On the Medical Frontier: The 2006 KDD Cup Competition and Results

Terran Lane
Department of Computer Science, University of New Mexico, Albuquerque, NM 87131 USA
terran@cs.unm.edu
Bharat Rao, Jinbo Bi, Jianming Liang, Marcos Salganicoff
Computer Aided Diagnosis and Therapy, Siemens Medical Solutions, Inc. Malvern, PA 19355 USA
{bharat.rao,jinbo.bi,jianming.liang,marcos.salganicoff}@siemens.com

ABSTRACT

The 2006 KDD Cup Competition featured three data mining tasks drawn from a medical imaging domain. At the core, all of these tasks were concerned with identifying pulmonary embolisms (PEs) from pre-processed computed tomography (CT) images of human lungs. However, these tasks were complicated by features such as multi-instance learning, stringent performance standards, hard-threshold evaluation functions, spatial correlations, and small training sets. This paper gives an overview of the data and medical imaging tasks, the competition and evaluation, and the competition victors.

1. INTRODUCTION

The Tenth KDD Cup competition was held between May and August, 2006. Participants from around the world worked to design cutting-edge data mining methods for a medical image detection problem. This paper describes the Computer-Aided Detection problem domain and the Pulmonary Embolism data sets that were the subject of the competition. We also describe the data mining and evaluation challenges involved in this type of data.

This domain is characterized by stringent performance standards necessary to meet FDA approval and to ensure physician confidence. The three performance tasks of the KDD Cup competition (Section 3) were formulated to reflect these hard, real-world constraints. We describe the performance tasks and the evaluation procedures. The hard constraints imposed by the scoring criteria turned out to be, in some sense, the most challenging aspect of this year's competition, and the strongest teams were often those who focused on scoring criteria. We examine the difficulties presented by the data, tasks, and scoring criteria.

The methods used by the winning teams are described in three companion reports to this one.

1.1 The Computer-Aided Detection Domain

Over the last decade, Computer-Aided Detection (CAD) systems have moved from the sole realm of academic publications, to robust clinical systems that are used by physicians in their clinical practice to help detect early cancer from medical images. For example, CAD systems have been employed to automatically detect (potentially cancerous) breast masses and calcifications in X-ray images, detect lung nodules in lung CT (computed tomography) images, and detect polyps in colon CT images, to name just a few CAD applications.

CAD applications lead to very interesting data mining problems. Typical CAD training data sets are large and extremely unbalanced between positive and negative classes. Often, fewer than 1% of the examples are true positives. When searching for descriptive features that can characterize the target medical structures, researchers often deploy a large set of experimental features, which consequently introduces irrelevant and redundant features. Labeling is often noisy as labels are created by expert physicians, in many cases without corresponding ground truth from biopsies or other independent confirmations. In order to achieve clinical acceptance, CAD systems have to meet extremely high performance thresholds to provide value to physicians in their day-to-day practice. Finally, in order to be sold commercially and employed clinically (at least in the United States), most CAD systems have to undergo a clinical trial (in almost exactly the same way as a new drug would). Typically, the CAD system must demonstrate a statistically significant improvement in clinical performance, when used, for example, by community physicians (without any special knowledge of machine learning) on as yet unseen cases. That is, to be accepted, the sensitivity of physicians with CAD must be (significantly) above their performance without CAD, without a corresponding marked increase in false positives (which may lead to unnecessary biopsies or expensive tests). In summary, very challenging machine learning and data mining tasks have arisen from CAD systems.

2. THE PULMONARY EMBOLISM DATA SET

In this section, we describe the CAD data set that was used for the 2006 KDD Cup Competition.

2.1 Pulmonary Embolism

Pulmonary embolism (PE) is a highly lethal condition that occurs when an artery in the lung becomes completely or partially blocked. In most cases, the blockage is caused by one or more blood clots that travel to the lungs from other parts of the body (e.g., legs or pelvis). While PE is not always fatal, it is nevertheless the third most common

cause of death in the US, with at least 650,000 cases occurring annually [2]. The clinical challenge, particularly in an Emergency Room scenario, is to correctly diagnose patients that have a PE and then send them on to therapy. This, however, is not easy, as the primary symptom of PE is dysapnea (shortness of breath), which has a variety of causes, some of which are relatively benign. Thus, it is hard to separate out the critically ill patients suffering from PE from other patients who may require a different treatment or none at all.

The two crucial clinical challenges for a physician, therefore, are to diagnose whether a patient is suffering from PE and to identify the location of the PE. Computed Tomography Angiography (CTA) has emerged as an accurate diagnostic tool for PE. However, each CTA study consists of hundreds of images, each representing one slice of the lung. Manual reading of these slices is laborious, time consuming and complicated by various PE look-alikes (false positives) including respiratory motion artifacts, flow-related artifacts, streak artifacts, partial volume artifacts, stair step artifacts, lymph nodes, and vascular bifurcation, among many others.

Additionally, when PE is diagnosed, medications are given to prevent further clots, but these medications can sometimes lead to subsequent hemorrhage and bleeding since the patient must stay on them for a number of weeks after the diagnosis. Thus, the physician must review each CAD output carefully for correctness in order to prevent over-diagnosis. Because of this, the CAD system has to produce only a small number of false positives per patient scan. The goal of a PE CAD system, therefore, is to automatically identify PEs with as few false positives as possible. To handle any residual false positives from CAD, in practice, each CAD PE finding must be finally reviewed and accepted by the radiologist before it is reported.

2.2 The Data Processing Pipeline

The PE CAD system developed at Siemens consists of the following three consecutive components:

1. Candidate Generation: Identify potential candidate regions of interest (ROI) from a medical image.

2. Feature Computation: Compute a number of descriptive features for each generated candidate.

3. Classification: Classify each candidate (in this case, whether it is a PE or not) based on its features.

The first stage is a "focus of attention" (FOA) stage that identifies *candidates*: image regions that stand out from the background and are more likely to contain the target object (PEs). The second stage converts from image space to a feature vector space that is more amenable to standard classification techniques. The final stage filters out non-PE candidates and returns PEs to the physician. This basic three-stage architecture is a common approach to identifying infrequent elements in image data [1].

There are significant real-world data mining challenges in the first two steps but, because of medical privacy considerations, the primary image data cannot be made publicly available. Thus, for the 2006 KDD Cup data, Steps 1 and 2 have been performed at Siemens and only feature values for every candidate ROI are provided. The goal of the KDD Cup is to design a series of classifiers related to Step 3.

2.3 Competition Data Set and Features

In the 2006 KDD Cup, a total of 67 cases were collected and labeled by expert chest radiologists, who reviewed each case and marked the PEs. The cases were randomly divided into training and test sets. The training set includes 46 cases, while the test set contains the remaining 21 cases. (Originally 23 test cases were provided, but 2 duplicate patient cases were identified and hence removed.) The test group was sequestered during the competition and was only used to evaluate the performance of the final system.

2.3.1 Candidate generation and labeling

All the 67 cases were processed with a prototype version[1] of the Siemens PE CAD system, which generated a total of 4424 candidates: 3033 candidates appear in the training set and 1391 candidates in the test set. Each candidate is a cluster of voxels (the 3-D analog of pixels), and represented by a representative point with a 3-D coordinate derived from the cluster of voxels.

Each candidate was then labeled as a PE or not based on 3-D landmark ground truth provided by the experts. In order to automatically label each candidate, each PE pointed out by an expert landmark is semi-automatically extracted and segmented. Therefore, the ground truth for each PE is also a cluster of voxels (i.e., the segmented PE). Any candidate that was found to be intersected with any of the segmented PEs in the ground truth was labeled as a PE. However, it should be noted that the PE segmentation process is semi-automatic, involving a manual process, and consequently the segmentation of a PE might not be perfect. In other words, the labeling may be noisy. Moreover, multiple candidates may intersect with the same segmented PE, that is, multiple candidates may correspond to a single PE. Since each PE has a unique identifier, there may exist multiple candidates labeled with the same PE identifier. This type of problem is sometimes referred to as a multiple-instance problem where each positive example has multiple instances.

2.3.2 Feature computation

For each candidate, a set of 116 features were calculated within the Siemens PE CAD system. Three of the features were the x, y, and z locations (a representative point in 3D) of the candidate. The remaining features were image-based features and were normalized to a unit range, with a feature-specific mean. The features can be categorized into those that are indicative of voxel intensity distributions within the candidate, those those summarizing distributions in neighborhood of the candidate, and those that describe the 3-D shape of the candidate and enclosing structures. When combined these features can capture candidate properties that can disambiguate typical false positives such as dark areas that result from poor mixing of bright contrast agents with blood in veins, and dark connective tissues between vessels, from true emboli. These features are not necessarily independent, and may be correlated with each other, especially with features in the same group.

2.3.3 Data format

Two text files are provided and they contain the training and test feature matrices, respectively, where each row represents an example, each column represents a feature. The

[1]Not commercially available.

first two columns supply the patient identifier and the PE identifier. The PE identifier is also our target label variable which tells whether or not the corresponding example is a PE. If it is a PE, the label is the PE identifier (a positive number); if it is not a PE, the label is set to 0. In the test data, all labels are set to -1 (which means unknown).

3. THE KDD CUP TASKS

The KDD Cup competition this year comprised three tasks: PE identification, patient classification, and patient negative predictive value. Teams could participate in any or all of these tasks, and separate winners were awarded in each task category. In addition, Tasks 1 and 2 each comprise three sub-tasks. Competitors had the option of submitting a single solution for all three sub-tasks or separate solution vectors for each sub-task. Typically, competitors chose to submit different solution vectors for each of the three sub-tasks in an effort to improve sensitivity as the false positive threshold was relaxed.

A *submission* for a sub-task is a vector containing a single binary value for each data point in the test data. The competition demanded a hard classification, rather than, say, a confidence value to reflect the medical reality that the customers (physicians) want to see a hard classification. A soft classification, such as a probability or confidence, would require additional interpretation and training on the part of the physician.

3.1 Task 1

The first classification task is to classify individual PEs. For clinical acceptability, it is critical to control false positive rates – A system that "cries wolf" too often will be rejected out of hand by clinicians. Thus, the goal is to detect as many true PEs as possible, subject to a constraint on false positives. For this task, we make the following definitions:

PE sensitivity is defined as the number of PEs correctly identified in a patient. A PE is correctly identified if *at least one* of the candidates associated with that PE is correctly labeled as a positive. In this case, identifying 2 or more candidates for the same PE makes no impact on the sensitivity.

False positives are defined as the number of candidates which are not true PEs but incorrectly classified by a classifier as PEs in a patient. That is, the false positive rate is the total of all negative candidates classified as PEs in the patient.

Average FP rate for a test set is the average number of FPs produced across all patients in that test set.

Example 1: Consider a patient with 2 PEs marked by a physician and a total of 17 candidates. Assume the first PE has 5 candidates associated with it, and the second PE has 3 candidates (8 positive labels with two PE-ids for this patient). If the classifier labels 3 of the candidates associated with the first PE correctly, none of the candidates associated with the second PE correctly, and marks 4 other candidates not associated with either PE as positive, then the classifier would have marked one PE correctly out of two possible (sensitivity=50%), with 4 false positives.

Example 2: Suppose that a test set has a total of 10 patients. Two classifiers are applied to that test set. Classifier A produces 2 FPs on each of the first nine patients and 3 FPs on the last patient. Classifier B produces zero FPs on each of the first nine patients and five FPs on the final patient. Then classifier A has an average FP rate of 2.1, while classifier B has an average FP rate of 0.5.

In this set of tasks, the goal is to produce a classifier that maximizes sensitivity, subject to a threshold on maximum allowable FPs (i.e., to maximize a Neyman-Pearson criterion.) If, in any test set, a classifier exceeds the maximum allowable average FP rate for that sub-task, the results are completely disqualified for the entirety of Task 1. Classifiers must meet the specified average FP threshold for all three sub-tasks, or the entire submission is disqualified from Task 1.

Example 3: The FP threshold for Task 1a (defined in the following) is 2 per patient. Classifier A from Example 2 produces an average of 2.1 FP per patient on the test set during this task and will be disqualified, regardless of its sensitivity. Classifier B, however, passes the FP threshold, so its sensitivity will be evaluated.

No extra credit was given for predictors that performed better than this FP metric. Thus, say, producing 1 false positive per patient, instead of 2, did not impact the final score. (Though achieving 1 FP/patient at a high sensitivity, would be extremely valuable, clinically speaking!).

Competitors were allowed to use different classifiers for each of the following sub-tasks, and to submit different solution vectors for each.

Task 1a: Build a system where the false positive rate is at most 2 per patient.

Task 1b: Build a system where the false positive rate is at most 4 per patient.

Task 1c: Build a system where the false positive rate is at most 10 per patient.

In each task, the submissions were ranked based on PE sensitivity, subject to the FP thresholds. Submissions that did not meet the FP thresholds were disqualified for that subtask and for the task as a whole.

3.2 Task 2

The second classification task is to classify each patient as having a PE or not. The reason why this is important is that patient treatment for PE is systemic, i.e., many aspects of the treatment are the same whether the patient has one or many PE. For this task, we make the following definitions:

Patient sensitivity is defined as the number of patients for whom at least one true PE is correctly identified. As above, a PE is identified if any one of the candidates associated with that PE is correctly classified, and multiple correct identifications in a single patient do not increase the sensitivity score.

False positives are defined as the number of candidates falsely labeled as a PE in a patient.

Average FP rate for a test set is the average number of FPs produced across all patients in that test set.

Example 1: Consider a patient with 2 PEs marked by a physician. Assume the first PE has 5 candidates associated

with it, and the second has 3 candidates associated with it. If the classifier labels 2 of the candidates associated with the first candidate correctly, none of the candidates associated with the second PE, correctly, and four other candidates not associated with either PE, then the classifier would have labeled the patient correctly, with 4 false positives.

Again, for this task, 3 classifiers should be built, and any classifier that yields an average FP rate above the specified FP threshold on any sub-task will be disqualified.

Task 2a: Build a system where the false positive rate is at most 2 per patient.

Task 2b: Build a system where the false positive rate is at most 4 per patient.

Task 2c: Build a system where the false positive rate is at most 10 per patient.

In each task, the submissions were ranked based on PE sensitivity, subject to the FP thresholds. Submissions that did not meet the FP thresholds were disqualified for that subtask and for the task as a whole. Competitors were free to use the same classifier(s) as in Task 1, or to build different classifiers for this task.

3.3 Task 3

The third classification task is to identify negative patients while producing no false negatives. The reason for this task is that one of the most useful applications for CAD would be a system with very high (nearly 100%) Negative Predictive Value. In other words, if the CAD system generated zero positive candidates for a given patient, we would like to be very confident that the patient was indeed free from PEs.

For this task, we make the following definitions:

A positive patient is defined as a patient who has at least 1 PE. Otherwise, it is a **negative patient**.

Identified as negative A patient is identified as negative when the CAD system produces no positive labels for any of that patient's candidates.

Negative predictive value (NPV) for a classifier is

$$\mathrm{TN}/(\mathrm{TN}+\mathrm{FN})$$

(i.e., number of true negative patients divided by the total of true and false negatives).

Note that the NPV is maximized by a classifier that correctly identifies some negative patients but produces no false negatives (no positive patients identified as negative). To qualify for this task, a classifier must have 100% NPV (i.e., when it says a patient has no positive marks, the patient must have no true PEs). The primary criterion is the highest number of negative patients identified in the test set (largest TN), subject to a minimum cut-off of identifying 40% of the negative patients on the test set. The first tie breaker is the sensitivity on PEs (as defined in Task 1), followed by the false positive rate on the entire test set.

3.4 Challenges

There were two broad classes of challenges in this competition: challenges arising from the data itself, and challenges arising from the evaluation criteria. The data challenges stemmed from the complex source (human medical imaging), sparse access to domain knowledge, and small data set sizes. The medical and anatomical aspects of the data raise a number of interesting issues, only some of which were exploited by any competitors.

Noisy ground truth Candidate labeling may be noisy, as candidates are labeled as PEs according to segmented PEs based on ground truth marks given by experts. Since the segmented PEs are not perfect, some candidates may be falsely labeled as PEs due to PE segmentation errors.

Feature correlation Many of the features computed for each candidate are correlated.

Imbalanced data The data is very imbalanced between positive and negative classes. Commonly, only around 1-5% of the candidates are true positives.

Sparse data The data is relatively sparse – even though, from a machine learning point of view, this data has relatively few positive examples, in real-life it costs several million dollars to collect, label, and build the features, all while maintaining strict patient confidentiality as per legal and ethical requirements.

Spatial correlation The PE candidates are strongly spatially correlated by their proximity to legitimate PEs and to anatomical artifacts that are likely to produce false PE candidates. For example, certain arteries are likely to produce false PE candidates and arteries tend to run vertically. In principle, it is possible to exploit the vertical alignment of arterial PE candidates to filter out such false positives, though it does not appear that any of the competitors attempted to do so.

Symmetry Some teams observed that lungs are bilaterally symmetric and that false PEs arising from anatomical structure should, therefore, also be likely to cluster symmetrically. Thus, statistics about anatomy can be gathered from both lungs, while PEs (presumably independent of anatomy) should stand out due to asymmetry.

The evaluation criteria in this year's KDD Cup were quite strict and created a much more difficult task than many data mining practitioners are used to working with. Many textbook and real-world data mining tasks are expressed in terms of relatively "soft" objective functions such as squared-error. In such tasks, small changes to the predicted value of a datum make only small changes to the resulting score. Even apparently "hard" objectives, such as 0-1 loss, write the complete data-set score as a linear function of independent hard-threshold evaluations for each data point. Thus, the classification of any single datum makes only a small impact on the global score. Further, we often "soften" even these evaluations by employing tools such as ROC curves that allow us to be agnostic about operating points and acceptable performance levels. Altogether, these tools allow us to tune our data mining pipelines fairly aggressively, attempting to eke out the maximum possible positive prediction performance.

By contrast, the evaluation criteria for the 2006 KDD Cup were very strict and highly nonlinear. PE sensitivity, patient

sensitivity, and false positives are used for evaluation. These metrics are more tuned to the clinical needs of physicians for decision support. Further, there were three separate hard-threshold decisions made in the scoring process:

Multiple instance scoring In all three tasks, the entities being classified (PEs or patients) were represented by *sets* of candidates. That is, a single PE (for example) might correspond to a dozen feature vectors. To receive credit for identifying the PE, a competitor had to correctly identify *any one* of this set. But no additional credit was received for identifying more than one candidate from this set. Thus, there is a sharp no credit/credit threshold. Compare this to the standard supervised learning framework in which most scores are linear in the number of correctly classified instances.

Acceptable false positive rate The FDA and physicians will reject a system with an unacceptable false positive (FP) rate. Therefore, it is useless to provide a system with high sensitivity that "almost" meets the false positive rate. Thus, a competitor's solution for some subtask was disqualified if it exceeded the allowable FP threshold for that sub-task (see Section 3).

Boolean `AND` aggregation Because Tasks 1 and 2 each had three sub-tasks (Section 3), it was necessary to aggregate the competitors' solutions across sub-tasks. Following the description document [4], Boolean `AND` was used for this aggregation. That is, to qualify for a task, a competitor's set of solution vectors had to meet the FP thresholds for *all three* sub-tasks.

These hard thresholds increase the impact of classification variance: small changes in labellings of the test set can yield dramatic changes in the score. In the worst case, changing a single label of one feature vector in one sub-task can shift a submission's FP rate enough to move it from qualification to disqualification. The sensitivity of a disqualified submission is defined to be zero, so the change in a single label can reduce a high sensitivity to zero.

Thus, controlling the variance of learned models is a key consideration in these tasks. This is very different than the "textbook" approach to data mining, in which the emphasis is often on the expectation of a model's performance. As it turned out, many of the high-performing submissions in this year's KDD Cup were from teams who devoted substantial attention to this issue.

4. OVERVIEW OF PARTICIPANTS AND APPROACHES

Overall, sixty-eight groups submitted solutions to at least one of the tasks. Participants spanned at least eighteen countries, judging from their email address domains.[2] There were twenty-four student-led teams from at least eight countries. Overall, the student-led teams gave strong performances, with student-led teams finishing in the top five places in Tasks 1 and 2, and in the top ten places in Task 3.

[2]This is almost certainly a conservative estimate, as there were a number of registrants who listed no affiliation or provided only globally accessible email addresses, such as `gmail.com`.

The space of techniques attempted by the competitors ranged across a similarly wide spectrum of provenances. Most teams brought to bear some combination of three critical steps of data mining: data cleaning/preprocessing, classifier training, and validation. The emphasis among these three varied widely however. Some teams viewed the problem almost exclusively as a dimensionality reduction problem (almost in an unsupervised fashion) and focused substantial effort on the data cleaning/preprocessing phase, while applying only very simple classifiers and validation thereafter. Others focused heavily on validation, attempting to design their validation methods to account for the stringency of the evaluation criteria. And, perhaps unsurprisingly, a majority of the teams focused most of their effort on creative variations in the classification space.

5. EVALUATION

The testing data for the KDD Cup were, unfortunately, quite limited. While the organizers had hoped to have more data available during the scope of the competition, this proved to be impossible. For both practical and privacy reasons, human medical imaging data are extremely expensive to collect and pre-process. In the end, only 1391 PE candidates (feature vectors), representing 67 patients, were available for testing. And, of those, two patients (112 PE candidates) turned out to be inadvertent duplicates from the training data set and had to be dropped during the evaluation phase. Thus, the final competition evaluation was on only 1279 PE candidates.

5.1 Bootstrap Sampling

To partially offset this small evaluation set, we employed bootstrap resampling [3] of the test set. This procedure provides an improved estimate of the expected sensitivity for a submission, at the cost of increased variance. In light of the sensitivity of the KDD Cup scoring procedure to variance (Section 3.4), this can have a large impact on classifiers (Section 5.2). In some sense, the bootstrap samples simulate random draws of patients from the same population as the original testing data set. The variance impact on a submission can then be interpreted as a measure of the robustness of the submission to variations in the patient pool. Because of the hard thresholds in the scoring function, it is important that a fielded classifier system for this kind of data not be overly sensitive to such variations.

An interesting problem arose in the sampling phase. The usual implementation of bootstrap sampling is to draw a new data set by uniformly sampling, with replacement, from the original data. However, this procedure is based on the assumption that the data are originally generated IID so that a uniform sample correctly captures marginal distributions. This assumption is violated in the 2006 KDD Cup data: candidates are correlated both within patients and by PE. Given these dependencies, it is not immediately clear how best to generate a bootstrap sample.

We examined two different sampling strategies. The *flat sampling* strategy ignored data correlations and simply drew uniformly with replacement from the test data. The *hierarchical sampling* strategy, on the other hand, first sampled uniformly with replacement from patients, and then drew candidates conditioned on patient. The goal was to mimic the generative process of the original data, in which PEs are

a function of a patient's health and do not simply occur IID in the population, independent of patient.

	MSE		KL Div	
	Flat	Hier	Flat	Hier
Pat	5e-4	0.03	9e-5	0.26
PE	9e-5	2e-3	2e-4	0.03
PE|Pat	0.041	0.034	0.0042	0.0038

Table 1: Comparison of flat versus hierarchical bootstrap sampling on the KDD Cup data. Columns give measurements of mean-squared error (MSE) and KL divergence (KL Div) between the bootstrapped sample and the original test data for both the flat and hierarchically sampled bootstraps. The rows give the values for the marginal distribution of patients, the marginal distribution of PEs, and the conditional distribution of PEs given patients. Note that while the hierarchical sample has larger MSE/KL than the flat sample for the marginals, it has smaller MSE/KL than flat for the conditional.

To understand the difference between the flat and hierarchical bootstrap samples, we compared the distribution of patients and PEs in the bootstrap to that in the original testing data. Table 1 gives the mean-squared error and KL divergence between the bootstrapped distribution and the original data distribution. The first two rows show the measured values for the marginal distributions of the patients and PEs, respectively, in the flat and hierarchically bootstrapped samples. The values for the flat samples are essentially zero, while the hierarchically bootstrapped samples have small, but non-negligible, errors. The last row, however, shows the measured errors for the conditional distributions of PEs, given patient. Here we see that the flat sampled data has worse error than the hierarchically sampled data. This finding affirms that the hierarchically bootstrap samples are more nearly preserving the important patient-PE dependency than the flat sample is. Thus, we believe that the hierarchical bootstrap better represents the distribution of patients and PEs that might be seen in a clinical setting than a flat sampled bootstrap would.

We also found that both sampling strategies decreased competitors' average scores, compared to non-bootstrapped scores. Most of the score decrease can be attributed to an increase in average FP rate, which pushed the score of many samples into the disqualified region. This is not surprising, given the increased variance due to bootstrapping. However, the relative rankings of competitors was less affected. In particular, the first place competitors in both Tasks 1 and 3 remained unchanged by bootstrapping, and most of the top competitors remained in the top decile. Furthermore, the winning competitors won under both flat and hierarchical bootstrapping. (Though the choice of bootstrap strategy did have a substantial impact on relative rankings lower down in the pool – especially below the top-ten competitors.)

For the final competition evaluation, we drew 200 hierarchical bootstrap samples from the testing data, each of 1279 candidates. A submission's score was the mean score over all bootstrap samples, where a sample on which the submission was disqualified was assigned a score of 0. The same 200 bootstrap vectors were used to evaluate all competitors.

5.2 The Effects of Boosting and Stringent Evaluation Criteria

The increase in average FP rate due to bootstrapping had dramatic impact on some competitors, while others were relatively less affected. Those who emerged at the top tier of the bootstrap sample were, overall, less affected by the sampling than were those in lower tiers. For example, on Task 1 the top one-third of competitors' scores changed by an average of 0.62 due to bootstrapping, while the middle third changed by an average of 1.02. (The bottom third were those who, largely, were disqualified and scored zero on both the original test data and the bootstrapped sample.) For Task 2, the effect was even more dramatic, with the top tier of competitors changing score by 2.71, while the middle tier shifted by 8.02.

In Task 3, it turns out that only two competitors even qualified on the original testing data. In this case, bootstrapping was a great boon to many competitors, as many of them qualified on some of the bootstrap samples and, therefore, received a non-zero score in the final evaluation. This allowed us to, for example, select a best student submission for Task 3 – a decision that would have been impossible under the original data.

Some competitors' scores were quite drastically impacted on the bootstrap data. For example, some slid from a non-zero score on the original test data to a zero score on the bootstrapped data. To understand this counterintuitive outcome, we examined the scores of individual bootstrap samples for a number of competitors. Tables 2 and 3 illustrate this effect for Task 1. Each of these tables shows a few rows from the bootstrap sample results. The first three columns show the competitors' raw FP rates on the corresponding sub-task. The next three columns indicate whether that FP rate qualified for the sub-task, and the final column shows the overall qualification for Task 1 on that draw of the bootstrap data.

Subtask FP Rate			Subtask Qualification?			Task
S-1	S-2	S-3	Q-1?	Q-2?	Q-3?	Qual?
1.19	2.81	8.00	T	T	T	T
1.29	3.71	10.57	T	T	F	F
0.90	3.00	9.00	T	T	T	T
1.43	3.76	10.05	T	T	F	F
0.71	2.95	8.24	T	T	T	T

Table 2: Example of a few results from the bootstrap evaluation of the winning submission for Task 1. Each row gives results from one bootstrap sample of the test data. The first three columns display the FP rate measured from the competitor's submission on that sample for each of the three sub-tasks of Task 1. The next three columns show whether or not the FP rate met the required FP threshold for a given sub-task on that sample. The final column displays whether or not that sample qualified to compete (Boolean `AND` of the previous three columns). We see that this competitor qualified on three of the five samples shown here. Altogether, this submission qualified on 179 of the 200 bootstrap samples.

Table 2 is data drawn from the winning submission for this task. We see that, even in the presence of bootstrapping, this submission qualifies on many samples. While the bootstrap impacted the average FP rates for this competi-

Subtask FP Rate			Subtask Qualification?			Task
S-1	S-2	S-3	Q-1?	Q-2?	Q-3?	Qual?
2.24	5.48	9.86	F	F	T	F
2.14	5.19	11.05	F	F	F	F
1.62	5.62	12.14	T	F	F	F
2.00	5.57	10.29	T	F	F	F
1.71	5.48	11.24	T	F	F	F

Table 3: Example of a few results from the bootstrap evaluation of another competitor's submission for Task 1. Each row gives results from the same bootstrap sample as displayed in Table 2. The columns are the same as in Table 2. While this submission qualified in a number of sub-tasks individually, in every sample showed here the submission was disqualified by the AND criterion. Altogether, this submission did not qualify on any of the bootstrap samples.

tor, the design was fairly robust to variance and maintained a passing score on 179 of the 200 bootstrap samples.

In Table 3, we see samples from a competitor whose submission moved from a non-zero score on the original testing data to zero on the bootstrapped data. In this case, although the submission qualified on a number of sub-tasks, it missed the target FP threshold on at least one sub-task and was disqualified in all cases. This competitor was much more deeply affected by the variance imposed by the bootstrap samples. In a more traditional data mining setting, in which the scoring function changes linearly with the number of incorrectly labeled instances, this would have had significantly less impact on this competitor's overall score. Under the KDD Cup's much more stringent scoring function, however, the relatively modest FP increases were dramatically amplified.

Interestingly, some competitors' scores actually *improved* after the bootstrapping. This was primarily in the case where a competitor was disqualified on the original test data, but qualified on some samples from the bootstrap. It appears that for these submissions, the original test data was a statistically poor sample, and that under a slightly different data set, they would have performed better. Interestingly, the number of competitors whose score improved from zero after bootstrapping turned out to offset those whose score was decreased to zero. In Task 1, for example, twenty-eight of the sixty-four competitors scored zero on the original test data, and twenty-eight scored zero on the bootstrapped data. However, four competitors had shifted from zero to non-zero, while four others shifted in the opposite direction. The notable exception, as mentioned above, was Task 3, in which twenty-one competitors received a non-zero score only after bootstrapping.

Overall, it appears that, while a mixed blessing in some cases, the combination of bootstrapping with a sharply nonlinear scoring function allowed us to identify competitors whose methods were most robust to data variance. Given the sensitivity of the target domain, in which evaluation criteria are quite strict, these are the competitors that would most likely win acceptance in the market.

6. VICTORS

For each task, we identified a winning team and one or more runners-up teams. Each task also had a "best student entry" prize, awarded to the best-performing student-led team.

6.1 Task 1: PE Identification

The winners in Task 1 were:

First Place Robert Bell, Patrick Haffner, and Chris Volinsky (AT&T Research).

First Runner Up Dmitriy Fradkin (Ask.com).

Second Runner Up Domonkos Tikk (Budapest University of Technology & Economics), Zsolt T. Kardkovács (Budapest University of Technology & Economics), Ferenc P. Szidarovszky (Szidarovszky Ltd. and Budapest University of Technology & Economics), György Biró (TextMiner Ltd.), and Zoltán Bálint (Budapest University of Technology & Economics).

Best Student Entry Karthik Kumara (team leader), Sourangshu Bhattacharya, Mehul Parsana, Shivramkrishnan K, Rashmin Babaria, Saketha Nath J, and Chiranjib Bhattacharyya (Indian Institute of Science).

6.2 Task 2: Patient Classification

The winners in Task 2 were:

First Place Domonkos Tikk (Budapest University of Technology & Economics), Zsolt T. Kardkovács (Budapest University of Technology & Economics), Ferenc P. Szidarovszky (Szidarovszky Ltd. and Budapest University of Technology & Economics), György Biró (TextMiner Ltd.), and Zoltán Bálint (Budapest University of Technology & Economics).

First Runner Up Ruiping Wang, Yu Su, Ting Liu, Fei Yang, Liangguo Zhang, Dong Zhang, Shiguang Shan, Weiqiang Wang, Ruixiang Sun, and Wen Gao (Institute of Computing Technology, Chinese Academy of Sciences).

Second Runner Up Cas Zhang, Y. Zhou, Q. Wang, and H. Ge (Joint R&D Lab, Chinese Academy of Sciences).

Third Runner Up Dmitriy Fradkin (Ask.com).

Best Student Entry Zhang Cas (IA, PKU).

6.3 Task 3: Negative Predictive Value

The winners in Task 3 were:

First Place William Perrizo and Amal Shehan Perera (DataSURG Group, North Dakota State University)

Runner Up Nimisha Gupta and Tarun Agarwal (Strand Life Sciences Pvt. Ltd.)

Best Student Entry Karthik Kumara (team leader), Sourangshu Bhattacharya, Mehul Parsana, Shivramkrishnan K, Rashmin Babaria, Saketha Nath J, and Chiranjib Bhattacharyya (Indian Institute of Science).

7. RESOURCES

The 2006 KDD Web site will continue to be active at `http://www.cs.unm.edu/kdd_cup_2006`. This site contains the full KDD Cup training and testing data and the full competition rules. This site also hosts a downloadable archive, containing all of the above, as well as the bootstrap vectors used for the competition, the list of excluded testing data (Section 5), and the scoring software. The three winning teams have provided reports on their approaches to the competition, appearing as companions to this one.

8. CONCLUSIONS

The 2006 KDD Cup was an exciting competition, with a rich problem domain and an excellent field of competitors. The data set was drawn from an important and high-profile real-world medical domain. The real-world facets of the data and domain contributed a number of important challenges to the competition, ranging from inter-point correlations to an extremely difficult scoring criterion. A broad field of teams competed in three tasks on this data. In the end, it appears that the strongest competitors were those who planned for robustness to sample variance.

Overall, the competition was quite strong and we had a number of truly excellent submissions. It is clear from the quality of results that all of the competitors poured enormous creativity and effort into the competition. We hope that this will spur continued research into the many challenges raised by this type of data and evaluation criteria.

9. ACKNOWLEDGMENTS

The organizers would like to thank Siemens Medical Solutions for their generous donation of data sets, expertise, and time to the KDD Cup. We also thank the KDD organizers who gave us the opportunity to work with the KDD Cup and who provided important assistance during the competition. Thanks also to Rich Caruana for thoughts on evaluation methods. We would like to thank Chad Lundgren for his invaluable support in developing the KDD Cup competition web site, shepherding submissions, and generally handling large amounts of infrastructure. But most of all, we thank all of the competitors for the great effort and creativity they put forth to compete.

Terran Lane's work was supported in part by NIMH grant number 1R01MH076282-01 as part of the NSF/NIH Collaborative Research in Computational Neuroscience Program.

10. REFERENCES

[1] M. C. Burl, L. Asker, P. Smyth, U. Fayyad, P. Perona, L. Crumpler, and J. Aubele. Learning to recognize volcanoes on venus. *Machine Learning*, 30(2–3):165–194, February 1998.

[2] C. Fried and J. Handler. Pulmonary embolism. `http://www.emedicine.com`. Accessed Oct 25, 2006.

[3] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[4] T. Lane, B. Rao, J. Bi, and M. Salganicoff. 2006 KDD Cup task: Computer aided detection of pulmonary embolism. `http://www.cs.unm.edu/files/kdd-cup-2006-task-spec-final.pdf`, October 2006.