International Journal of Computer and Electrical Engineering, Vol. 1, No. 2, June 2009
1793-8163

An Approach for Discretization and Feature Selection Of Continuous-Valued Attributes in Medical Images for Classification Learning

Jaba Sheela L [1], Dr.V.Shanthi [2]

***Abstract*— Many supervised machine learning algorithms require a discrete feature space. In this paper, we review previous work on continuous feature discretization and, identify defining characteristics of the method. We then propose a new supervised approach which combines discretization and feature selection to select the most relevant features which can be used for classification purpose. The classification technique to be used is Associatve Classifiers. The features used are Harlick Texture features extracted from MRI Images. The results show that the proposed method is efficient and well-suited to perform preprocessing of continuous valued attributes.**

***Keywords*— Classifier, Discretization, Feature Selection, MRI.**

I. INTRODUCTION

Medical images are a fundamental part of medical diagnosis and treatment. These images are different from typical photographic images primarily because they reveal internal anatomy as opposed to an image of surfaces. They include both projection x-ray images and cross-sectional images, such as those acquired by means of computed tomography (CT) or magnetic resonance imaging (MRI), or one of the other tomography modalities (SPECT, PET, or ultrasound, for example). Medical image processing is a branch of image processing that deals with such images. It is driven both by the peculiar nature of the images and by the medical applications that make them useful. Medical images contain a wealth of hidden information that can be exploited by physicians in making reasoned decisions about a patient. However, extracting this relevant hidden information is a critical initial step to their use. This motivates the use of data mining techniques for efficient knowledge extraction.

Mining medical images involves many processes. The process to be used depends on the type and complexity of image to be mined. For instance, it is simpler to mine 2-dimentional x-rays as compared to 3- dimensional *CT* scans of the brain. However, some processes are fundamental to the task of medical image mining, regardless of the complexity of image. We briefly discuss these processes below.

Data Preprocessing: 'This stage consists of several processes. These processes include data normalization, data preparation, data transformation, data cleaning, and data formatting. Normalization techniques are required to integrate the different image formats to a common format. Data preparation alters images to present them in a format suitable for transformation techniques, Next, the image is transformed in order to obtain a compressed (lossless) representation of it, e.g., using wavelet transforms. Segmentation is done to identify regions of interest (ROI) for the mining task, usually achieved using classifier systems. The segmentation step finds corresponding regions within an image, since item sets are extremely large.

Feature Extraction: Images have a large number of features. It is important to identify and extract interesting features for a particular task in order to reduce the complexity of processing. These are attributes or portion of the image being analyzed that is most likely to give interesting rules for that problem. Not all the attributes of an image are useful for knowledge extraction. This stage increases the overall efficiency of the system. Image processing algorithms are used, which automatically extract image attributes such as local color, global color, texture, and structure. Texture is the most useful description property of an image and it specifies attributes, such as resolution, which can be used in image mining. An image can be adequately represented using the attributes of its features. The extraction of the features from an image can be done using a variety of image processing techniques. We localize the extraction process to very small regions in order to ensure that we capture all areas.

Rule Generation: Since this is a highly knowledge based domain, associated domain knowledge can be used to improve the data-mining task. This data integration is an important concept because medical images are not self contained, and are often used in conjunction with other patient data in the process of diagnosis. We expect association rules of two forms: (i) Image contents unrelated to spatial relationships, e.g., if an image has a texture X, it is likely to contain protrusion Y and (ii) Image contents related

Manuscript received January 6, 2009.

L. Jaba Sheela M.C.A, M.Phil, (Ph.D) is with the Department of Master of Computer Applications of Panimalar Engineering College, Chennai. India which is affiliated to the Anna University and accredited by the AICTE, New Delhi, India. (Phone: 91-044-9789043623)

Dr.V.Shanthi, M.Sc.M.Phil, Ph.D is woking as Professor in the Department of Master of Computer Applications of St.Joseph's Engineering College, Chennai. India which is affiliated to the Anna University and accredited by the AICTE, New Delhi, India. (Phone: 91-044-9884126186)

to spatial relationships, e.g., if X is between Y and Z it is likely there is a T beneath. A low minimum support and high minimum confidence is desirable, since few image data sets have high support.

Interpretation of patterns and knowledge extraction: Not all the interesting rules are medically important. To make our technique relevant, the rule presented must be significant and meaningful.

In this paper we focus on one of the preliminary process, that of Feature selection & discretization. This can eliminate some irrelevant and/or redundant attributes. By using relevant features, classification algorithms can in general improve their predictive accuracy, shorten the learning period, and form simpler concepts.

II. Feature Extraction

The Medical images used for the study were MRI images, obtained from Aarthi MRI & CT SCAN Center, Chennai. MRI images in DICOM format are converted into jpg format with a tool called syngo FastView. They had 256 gray levels which were reduced to 16 gray levels in the first step of the algorithm. In that way the computing time were decreased. After that, one should compute the Co-occurrence matrices, calculated for the directions of 0°, 45°, 90°, and 135°, and for the distances 1, 2, 3, 4, and 5. Twenty Matrices of 16 X 16 integer elements per image are produced. For each matrix, 7 features proposed by Harlick [9] are calculated producing a feature vector of 140 elements to represent each image.

III. Review of Existing Methods for Discretization

A large number of machine learning and statistical techniques can only be applied to datasets composed entirely of nominal variables. However, a very large proportion of real datasets include continuous variables: that is variables measured at the interval or ratio level. One solution to this problem is to partition numeric variables into a number of sub ranges and treat each such sub-range as a category. This process of partitioning continuous variables into categories is usually termed as *discretization.*

A variety of discretization methods have been developed in recent years. Dougherty, Kohavi and Sahami [1] have provided a valuable systematic review of this work in which discretization techniques are located along two dimensions: unsupervised vs. supervised, and global vs. local.

Unsupervised discretization procedures partition a variable using only information about the distribution of values of that variable: in contrast, supervised procedures also use the classification label of each example. Typical unsupervised techniques include:

Equal interval width methods in which the range of values is simply divided into sub-ranges of equal extent, and equal frequency width methods in which the range is divided into sub-ranges containing equal number of examples. More sophisticated unsupervised methods draw on techniques of cluster analysis, to identify partitions that maximize within group similarity while minimizing between groups similarity.

Supervised techniques normally attempt to maximize some measure of the relationship between the partitioned variable and the classification label. The methods include:

1. X^2 test to determine which groups should be merged. ChiMerge [2], [3] algorithm uses this method.
2. Entropy or information gain which measure the strength of the relationship [4].

Supervised techniques might reasonably be expected to lead to more accurate classification trees since the partitions they produce are directly related to the class to be predicted. On the other hand one might expect most of the unsupervised techniques to be considerably faster since they involve little more than sorting the data, an operation which is common to all discretization methods.

Global discretization procedures are applied once to the entire dataset before the process of building the decision tree begins. Consequently a given variable will be partitioned at the same points whenever it is used in the tree. In contrast, local discretization procedures are applied to the subsets of examples associated with the nodes of the tree during tree construction: consequently the same variable may be discretized many times as the tree is developed and the final tree may include several partitionings of the same variable.Since local discretization techniques can develop alternative partitionings for different parts of the sample space, one would expect them to be superior to global methods in producing accurate classification trees. However one would also expect to pay a considerable price in execution speed for this improved accuracy since the discretization process may be repeated many times as the tree is built.

In this section we describe methods used in some of the more popular data mining algorithms.

A. Manual Approach

Discretization the values of continuous features into small number of intervals is the task of feature discretization process wherein each interval is mapped to a discrete symbol. Priori knowledge about the feature is used in this case. Without any knowledge about the feature, a discretization is much more difficult and, in many cases, arbitrary. A reduction in feature values usually is not harmful for real-world data mining applications, and it leads to a major decrease in computational complexity. Therefore, we will introduce, in the next two sections, several automated discretization techniques.

B. Binning

Binning is applied to each individual feature (or attribute). It does not use the class information. Suppose we have the following set of values for the attributes: Age: 0, 4, 12, 16, 16, 18, 24, 26, 28. Two possible ways in which Binning can be applied are: Equi-width binning or Equi-frequency binning.

C. Entropy- based Discretization

Entropy based methods use the class-information present in the data. The entropy (or the information content) is calculated on the basis of the class label. Intuitively, it finds the best split so that the bins are as pure as possible, i.e., the majority of the values in a bin correspond to having the same class label. Formally, it is characterized by finding the split

with the maximal information gain.

The entropy (or the information content) for S is obtained as:

Entropy(S) = -p.log (p) – n.log (n).

Let X = v be a possible split, dividing the set S into two sets, S_1 and S_2 where, S_1 is set of value of $X \leq v$ and S_2 is set of value $X > v$.

Information of the split,

Info (S_1, S_2) = ($|S_1|$ / | S |). Entropy (S_1) + ($|S_2|$ / $|S|$). Entropy (S_2)

Information gain of the split,

Gain (v, S) = Entropy(S) – Info (S_1, S_2)

Here | S | represents the cardinality (number of data points) of the set S.

For example, if we want to split on the attribute-value, X = 14

S_1 = {(0, P), (4, P), (12, P)} and S_2 = {(16, N), (18, P), (24, N), (26, N), (28, N)}

Info (S_1, S_2) = (3/9). Entropy (S_1) + (6/9). Entropy (S_2) = 0 + (6/9). 0.1963 = 0.1308

Gain (14, S) = Entropy(S) – 0.1308

The goal of this algorithm is to find the split with the maximum information gain. Maximal gain is obtained when Info (S_1, S_2) is minimal. The best split(s) are found by examining all possible splits and then selecting the optimal split.

In practice it is not necessary to consider every possible cut point. Fayyad and Irani [4] have shown that optimal cut points for entropy minimization must lie between examples of different classes.

IV. PROPOSED PEAR ALGORITHM

PEAR (**Pr**E**p**rocessing Solution for **A**ssociation **R**ules) is a novel supervised algorithm that performs discretization of the continuous values of the features. The following definitions are necessary before detailing the PEAR algorithm.

Definition 1: ***Class*** is the most important keyword of a diagnosis given by a specialist.
Definition 2: ***Cut Points*** are the limits of an interval of values.
Definition 3: ***Majority Class*** is the most frequent class of an interval.

PEAR processes each feature separately. Let R be the set of training image transactions. Let *f* be a feature of the feature vector F. Let f_i be the value of the feature *f* in a transaction *i*. PEAR uses a data structure that links f_i to the class c_i, for all i $\in$ R, where c_i is the class of the transaction *i*. Each line in the data structure is called an instance. An **instance** I_i has information about the image Img_i.

Definition 4: An instance I_i **belongs** to an interval T_k if its value f_i is between two consecutive cut points u_p and u_{p+1}, i.e., $f_i \in T_k = [u_p, u_{p+1}]$

The algorithm PEAR uses two input thresholds:

• **minperint**: restricts the minimal number of occurrences of the majority class allowed in an interval;
• **mintofuse**: restricts the minimum occupancy of the majority class in an interval.

Let M_k be the majority class of interval T_k, and | M_k | the number of occurrences of M_k in the interval. When determining the data intervals, the algorithm PEAR creates a cut point u_p if

• **Condition 1**: The class label of the current instance I_i, , $i \geq 1$, is different from the class label of the previous instance, i.e., $c_i \neq c_{i-1}$.

Condition 1 generates too many cut points, especially when working with noisy data. The larger the number of cut points, the larger the number of intervals. Each interval represents an item in the process of mining association rules. The use of many items potentially generates a huge number of irrelevant rules, with low confidence. Hence, it is important to keep the number of cut points small and, consequently, generating a small number of items. The next two conditions are used to remove unnecessary cut points

• **Condition 2**: The number of occurrences of the majority class in an interval must be equal or greater than the *minperint* threshold, i.e. | M_k | $\geq$ *minperint*,
• **Condition 3**: The middle cut point u_{p+1} of two consecutive intervals $T_k = [u_p, u_{p+1}]$ and $T_{k+1} = [u_{p+1}, u_{p+2}]$ is removed if $M_k = M_{k+1}$ and ($|M_k|$ / $|T_k|$) $\geq$ *mintofuse* and ($|M_{k+1}|$ / $|T_{k+1}|$) $\geq$ *mintofuse*, where $|T_k|$ is the number of instances belonging to the interval T_k.

Algorithm PEAR

Input : Image Feature Vectors F, Image classes C, *minperint, mintofuse, valreduct*
Output : Processed Feature Vector V

1: for each feature $f \in$ F do
2: Sort f values
3: For each transaction I, create an instance I_i of the form c_i, f_i where $c_i \in$ C
4: Use Condition 1 to create the vector U of cut points u_p
5: end for
6: for each $u_p \in$ U do
7: Remove u_p according to Condition 2
8: Remove u_p according to Condition 3
9: Save the remaining cut points in a vector U_f
10: end for

11: Rank the features *f* according to the number of cut point in U*f*
12: Select the 1-valreduct * | F | features that generate the least number of cut points
13: Write the selected feature discretized in V
14: Return V

PEAR is also employed to select the most relevant features, according to the following criterion:

Criterion 1: *The features that generate the smallest number of cut points are selected as the most relevant ones.*

Since the cut points are found according to the variation of the class label, the most discriminative features are those which present the smallest class variation, i.e., the ones which generate fewer cut points. The PEAR algorithm returns a list of features ranked by the number of cut points generated. A threshold valreduct is used to state the percentage of reduction of the original number of features. The number of features returned are (1- valreduct) * | F |, where | F | is the original number of features of the feature vector.

Algorithm PEAR is employed to solve two problems: feature discretization and selection. An important reduction of irrelevant features is achieved using the PEAR algorithm to select the most relevant ones, speeding up the whole process.

V. EXPERIMENTAL RESULTS

We have performed several experiments to validate our proposed method. We measured the accuracy of the proposed method. The datasets are composed of ROIs identified in Hippocampus, taken from MRI of the brain. We use the Harlick feature extractor algorithm based on texture for feature extraction.

Each image was represented by a feature vector composed of 140 features. The image features were submitted to the PEAR algorithm, using the following input parameters: *minperint = 6, mintofuse = 0.8* and valreduct =17 %, which are tuning parameters set by the user. PEAR selects 24 features as the most relevant ones, obtaining a reduction of 83% in the feature vector size.

To measure the effectiveness of the feature selection task performed by the PEAR algorithm, we employed an approach based on the well-known precision and recall (P&R) graphs. The measures of precision and recall are defined as

$$\text{Precision} = \frac{TRS}{TR}$$

$$\text{Recall} = \frac{TRS}{TS}$$

where TR is the number of relevant images in the dataset; TRS is the number of relevant images in the query result; TS is the number of images in the query result.

We measured the effectiveness of PEAR in selecting features. For comparison purposes we also applied Relief [6], a well-known feature selection algorithm. The 24 most relevant features returned by Relief were also To build the Precision versus Recall graphs, we considered three cases of feature vectors to represent the images: a) using 140 original features, b) using the 24 features selected by PEAR, and c)using the 24 features selected by Relief. Similarity queries were executed and the P&R graphs were constructed.

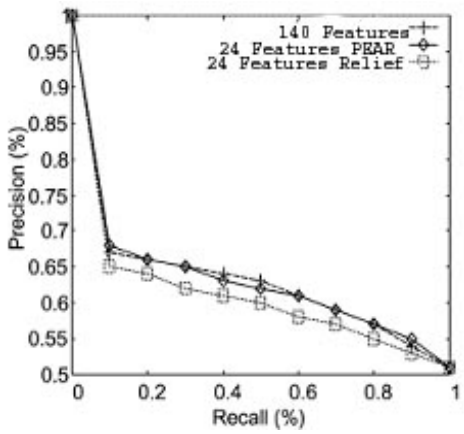

Figure. 2 shows the P&R graph obtained - 140 orginal features, 24 features selected by PEAR, 24 features selected by Relief.

Fig 2. shows the P & R Graph obtained. It shows that, even with a reduction of 83% of the feature vector size, the precision values are maintained. Moreover, PEAR reaches higher values of precision than Relief. Relief took 4.3 s to select the features and PEAR took 3.4 s (21% less time). This can represent a significant difference for larger datasets. While Relief executes several distance calculations, PEAR scans each feature value only once when performing the feature selection task. Indeed, recall that PEAR performs simultaneously feature selection and discretization.

VI. CONCLUSION

The increasing use of image exams in the last 25 years has greatly contributed to improve the diagnosing of diseases as well as to enhance the health care of patients. However, the volume of images has grown at a fast pace and the specialists have been unable to keep up with diagnosing. The feature discretization and selection process speeds up and reduces the complexity of the whole diagnosis method, making it faster and more accurate than traditional approaches. The results show that the proposed method is efficient and well-suited to perform the combined task of feature selection and discretization for images.

REFERENCES

[1] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and Unsupervised Discretization of Continuous Features, *Machine Learning: Proceedings of the Twelfth International Conference*, 1995.
[2] R. Kerber. Chimerge: Discretization of numeric attributes. *AAAI-92, Proceedings Ninth National Conference on Artificial Intelligence*, pages 123-128. 1992.
[3] H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. *Proceedings of the IEEE 7th International Conference on Tools with Artificial Intelligence*.
[4] H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6, pages 393-423. 2002.
[5] U. Fayyad and K. Irani. On the handling of continuous-balued attributes in decision tree generation, *Machine Learning*, v.8 n.1, p. 87-102, Jan 1992.
[6] K. Kira and L. A. Rendell, "A practical approach for feature selection, " in *Proc. 9th Int. Conf. Mach. Learning*, Aberdeen, Scotland, 1992, pp. 249–256.

[7] Holmes G., Nevill-Manning C.: Feature Selection via the Discovery of Simple Classification Rules, *Proc. of the International Symposion on Intelligent Data Analysis (IDA-95*), Baden-Baden, Germany, 1995.
[8] Family A., Wei-Min Shen, Weber R., Simoudis E.: Data Preprocessing and Intelligent Data Analysis. *Intelligent Data Analysis, Eslevier* 1996
[9] R. M. Haralick, K. Shanmugam, and I. Distein, "Textural features for image classification, " *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, pp. 610–621, 1973.
[10] L.Jaba Sheela Dr.V.Shanthi , Image Mining Techniques for Classification and Segmentation of MRI data. – *International Journal of Theoretical and Applied Information Technology, JATIT*. Vol. 3 No. 4, pp 115-121, 2007.