Medical Image Classification Using an Efficient Data Mining Technique

Safwan Mahmud Khan
Department of Computer Science & Engineering
University of Dhaka, Bangladesh
sainankhan@yahoo.com

Md. Rafiqul Islam
Department of Computer Science & Engineering
University of Dhaka, Bangladesh
rafik@udhaka.net

Morshed U. Chowdhury
School of Information Technology,
Deakin University, Melbourne,Australia
muc@deakin.edu.au

Abstract

Data mining refers to extracting or "mining" knowledge from large amounts of data. It is an increasingly popular *field that uses statistical, visualization, machine learning, and other data manipulation and knowledge extraction techniques aiming at gaining an insight into the relationships and patterns hidden in the data. Availability of digital data within picture archiving and communication systems raises a possibility of health care and research enhancement associated with manipulation, processing and handling of data by computers. That is the basis for computer-assisted radiology development. Further development of computer-assisted radiology is associated with the use of new intelligent capabilities such as multimedia support and data mining in order to discover the relevant knowledge for diagnosis. It is very useful if results of data mining can be communicated to humans in an understandable way. In this paper, we present our work on data mining in medical image archiving systems. We investigate the use of a very efficient data mining technique, decision tree, in order to learn the knowledge for computer-assisted image analysis. We apply our method to classification of x-ray images for lung cancer diagnosis. . The proposed technique is based on an inductive decision tree learning algorithm that has low complexity with high transparency and accuracy. The results show that the proposed algorithm is robust, accurate, fast, and it produces a comprehensible structure summarizing the knowledge it induces.*

1. Introduction

In technology radiology departments are at the center of a massive change. The ubiquitous radiographic film that has been the basis of image management for almost 100 years is being displaced by new digital imaging modalities such as 1. computed tomography (CT); 2. magnetic resonance (MR); 3. nuclear medicine (NM); 4. ultrasound (US); 5. digital radiography (DF); 6. computed radiography (CR) using storage phosphor imaging plates or film digitizers; 7. digital angiography (DA); 8. MR spectoscopy (MRS); 9. electron emission radiography (EMR).

Digital image management systems are under development now to handle images in digital form. These systems are termed Picture Archiving and Communication Systems (PACS) [1,2]. The PACS are based on the integration of different technologies that form a system for image acquisition, storage, transmission, processing and display of images for their analysis and further diagnosis. The main objective of such systems is to provide a more efficient and cost-effective means of examining, storing, and retrieving diagnostic images. These systems must supply the user with easy, fast, reliable access to images and associated diagnostic information.

Data mining is an increasingly popular field that exploits statistical, visualization, machine learning, and other data manipulation and knowledge extraction techniques. It is aimed at gaining an insight into the relationships and patterns hidden in the data. Often databases grow so large that human interpretation of the data is not feasible and accordingly, there is a gap between data generation and data understanding and usage. One of the main goals of applying data mining learning algorithms in medicine is to uncover new relations among data and reveal new patterns that identify diseases, determine prognoses, or indicate certain treatments. Many factors affect the success of learning algorithms on a given task. The quality of the data is one such factor. For instance, if information is irrelevant or redundant, or the data is noisy and unreliable, then knowledge discovery during training is more difficult. Feature subset selection [3, 4, 5], on the

other hand, is the process of identifying and removing as much of the irrelevant and redundant information as possible. It reduces the dimensionality of the data and allows learning algorithms to operate faster and more effectively. Learning algorithms differ in the degree of emphasis they put on feature selection. There are algorithms that classify novel examples by retrieving the nearest stored training example, using all the available features in its distance computations. On the other hand, there are algorithms that try to focus on relevant features and ignore irrelevant ones. Decision tree inducers are examples of this approach. By testing the values of certain features, decision tree algorithms attempt to divide training data into subsets containing a strong majority of one class [6].

In this paper, we present our work on data mining for image archiving systems in medicine. We explain our methodology for performing data mining in these systems. Here we introduce an efficient machine learning algorithm, one of the data mining techniques, to learn the important lung cancer attributes needed for interpretation. The proposed technique is based on an inductive decision tree learning algorithm that has low complexity with high transparency and accuracy. In addition, among all features, we use only the subset of features that leads to the best performance. The evaluation of our result is based on cross validation approach [7].

This paper is organized as follows: Section 2 presents the application itself and the patient dataset. The whole process and the proposed inductive machine learning decision tree algorithm and explanation of how to select the best features are described in section 3. The results and error analysis of the proposed algorithm is introduced in section 4. Conclusion with a discussion of the results and the prospects for future work are given in section 5.

2. Application and data collection

Knowledge acquisition [11] is the first step in developing an image interpretation system. There are a variety of knowledge acquisition methods. The kind of method used for this depends on the inference method the image interpretation system is based on.

We want to develop a knowledge acquisition method for such applications where no generalized knowledge about the domain is available but a large data base of images associated with expert description and interpretation. If we think of the recent trend to picture archiving systems in medicine and other domains, tasks such as these become quite important. Therefore, the aim of our project is development of knowledge acquisition method such as an inductive decision tree machine learning algorithm for medical image classification, which can help to solve some cognitive, theoretical and practical problems. The application of data mining technique will help to get some additional knowledge about specific features of different classes and the way in which they are expressed in the image which can help to find some inherent non-evident links between classes and their imaging in the picture. It can help to get some nontrivial conclusions and predictions.

For our experiment, we used a database of tomograms of 250 patients with verified diagnoses (80 cases with benign disease and 170 cases with cancer of lung). Patients with small pulmonary nodules (up to 5 cm) were selected for this test. Conventional (linear) coronal plane tomograms with 1 mm thickness of section were used for specific diagnosis.

Original linear tomograms were digitized with step of 100 micron (5,0 line pairs per millimeter) to get 1024 x 1024 x 8 bits matrices with 256 levels of gray. The use of linear tomograms and such a digitization enabled an acquisition of high spatial resolution of anatomical details that were necessary for the specific diagnosis of lung nodules.

To improve results of specific diagnosis of small solitary pulmonary nodules we used optimal digital filtering and analysis of post-processed images. The processing emphasized diagnostically important details of the nodule and thus helped to improve the reliability of image analysis: the physician was more certain in feature reading and interpretation. The radiologist worked as an expert on this system.

3. The process

3.1. Pre-processing

First, an attribute list was set up together with the expert, which covered all possible attributes used for classification by the expert as well as the corresponding attribute values, see Table 1. We learned our lesson from another experiment and created an attribute list having no more than three attribute values. Otherwise, the resulting decision tree is hard to interpret and the tree building process stops very soon because of the splitting of the data set into subsets according to the number of attribute values.

Then, the expert collected the database and communicated with a computer answering to its requests. He determined whether the whole tomogram or its part had to be processed and outlined the area of interest with overlay lines and he also outlined the nodule margins. The parameters of optimal filter were then calculated automatically. A radiologist watched the processed image displayed on-line on a TV monitor, evaluated its specific features (character of boundary, shape of the nodule, specific objects, details

and structures inside and outside the nodule, etc.), interpreted these features according to the list of attributes and inputted the codes of appropriate attribute values into the database program. Hard copies of the previously processed images from the archive have been used in this work as well. The collected data set was given as a dBase-file to the inductive machine learning algorithm.

Table 1. Attribute list

Attribute	Short Name	Attribute Values
Class	*CLASS*	1 malignant 2 benign
Structure inside the nodule	*STRINSNOD*	1 Homogeneous 2 Inhomogeneous
Regularity of Structure inside the nodule	*REGSTRINS*	1 Irregular Structures 2 Regular orderly
Cavitation	*CAVITATIO*	0 None 1 Cavities
Areas with calcifications inside he nodule	*ARWCAL*	0 None 1 Areas with calcifications
Scar-like changes inside the .nodule	*SCARINSNOD*	0 None 1 Possibly exists 2 Irregular fragmentary dense shadow
Shape	*SHAPE*	1 Nonround 2 Round 3 Oval
Sharpness of margins	*SHARPMAR*	1 NonSharp 2 MixedSharp 3 Sharp
Smoothness of margins	*SMOMAR*	1 NonSmooth 2 MixedSmooth 3 Smooth
Lobularity of margins	*LOBMAR*	0 NonLobular 1 Lobular
Angularity of margins	*ANGMAR*	0 Nonangular 1 Angular
Convergence of vessels	*CONVVESS*	1 Vessels constantly 2 Vessels are forced away the nodule 3 None
Vascular Outgoing Shadows	*VASCSHAD*	0 None 1 Chiefly vascular
Outgoing sharp thin tape-lines	*OUTSHTHIN*	0 None 1 Outgoing sharp thin tape-lines
Invasion into surrounding tissues	*INVSOURTIS*	0 None 1 Invasion into surrounding tissues
Character of the lung pleura	*CHARLUNG*	0 No Pleura 1 Pleura is visible
Thickening of lung pleura	*THLUNGPL*	0 None 1 Thickening
Withdrawing of lung pleura	*WITHLUPL*	0 None 1 Withdrawing
Size of Nodule	*SIOFNOD*	Numbers (eg, 1.2)in cm

3.2. Feature subset selection

The problem of feature subset selection involves finding a "good" set of features under some objective function [8]. The number of features (i.e. attributes) and number of instances in the raw dataset can be enormously large. In general, the larger the dataset used for training these models, the higher the chances that all representative cases be included. But, this enormity of data may cause the process of discovery the knowledge more difficult. The benefits of feature selection for learning can include a reduction in the amount of data needed to achieve learning, improved predictive accuracy, learned knowledge that is more compact and easily understood, and reduced execution time. The last two factors are of particular importance in the area of data mining. The objective of feature selection is to select a minimal subset of features according to some reasonable criteria so that the original task can be achieved equally well. By choosing a minimal subset of features, irrelevant and redundant features are removed.

We consider our feature subset selection problems as a problem of finding the set of features which are most dissimilar to each other. Ideally, decision tree induction should use only the subset of features that leads to the best performance. In this paper, we adapt the decision tree induction algorithm to select features for use by learner. Induction decision tree algorithm was applied to lung cancer dataset. Only the features that appeared in the final decision tree were used with the classifier.

3.3. The algorithm

In this paper, for our implementation we adapt the ID3 recursive partitioning as a decision tree algorithm [6]. The idea of ID3 decision tree algorithm is to divide the patients into ever smaller groups until creating the groups with all or majority of patients corresponding to the same class. It generates a decision tree from a given set of attribute-value tuples. Each of the interior nodes of the tree is divided by two parts: one is labeled by an attribute

with the number of Datasets (DS), and the other contains the attribute value depending which we come from the parent node to this node. The leaves of the tree correspond to the classes. The tree construction process is guided by choosing the most informative attribute at each step, aimed at minimizing the expected number of tests needed for classification. Let S be the current set of training examples, and C_j, $j=1,2,.....n$ be the set of decision classes. A decision tree is constructed by repeatedly calling the decision tree algorithm in each generated node of the tree. Tree construction stops when all examples in a node are of the same class. This node, called a leaf, is labeled by a value of the class variable. Otherwise, the most informative attribute A_i, is selected as the root of the subtree, and the current training set S is split into subsets S_i according to the values of the most informative attribute. Recursively, a subtree T_i is built for each subset S_i. Each leaf is labeled by exactly one class name. However, leaves can also be empty, if there are no training examples having attribute values that would lead to a leaf or can be labeled by more than one class name if there are training examples with same attribute values and different class name [9]. Figure1 shows the decision tree algorithm for learning our datasets. We have to note that the derived decision trees can easily be converted into classification rules. In order to classify a sample, the attribute values of the sample are tested against the decision tree. A path is created from the root to a leaf node that holds the class prediction for the sample.

Input: - The set of attributes $A_1, A_2,.., A_n$
The Class label C_i
The number of classes P_i
Training set S of examples
Output: Decision tree and set of rules
Processing:
1. Create a root node
2. If all examples have the same class label value, give the root this label
Else If attributes is empty label the root according to the most common value; Else
3. Calculate the information gain for each attribute
4. Select the attribute, A, with the lowest average entropy (highest information gain) and make this the attribute tested at the root
5. For each possible value, v, of this attribute do
5.1- Add a new branch below the root, corresponding to $A = v$
5.2- Let Examples (v) be those examples with $A = v$
5.3- If Examples (v) is empty, make the new branch a leaf node labeled with the most common value among Examples Else
5.4- Let the new branch be the tree created by ID3 (Examples (v), class label, Attributes - {A})
6. End

Figure 1. Decision tree algorithm

In this paper, we incorporated entropy into the proposed algorithm. The entropy determines noise and chaos in the data. It will be used to find the best attribute, as well as computing the information gain of the attributes. Both entropy and information gain adapted as criteria for finding the most significant attributes for step-2 in the proposed algorithm given in Figure 1. The idea is based on the number of Benigns' and Malignants' in each of the subsets generated by the different attribute values. In the general case where S consists of C classes, the entropy is defined as follow:

$$Entropy(S) = -\sum_{i=1}^{C} \frac{P_i}{|S|} \log_2 \left(\frac{P_i}{|S|}\right)$$

where $|S|$ and P_i are number of examples in S and classes C_i, respectively. The information gain measures the effectiveness of an attribute A, by calculating the reduction of entropy caused by partitioning the set S using the attribute A. We can define the information gain as follow:

Information-Gain(S,A)=Entropy(S) - H(S,A)

Where, H(S, A) is defined as follow:

$$H(S, A) = \sum_i (|S_i| / |S|) . H(S_i)$$

Where S_i is a subset of S for which attribute A takes on value i. Where H(S_i) is the entropy of the system of subsets S_i, which takes the following form:

$$S_i = \{s \in S / A(s) = i\}$$

4. Results

Typical applications of learning algorithms require two sets of examples: training examples and test examples. The set of training examples is used to produce the learned concept descriptions and a separate set of test examples is needed to evaluate the accuracy. In this experiment, we used 175 dataset among the 250 as training examples and the rest was used as test datasets. The resulting decision tree we get is as shown in Figure 2 (mentioned in the next page).

From the all datasets attributes we created different subsets of features with 10, 13, 15, and 17 features by selecting the most dissimilar features. The first subset included the following 10 features: REGSTRINS, ARWCAL, LOBMAR, CONVVESS, VASCSHAD, OUTSHTHIN, INVSOURTIS, SIOFNOD, SPICMAR, and SHAPE. The next subset of features included three more features with high dissimilarity value and so on. From these subsets our algorithm induced decision trees and calculated the error rate based on cross validation,

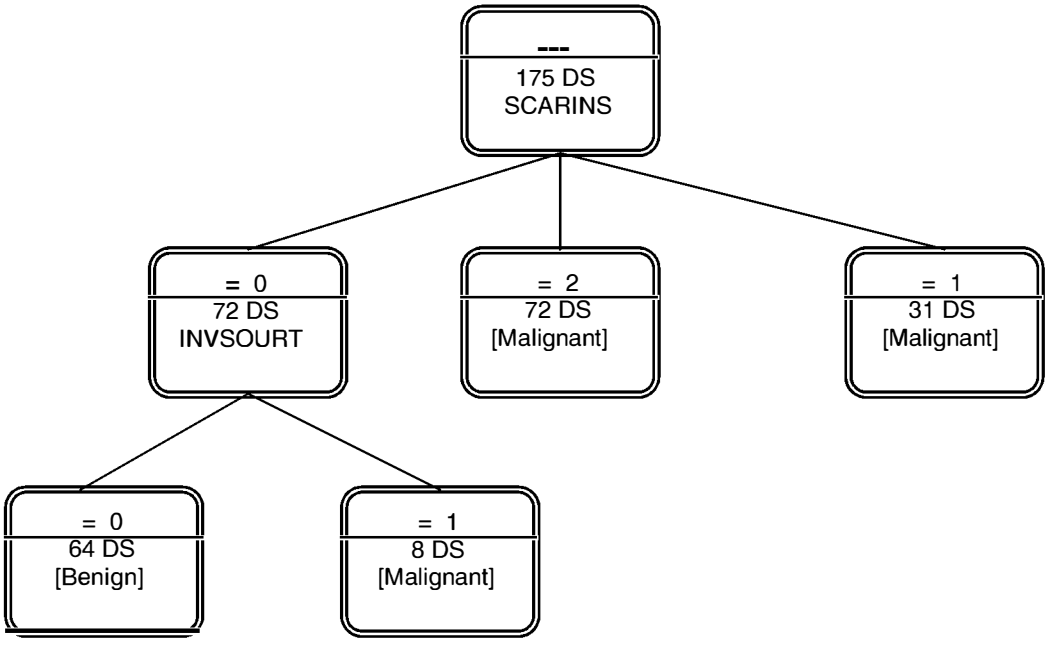

Figure 2. Decision tree with feature subset selection

see Table 2. Here we calculate error rate,
E = (M/N).100%

Where M is the number of misclassified samples and N is the number of all samples.

Table 2. Error rate for different feature subsets

Feature Number	Decision Tree Error Rate
10	14.85
13	4.5714
15	4.5714
17	7.42

5. Conclusion and future work

In this paper, the basis for our study is a sufficiently large database with images and expert descriptions. We were able to learn the important attributes needed for interpretation and the way in which they were used for decision making from this database by applying an efficient data mining technique, inductive decision tree learning algorithm.

The explanation capability of the induced tree was reasonable. The attributes included in the tree represented the expert knowledge. The technique showed very good results. Finally, we can say that picture archiving systems in combination with data mining methods open the possibility of advanced computer-assisted medical classification systems. However, it will not give the expected result if the PACS have not been set up in the right way. Pictures and experts descriptions have to be stored in a standard format in the system for further analysis. Since standard vocabulary and very good experts are available for many medical diagnosis tasks, this should be possible. If the vocabulary is not a priori available, then vocabulary can be determined by a methodology based repertory grid. What is left is to introduce this method to the medical community, which we are going on to do recently for mammography image analysis and lymph nodule diagnosis. Unfortunately, it is not possible to provide an image analysis system, which can extract features for all kind of images. Often it is the case that it is not clear how to describe a particular feature by automatic image feature extraction procedures. The expert's description will still be necessary for a long time. However, once the most discriminating features have been found, the result can lead in the long run to fully automatic image diagnosis system which is set up for specific type of image classification.

6. References:

[1] D.M. Chaney, Coordinate your plans for EMR and PACS investments, Health Management Technology (Aug. 199) vol.20, no.7, p.24, 26-7.

[2] K. Adelhard, S. Nissen-Meyer, C. Pistitsch, U. Fink, M.Reiser, Funtional requirements for a HIS-RIS-PACS-interface design, including integration of "old" modalities, Methods of Information in Medicine (March 1999) vol. 38, no. 1, p. 1-8.

[3] LIU H. SETIONO R., Some Issues on Scalable Feature Selection, Expert systems with Application, vol. 15, Elsevier. pp.333-339, 1998.

[4]. LIU H. SETIONO R., Incremental feature selection, Applied Intelligence, vol. 9, no. 3,pp.217-230, Kluwer Academic Publishers, November, 1998.

[5]. LIU H. MOTODA H., Feature Transformation and Subset Selection, IEEE Intelligent Systems, vol. 13, no. 2, pp.26-28, March/ April, 1998.

[6] SETIONO R. LIU H., A Connectionist Approach to Generating Oblique Decision Trees, IEEE Transactions on Systems, Man, and Cyberneties (Part B: Cyberneties), vol. 29, no. 3, pp. 440-444, June 1999.

[7] USAMA M. FAYYAD, Mining Databases: Towards Algorithms for knowledge Discovery, Data Engineering Bulletin, vol. 2, no. 1, pp. 39-48, 1998.

[8] HALL, M.A., Correlation-based feature selection for discrete and numeric class machine learning, Proceedings of the Seventeenth International Conference on Machine Learning, Stanford University, CA. Morgan Kaufmann Publishers, 2000.

[9] WOLF S., OLIVER H., HERBERT S., MICHAEL M., Intelligent Data Mining for Medical Quality Management Proceedings of the Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2000), Berlin, Germany, August 2000.

[10] S. Heywang-Köbrunner, P. Perner, Optimized Computer-Assisted Diagnosis based on Data Mining, Expert Knowledge and Histological Verification, IBaI Report August 1998 ISSN 1431-2360.

[11] J.H. Boose, D.B. Shema, and J.M. Bradshaw, Recent progress in Aquinas: a knowledge acquisition workbench, Knowledge Acquisition 1 (1989): 185-214.