Indian Journal of Science and Technology, Vol 9(6), DOI: 10.17485/ijst/2016/v9i6/84772, February 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Medical Image Classification using Hybrid classifier by Extending the Attributes

S. Padmapriya[1]*, E. Kirubakaran[2] and N. M. Elango[3]

Department of Information Technology, AVCCE, India; padmapriya.anand14@gmail.com
BHEL, India; ekirubakaran@gmail.com
Madanapalle Institute of Technology and Science, India; nmeoxford@yahoo.com

Abstract

Background/Objective: To create a Computer Aided Diagnosis system to detect the abnormalities in the human tissue images by extending the attributes**. Methods/Statistical Analysis:** An efficient and hybrid classifier using "K-Ratio Super Item set Finding-Nearest Neighborhood Classifier (KRSIF-NNC) Algorithm" is proposed. It classifies the tumor cells in an effective manner by adopting extended attributes from small datasets. The glioblastoma and lung cancer tissue image samples are keyed in to the algorithm which classifies them into four grades. **Findings:** From the histopathology (tissue) images the pathologists will be able to diagnose the abnormalities in the tissues. Examination and judgments are based on the pathologist's personal experience. The problem is during the manual diagnosis, there is a chance of missing some cancerous cells in the tissue images. This is solved by adopting the proposed classifier which automatically do the diagnostic process and classify it into proper grade. Thus the proposed classifier improves the classification process. **Applications/Improvements:** Improved classification is required to detect the cancer grades. This hybrid approach has better classification accuracy than other approaches with 4% improvement which is very essential.

Keywords: Extending Attributes, Hybrid Classifier, KRSIF-NNC, Tissue Images Naïve Bayes Classifier

1. Introduction

Millions of medical images are generated by healthcare centres and hospitals on an everyday base. Imaging is employed as a preferred diagnostic tool by more and more medical procedures. So, there arises a demand to develop methods for efficient mining in databases of images, which are more difficult than mining in purely numerical databases. As an instance, imaging techniques like MRI, PET and a collection of ECG signals generates gigabytes of information per day. This requires high capacity data storage devices and novel instruments to analyze such information.

Data mining[1–2] plays a critical part in studying patterns and guides to effective decision. Association rule mining[3–4] has been used in most of the research for finding the rules for diagnosis in large and small databases. This work starts with pathological images which are very complex at the same time crucial for tumour analysis[5]. The pathologists use these biopsy samples which are removed from patients and the process involved in the preservation of tissue slides[6–7]. Finally they need enough good quality of tissues to achieve a diagnosis. The cross sections of the tissues are made with wax and dyed with one or more stains to separate cellular components for structural as well as architectural analysis[8–9]. Most commonly Haematoxylin and Eosin stain is used which separates cell nuclei and cytoplasm. The examination of these tissues are based on pathologists personal experience. But the proposed work, automates the diagnosis process. It involves image processing[10–11] concepts and data mining techniques. Different filtering techniques[12] are used to remove noise in different medical images. In the proposed work, it is observed that median filter

*Author for correspondence

works good for cancer images. The classification oriented membership degrees is computed to construct new attributes[13] to increase the amount of information for small data set analysis. The uses of Support Vector Machines (SVMs) and Decision Tree (DT) classification are showed as a possible methodology for the characterization of the degree of malignancy of brain tumours. But the proposed method uses a novel hybrid classifier to improve the classification performance in analyzing the malignancy level.

2. Proposed Work

Figure 1 symbolizes the proposed method. In the proposed method[14], automated categorization of images are processed and classified using two different phases, such as Training Phase and Testing Phase. Objective of training phase is to construct a classification model using attributes extracted from the images, then assessing the effectiveness of the model by using new images (testing stage). The process of building the classification model (classifier) includes image preprocessing and extraction of features from images (i.e. training set).

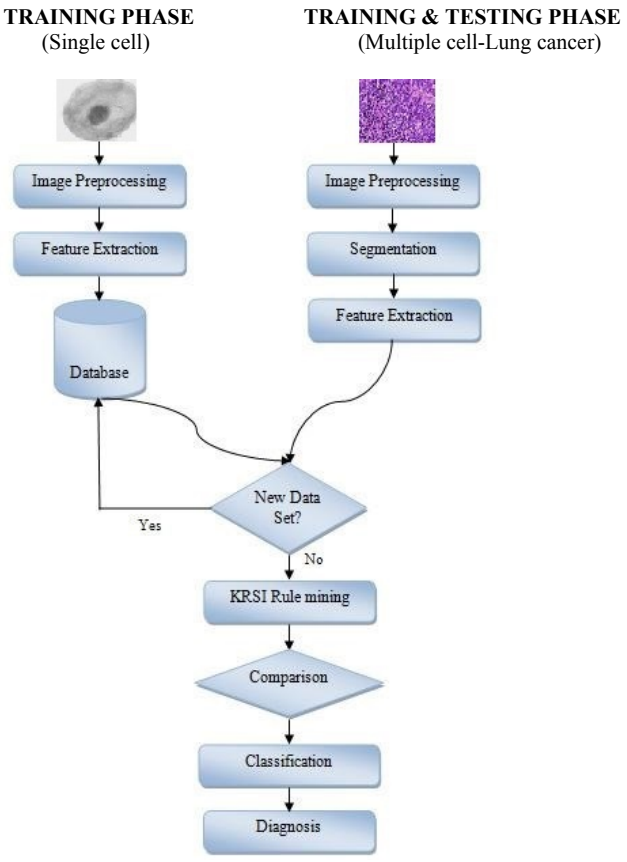

Figure 1. Overview of proposed system.

2.1 Preprocessing Phase

A sequential step of image pre-processing, feature extraction and classification followed. The essential component in image mining is the identification of similar objects in different images. Here the analysis focused on morphological feature of cell. To improve the quality of the images the works begin with preprocessing phase. The image is converted to gray scale using histogram equalization. Then Gaussian filter is applied to remove the noise in the image. The filtered image is given as input to Median filter. This filter minimizes the salt and pepper noise and the main advantage is, it preserves the edges in the image and provides a quality image. After preprocessing, the features or essential properties found within the images have to be extracted. This process identifies cell area, cell perimeter, nucleus area and nucleus perimeter. The attributes are measured and stored.

Image segmentation finds its best usage in medical applications. Thresholding technique has been used for segmentation in this work. Cell nuclei are darker than the surrounding cytoplasm. All the cell nuclei belonging to a class tend to have same gray level. By applying dilation and erosion, extra parts which were not part of the nuclei were removed and the boundary of the cell nuclei became prominent. Then filling operation is done to create a uniform intensity level inside the cell nuclei.

2.2 Extend the Small Datasets

Extracted features from trained images are stored into the database. Although the tissue images has been obtained for research from hospitals and academic laboratories, due to its limited experimentally determined biological activity, most studies for lung/brain cancerous cells has been performed on small datasets. Undesired characteristics of our dataset - it is small. The performance cannot be accomplished when it comes to small training sets as small datasets cannot provide enough information due to the gaps that exist between samples, even the domain samples cannot be ensured. So the algorithm extends the data set by taking k-ratios of these morphological structure ratio of Nucleus Area to Cell Area (NA/CA), ratio of Cell Area to Cell Perimeter (CA/CP), ratio of Nucleus Perimeter to Cell Perimeter (NP/CP), ratio of Nucleus Area to Nucleus Perimeter (NA/NP), ratio of Nucleus

Area to Cytoplasm Area (NA/CA), percentage variation of NA/CA ratio, percentage variation of CA/CP ratio, percentage variation of NP/CP ratio, percentage variation of NA/NP ratio, percentage variation of NA/CA ratio.

2.3 Association Rule Mining

In the data mining literature, association rule mining has been extensively investigated. Among the many efficient algorithms proposed, the most popular being apriori and FP-Tree growth. Aim of association rule mining is to discover associations between the features in a database.

2.3.1 KRSI Algorithm

In our approach we have used the "K-Ratio Super Itemset Finding" (KRSIF) rule mining algorithm is to discover association rules among the features extracted and the category to which each cell belongs. The antecedent of the rules is composed of a conjunction of features from the cell while the consequent of the rule is always the category to which the cell belongs. i.e., a rule would describe frequent sets of features as per category Grade 1, Grade 2, Grade 3 and Grade 4.

2.3.1.1 Input

Item set fetched from transaction database; data from frequent k-ratio super item set table, the minimum threshold defined by user and the least of frequent item sets.

2.3.1.2 Output

K–ratio super item sets.

Retrieving ratio values for normal/cancerous cells and storing it in an array.

$l = l_n(\vec{Nr}), l_n \rightarrow$ returns length of the array coefficient $\vec{Nr}$.

For i = 1:l

$\vec{Na}(\mathrm{i}) \leftarrow \vec{Nr}(\mathrm{i},1).\mathrm{A}$

$\vec{Np}(\mathrm{i}) \leftarrow \vec{Nr}(\mathrm{i},1).\mathrm{P}$

$\vec{Rap}(1,\mathrm{i}) \leftarrow \vec{Nr}(\mathrm{i}) / \vec{Np}(\mathrm{i})$

end

Rlb ← min($\vec{Rap}$);

//Rlb → minimum lower boundary of ratio.

Rub ← max($\vec{Rap}$);

//Rub → maximum upper boundary of ratio.

Where A → Area and P → Perimeter.

Calculating the percentage variation for each ratios.

$\vec{R} \times \vec{C} = \mathrm{fsz}(\vec{Rap})$, $fsz \rightarrow$ size of the array coefficient and returns a $\vec{R} \times \vec{C}$ matrix. ie., returns the number of $\vec{R}$ and $\vec{C}$ in $\vec{Rap}$ as separate output variables.

For i ← 1: $\vec{C}$

$\vec{Rpv}$ {1,1}{1,i}←(100 / ((Rub – Rlb)/($\vec{Rap}$ (i) – Rlb)))

$\vec{Cpr}$ {i,1}{1,1} ← $\vec{Rpv}$ {1,1}{1,i}

end

2.3.2 KRSI-NN Classifier

Features relevant to the classification are extracted from the cleaned images after pre-processing and enhancing. Extracted features from trained images are stored after mining and are then used by the classifier during classification process. Proper classification is done through image mining techniques by matching extracted features with trained data set. K-Ratio Super Item set Finding-Nearest Neighborhood Classifier (KRSIF-NNC) algorithm used in the proposed system classifies the rules generated into 4 different grades: Grade 1, Grade 2, Grade 3, Grade 4.

3. Data Analysis

The proposed classifier is implemented using MATLAB 7.10 software. In the experiment, the large data collections are used to analysis the performance of classifiers. In order to measure the performance, a set of medical image data set is given as input. The data set used in this research is acquired from 'The Cancer Genome Atlas' data repository. Two different data sets (Glioblastoma, Lung cancer) are used in this research. The tissue image (slide) is given as input. Around 20 slides were used for training set and remaining 10 slides were used as test set. This experiment was repeated each time using different test set. The performance of the classification is calculated using the trained KRSIF-NN classifier. Then the performance is analyzed using ROC curve.

4. Experimental Results

The use of ROC (Receiver Operating Characteristics) is used as a tool to evaluate the performance of classification models in machine learning. ROC is obtained by plotting false positive rate as the X axis and true positive

rate as the Y axis. With an ROC curve of a classifier, the evaluation metric will be the area under the ROC curve. The larger the area under the curve (the more closely the curve follows the left-hand border and the top border of the ROC space), the more accurate the test. Thus, the ROC curve for a perfect classifier has an area of 1. An ROC curve, which lies towards the upper left corner of the graph (high true positive and low false positive rate) is the desirable position.

To find the effect of proposed hybrid classifier (KRSIF-NNC) with single classifier (Naïve Bayes) using True Positive Rate and False Positive Rate as a parameter–ROC curve.

Figure 2 and Figure 3 shows the ROC curve evaluating the performance of Naïve Bayes classifiers and KRSIF-NNC hybrid classifier for brain/lung cancer data

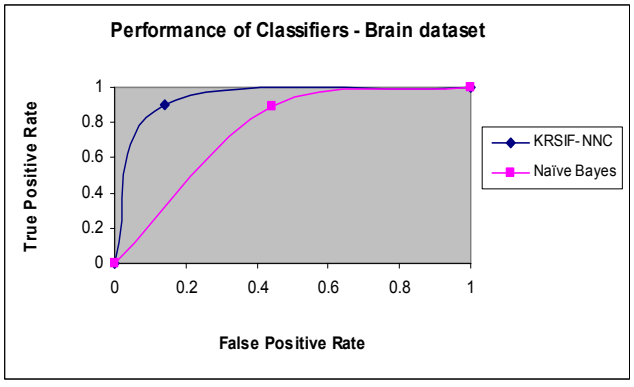

Figure 2. Performance of classifiers on brain cancer dataset.

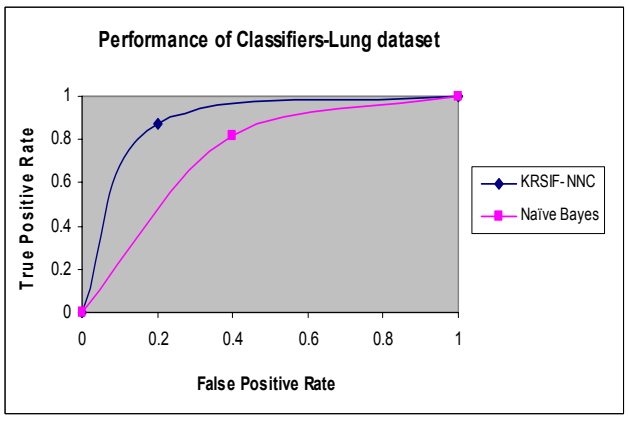

Figure 3. Performance of classifiers on lung cancer dataset.

sets. From the plot of area under the ROC curve, it is clear that KRSIF-NNC (hybrid classifier) is closer to the perfect point (0, 1) than the other classifier. This shows that KRSIF-NNC (hybrid classifier) is best when compared with Naïve Bayes (single classifier).

5. Conclusion

Our proposed approach for nuclear segmentation addresses technical variations by utilizing global information from a set of reference images. In the traditional classification approach, single classification methods like ARM or Naïve Bayes Classifier are used; whereas in this proposed method, combination of efficient hybrid mining using nearest neighborhood decision technique features were used for the medical image classification. The proposed classifier performs well with the existing classifier. So this will assist physicians as a "second option" in clearly diagnosing the cancerous cells.

6. References

1. Ordonez C, Ezquerra N, Santana CA. Constraining and summarizing association rules in medical data. Knowledge and Information Systems. 2006 Mar; 9(3):259–83.
2. Sudha M, Kumaravel A. Performance comparison based on attribute selection tools for data mining. Indian Journal of Science and Technology. 2014 Nov; 7(S7):61–5.
3. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. Proceedings of ACM-SIGMOD International Conference on Management of Data. 2000 Jun; 29(2):1–12.
4. Ribeiro MX, Traina A, Traina C, Azevedo-Marques PM. An association rule-based method to support medical image diagnosis with efficiency. IEEE Transactions on Multimedia. 2008 Feb; 10(2):277–85.
5. Rohde GK, Ribeiro AJS, Dahl KN, Murphy RF. Deformation-based nuclear morphometry: Capturing nuclear shape variation in HeLa cells. Cytometry. Journal of the International Society for Analytical Cytology. 2008 Apr; 73(4):341–50.
6. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: A review. IEEE Reviews in Biomedical Engineering. 2009; 2:147–71.
7. Demir C, Yener B. Automated cancer diagnosis based on histopathological images: A systematic survey. Department of Computer Science. Rensselaer Polytechnic Institute; 2009 May. p. 1–16.
8. Chang H, Fontenay GV, Han J, Cong G, Baehner FL, Gray JW, Spellman PT, Parvin B. Morphometic analysis of TCGA

Glioblastoma multiforme. BMC Bioinformatics. 2011; 12(484):1–12. Doi no: 10.1186/1471-2105-12-484.

9. Doyle S, Hwang M, Shah K, Madabhushi A, Feldman M, Tomaszeweski J. Automated grading of prostate cancer using architectural and textural image features. IEEE Explore 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI'07; Arlington, VA. 2007 Apr 12-15. p. 1284–7.
10. Vaidehi K, Subashini TS. Breast tissue characterization using combined K-NN classifier. Indian Journal of Science and Technology. 2015 Jan; 8(1):23–6.
11. Bharathi K, Karthikeyan S. A novel implementation of image segmentation for extracting abnormal images in medical image applications. Indian Journal of Science and Technology. 2015 Apr; 8(S8):333–40. Doi no: 10.17485/ijst/2015/v8iS8/61920.
12. Shinde B, Mhaske D, Patare M, Dani AR. Apply different filtering techniques to remove the speckle noise using medical images. International Journal of Engineering Research and Applications. 2012 Jan-Feb; 2(1):1071–9.
13. Li DC, Liu CW. Extending attribute information for small data set classification. IEEE Transactions on Knowledge and Data Engineering. 2012 Mar; 24(3):452–64.
14. Padmapriya S, Kirubakaran E, Elango NM. Advanced medical image mining technique using efficient hybrid classifier for small dataset. International Journal of Applied Engineering Research. 2014; 9(23):19355–76.