untoward effects on sensitivity thereby providing a positive economic and system benefit over existing clinical methods.

Methods

A data set from the Hospital of the University of Pennsylvania in the Department of Emergency Medicine was modified for this study.

Data Labeling: Data was labeled based upon a final WHO diagnosis. The number of possible diagnoses considered was simply collapsed into a binary classification problem of:

i) patients with pathology that required admission, ii) patients who did not require admission. Patients who "required admission" were ones who either had a final diagnosis of "Acute MI", "USA", "Aortic Dissection"(2), "Pulmonary Embolism"(10), or experienced a major complication within 30 days of being seen in the ED.

Pre-Processing and Incomplete Data: Incomplete data was replaced with maximum likelihood estimates of data among examples that share a final label. Data categories such as blood pressure and heart rate demonstrate that demonstrate non-monotonic behavior were divided into n variables with binary labels. Specifically, a very high/low blood pressure or heart rate denote abnormalities of different type and even degree.

Scaling of Data: To avoid domination of other parameters by one parameter in the computation of the inner products, some feature values had to be "scaled". Several authors have commented on this, but this is not a formal requirement of SVM application[30]. In the chest pain data set, this is potentially an issue with a generally very sparse matrix with mostly binary indicator variables and a several data points with routine values in excess of 5000. Consequently, all data values were linearly scaled on the interval [0, 1] by dividing by the maximum value for each parameter.

Asymmetric Cost Regularization: C_2 is the penalty incurred when a positive class (a data example justifying admission) is labeled for discharge (a "false negative") while C_1 is the penalty incurred when a negative class (a data example justifying discharge) is labeled for admission (a "false positive"). The form of the dual is then:

$$\max_{\alpha} \; W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \left(x^{(i)}\right)^T x^{(j)}$$

$$s.t.: \quad \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$0 \le \alpha_i \le C_1 \;\; for \;\; y^{(i)} = +1$$

$$0 \le \alpha_i \le C_2 \;\; for \;\; y^{(i)} = -1$$

Choosing a Kernel: The data is not completely separable with a linear kernel for any value of C_1 or C_2. Non-linear feature spaces were considered including radial basis and Polynomial Kernels. Two methods were used to select and construct kernels: Kernels were selected to minimize "alignment" to each other while maximizing "alignment" of the component kernels to the data. The definition of alignment used is attributable to Cristianini et al[31]:

$$A\left(K_i, YY^T\right) = \frac{\left\langle K_i, YY^T \right\rangle}{\sqrt{\left\langle K_i, K_i \right\rangle \left\langle YY^T, YY^T \right\rangle}}$$

$$A\left(K_i, K_j\right) = \frac{\left\langle K_i, K_i \right\rangle}{\sqrt{\left\langle K_i, K_i \right\rangle \left\langle K_i, K_i \right\rangle}}$$

where K_i and K_i are Kernel Matrices

$$Linear \;\; Kernel : XX^T$$

$$Radial \;\; Basis \;\; Kernel : e^{-\left(\gamma\left(x_i - x_j\right)\left(x_i - x_j\right)^T\right)}$$

$$Polynomial \;\; Kernel : K(X) = \left(XX^T + 1\right)^p \; XX^T$$

In performing this task, three kernels were considered (with the understanding that the linear kernel is just a degenerate case of the radial basis kernel). The second method of kernel selection is detailed below. The alignment method was well suited to selecting the best value for p of the polynomial kernel.

Setting Parameter Values: The values of C_1, the linear function $C_2 = f(C_1)$, and the value gamma for the radial basis kernel are not known a priori. Furthermore, of the 3 fundamental kernel forms considered, nothing clearly identified one kernel as better than

another. The following maximization problems were therefore posed and solved with "hold-out cross validation" (HOCV) using 7.5% of the data set:

$$\arg\max_{\theta}\left\{\begin{array}{l}\arg\max_{C_1,\ f(C_1),\ \gamma,\ K(x)}\left\{(\mu(sensitivity)+specificity)\ \middle|\ \max_{\alpha}\ W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j\left(x^{(i)}\right)^T x^{(j)}\right\}\\ \arg\max_{C_1,\ f(C_1),\ \gamma,\ K(x)}\left\{(1-error)\ \middle|\ \max_{\alpha}\ W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j\left(x^{(i)}\right)^T x^{(j)}\right\}\end{array}\right\}$$

$$s.t.:\ \sum_{i=1}^{m}\alpha_i y^{(i)}=0$$

$$0\le\alpha_i\le C_1\ \ for\ y^{(i)}=+1$$

$$0\le\alpha_i\le C_2\ \ for\ y^{(i)}=-1$$

$$\mu\in\{\Re:\ \mu\in[1,1.3]\}$$

The variable μ above formalizes the notion that the model is intended to emphasize sensitivity over specificity in the non-separable case. No mixed kernels were considered in the course of this optimization. Binary search over the value space was used. While not ideal, in the interests of time, if the algorithm in the course of optimization did not converge within 2000 runs of the main smo_train loop, then the decision function was assessed with the parameter estimates at forced termination.

Principal Components Analysis: The data was analyzed with and without dimensional reduction with Principal Components Analysis (PCA). Due to time limitations, parameters were not optimized in the reduced feature spaces.

Results

Descriptive Statistics of Data Set: 4356 patients are compiled in the data set from an urban academic Emergency Department from July, 1999 through December, 2002. 20% (873) were diagnosed with ischemic heart disease or a related disease process benefiting from admission. Of these 873, less than 1% had an acute process such as aortic dissection or pulmonary embolism as the putative non-ischemic cause for the patient's discomfort. Demographics are presented in Table 1. After pre-processing of the data set, we identified 88 total parameters in the data set (Table 2).

Optimal Penalties (C_1 and C_2): C_1^* = 2.0 from the parameter optimization. The functional relationship between C_2^* and C_1^* was: $C_2^* = f^*(C_1^*) = 32*C_1^*$. As a validation method for the value of C_1^* selected, a Receiver Operator Curve (ROC) curve was constructed over different training sample sizes for a set value of C_1 and $C_2 = f(C_1)$. The largest area under the curve (AUC) was achieved for values approximately in the range of $C_1 = 2$, $C_2 = 32*C_1$. The match was not exact, but the proximity of the estimates validated the optimization results.

Performance: The best test set performance achieved a sensitivity and specificity of 93.5% and 43.5% respectively after training with 30% of the data set. Training set sensitivity and specificity were 94.6% and 46.4% respectively. With the asymmetric cost SVM, we achieved a higher sensitivity along with a lower specificity and overall error rate of 27.5%. There was no systematic pattern to the errors. However, the algorithm was designed to

bias towards a lower false negative rate. This is predicated on the tremendous costs of an untreated cardiac ischemic event in terms of QALYs lost and costs incurred to the system from complications due to delays in therapy and the cost of litigation resulting from the error in question (Appendix A: The Argument for a "Biased Classifier").

Feature Selection: The true impact of feature selection could not be accurately assessed as parameters were not optimized prior to applying PCA to the data. As it stands, feature selection performed best when the top ten or eleven features were selected.

Conclusions

Insights: First and foremost, the data is highly non-separable which is consistent with the initial hypothesis and comports with the observations of experienced clinicians and the published literature on the topic. This characteristic of the sample space more than any other may be what defines it as one of the most difficult decision making problems in medicine. Despite this, considerable insight is obtained.

Comparison to the Status Quo: The asymmetric cost SMO was able to classify patients into a category for discharge at a rate higher than current clinical practice. The associated drop in sensitivity may not (depending on the prior) meet the rubric of a posterior probability of disease that is less than 1%. The sensitivity achieved out-performs many machine learning applications applied to this decision space, but does so with a drop in specificity.

Limitations and Future Directions: There are several aspects of the methodology that were not fully optimized prior to analysis. As stipulated above, in the interests of time, when optimizing parameters the smo_train loop was terminated if there was no convergence within 2000 runs. The algorithm should be allowed to converge over time thereby avoiding an incomplete/inaccurate optimization. One way to facilitate this is to incorporate an optimization to Platt's algorithm identified by Keerthi et al that is particularly helpful for large values of C[32]. Furthermore, more analysis on the impact of different types of kernels, including compound kernels, on the performance of the algorithm needs to be assessed.

Generalizability of Results: It's entirely possible that the feature space over which data was collected was still inadequate. A consensus approach among experienced clinicians should be used to establish the data collection feature space for any future work. The demographics of this data set are not representative of most Emergency Departments in this country. The parameter estimates derived in this research are likely not applicable to a suburban Emergency Department in the Bay Area without risking high generalization error. This hypothesis is, of course, amenable to exploration as such data sets become available.

References: Reference, tables, and Appendix available upon request -- too long for restricted posting format