

Learning Cue Phrase Patterns from Radiology Reports Using a Genetic Algorithm

Robert M. Patton, Barbara G. Beckerman, and Thomas E. Potok

Abstract— 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. This work describes an approach to learning cue phrase patterns in radiology reports that utilizes a genetic algorithm (GA) as the learning method. The approach described here successfully learned cue phrase patterns for two distinct classes of radiology reports. These patterns can then be used as a basis for automatically categorizing, clustering, or retrieving relevant data for the user.

I. INTRODUCTION

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), 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 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. 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.

Therefore, what is needed is an automated means of learning the characteristic cue phrase patterns of the natural language used in the radiology reports and using those learned patterns as a basis for automatically categorizing, clustering, or retrieving relevant data for the user. This paper describes preliminary work being performed to address the learning aspect of this approach. Section 2 will discuss the background of the radiology reports being addressed by this work. Section 3 will describe the learning approach, while section 4 discusses results. Section 5 will discuss future work.

II. BACKGROUND

This work focuses on the language domain of mammography reports. In the report, the radiologist describes the features or structures that they see or do not see in the image. Essentially, this report is meta-data that is written by a human subject matter expert about the image. In order to effectively train a computer-assisted detection (CAD) system, these reports could be mined and used as supplemental meta-data. Unfortunately, little work has been done to utilize and maximize the knowledge potential that exists in these reports.

In this preliminary study, unstructured mammography reports were used. These reports represented 12,809 patients studied over a 5-year period from 1988 to 1993. There are 61,064 actual reports in this set. Each report generally consists of two sections. The first section describes what features the radiologist does or does not see in the image. The second section provides the radiologist's formal opinion as to whether or not there are suspicious features that may suggest malignancy (i.e., or the possibility that the patient has cancer). The set of reports also includes

Manuscript received March 13, 2009. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

R. M. Patton is with the Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA (phone: 865-576-3832; fax: 865-241-0337; e-mail: pattonrm@ornl.gov).

B. G. Beckerman is with the Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA (phone: 865-576-2681; fax: 865-241-3191; e-mail: beckermanbg@ornl.gov).

T. E. Potok is with the Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA (phone: 865-574-0834; fax: 865-241-0337; e-mail: potokte@ornl.gov).

a number of reports that simply state that the patient canceled their appointment.

To provide a better perspective of the challenge of mining these reports, consider the following question. Given a database of these reports, how does one retrieve those reports that represent abnormalities in the patient? In mammography, most patient reports will represent “normal” conditions in the patient. Consequently, the reports with “abnormal” conditions are rare (defining the difference between what is “normal” and “abnormal” is beyond the scope of this paper).

The main problem of trying to find abnormal reports lies in the language that is used in mammograms. As discussed in [5], abnormal reports tend to have a richer vocabulary than normal reports. In addition, normal reports tend to have a higher number of “negation” phrases. These are phrases that begin with the word “no” such as in the phrase “no findings suggestive of malignancy.” These negation phrases generally occur in normal reports.

The goal, then, is to develop an automated approach to learning the skip bigrams (or s-grams) of cue phrases in the mammography language that sufficiently characterize the reports such that information retrieval becomes both more accurate and simplistic while, at the same time, not being computationally intensive [1],[2],[6]. S-grams are word pairs in their respective sentence order that allow for arbitrary gaps between the words. For a phrase such as “no findings suggestive of malignancy”, an s-grams would be the words “no” and “malignancy.” This s-gram uniquely identifies a particular semantic in the language of mammography reports and enables the identification of all possible variations of such phrases. Higher-level patterns may then be formed from these s-grams.

The work here describes a possible approach toward this goal of automatically learning s-grams that can provide meaningful retrieval on domain-specific data and the results achieved thus far.

III. LEARNING APPROACH

As discussed in section 2 and in [5], mammography reports exhibit two characteristics. First, abnormal reports tend to have a wider variation in the language that is used. Consequently, these reports tend not to cluster with other reports. The second characteristic is that normal reports use more negation phrases than abnormal reports. It is these two characteristics that we seek to exploit in this approach.

To exploit the first characteristic, an enhancement of the maximum variation sampling technique [5] is developed. This technique is implemented via a genetic algorithm (MVS-GA) and is discussed in the next section along with the enhancements. In addition, the work described here differs from [5] in that the MVS-GA is used to learn common phrase patterns among diverse documents and not

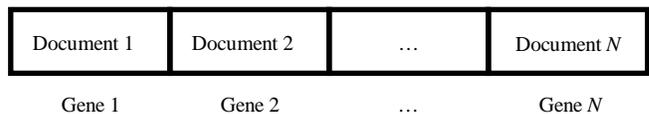
explicitly for sampling. To exploit the second characteristic, the MVS-GA is augmented with a simple memory that stores the common phrase patterns of samples that failed to survive in the MVS-GA. This will be discussed later.

A. Learning from Maximum Variation Sampling

Maximum variation sampling is a nonprobability-based sampling. This form of sampling is based on purposeful selection, rather than random selection. Since abnormal mammography reports are not as common as normal ones, random sampling would make it difficult to find them. Within nonprobability-based sampling, there are several categories of sampling [4], one of which is maximum variation sampling (MVS) [4]. This particular sampling method seeks to identify a particular sample of data that will represent the diverse data points in a data set. In this case, the diverse data points will represent abnormal mammograms. The MVS is naturally implemented as a genetic algorithm (MVS-GA).

Before applying a GA to the analysis of mammography 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 [3],[7],[8]. Once this has been done, the articles are then transformed into a vector-space model (VSM) [9],[10]. In a VSM, a frequency vector of word and phrase 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.

Fig. 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 for the fitness function was to achieve a 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.

The data set for this current work utilizes a set of 61,064 reports. Within this data are numerous reports that simply state that the patient canceled their appointment. These reports are very small in length and are exceptionally distinct from all other reports (similarity values approaching zero). Unfortunately, the MVS-GA from [5] gravitates toward these cancellation reports as the best solution for a maximum variation sample.

In an effort to effectively characterize the phrase patterns of the mammography reports, it is necessary to examine reports that are longer in length, so that more language can be examined for patterns. In addition, abnormal reports tend to be longer in length than normal reports since the radiologist is describing the abnormalities in more detail. Consequently, the fitness function of the MVS-GA was enhanced to incorporate penalty functions as shown in equations 1 – 3.

$$Fit(i) = \sum_{j=0}^N \sum_{k=j+1}^N \alpha_j + \beta_k + Sim(G_{ij}, G_{ik}) \quad (1)$$

$$\alpha_j = e^{-\left(\frac{\|j\|}{100}\right)} \quad (2)$$

$$\beta_k = e^{-\left(\frac{\|k\|}{100}\right)} \quad (3)$$

In Eq. 1, 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 indicating that the two reports are identical and 0 indicating that they are completely different in terms of the words used in that report. Therefore, in order to find a sample with the maximum variation, Eq. 1 must be minimized (i.e., lower fitness values are better). In this fitness function, there will be $(N^2 - N) / 2$ comparisons for each sample to be evaluated.

The penalty functions are incorporated into the fitness function in order to penalize individuals in the MVS-GA based on the length of the documents they represent. Shorter documents receive higher penalties while longer documents receive much lower penalties. The penalty functions also return values that are between 0 and 1, inclusive. As a result of the penalty functions, the

cancellation reports will receive the highest fitness values, while lengthy, abnormal reports will receive the lowest fitness values.

After the MVS-GA is executed, the end result is a best sample of mammography reports that are as diverse from each other as possible. Once this sample is achieved, then phrases are extracted from each document in the sample. For each phrase in the document, s-grams are extracted. Next, the s-grams are counted across the sample of documents. S-grams that are common across the sample will have higher frequency counts while s-grams with a frequency of 1 uniquely identify a particular document in the sample. For this work, only those s-grams that are the most common in the best sample found are considered valuable. It is these s-grams that have the ability to uniquely retrieve abnormal documents from a large set.

B. Learning from Failures

For this work, the MVS-GA has been augmented to store the common s-grams of the individuals that failed to reproduce children. This will enable answering questions such as what characteristic phrases make failed individuals inferior to successful individuals. After each generation, s-grams and their frequencies from each failed individual are extracted from each individual and stored in memory. After the MVS-GA has completed, the memory now contains the most common s-grams that caused individuals to fail in the GA. The end result is that the MVS-GA learns the s-grams for both abnormal and normal classes of reports.

IV. RESULTS

The s-grams discovered by this learning algorithm on the data set are shown in Tables I and II. Table I shows the top

TABLE I
TOP TEN S-GRAMS FROM MVS-GA BEST SOLUTION

| Rank | S-gram | Example | Observed Variants |
|------|-----------------------|--|-------------------|
| 1 | magnification & views | magnification views requested | 660 |
| 2 | core & biopsy | stereotactic guided core biopsy of microcalcifications | 633 |
| 3 | needle & localization | ultrasound-guided needle localization procedure | 245 |
| 4 | nodular & density | showing questionable increased nodular density | 2726 |
| 5 | lymph & node | atypically located intramammary lymph node | 748 |
| 6 | needle & procedure | stereotactic needle core biopsy procedure | 57 |
| 7 | compression & views | right anterior compression views | 772 |
| 8 | spot & views | recommended utilizing spot views | 852 |
| 9 | spot & compression | spot compression image | 1123 |
| 10 | spot & magnification | medially exaggerated right cc spot magnification | 650 |

ten s-grams from the best solution obtained by the MVS-GA. These s-grams tend to uniquely define abnormal reports. Many of these s-grams refer to procedures that are performed in the event that a suspicious feature in the patient was observed by the radiologist. For example, the patient may be asked to return with a few weeks for additional imaging such as an ultrasound and magnification imaging. In addition, patients with suspicious features may undergo biopsy, and in some cases, may also have a needle localization performed to enhance the biopsy procedure.

TABLE II
TOP TEN S-GRAMS WITH THE WORD "NO"

| Rank | S-gram | Example | Observed Variants |
|------|---------------------|--|-------------------|
| 1 | no & suspicious | no finding strongly suspicious | 1225 |
| 2 | no & calcifications | no clear cut clustered punctate calcifications | 137 |
| 3 | no & evident | no mass lesions evident | 46 |
| 4 | no & masses | no new focal masses | 365 |
| 5 | no & malignancy | no specific evidence of malignancy | 286 |
| 6 | no & residual | no residual microcalcifications | 56 |
| 7 | no & skin | no skin abnormalities noted | 68 |
| 8 | no & thickening | no skin thickening seen | 42 |
| 9 | no & complications | no apparent complications | 16 |
| 10 | no & change | no apparent interval change | 384 |

Furthermore, since breast cancer often affect the lymph nodes, radiologist look for abnormalities relating to the lymph nodes as well. As can be seen in Table I, the MVS-GA successfully learned key s-grams that would significantly enhance automated retrieval and analysis of abnormal reports.

Table II show the top ten s-grams that begin with the "no" and were learned from the failed individuals in the MVS-GA. As discussed previously, most normal reports contain some form of a "negation" phrase. These phrases refer to the non-existence of a particular feature or condition in which the radiologist was searching. Abnormal reports may contain such negation phrases, however, abnormal reports tend to be more focused on the abnormalities that were found and not the abnormalities that were not found. Consequently, the MVS-GA successfully learned from the failed samples the common s-grams of normal reports.

One of the most significant aspects of these results is that the learning algorithm did not require any specialized ontology or dictionary or feedback from a subject matter expert. This approach utilized an unsupervised, domain independent learning algorithm to achieve these results. Now that the s-grams have been learned, relevant documents can now be retrieved and analyzed. Future work

will examine the retrieval quality of this approach.

V. FUTURE WORK

While the work described here focuses primarily on the learning aspect of mining radiology reports, there are many avenues for future research. First, this work uniquely identified s-grams that defined two classes of mammography reports (abnormal and normal). Other data sets may have more than two classes of data, and so future work will investigate the expansion of this approach to identify n classes of data. Secondly, the work focused on a single learning algorithm that could be used for an intelligent software agent. However, intelligent agents have additional capabilities that can be utilized. To further enhance the learning capability and domain flexibility, future work will investigate cooperative agent learning to enhance this approach. Finally, the current approach used a very rudimentary memory. A more advanced cognitive memory model will be explored in the future.

ACKNOWLEDGMENT

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

REFERENCES

- [1] Abdalla, R. M., and Teufel, S. 2006. A bootstrapping approach to unsupervised detection of cue phrase variants. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (Sydney, Australia). COLING 2006. ACM Press, New York, NY, 2061-2064. DOI= <http://dx.doi.org/10.3115/1220175.1220291>
- [2] Cheng, W., Greaves, C. and Warren, M. 2006. From n-gram to skipgram to concgram. International Journal of Corpus Linguistics 11/4: 411-33.
- [3] Fox, C. 1992. "Lexical analysis and stoplists." In Information Retrieval: Data Structures and Algorithms (ed. W.B. Frakes and R. Baeza-Yates), Englewood Cliffs, NJ: Prentice Hall.
- [4] Patton, M.Q. 1990. Qualitative Evaluation and Research Methods, Second Edition. Newbury Park, CA: Sage Publications, Inc.
- [5] Patton, R.M., Beckerman, B., and Potok, T.E. 2008. Analysis of mammography reports using maximum variation sampling. In Proceedings of the 2008 GECCO conference companion on Genetic and Evolutionary Computation (Atlanta, GA). GECCO 2008. ACM Press, New York, NY, 2061-2064. DOI= <http://doi.acm.org/10.1145/1388969.1389022>
- [6] Pirkola, A, Keskustalo, H., Leppänen, E., Käsälä, A. and Järvelin, K. 2002. "Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants." Information Research, 7(2) [Available at <http://InformationR.net/ir/7-2/paper126.html>]
- [7] Porter, M. 1980. "An algorithm for suffix stripping." Program vol. 14, pp. 130-137.
- [8] Porter Stemming Algorithm. Current Feb. 5, 2009. <http://www.tartarus.org/~martin/PorterStemmer/>
- [9] Raghavan, V.V., and Wong, S.K.M. 1986. "A critical analysis of vector space model for information retrieval." Journal of the American Society for Information Science, Vol.37 (5), p. 279-87.
- [10] Salton, G. 1983. Introduction to Modern Information Retrieval. McGraw-Hill.