2006 KDD Cup Task: Computer Aided Detection of Pulmonary Embolism

Terran Lane
Bharat Rao
Jinbo Bi
Marcos Salganicoff

Document date: Oct 10, 2006
Document version: 1.1

Description of CAD systems

Over the last decade, Computer-Aided Detection (CAD) systems have moved from the sole realm of academic publications, to robust commercial 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 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 (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 – i.e., the sensitivity of physicians with CAD must be (significantly) above their performance without CAD, and 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

Challenge of Pulmonary Emboli Detection

Pulmonary embolism (PE) is a condition that occurs when an artery in the lung becomes blocked. In most cases, the blockage is caused by one or more blood clots that travel to the lungs from another part of your body. 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.[1] 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, making it hard to separate out the critically ill patients suffering from PE.

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 overdiagnosis. Because of this, the CAD system must provide only a small number of false positives per patient scan.

The goal of a CAD system, therefore, is to automatically identify PE's. In an almost universal paradigm for CAD algorithms, this problem is addressed by a 3 stage system:

1. Identification of candidate regions of interest (ROI) from a medical image,
2. Computation of descriptive features for each candidate, and
3. Classification of each candidate (in this case, whether it is a PE or not) based on its features.

In this year's KDD Cup data, Steps 1 and 2 have been done for you. Your goal is to design a series of classifiers related to Step 3.

The PE Data

For this task, a total of 69 cases were collected and provided to an expert chest radiologist, who reviewed each case and marked the PEs. The cases were randomly divided into training and test sets. The training set includes 38 positive cases and 8 negative cases, while the test set contains the remaining 23 cases. The test group is sequestered and will only be used to evaluate the performance of the final system. (See **Rules**, below.)

[1] C. Fried and J. Handler, "Pulmonary Embolism", www.emedicine.com

Additional train/test data may become available during the course of the competition. Any additions to the data will be announced to all registered participants via the competition mail list and will be posted online.

Note: It turns out that patients numbers 3111 and 3126 in the test data duplicate patients 3103 and 3115, respectively, in the training data. Therefore, 3111 and 3126 were dropped from the final evaluation for the competition. These patients should be discarded from testing for any future comparisons to the KDD Cup results. The file `drop_duplicate_patients_mask.dat`, in the KDD Cup archive file, can be used to identify the excluded rows in the `KDDPETest.txt` file.

Candidate generation and labeling

Each case was processed through a candidate generator to identify potential PE candidates. A total of 4429 candidates were identified in the candidate generation process: 3038 candidates appear in the training set, and 1391 appear in the test set. Each candidate is a cluster of voxels (the 3-D analog of pixels) with gray values for each voxel in the cluster.

Each candidate was then labeled as a PE or not based on proximity to a 3-D landmark provided by the expert. That is, any candidate that was found to be within a certain distance of the expert's mark was labeled as a PE. Since PEs are not perfect spheres, but rather irregular objects, candidates that are not located on a PE, but are in close proximity, may be (incorrectly) labeled as a PE simply based on its location. In other words, the labeling may be noisy. Also, note that multiple candidates often correspond to a single PE (the same mark from the expert). Since each PE has a unique identifier, there may exist multiple candidates with the same PE identifier. In other words, this problem is a multiple-instance problem, where each positive example has multiple instances.

Feature generation

For each candidate, a set of 116 features are calculated. Three of the features are the *x*, *y*, and *z* locations of the candidate. The remaining features are image-based features and are normalized to a unit range, with a feature-specific mean. Note that these features are not necessarily independent, and may be correlated with other features.

The features can be categorized into three groups: those that are indicative of voxel intensity distributions within the candidate, those that measure intensity distributions in the neighborhood of the candidate, and those that describe the 3-D shape of the candidate.

Data formatting

We provide two text files that contain the training and test feature matrices, respectively, where each row represents an example, each column represents a feature. The 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), and 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). The test data will be made available by July 10, 2006.

Classification Tasks

Task 1: The first classification task is to label individual PE's. 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 PE's as possible, subject to a constraint on false positives.

For this task, we make the following definitions:

a) **PE sensitivity** is defined as the number of PE's 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*. **Note:** identifying 2 or more candidates for the same PE makes no impact on the sensitivity.
b) **False positives** are defined as *the number of candidates falsely labeled as a PE in a patient* – i.e., the total of all negative candidates labeled as PEs in the patient.
c) The **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 PE's marked by a physician and a total of seventeen candidates. Assume the first PE has 5 candidates associated with it, and the second PE has 3 candidates (8 positive labels with two PE-id's 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 ten 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, your job is to *produce a classifier that maximizes sensitivity, subject to a threshold on maximum allowable FPs.* (I.e., maximize a Neyman-Pearson criterion.) If, in any test set, your classifier exceeds the maximum allowable *average FP rate* for that sub-task, your results will be completely disqualified for the entirety of Task 1. *Your classifiers* must *meet the specified average FP threshold for* all three sub-tasks*, or your entire submission will be disqualified from Task 1.*

Example 3: The FP threshold for Task 1a 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 will be given for predictors that perform better than this FP metric, say 1 false positive per patient (though achieving, say 1 FP/patient at a high sensitivity, would be extremely valuable clinically speaking!).

You may (probably should) use different classifiers for each sub-task below:

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 classifiers will be ranked based on PE sensitivity, as long as the false positive rate meets the specified threshold.

Task 2: The second classification task is to label each *patient* as having a PE or not. The reason 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's. For this task, we make the following definitions:

a) **Patient sensitivity** is defined as *the number of patients for whom **at least one** true PE is correctly identified*. As above, a PE is correctly identified if any one of the candidates associated with that PE is correctly labeled, and multiple correct identifications in a single patient do not increase the sensitivity score.
b) **False positives** are defined as *the number of candidates falsely labeled as a PE in a patient*.
c) The **average FP rate** for a test set is the *average number of FPs produced across all patients in that test set*.

Example 1: Again consider a patient with 2 PE's 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 classifiers will be ranked based on patient sensitivity as long as the false positive rate obeys the specified FP rate. You may use the same classifier(s) as in Task 1, or build different classifiers for this task.

Task 3: One of the most useful applications for CAD would be a system with very high (100%?) Negative Predictive Value. In other words, if the CAD system had zero positive candidates for a given patient, we would like to be *very* confident that the patient was indeed free from PE's. In a very real sense, this would be the "Holy Grail" of a PE CAD system.

Unfortunately, as our training set contains relatively few negative PE cases (20 in all), building such a classifier may be a very hard task. However, we anticipate that we will have a larger number of negative cases by the time the test data set is released, allowing a better measure of the performance of the system on this task.

For this task, we make the following definitions:

a) A patient is identified as ***negative*** *when the CAD system produces no positive labels for any of that patient's candidates*.
b) The ***negative prediction value*** *(NPV) for a classifier is TN/(TN+FN)* (i.e., number of true negatives 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 PE's). The primary criterion will be 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 will be the sensitivity on PE's (as defined in Task 1), followed by the false positive rate on the entire test set.

Challenges and Considerations

In addition to the multiple instance labeling problem (more than one mark / PE), this classification task contains a number of different challenges, including the following:

1. The metrics used for evaluation (PE sensitivity, patient sensitivity, and false positives) are not the same as "traditional" sensitivity/specificity measures used in other classification problems. These metrics are more tuned to the clinical needs of physicians for decision support.
2. Candidate labeling may be noisy, as candidates are labeled as PEs based on proximity to a ground truth mark by an expert. Since PEs are not spherical, but

rather irregularly shaped, some candidates may be labeled as PEs even if they do not lie on the PE.
3. Positive candidates are spatially correlated and should not be considered to be IID.
4. Many of the features generated for each classifier are correlated.
5. The data is very imbalanced.
6. 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 cost several million dollars to collect, label, and build the features, all while maintaining strict patient confidentiality as per legal and ethical requirements.

Many of these challenges are true of other machine learning problems in the medical domain. Appendix A describes some other specific challenges for machine learning problems that are being tackled.

Rules

Eligibility

The contest is open to any party planning to attend KDD 2006. Each of the three tasks will be evaluated separately; you can enter as many tasks (or as few tasks) as you like. A person can participate in only one group per task.

Registration

Each participating group must register with the competition in order to gain access to the training data. The registration must indicate a single "group lead" who will be point of contact for the group. Each registered group lead will be subscribed to the KDD Cup 2006 mail list. The mail list will be used for contact with the participating groups and to announce rule clarifications, availability of additional data, etc. Groups are also encouraged to use this list to post questions and hold discussions.

Participation in tasks

This year's KDD Cup consists of three different tasks. A group may choose to submit to any or all of these tasks. Performance in one task will not positively or negatively impact the evaluation of performance in a different task. If a group chooses to participate in Tasks 1 or 2, they *must* submit results for all of the sub-tasks. The decision to participate in a given task will be made during results submission; a group does *not* have to commit to any particular tasks when registering to receive the training data.

Test data

The same testing data set will be shared among all three tasks. Groups will simply provide different labelings of that test data, depending on which task they are submitting to.

The testing data is sequestered and will be made available nearer the end of the competition. Details on the submission and evaluation process will be posted and announced soon.

Data format

The training data will consist of a single data file plus a file containing field (feature) names. Each line of the file represents a single *candidate* and comprises a number of whitespace-separated fields:

Field 0: Patient ID (unique integer per patient)

Field 1: Label – 0 for "negative" – this candidate is not a PE; >0 for "positive" – this candidate is a PE

Field 2+: Additional features

Semantic information on additional features may become available during the contest.

The testing data will be in the same format data file, except that Field 1 (label) will be "-1" denoting "unknown".

Evaluation

Each submission will be evaluated according to the criteria set forth under each task, above. The winner for each task will be the group with the best score according to the specified metric for that task. In the event of a tie, multiple winners may be awarded or, at the chair's option, a tie-breaking metric may be employed. Results of the competition will be announced individually to participants in advance of KDD; public announcement of results will be during the opening ceremony of KDD.

Timeline

May 15: Release of KDD Cup Specification v. 1.0; availability of training data.
July 10: Submission mechanism open.
July 17: End of results submission phase.
Aug 1: Results announced.
Aug 23-26: KDD

Appendix A: Other Machine Learning Applications and Challenges at Siemens

Virtually all the problems listed below share most of the challenges (1-6) listed above. Below, we share more high-level information from just four of the many interesting machine learning and data mining problems that are being tackled at the Computer-Aided Diagnosis & Therapy Group, Siemens Medical Solutions, Malvern, PA.

1. Colon: We are developing CAD algorithms to identify polyps in the colon with a high sensitivity, while the producing fewer than 2 false positives per patient.

Challenge: Since the CG Sensitivity has to be high, many candidates are generated from each volume. Further, due the spatial proximity, both the features and the labels of the candidates are highly correlated. While traditional classification algorithms tend to assume that the samples are drawn independently and identically from an underlying distribution, explicitly accounting for these inter sample correlations can improve the diagnostic accuracy. Furthermore, each patient often has two CT's taken in different positions, and combining information from these two sources taking into account the movement of the colon and within the colon is very complex.

2. Mammography: In order to improve the detection of early stage breast cancer, we are designing CAD algorithms to identify lesions containing masses or clusters of micro-calcifications.

Challenge: The training set for learning micro-calcification classifiers contains several hundred thousand samples, presenting computational problems for most modern algorithms for training classifiers. Secondly, the radiologist is able to correlate information across multiple mammographic views (CC/MLO) obtained from different orientations. However, building registration algorithms to perform this remains a difficult challenge for computer vision algorithms since the soft breast tissue is compressed from different orientations in each view. Building statistically rigorous classification algorithms that can optimally combine multiple X-ray views and data from other modalities such as MRI etc remains a significant challenge.

3. Identifying patients for clinical trials: We develop algorithms to identify a small set of patients who satisfy a variety of medical criteria (e.g. does the patient have high cholesterol levels, but NOT high blood pressure, and NOT taking any of a set of interacting medicines). These patients be eligible to participate in the clinical trial conducted by manufacturers of drug or medical device in order to obtain FDA approvals.

Challenge: Medical information is available from a variety of structured sources such as billing records, records of drug purchases, demographic databases. However much of the information can only be accessed from doctors' notes in the form of unstructured text. Thus we are presented with several challenges. First, the accuracy of NLP algorithms for extracting information from text are inherently limited. Second, the medical data from several different data sources needs to be combined, in order to address clinical questions (eg does the patient have high blood pressure?). Third, the data is often incomplete, lacking several pieces of information that has to be inferred from other statements.

Fourth, the data may be internally inconsistent due to temporally non-stationary: the patient may have been healthy historically, then developed an infection, received treatment, and been cured; the statistically information about him changes as a function of time. Finally, the number of patients who satisfy all the eligibility criteria for these trials is extremely small: often no more than a handful of patients are eligible out of more than 5 Million patients at a large hospital's database.
However, we are able to utilize the large redundancy of information from multiple data sources over a period of time in order to accurately identify the small list of eligible patients from a large pool of patients in the database with greater > 50% accuracy.

Appendix B: Definitions

Computed tomography (CT) - a medical imaging method to generate a three-dimensional image of the internals of an object from a large series of two-dimensional X-ray images taken around a single axis of rotation.

CT Angiography (CTA) - a procedure that uses specially designed x-rays and intravenous contrast to see the detailed anatomy of the blood vessels throughout the body. It is most frequently utilized in the evaluation of arteries in the head, neck, chest, abdomen and legs.