Analysis of Mammography Reports Using Maximum Variation Sampling

Robert M. Patton
Oak Ridge National Laboratory
P.O. Box 2008 MS 6085
Oak Ridge, TN 37831-6085
1-865-576-3832
pattonrm@ornl.gov

Barbara Beckerman
Oak Ridge National Laboratory
P.O. Box 2008 MS 6085
Oak Ridge, TN 37831-6085
1-865-576-2681
beckermanbg@ornl.gov

Thomas E. Potok
Oak Ridge National Laboratory
P.O. Box 2008 MS 6085
Oak Ridge, TN 37831-6085
1-865-574-0834
potokte@ornl.gov

ABSTRACT
A genetic algorithm (GA) was developed to implement a maximum variation sampling technique to derive a subset of data from a large dataset of unstructured mammography reports. It is well known that a genetic algorithm performs very well for large search spaces and is easily scalable to the size of the data set. In mammography, much effort has been expended to characterize findings in the radiology reports. Existing computer-assisted technologies for mammography are based on machine-learning algorithms that must learn against a training set with known pathologies in order to further refine the algorithms with higher validity of truth. In a large database of reports and corresponding images, automated tools are needed just to determine which data to include in the training set. This work presents preliminary results showing the use of a GA for finding abnormal reports without a training set. The underlying premise is that abnormal reports should consist of unusual or rare words, thereby making the reports very dissimilar in comparison to other reports. A genetic algorithm was developed to test this hypothesis, and preliminary results show that most abnormal reports in a test set are found and can be adequately differentiated.

Categories and Subject Descriptors
H.4.2 [**Types of Systems**]: Decision Support

General Terms
Algorithms

Keywords
Unstructured radiology reports, text analysis, genetic algorithms, maximum variation sampling

1. INTRODUCTION
Currently, no automated means of detecting abnormal mammograms exist. While knowledge discovery capabilities through data mining and data analytics tools are widespread in many industries, the healthcare industry as a whole lags far behind. Providers are only just beginning to recognize the value of data mining as a tool to analyze patient care and clinical outcomes [4]. Other work is being done in the medical environment to use automated software tools to extract knowledge from unstructured radiology reports [3]. Preliminary findings demonstrate that automated tools can be used to validate clinically important findings and recommendations for subsequent action from unstructured radiology reports. Commercially available software is also being tested to automate a method for the categorization of narrative text radiology reports, in this case dealing with the spine and extremities [20]. The research conducted by the authors investigates the use of unstructured reports for mammography to test the hypotheses that genetic algorithms can differentiate between implied assessments and radiological interpretation in radiology reports, which can be later correlated to the images for extraction and testing.

In mammography, much effort has been expended to characterize findings in the radiology reports. Various computer-assisted technologies have been developed to assist radiologists in detecting cancer; however, the algorithms still lack high degrees of sensitivity and specificity, and must undergo machine learning against a training set with known pathologies in order to further refine the algorithms with higher validity of truth. In a large database of reports and corresponding images, automated tools are needed just to determine which data to include in the training set. Validation of these data is another issue. Radiologists disagree with each other over the characteristics and features of what constitutes a normal mammogram and the terminology to use in the associated radiology report. Abnormal reports follow the lexicon established by the American College of radiology Breast Imaging Reporting and Data System (Bi-RADS) [1], but even within these reports, there is a high degree of text variability and interpretation of semantics. The focus has been on classifying abnormal or suspicious reports, but even this process needs further layers of clustering and gradation, so that individual lesions can be more effectively classified. The tools that are needed will not only help further identify problem areas but also

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-131-6/08/07.

support risk assessment and other knowledge discovery applications.

The knowledge to be gained by extracting and integrating meaningful information from radiology reports will have a far-reaching benefit, in terms of the refinement of the classifications of various findings within the reports. This will support validation, training and optimization of these and other machine learning and computer-aided diagnosis algorithms to work both in this environment and with other medical and imaging modalities. In the near-term, the overall goal of this work is to accurately identify abnormal radiology reports amid a massive collection of reports. The challenge in achieving this goal lies in the use of natural language to describe the patient's condition. The premise of this work is that abnormal radiology reports consist of words and phrases that are statistically rare or unusual. If this is true, then it is expected that abnormal reports will be significantly dissimilar in comparison to normal radiology reports.

Our goal, then, is to find an ideal sample of mammography reports that represents the diversity without applying clustering techniques or without prior knowledge of the population categories. To achieve this objective, our approach employs adaptive sampling [21][22]. This sampling technique continues to draw samples from the population based on previous samples until some criteria have been met. Previous results indicated that an ideal sample could be found very quickly using this approach [11][12].

2. BACKGROUND

The focus of this work is on text analysis of radiology reports, specifically mammography reports. There has been considerable work in a variety of areas in the text analysis community and a wide array of problems with processing and analyzing text data. Some of these areas of text analysis include retrieval, categorization, clustering, syntactic and semantic analysis, duplicate detection and removal, and information extraction to name a few [2][16][23]. These areas range from analyzing entire datasets to analyzing a single document. In general, as the size of the dataset increases, many of these approaches begin performing poorly, or the value of their results begins to diminish. For example, clustering usually requires comparing every document with every other document. Obviously, as the dataset size increases, performance will noticeably suffer. However, with categorization, the performance may not suffer considerably, but the quality of the results will be diminished if a sufficient number of categories are not identified or if the categories are not clearly or accurately identified [18].

What is needed then is a means of finding a characteristic subset from a large data set. However, there are a variety of issues in simply identifying the content and creating the actual subset to be used. Naturally, this leads into sampling techniques. Sampling can be divided into two main categories: probability-based and nonprobability-based. Probability-based sampling is based on probability theory and the random selection of data points from the dataset. Nonprobability-based sampling is based on purposeful selection, rather than random selection. The advantage of this form of sampling is that it allows the analyst to look at data that may not otherwise be visible via the random selection process. Within nonprobability-based sampling, there are several categories of sampling [10], one of which is maximum variation sampling (MVS) [10]. This particular sampling method seeks to identify a particular sample of data that will represent the diverse data points in a data set. According to Patton [10], "This strategy for purposeful sampling aims at capturing and describing the central themes or principle outcomes that cut across a great deal of [data] variation." In a large text corpus, this form of sampling provides the ability to quickly characterize the different topics, or "threads" of information that are available.

A genetic algorithm (GA) was developed to implement the maximum variation sampling technique. It is well known that a genetic algorithm performs very well for large search spaces and is easily scalable to the size of the data set. In addition, GAs are also particularly suited for parallelization [7][19]. To better understand the need for scalability and the size of the search space in this problem domain, consider a set of 10,000 radiology reports. Now, suppose an analyst needs to reduce this data set to 200 representative reports (only 2% of the entire data set). In that case, there are approximately 1.7×10^{424} different combinations of reports that could be used to create a single sample. Clearly, a brute force approach is unacceptable. In addition, many of the combinations would consist of duplicate data that would lower the quality of the result for the analysts. Ultimately, an intelligent and scalable approach such as a genetic algorithm is needed. As demonstrated by Mutalik [8], a parallel genetic algorithm is well suited to a combinatorial optimization problem.

The following sections will describe the implementation and results of implementing an MVS technique using a GA and the application to radiology reports.

3. MVS-GA DESIGN

Before applying a GA to the analysis of radiology reports, the reports must be prepared using standard information retrieval techniques. First, reports are processed by removing stop words and applying the Porter stemming algorithm [5][13][14]. Once this has been done, the articles are then transformed into a vector-space model (VSM) [15][17]. In a VSM, a frequency vector of word occurrences within each report can represent each report. Once vector-space models have been created, the GA can then be applied.

Two of the most critical components of implementing a GA are the encoding of the problem domain into the GA population and the fitness function to be used for evaluating individuals in the population. To encode the data for this particular problem domain, each individual in the population represents one sample of size *N*. Each individual consists of *N* genes where each gene represents one radiology report (each report is given a unique numeric identifier) in the sample. For example, if the sample size were 10, each individual would represent one possible sample and consist of 10 genes that represent 10 different reports. This representation is shown in the following figure.

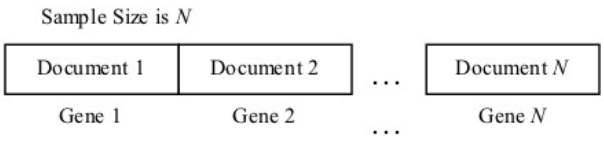

Figure 1. Genetic representation of each individual

The fitness function evaluates each individual according to some predefined set of constraints or goals. In this particular application, the goal was to achieve an ideal sample that represents the maximum variation of the data set without applying clustering techniques or without prior knowledge of the population categories. To measure the variation (or diversity) of our samples, the summation of the similarity between the vector-space models of each document (or gene) in the sample is calculated as shown in the following equation.

$$Fitness(i) = \sum_{j=0}^{N} \sum_{k=j+1}^{N} Similarity(Gene(i,j), Gene(i,k))$$

Equation 1. Fitness Function

In this equation, the Similarity function calculates the distance between the vector space models of gene *j* and *k* of the individual *i*. This distance value ranges between *0* and *1* with *1* meaning that the two reports are identical and *0* meaning they are completely different in terms of the words used in that report. Therefore, in order to find a sample with the maximum variation, Equation 1 must be minimized. In this fitness function, there will be $(N^2 - N) / 2$ comparisons for each sample to be evaluated.

The defined fitness function can be computationally intensive for large sample sizes or for data sets with lengthy news articles. To compensate for this, the GA developed for this work was designed as a global population parallel GA [9]. To implement the parallel GA, four slave threads were developed to perform fitness evaluation of the individuals in the population. The results of the evaluations were then returned to the master thread, which continued operating the GA in order to apply selection, crossover, and mutation. The use of four slaves was done to fully utilize a four-processor machine for use in this work. This parallelization made a significant reduction in the runtime of the algorithm.

To create children from a given population, genetic operators such as selection, crossover, and mutation are applied to the individuals. For each generation, an average fitness value is calculated for the population. Individuals with fitness values that are above this average are selected as parents, while the other individuals are discarded. This can be a very aggressive selection process if there are extremely fit individuals that are far above the average. Once parents are selected, crossover and mutation operators are applied to the parents to create children. The crossover and mutation operators are 1-point operators [6].

The population size was defined as 2,000 and the number of generations was set to 250. The crossover rate was set to 0.7 and the mutation rate was set to 0.03. The sample size was set to 15 and the data set size was 100 radiology reports. In this case, there are approximately 2.53×10^{17} different combinations of reports that could be used to create a single sample.

4. DATA

In this preliminary study, unstructured mammography reports from a large data set were used. These reports consisted of 9,000 patients studied over a 5-year period from 1993 to 1998. Of this large data set, a human expert manually classified 100 reports as being normal and 100 reports as being abnormal. From the normal set of reports, a random sample of 90 reports was selected. From the abnormal set, a random sample of 10 reports was selected. The two samples were then merged to create a third set of 100 reports. This third set was used to test the GA. If the premise that abnormal radiology reports consist of words and phrases that are statistically rare or unusual, then the expected outcome of the GA will be a sample of reports consisting predominantly of abnormal reports.

5. RESULTS & DISCUSSION

Thirty runs of the GA were performed. Based on these runs, the GA consistently found 8 out of 10 abnormal reports. The remainder of the sample consisted of 7 normal reports. Upon further analysis of the 10 abnormal documents, it was found that 4 of the reports were very similar to each other, while the other 6 were very distinct. Consequently, 2 of the 10 abnormal reports were consistently absent from the final sample.

Upon further analysis of the normal documents that were included in the final sample, it was determined that several of the reports represented "boundary" cases. These were reports that, while considered normal, represented situations where a patient had either already undergone a lumpectomy or had a family history of breast cancer and showed high potential for breast cancer. Other normal reports that were in the final sample consisted of patients that needed further examination and therefore underwent spot magnification for further confirmation. Another report represented a patient where the radiologist had difficulty in determining a nodule in the image and suggested that it was a "small deformable cyst." Overall, these preliminary results from the GA showed encouraging performance to find both abnormal reports and potentially unusual normal reports without prior categorization or a predefined vocabulary of terms to search.

Additional analysis of the final sample revealed another characteristic of the reports. For each report, word phrases unique to that specific report were extracted. In this case, unique word phrases are those phrases that only appear in one report in the sample. As shown in Table 1, normal reports tended to have fewer unique word phrases as compared to abnormal reports. In addition, abnormal reports tended to have more variability in the number of unique word phrases.

Table 1. Number of unique phrases for each report

Normal Reports	Abnormal Reports
18	26
15	63
11	38
14	43
16	29
0	45
23	22
--	27
Avg: 13.857	**Avg:** 36.625
Std Dev: 7.151	**Std Dev:** 13.553

Further investigation into the word phrases of the abnormal reports revealed a wide-ranging vocabulary and semantics. **Error! Reference source not found.** shows example word phrases from both normal and abnormal reports.

Table 2. Sample word phrases from reports

Normal Reports	Abnormal Reports
benign biopsy	intraductal carcinoma
breasts unchanged	rod shaped calcifications
microcalcifications identified	defined hyperdense nodule
remain unchanged	hypoechoic lesion
small deformable cyst	recommend excisional biopsy
benign macrocalcification	lobulated hypoechoic mass

Analysis of the word phrases provides further evidence to support our hypothesis that abnormal reports consists of statistically rare or unusual words, and thereby making them easier to identify in a large collection of reports.

6. SUMMARY

Currently, text analysis of mammography reports remains a significant challenge. However, solving this issue would provide numerous benefits. The work described here represents preliminary results in applying a GA to assist with identifying abnormal mammography reports from a large set of reports. Initial results were very encouraging and show tremendous potential for future work. Future work will seek to leverage this technique to develop a more advanced and specific training set of images to further enhance image-based algorithms.

7. ACKNOWLEDGMENTS

Our thanks to Robert M. Nishikawa, PhD, Department of Radiology, University of Chicago for providing the large dataset of unstructured mammography reports, from which the test subset was chosen.

8. REFERENCES

[1] American College of Radiology (ACR). ACR BI-RADS® - Mammography. 4th Edition. In: ACR Breast Imaging Reporting and Data System, Breast Imaging Atlas. Reston, VA. American College of Radiology; 2003.

[2] X. Cui and T. E. Potok, A distributed agent implementation of multiple species Flocking model for document partitioning clustering, in *Lecture Notes in Computer Science*. vol. 4149 NAI Edinburgh, United Kingdom: Springer Verlag, Heidelberg, D-69121, Germany, 2006, pp. 124-137.

[3] Dreyer, K.J., Kalra, K.M., Maher, M.M., et al, Application of Recently Developed Computer Algorithm for Automatic Classification of Unstructured Radiology Reports: Validation Study, Radiology 324, 2 (Feb. 2005), 323-329.

[4] Fickenscher, K.M., The New Frontier of Data Mining, Health *Management Technology* (26) 10:32-36, October 2005.

[5] C. Fox, "Lexical analysis and stoplists." In Information Retrieval: Data Structures and Algorithms (ed. W.B. Frakes and R. Baeza-Yates), Englewood Cliffs, NJ: Prentice Hall, 1992.

[6] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, 1989.

[7] H. Muehlenbein, "Parallel Genetic Algorithms, Population Genetics, and Combinatorial Optimization", Proc. of the Third International Conference on Genetic Algorithms, Morgan Kaufmann, 1989.

[8] P.P. Mutalik, et al., "Solving Combinatorial Optimization Problems Using Parallel Simulated Annealing and Parallel Genetic Algorithms", Proceedings of the 1992 ACM/SIGAPP symposium on Applied computing: technological challenges of the 1990's, 1992, pp 1031 – 1038.

[9] M. Nowostawski, and R. Poli, "Parallel Genetic Algorithm Taxonomy", Proc. of the Third International Conference on Knowlege-based Intelligent Information Engineering Systems, 1999, pp. 88 – 92.

[10] M.Q. Patton, Qualitative Evaluation and Research Methods, Second Edition. Newbury Park, CA: Sage Publications, Inc., 1990

[11] R.M. Patton, and T.E. Potok, "Adaptive Sampling of Text Documents," Proc. of the 13th International Conference on Intelligent and Adaptive Systems and Software Engineering, July 2004.

[12] R. M. Patton and T. E. Potok, Characterizing large text corpora using a maximum variation sampling genetic algorithm, in *Proc. of the 8th annual conference on Genetic and Evolutionary Computation* Seattle, Washington, USA: ACM Press, 2006.

[13] M. Porter, "An algorithm for suffix stripping." Program vol. 14, pp. 130-137, 1980.

[14] Porter Stemming Algorithm. Current Jan. 30, 2004. http://www.tartarus.org/~martin/PorterStemmer/

[15] V.V. Raghavan and S.K.M. Wong, "A critical analysis of vector space model for information retrieval." Journal of the American Society for Information Science, Vol.37 (5), p. 279-87, 1986.

[16] J.W. Reed, Y. Jiao, T.E. Potok, B.A. Klump, M.T. Elmore, and A.R. Hurson, TF-ICF: A new term weighting scheme for clustering dynamic data streams, In *Proc. of the 5th International Conference on Machine Learning and Applications (ICMLA'06)*, pp.258-263, 2006.

[17] G. Salton, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[18] F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Volume 34 , Issue 1 (March 2002) , pp 1 – 47.

[19] R. Tanese, Distributed Genetic Algorithms for Function Optimization, Ph.D. thesis, University of Michigan, 1989, Computer Science and Engineering.

[20] Thomas, BJ; Ouellette, H; Halpern, EF; Rosenthal, DI. Automated Computer-Assisted Categorization of Radiology Reports, AJR: 184, 687-690. February 2005.

[21] S.K. Thompson, Sampling. John Wiley and Sons, Inc., New York, 1992.

[22] S.K. Thompson and G.A.F. Seber, Adaptive Sampling. John Wiley and Sons, Inc., New York, 1996.

[23] P. Yan, Y. Jiao, A.R. Hurson, and T.E. Potok, Semantic-based information retrieval of biomedical data, In *Proc. of the 21st Annual ACM Symposium on Applied Computing – Semantic-Based Resource Discovery, Retrieval and Composition (RDRC)*, pp. 1700-1704, 2006.