Using the Apriori Algorithm for Medical Image Classification

SORINA GHITA

Introduction

Today, the analysis of an image is done by a radiologist and is time consuming. Also, the amount of images is growing faster then the number of radiologists can analyze.

The number of images (multimedia data) that are being collected every day is growing especially in radiology and this brings us to the problem of extracting meaningful information from such collections of raw image data without the need of human intervention.

Because of the need to analyze more images with a very high accuracy and reliability there is a need for software, which can help to reduce the workload of the radiologist.

New machine learning based algorithms might be used to learn on a small set of training images to classify a large collection of images.

Classification is an important part of any knowledge-retrieval system and especially significant in these applications, where images are the main source of information in the decision making process.

The goal of this research is to increase diagnostic accuracy and optimize decision time allowing a detailed analysis of a large number of images in the shortest time.

The method used

The image cclassification process that will be used is the following:

1. Image acquisition
2. Image enhancement
3. Feature extraction
4. Classification

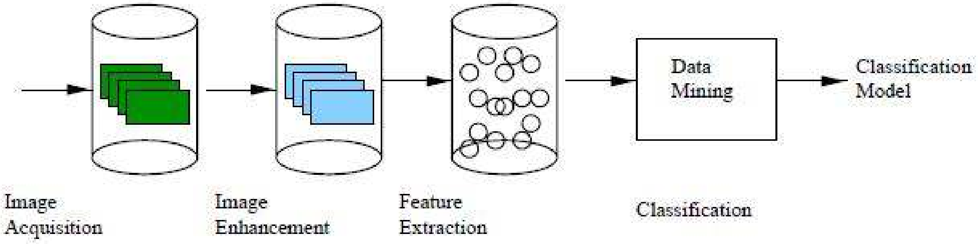

Figure 1. Image cclassification process

Image acquisition

To have access to real medical images for experimentation is a very difficult undertaking due to privacy issues and heavy bureaucratic hurdles. The data collection that will be used in my experiment will be taken from the Institute for Mother and Childcare Alfred D. Rusescu (IOMC) that will collaborate with me in this research.

The approach envisaged may be used in any medical field but this research will focus on various congenital malformations foud in ultrasound screening of infants / children or for detection of ultrasound fetal Down syndrome.

For example, the fetal ultrasound may detect signs of Down Syndrome in the first half of pregnancy. A fetal ultrasound image can show a more prominent position (than normal) in the back of the neck of the unborn child. This prominent position is detected by measuring the distance between skin surface and neck bones.

The first step is to create a database for storing the medical images.

For this I used the Oracle Database 11g Enterprise Edition. With the release of Oracle Database 11g, Oracle brings the benefits of a modern, high performance relational database to the problem of storing vast quantities of data required for medical imaging.

This technology, part of Oracle Multimedia, a feature of Oracle Database, can be used to increase developer productivity and improve security and reliability of storage of diagnostic images and other DICOM (Digital Imaging and Communications) content.

The advantages of using Oracle Multimedia to store media objects are:

1. Both the descriptions of an image and the image itself can be stored using industry standard formats.
2. The Oracle Multimedia object model and methods make application programming simple and application maintenance far easier.
3. Support for standard streaming output technology enables convenient retrieval and easy delivery to media players.
4. Extraction and indexing of all image metadata is greatly simplified.

To load the images in the Oracle Database 11g Enterprise Edition I'm using Oracle Application Express (formerly Oracle HTML DB), which is a web-based application development and deployment tool, bundled with Oracle Database 11g.

The application created was called DICOM image archive.

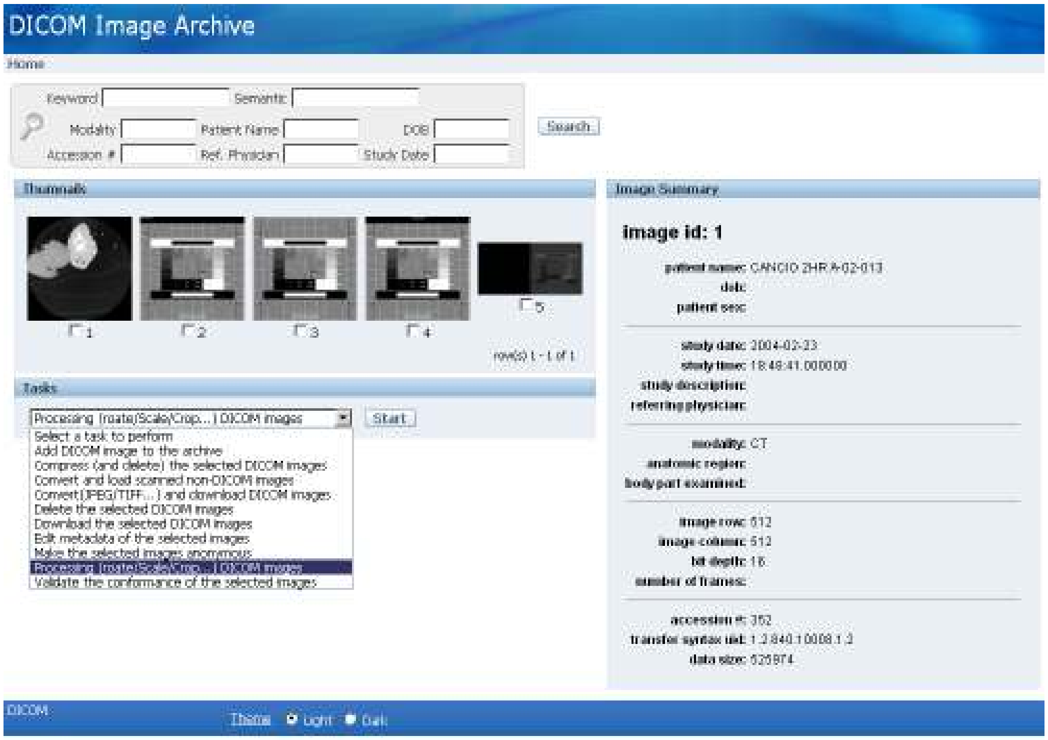
DICOM Image Archive
Home
Keyword
Semantic
Modality
Patient Name
DOB
Accession #
Ref. Physician
Study Date
Search
Thumbnails
□ 1
□ 2
□ 3
□ 4
□ 5
row(s) 1 - 1 of 1
Tasks
Processing (rotate/Scale/Crop...) DICOM images
Start
Select a task to perform
Add DICOM image to the archive
Compress (and delete) the selected DICOM images
Convert and load scanned non-DICOM images
Convert(JPEG/TIFF...) and download DICOM images
Delete the selected DICOM images
Download the selected DICOM images
Edit metadata of the selected images
Make the selected images anonymous
Processing (rotate/Scale/Crop...) DICOM images
Validate the conformance of the selected images
Image Summary
DICOM
Theme
● Light
● Dark

image id: 1

patient name: CANCIO 2HR A-02-013
dob:
patient sex:

study date: 2004-02-23
study time: 18:48:41.000000
study description:
referring physician:

modality: CT
anatomic region:
body part examined:

image row: 512
image column: 512
bit depth: 16
number of frames:

accession #: 352
transfer syntax uid: 1.2.840.10008.1.2
data size: 525974

Figure 2. DICOM image archive

In the DICOM image archive, a user can import and export DICOM images; retrieve the full (original) size DICOM image in JPEG format, the JPEG thumbnail of the DICOM image and the metadata for each DICOM image; edit the attributes of each DICOM image; process the content of the DICOM images; validate each DICOM image based on the constraint rules and make the selected DICOM images anonymous.

The application also provides keyword search, field search and semantic search to search within the DICOM image archive.

Although this application is the easiest way to load images into the database, this operation is manual and is timeconsuming.

The ideal way to load images into the database would be capturing the images directly from the medical devices.

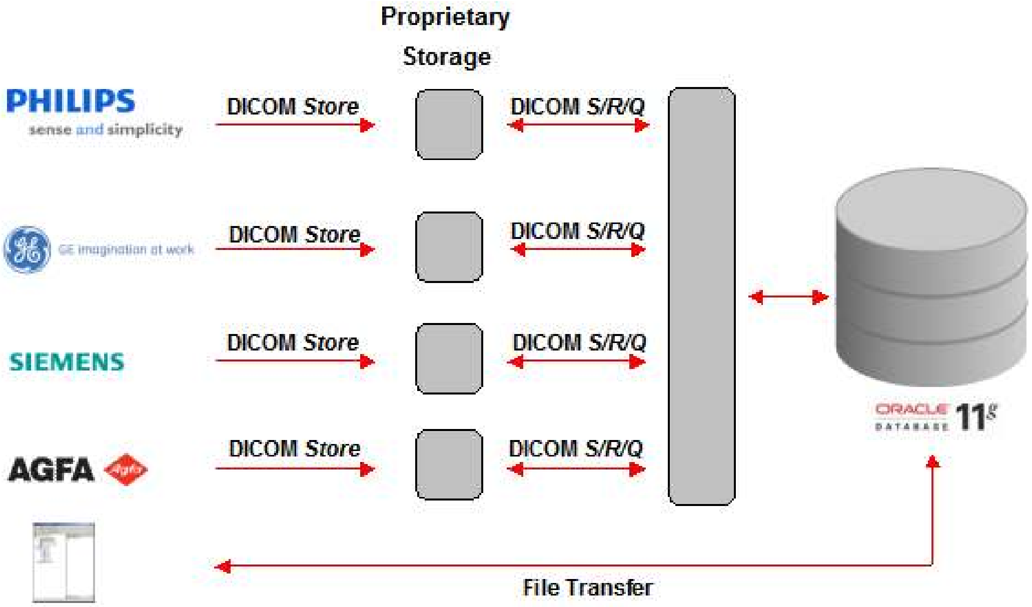

Figure 3. Image acquisition – preferred situation

Since a key component of any medical record is the diagnostic images made of the patient and that is why efficient storage and rapid delivery of those images is a very important.

For the next step of my research I will integrate the DICOM image archive application (that gives access to the Oracle Database 11g) with the medical equipment used for acquiring the data in order to automate the process.

Image enhancement

The images are difficult to interpret, and a preprocessing phase of the images is necessary to improve the quality of the images and make the feature extraction phase more reliable.

Pre-processing is always a necessity whenever the data to be mined in noisy, inconsistent or incomplete and pre-processing significantly improves the effectiveness of the data mining techniques

The pre-processing techniques are applied to the images before the feature extraction phase.

In the digitization process, noise could be introduced that needs to be reduced by applying some image processing techniques At the time that the images were taken, the conditions of illumination are generally different.

I will apply to the images two techniques: a cropping operation and an image enhancement one.

The first one will be used in order to cut the black parts of the image as well as the existing artifacts such as written labels etc. For most of the images almost 50% of the whole image comprised of a black background with significant noise. The cropping to eliminate noise will be done first before the image enhancement to avoid enhancing noise and hindering the cleaning phase. This means eliminating the unwanted parts of the image (usually peripheral to the area of interest). The cropping operation will be done automatically by sweeping through the image and cutting horizontally and vertically the image those parts that had the mean less than a certain threshold.

Image enhancement helps in qualitative improvement of the image with respect to a specific application. In order to diminish the effect of over brightness or over darkness in the images and accentuate the image features, I will apply a widely used technique in image processing to improve visual appearance of images known as Histogram Equalization. Histogram equalization increases the contrast range in an image by increasing the dynamic range of grey levels (or colors). This improves the distinction of features in the image. This means widening the peaks in the image histogram and compressing the valleys. This process equalizes the illumination of the image and accentuates the features to be extracted.

Feature extraction

After cropping and enhancing the images, which represents the data cleaning phase, features relevant to the classification will be extracted from the cleaned images.

The extracted features will be organized in a database in the form of transactions, which in turn constitute the input for the classification algorithms used.

The transactions are of the form $\{$ImageID, Class Label, $F1; F2; :::; Fn\}$ where $F1:::Fn$ are n features extracted for a given image.

The database will be constructed by merging some already existing features in the original database with some new visual content features that we extracted from the medical images using image processing techniques.

Classification

For the classification of these images I'm proposing association rule mining using the apriori algorithm. The approach is suitable for analyzing large sets of photograph.

I'll use the concept of association rules with recurrent items for classification with information specific to image analysis, such as the number of occurrences of a particular feature on the image with uniform feature characteristics.

Association rule mining typically aims at discovering associations between items in a transactional database. Given a set of transactions $D = \{T1; ::; Tn\}$ and a set of items $I = \{i1; ::; im\}$ such that any transaction T in D is a set of items in I, an association rule is an implication *A*=> *B* where the antecedent *A* and the consequent *B* are subsets of a transaction *T* in *D*, and *A* and *B* have no common items. For the association rule to be acceptable, the conditional probability of B given A has to be higher than a threshold called minimum confidence.

Association rules mining is normally a two-step process, wherein the first step frequent item-sets are discovered (i.e. item-sets whose support is no less than a minimum support) and in the second step association rules are derived from the frequent item-sets.

In my approach, I will use the apriori algorithm in order to discover association rules among the features extracted from the image database and the category to which each image belongs.

I will constrain the association rules to be discovered such that the antecedent of the rules is composed of a conjunction of features from the image (color, texture) while the consequent of the rule is always the category to which the image belongs.

In other words, a rule would describe frequent sets of features per category normal and abnormal based on the apriori association rule discovery algorithm.

After all the features will be merged and put in the transactional database, the next step is applying the apriori algorithm for finding the association rules in the database constrained as described above with the antecedent being the features and the consequent being the category.

Once the association rules will be found, they will be used to construct a classification system that categorizes the images as normal or abnormal. The most delicate part of the classification with association rule mining will be the construction of the classifier itself. Although the knowledge will be extracted from the database by finding the existing association rules, the main question is how to build a powerful classifier from these associations. The association rules that will be generated from the database could imply either normal or abnormal. When a new image will have to be classified, the categorization system will return the association rules that will apply to that image. The first intuition in building the classification system is to categorize the image in the class that has the most rules that apply. This classification would work when the number of rules extracted for each class is balanced. In other cases, a further tuning of the classification system will be required. The tuning of the classifier is mainly represented by finding some optimal intervals of the confidence such as both the overall recognition rate and the recognition rate of abnormal cases are at its maximum value. In dealing with medical images it is very important that the false negative rate be as low as possible. It is better to misclassify a normal image than an abnormal one. That is why in the tuning phase I'll take into consideration the recognition rate of abnormal images. It is not only important to recognize some images, but to be able to recognize those that are abnormal. By applying the apriori algorithm with additional constraints on the form of the rules to be discovered a relatively small set of association rules will be generated associating sets of features with class labels. These association rules will constitute the classification model.

The discovery of association rules in the images feature database will represent the training phase of the classifier. To classify a new image, it enough to extract the features from the image as was done for the training set, and applying the association rules on the extracted features to identify the class the new image falls into.

Conclusion

For many decades, the healthcare system operated on paper and film. Registration and billing were paper based and diagnostic imaging was film based. Film was the norm in the diagnostic arena, and light boxes were the "browser" of choice. Information technology was first introduced in administrative systems, not in diagnostic or medical records systems. Over the past decade, medical records systems were introduced to manage patient records electronically and medical imagery systems were introduced to manage the new digital radiology modalities.

The big players in CAD (computer aided diagnosis) focus on helping radiologist to extract many features by hand from patients' images. This still helps one radiologist with one patient and does not scale to analyzing large numbers of patients and determining if there is a malformation or not. All players can identify the obvious cases of malformations but the biggest problem is the large number of false positive (saying the images contains a malformation but it doesn't). It is much easier to detect malformation in a late stage but treatment is then much more difficult.

With this research I'm focusing on automatic classification of medical images. New machine learning based algorithms can be used to learn on a small set of training images to classify a large collection of images. By analyzing a large collection of images and thus reduce workload one can focus on early screening and safe life and safe money. This solution will facilitate digital storage and distribution of patient images across healthcare network, will increase diagnostic accuracy and optimize decision time allowing a detailed analysis of a large number of images in the shortest time. This will increase efficiency and quality of care.

References

Annamalai M, Zwanenburg E. Multimedia Medical Archiving – Introduction and Business Opportunities, Oracle EMEA SalesKickoff 2008.

Antonie M.L., Zaıane O. R., Coman A.: Associative classifiers for medical images, Revised Papers from MDM/KDD and PAKDD/KDMCD, (2002) 68–83.

Ardizzone E., Daurel T., Maniscalco U., Rigotti C.: Extraction of association rules between low-level descriptors and semantic descriptors in an image database, In Proc. 1st Int. Workshop on Multimedia Data and Document Eng., (2001).

Li W., Han J., Pei J.: CMAR: Accurate and efficient classification based on multiple class-association rules, IEEE International Conference on Data Mining, (2001).

Liu B., Hsu W., Ma Y.: Integrating classification and association rule mining, In Proc. 4th Int. Conf. On Knowledge Discovery and Data Mining, (1998) 80–86.

Managing Unstructured Data With Oracle Database 11g— Oracle White Paper, July 2007.

Miller J, Thrall J. Clinical molecular imaging. Journal of the American College of Radiology. 2004;1(suppl 1): 4–23.

Oracle Multimedia: Managing Multimedia Content — Oracle White Paper, July 2007.

http://www.oracle.com/technology/products/intermedia/index.html

Oracle Database 11g DICOM Medical Image Support — Oracle White Paper, July 2007.

Tesic J., Newsam S., Manjunath B. S.: Mining image datasets using perceptual association rules, In Proc.SIAM International Conference on Data Mining, 6th Workshop on Mining Scientific and Engineering Datasets, (2003) 71–77.