KDD Cup 2008 and the Workshop on Mining Medical Data

R. Bharat Rao
Siemens Medical Solutions
Malvern, PA
bharat.rao@siemens.com

Oksana Yakhnenko
Iowa State University
Ames, IA 50010
oksayakh@cs.iastate.edu

Balaji Krishnapuram
Siemens Medical Solutions
Malvern, PA
balaji.krishnapuram@siemens.com

ABSTRACT

In this report we summarize the KDD Cup 2008 task, which addressed a problem of early breast cancer detection. We describe the data and the challenges, the results and summarize the algorithms used by the winning teams. We also summarize the workshop on Mining Medical Data held in conjunction with SIGKDD on August 24, 2008 in Las Vegas, NV that brought together researchers working on various aspects of applying machine learning and data mining to challenging tasks in medical and health care domains.

1. INTRODUCTION

KDD CUP 2008

The KDD Cup 2008 was held between May and July 2008. The focus of KDD Cup 2008 was early breast cancer detection from mammogram images using computer-aided detection (CAD). The focus of the challenge was the identification of malignant mass lesions in X-ray images of the breast.

In addition to being a life-saving task, early breast-cancer diagnosis from mammograms poses several theoretical challenges to the data mining and machine learning communities. The dataset is highly imbalanced—on average only about 5 out of every 1000 asymptomatic women undergoing routine annual screening have breast cancer. Since breast cancer is typically diagnosed using four images (Mediolateral Oblique and Cranio-Caudal views of each breast), CAD algorithms can utilize information from multiple images views simultaneously, but they have to figure out how to correlate information across these images. The measures of accuracy relevant for this problem are quite different from the usual $0-1$ loss that is minimized in the machine learning literature. The data is not independently and identically distributed (iid). The cost of acquiring medical images limits the amount of available training data, so transfer learning and other domain adaptation techniques become crucial to successful design of CAD systems. Hand-engineered systems are extremely labor intensive to develop, and the run time of CAD systems has to be low so architectures like the Viola-Jones cascade are very attractive.

In order to focus attention on some specific issues within a limited timeframe, the KDD Cup presented two challenges. The first challenge was to maximize area under the Free-response Receiver Operating Characteristics (FROC) curve in the clinically significant region of $0.2-0.3$ False positives per image. This task ensures that the proposed CAD systems maximize their sensitivity of cancer detection while maintaining a false positive rate that user-surveys have determined to be acceptable to radiologists. The second challenge was to reduce the workload for a radiologist by preparing lists of potentially cancerous patients for further review while ensuring that no patient with a malignancy is missed. We are grateful to Siemens Medical Solutions USA for generously providing data features and the labels. After learning from a training set, the participants were asked to provide results for a held out test dataset for which they were not provided ground truth labels. We received over 200 submissions for each challenge, and two teams were declared joint winners: the IBM Predictive Modeling Group, and National Taiwan University.

Workshop on Mining Medical Data

In addition to KDD Cup 2008 we solicited submissions for Workshop on Mining Medical Data, held in conjunction with SIGKDD 2008 on August 24, 2008. The goal of the workshop was to bring together researchers to discuss challenges that arise in mining medical data, as well as discuss directions for future research. We received over 20 high quality submissions out of which we selected 4 papers for presentation at the workshop.

2. KDD CUP 2008: BACKGROUND AND TASKS

We first describe the KDD Cup 2008 challenge, and then summarize the results received, introduce the winners and summarize the strategies used by the winning teams.

2.1 Breast Cancer background

Breast cancer is a disease in which malignant (cancer) cells form in the tissues of the breast. Breast cancer is the second leading cause of cancer deaths in women today (after lung cancer) and is the most common cancer among women, except for skin cancers. About 1.3 million women are expected to be diagnosed annually with breast cancer worldwide, and about 465,000 will die from the disease. In the United States alone, in 2007 an estimated 240,510 women were expected to be diagnosed with breast cancer, and 40,460 women are expected to have died from breast cancer [1].

2.2 Screening for breast cancer

Screening refers to the task of looking for cancer in asymptomatic people i.e., before a person has any symptoms of the disease. Cancer screening can help find cancer at an early stage. When abnormal tissue or cancer is found early, it is often easier to treat. By the time symptoms appear, cancer may have begun to spread. The good news is that breast cancer death rates have been dropping steadily since 1990, both because of earlier detection via screening and better treatments.

The most common breast cancer screening test is a mammogram. A mammogram is an x-ray of the breast. The ability of a mammogram to find breast cancer may depend on the size of the tumor, the density of the breast tissue, and the skill of the radiologist. The mammogram is considered the standard of care for most asymptomatic women. For instance, in the US , insurance companies routinely reimburse for an annual screening mammography examination, for all asymptomatic women over the age of 40. These exams are credited with reducing the breast cancer death rate by approximately 30% since 1990.

However, the reading of screening mammograms is challenging. Findings on a screening mammogram leading to further recall are identified in approximately 5%-10% of patients, even though breast cancer is ultimately confirmed in only three to ten cases in every 1,000 women screened. Perhaps even more importantly, there is compelling evidence that many breast cancers detected at screening mammography are, in retrospect, visible on the previously obtained mammograms but have been missed by the interpreting radiologist in the prior year. There are several reasons for this: The complex radiographic structure of breast tissue, particularly in dense breasts; the subtle nature of many mammographic characteristics of early breast cancer; human oversight; poor quality films and even fatigue or distraction are all reasons why cancer is not detected by mammography.

To overcome the known limitations of human observers, second (i.e. double) reading of screening mammograms by another radiologist has been implemented at many sites. Studies indicate a potential 4%-15% increase in the number of cancers detected with double reading. In a radiology practice that performs 10,000 screening examinations per year, generally between 30-100 cancers per year will be detected. Thus, double reading in this practice could contribute to the diagnosis of 1-15 additional cancers per year. However, this approach results in a doubling of the radiologist-effort so it is not financially viable.

2.3 Computer-aided detection

Rapid and continuing advances in computer technology, as well as the ready adaptation of radiology images to digital formats, have increased the interest in computer prompting to enable the attending radiologist to act as his or her own second reader. One very promising adaptation of computer-prompting technology is computer-aided detection (CAD) in screening mammography. Current CAD systems demonstrate a high rate of detecting cancerous features on mammograms, but further improvements in both sensitivity and specificity would lead to tremendous benefits both in terms of lives saved each year, and in terms of reduction n the workload of radiologists. For the last 8-10 years, US insurance companies have begun to provide additional reimbursement to mammographers who run CAD algorithms on the mammograms in other words, physicians are now reimbursed for running a machine learning algorithm to help them better detect cancer.

In an almost universal paradigm, the CAD problem is addressed by a 4 stage system:

1. candidate generation which identifies suspicious unhealthy candidate regions of interest (candidate ROIs, or simply candidates) from a medical image;

2. feature extraction which computes descriptive features for each candidate so that each candidate is represented by a vector x of numerical values or attributes;

3. classification which differentiates candidates that are malignant cancers from the rest of the candidates based on x;

4. visual presentation of CAD findings to the radiologist

In this KDD Cup, we focus on stage 3, learning the classifier to differentiate malignant cancers from other candidates.

2.4 Data

A breast cancer screen typically consists of 4 x-ray images; 2 images of each breast from different directions (these views are called MLO and CC). Thus, most (but not all) patients would have MLO and CC images of both their breasts, giving a total of 4 images per patient. For the purposes of the KDD Cup, each image is represented by several candidates (see stage 1 above). For each candidate, we provided the image ID and the patient ID, (x, y) location, several features, and a class label indicating whether or not it is malignant. We provide features computed from several standard image processing algorithms 117 in all but due to confidentiality reasons we were unable to provide some additional proprietary features. The labels indicated whether a candidate was malignant or benign (based on either a radiologists interpretation or a biopsy or both). Note that several candidates could correspond to the same lesion. Thus, we also provided a unique lesion-ID for the malignant lesions in the training data. The lesion-ID information was not included in the test data. The classification algorithm has to automatically learn how to correlate suspicious regions across the images for test patients.

Training Data:

To support this KDD Cup challenge, training information was provided for a set of 118 malignant patients (patients with at least one malignant mass lesion). We also included data from 1594 normal patients where all candidates are presumed to be benign. The training set consisted of a total of 102,294 candidate ROIs, each described by 117 features, but only an extremely small fraction of these 102,294 candidates was actually malignant. where all candidates are presumed to be benign. The training set consisted of a total of 102,294 candidate ROIs, each described by 117 features, but only an extremely small fraction of these 102,294 candidates was actually malignant.

Test Data:

Data from over 1000 patients was provided without any information about labels.

2.5 Challenges

We conducted two different yet closely related challenges based on this data. On the test data, the participants were asked to return scores for each candidate lesion (for challenge 1) and a score for each test patient (for challenge 2).

2.5.1 Challenge 1

The rate of prevalence of malignant patients in a screening environment is extremely low (on average only around 5-10 patients out of 1000 screening patients have breast cancer). Therefore, in the first challenge, the participating entries were judged in terms of the area under the FROC curve [2] in the clinically relevant region 0.2-0.3 False positives per image. To support this, the participants have to assign a confidence score for every candidate of the test set that indicates the confidence of their classifier that the candidate is malignant. A high score of corresponds to absolute confidence that the candidate region is malignant, and a low score indicates absolute confidence that the candidate region is benign.

2.5.2 Challenge 2

In the second challenge, our aim is to reduce the workload for radiologists, by asking them to only read a subset of cases that the algorithm deems at least somewhat unclear or suspicious. Thus our second challenge is evaluated in terms of the fraction of patients who are labeled as completely normal (not requiring radiologist review of images) such that the CAD algorithms have a 100% sensitivity of the malignant patients. CAD systems which failed to have a 100% sensitivity were disqualified from the challenge. To support this challenge, the participants had to indicate whether each patient should be reviewed by a radiologist, or not.

2.6 Hints provided

The obvious method of classification is to try to build classifiers that simply label each candidate independently. We also provided several ideas for the participants to potentially improve their algorithms.

Leverage two views of the same breast: Almost always, a cancerous lesion is visible in both views (MLO, CC) of the breast radiologists routinely try to correlate the two views while diagnosing the patient. In rare cases, however, some lesions may only be visible in one view, especially in certain areas of the breast. However, negative candidates may either be present in one view (e.g., for image artifacts) or in both views (e.g., if generated by the presence of benign cyst).

Unfortunately, since each view is a 2D image obtained from an orthogonal direction, it is not possible to perfectly register (i.e., correlate the locations across) the X-ray images using simple algorithms, e.g., using affine transformations. However, some of a lesions features are typically preserved across the two views; particularly, the distance of a lesion from the nipple, and perhaps some of the features themselves relating to size of the lesion, texture, etc. Thus the first idea that may be useful for this challenge is to develop algorithms that simultaneously classify candidates from a pair of images from the same breast. These algorithms could try to exploit correlations in classification decisions for the same region of a breast. To support this, training and testing data sets included features that identify the (x, y) location of the nipple as well as the (x, y) location of the candidate.

Various researchers have combined multiple views of a mammogram to improve detection accuracy [14; 12; 10; 8; 7; 3; 5; 4; 6; 11].

Class Imbalance: Participants will be able to leverage ideas from classifier design under extreme class imbalance (the vast majority of the regions are normal, and only a small fraction of the regions are actually malignant), and feature selection (a large number of features are proposed and several of them may not be very useful for the task). The prevalence rate (malignant patients as a fraction of all patients) may differ between the training and testing sets.

Exploit correlations within an image: Participants may develop novel algorithms for exploiting potential correlations between the diagnoses of suspicious regions within a single image (e.g. if they are spatially adjacent) [13; 9].

Optimize AUC only in narrow FP range: It may be useful to develop training algorithms to maximize the area under the ROC curve (AUC) in a clinically relevant false positive (FP) range, a problem that has not been adequately addressed in the machine learning/data-mining current literature.

2.7 Submission statistics and summary of the results

Around 700 groups registered for the Challenge and downloaded training and test data. Overall, we received very high quality submissions and results, comparable with several commercially available MammoCAD systems. For Challenge 1 we received around 240 submissions out of which 3 submissions obtained area-under-curve score of over > 0.09 – in other words they had an average sensitivity of over 90% for the clinically relevant FP range of 0.2-0.3 average false positives per image. Further, 66 submissions obtained AUC in the range of (0.08, 0.09), and 94 submissions obtained AUC in the range of (0.07, 0.08).

We received around 200 submissions for Challenge 2. This was a much more difficult challenge and out of these submissions, 22 submissions actually qualified by obtaining the necessary 100% sensitivity; however some submissions have achieved this by classifying all candidates as positive and thus failed to help radiologists. Out of the qualifying teams, 1 team obtained specificity of over 60% (IBM Predictive Modeling Group). 2 teams achieved specificity 10%. Out of other submissions, 6 teams achieved sensitivity of over 90% with specificity of over 60% (however they did not qualify as they did not achieve the needed perfect sensitivity).

2.7.1 Winners for Challenge 1 and Challenge 2

Two teams were able to achieve consistently high scores for both challenges and their results were very close. Both of these teams were well ahead of their competitors. One of the teams discovered a statistical correlation between the a subset of the patient ID range and malignancy of the patient. To our knowledge, this was the only team which was able to detect and take advantage of such correlation. Since using patient ID to predict cancer is unrealistic in real life, in order to be fair to all submissions, we declared 2 winning teams.

The winning teams were **Predictive Modeling Group, IBM Research** that consisted of Claudia Perlich, Prem Melville, Grzegorz Swirszcz, Yan Liu, Saharon Rosset and Richard Lawrence; and **National Taiwan University** whose members were Hung-Yi Lo, Chun-Min Chang, Tsung-Hsien Chiang, Cho-Yi Hsiao, Anta Huang, Tsung-Ting Kuo, Wei-Chi Lai, Ming-Han Yang, Jung-Jung Yeh, Chun-Chao Yen and Shou-De Lin.

The two winning teams are publishing a detailed report of their approaches in this issue.

3. WORKSHOP ON MINING MEDICAL DATA

The workshop on Mining Medical Data invited the submission of papers related to mining medical data. We also encouraged winners of the KDD Cup 2008 to submit papers to this workshop describing their entry. However, the workshop was broader in scope, and we also welcomed other submissions related to the mining of medical data from structured sources such as structured databases and from unstructured data sources such as medical images, textual notes, etc. We particularly invited papers describing systems that are able to combine all available patient information whether from structured sources or from unstructured sources, to support medical decision making.

All submitted papers were evaluated by the workshop program committee based on scientific merits and novelty as perceived by the committee. Accepted papers were presented on August 24, 2008 and appeared in the workshop proceedings.

We received over 20 submissions out of which 4 papers were accepted to be presented at the workshop. Although we received several additional submissions of a high quality, due to time constraints on our half day workshop we were unable to accept several intriguing and exciting papers.

3.1 Contributed talks at Workshop on Mining Medical Data

3.1.1 A Data-Mining Framework for Classification of High Resolution Magnetic Resonance Images

Magnetic resonance imaging (MRI) allows the display of brain structures with high resolution. To fully exploit the potential of this imagining modality, data mining methods are required to reveal subtle differences in brain structure caused by disorders such as Mild Cognitive Impairment (MCI) and early stage Alzheimers disease (AD). The paper by Christian Bhm, Annahita Oswald, Claudia Plant and Bianca Wackersreuther discussed a data mining framework which combined elements from feature selection, clustering, classification to provides a concise visualization of affected areas in the brain.

3.1.2 Large-Scale Regression-Based Pattern Discovery in International Adverse Drug Reaction Surveillance

The paper by Ola Caster, Niklas Norn, David Madigan and Andrew Bate demonstrated the first use of shrinkage logistic regression as a pattern discovery method for the international adverse drug reaction surveillance conducted by the World Health Organization (WHO). This novel method is compared to bivariate pattern discovery, the standard approach for detecting adverse drug reactions. The authors results showed that their approach can eliminate false positives and false negatives by using information from other covariates. It was impressive to see that their approach could have detected drug safety issues earlier than the standard approach (as validated on retrospective data). However, acknowledge that their approach cannot completely replace bivariate methods, for two reasons: it fails to identify some established drug safety concerns; and there is a loss of transparency which makes the model difficult to interpret for humans. They suggest that their method should be used in parallel with the existing method to detect adverse drug reactions.

3.1.3 Boosting Framework for Biomedical Image Retrieval

Understanding anatomical structure and automatically finding and extracting features is a challenging research problem. Chandan Reddy and Fahima Bhuyan discuss a principled boosting based framework that accomplishes this for biomedical image retrieval. Instead of hand-designing features based on domain knowledge, an adaptive boosting algorithm is employed to automatically identify features and create models for different categories of biomedical images. Experimental results on different biomedical image categories show the robustness and accuracy of the proposed system. Using some standard performance metrics, the paper also provides some insights about the complexity of these image categories.

3.1.4 CARE for Your Future: Prospective Disease Prediction Using Collaborative Filtering

The cost of health care, especially for chronic disease treatment, is quickly becoming unmanageable. This crisis has motivated the drive towards preventive medicine, where the primary concern is recognizing disease risk and taking action at the earliest signs. Darcy Davis, Nitesh Chawla, Nicholas Blumm, Nicholas Christakis and Albert-Laszlo Barabasi address this fascinating topic with novel solutions in their paper. Universal testing is neither time nor cost efficient. The authors propose CARE, a Collaborative Assessment and Recommendation Engine, which relies only a patients medical history using ICD-9-CM codes in order to predict future diseases risks. CARE combines collaborative filtering methods with clustering to predict each patients greatest disease risks based on their own medical history and that of similar patients. The authors also describe an Iterative version, ICARE, which incorporates ensemble concepts for improved performance. The authors present experimental results on a large Medicare dataset, demonstrating that CARE and ICARE perform well at capturing future disease risks.

4. RESOURCES

The KDD Cup 2008 and Workshop on Mining Medical Data website is available at `www.kddcup2008.com`. The website contains training and test features and labels for the training data; code for evaluation of FROC curve and proceedings to the KDD Cup 2008 and Workshop on Mining Medical Data.

5. CONCLUSION

The 2008 KDD Cup was an exciting competition that addressed an important and interesting problem with data from a real-world medical application. A large number of teams participated in the challenge, and both winning approaches developed novel theoretical approaches. Overall, the competition was quite strong and we had a number of truly excellent submissions that were comparable to commercially available CAD systems in terms of their diagnostic accuracy. It is clear from the quality of results that all of the competitors contributed a great amount of effort and creativity into the competition, and we thank them for their efforts.

The Workshop on Mining Medical Data was a successful event that brought together researchers in Data Mining who address important problems in medical domain (such as MRI image analysis and predictive modeling of diseases). At this event the KDD Cup winners also presented their results and algorithms. The workshop was attended by over 70 people and many of whom were actively involved in discussions. The discussions were interesting and often quite passionate, and suggested a number of interesting and important directions for future work. We hope that this workshop will spur continued research into the many challenges that need to be addressed by medical decision support systems.

6. REFERENCES

[1] American Cancer Society, Surveillance Research, 2007.

[2] D. P. Chakraborty. Maximum likelihood analysis of free-response operating characteristic (FROC) data. *Medical Physics*, 16:561–568, 1989.

[3] Walter F. Good, Bin Zheng, Yuan-Hsiang Chang, Xiao Hui Wang, Glenn Maitz, and David Gur. Multi-image cad employing features derived from ipsilateral mammographic views. *SPIE Vol. 3661*, 3661:474–485, 1999.

[4] Shalini Gupta, Priscilla F. Chyn, and Mia K. Markey. Breast cancer cadx based on bi-rads descriptors from two mammographic views. *Medical Physics*, 33:1810–1817, 2006.

[5] Shalini Gupta and Mia K. Markey. Correspondence in texture features between two mammographic views. *Medical Physics*, 32:1598–1606, 2005.

[6] Bei Liu, Charles E. Metz, and Yulei Jiang. Effect of correlation on combining diagnostic information from two images of the same patient. *Medical Physics*, 32:3329–3338, 2005.

[7] P. Lucas N. de Carvalho Ferreira, M.Velikova. Bayesian modelling of multi-view mammography. In *Proceedings to ICML workshop "Machine Learning for Health Care Applications"*, 2008.

[8] Wei Qian, Dansheng Song, Minshan Lei, Ravi Sankar, and Edward Eikman. Computer-aided mass detection based on ipsilateral multiview mammograms. *Academic Radiology*, 14:530–538, 2007.

[9] Vikas Raykar, Balaji Krishnapuram, Murat Dundar, Jinbo Bi, and R. Bharat Rao. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *Proceedings of the 2008 International Conference on Machine Learning (ICML)*, 2008.

[10] Berkman Sahiner, Heang-Ping Chan, Lubomir M. Hadjiiski, Mark A. Helvie, Chinatana Paramagul, Jun Ge, Jun Wei, and Chuan Zhou. Joint two-view information for computerized detection of microcalcifications on mammograms. *Medical Physics*, 33:2574–2585, 2006.

[11] Alain Tiedeu, Christian Daul, Pierre Graebling, and Didier Wolf. Correspondences between microcalcification projections on two mammographic views acquired with digital systems. *Computerized Medical Imaging and Graphics*, 29:543553, 2005.

[12] Saskia van Engeland and Nico Karssemeijera. Combining two mammographic projections in a computer aided mass detection method. *Medical Physics*, 34:898–905, 2007.

[13] Volkan Vural, Glenn Fung, Balaji Krishnapuram, and Jennifer Dy. Batch classification with applications to computer aided diagnosis. In *Proceedings of the 2006 European Conference on Machine Learning (ECML)*, 2006.

[14] Bin Zheng, Joseph K. Leader, Gordon S. Abrams, Amy H. Lu, Luisa P. Wallace, Glenn S. Maitz, and David Gur. Multiview-based computer-aided detection scheme for breast masses. *Medical Physics,*, 33:3135, 2006.