Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 3, March 2016, pg.604 – 609

PERFORMANCE ANALYSIS OF DATA MINING ALGORITHMS FOR MEDICAL IMAGE CLASSIFICATION

P.Keerthana[1], P.Thamilselvan[2], J.G.R. Sathiaseelan[3]

Research Scholar, Department of Computer Science, Bishop Heber College Tiruchirappalli, Tamilnadu, India[1]

Research Scholar, Department of Computer Science, Bishop Heber College Tiruchirappalli, Tamilnadu, India[2]

Head, Department of Computer Science, Bishop Heber College Tiruchirappalli, Tamilnadu, India[3]

pkeerthana.m.sc@gmail.com [1] *Thamilselvan1987@gmail.com* [2] *jgrsathiaseelan@gmail.com* [3]

Abstract- Image mining is one of the predominant researches in computer science field. It is a process of extracting valuable information from a huge amount of dataset. In image mining, there are several techniques are adopted such as image classification, image clustering, regression analysis and association rule mining. In this work, we have concentrated on image classification technique to analyze the performance of image mining algorithms based on classification accuracy, processing time, error rates, sensitivity and specificity. Image classification is an essential technique in image mining and it is most important part of medical image analysis. It has two level processes. In the first level, the typical model is built telling a determined collection of concept or data classes. In second level the model is used for classification. The main objective of this paper is to identify better method of image mining in medical image analysis based on the performance analysis. Some predominant image mining algorithms such as Classification, Regression Tree (CART), K-Means, Naive Bayes (NB), Decision Tree (DT) K-Nearest Neighbor and Support Vector Machine (SVM).These algorithms are used for the performance of medical image classification.

Keywords*: *Image Mining, Image classification, classification accuracy, performance analysis, medical images.

I. INTRODUCTION

Image mining is the process used to extract meaningful information from images. It deals with the embedded knowledge, extracting inherent data, image data relationship and other patterns that are not clearly found in the images [1]. There is a system which is Content Based Image Retrieval (CBIR) which aims at searching of images available in databases for any particular images so as to get a related image. The extracting images based on some features such as shape, colour, region and so on. On the other end, Retrieval of image is the fast developing and challenging research part in both unmoving and moving images. Especially, the medical image classification plays an important role in human diagnosis and treatment. It is

also used for healthcare students in the educational domain and studies by explaining with these images. Medical images are mainly used to detect specific diseases occur in the human body. In this, CAD act as supporting agent for the complete analysis of images and this system involves all cancer types as well as the coronary artery disease [2]. There are several proposed algorithms such as Gentle boost and Support Vector Machine (SVM) algorithms are used [3]. The retrieval of images in a huge collection is based on some projection colors and various mathematical representations were introduced and applied in the images. Further it is sub grouped as RGB combinations for retrieval process. The image retrieval process provides the most excellent solution in large image set comparing with 10000 images with the other various categories [4]. Consequently image mining is providing more attention for the researches in the field of information retrieval and multimedia databases. So, those researches can easily mine the relevant data from large amount of images are increasingly in order.

The algorithm called Support Vector Machine (SVM) which makes classification in a multidimensional space to separate different class labels. It also supports both the regression and classification tasks [5]. Image matching is more important in the field of mining images. Frequently used technique is nearest neighbourhood in which objects are represented as n-dimensional vectors. In [6] the visual queries are represented in the retrieval process. So that the images mainly based on the user request and the mechanism is considered as query-by-example used to compare the target images to find the image indices present in the image database. For ease of access digital medical images stored in huge databases as well as Content based image retrieval (CBIR) which is mainly used in diagnostic cases like query medical image. The CBIR images is based on some features such as edge, shape and colour which are extracted automatically [7]. If there is empty in the image set or less than the total images then the system randomly chosen the image for creating the association rules. This paper gives a survey on several techniques in image mining which was already proposed method they are Support Vector Machine, CART, Naive Bayes, Adaboost and Decision Tree. This paper provides best method in medical image classification based on the classification accuracy, processing time and error rates.

II. RELATED WORK

Deshpande et al [8] provides data mining approach which is used to identify the image content present in the association rules. The association rule algorithm helps to detect the regular item set with the help of some iterative methods. This algorithm helps to minimize the number of scans in Apriori algorithm. It is very essential to advance the image quality and make the extraction phase as simple and reliable.

Li-Hong Juang et al [9] focused on tracking tumor objects of (MRI) brain images by using K-means algorithm. The process which is also useful for detecting exact lesion objects in images. The main purpose of this algorithm is to resolve the MRI image by changing the gray-level image into colour image. S.L.A. Lee et al [10] concentrated on lung nodule detection which is used to spot the lung abnormalities in CT lung images with the help of Random forest algorithm. This algorithm provides hybrid random forest based nodule classification. It is also used to detect 32 patients with 5721 images. The accuracy in proposed system is noted as 97.11 whereas in the developed system the high receiver operator characteristic is given 97.86% accuracy.

Mahnaz Etehad Tavakol et al [11] provide the high infrared cameras to diagnose the vascular changes of breasts by using the adaboost algorithm. The algorithm is used to classify the invisible images into benign, malignant and normal. In this system the accuracy of 83% is given which gives better performance than the proposed system of 66%. Ming-Yih Lee et al [12] proposed an entropy based feature extraction and some other protocols for the breast cancer diagnosis using decision tree algorithm. The Morphological operations used in this system to detect the unified abnormal regions. This method gives 86% accuracy which is better than the proposed system of 59%.

Ye Chen et al [13] focused on the detection of brain structural changes from the Magnetic resonance images which helps to aid the treatment of neurological diseases with the help of Support Vector Machine algorithm. In addition the algorithm which helps to analyse the MR images from the various datasets. The accuracy range between 70% and 87% are noted. Wen-Jie Wu et al [14] suggested both the classification accuracy and the optimal classification model which helps to detect the ultrasound breast tumor images by using genetic algorithm. The algorithm is to calculate the near optimal parameters to differentiate the tumor as benign or malignant. The accuracy of proposed system is 95% which is improved better in the developing system by reducing the biopsies of benign lesions.

Daniel J. Evers et al [15] has given the study to evaluate whether the optimal spectroscopy improve the accuracy of transthoracic lung biopsies using Classification and regression tree (CART) algorithm. Based on the derived parameter the algorithm classifies the type of tissue present in the system. The overall accuracy is 91% sensitivity.

Daniel J. Evers et al [16] has given the study to evaluate whether the optimal spectroscopy improve the accuracy of transthoracic lung biopsies using Classification and regression tree (CART) algorithm. Based on the derived parameter the algorithm classifies the type of tissue present in the system. The overall accuracy is 91% sensitivity.

Min-Chun Yang et al [17] enhance the naïve bayes classification algorithm by separating the ultra sound images pixel-

by-pixel then the image measured by gray scale is converted to binary image which is then evaluated by two-phase criteria. So, the detection sensitivity can be further developed. Shengjun Zhou et al [18] suggested that in the medical applications the images are segmented. To manage the segmentation, fuzzy c-means clustering do the classification of pixels into some divisions. Then the algorithm assigns the membership values for those pixels to form the centroid.

Ravi Babu et al. [19] focused to determine the image classification rate for the purpose of digital image classification. The K-Nearest neighbor algorithm uses the leaming technique to find out the classification time of those images. The lazy based and instance based are the two leaming techniques. To compare the curves the algorithm is used which based on some comparison. Finally the nearest neighbor classifiers used to measure the distance of the two curves [20].

III. COMPARATIVE ANALYSIS

In this part, the comparative results and the datasets are listed for the data mining algorithms. The accuracy of various algorithms is clearly noted in this study.

1. Dataset Description

Various image datasets helps to find the classification performance of data mining algorithms. The used data sets are shown in table 1.

S. NO	ALGORITHM	DATASET
1	SVM	Brain Images
2	CART	Lung Images
3	K-Means	Brain Lesion image Dataset
4	Naive Bayes	Breast Lesion Images
5	Decision Tree	Breast Images

Table 1. Dataset Description

2. Comparison of Data Mining Algorithms

This part lists out the positive and negative aspects used in various algorithms present in this paper for the data mining algorithm.

S. NO	ALGORITHM	PURPOSE	LIMITATIONS
1	SVM	It is used to analyze the MR images from the heterogeneous dataset	Some of the features are not properly used in local image features.
2	CART	This algorithm is used to enable the accurate fraction estimation of the substances	Complex classification steps are followed.
3	K-Means	It is used to find exact lesion objects	Parameters are not sufficient for the detection process
4	Naive Bayes	It aims to enhance computer aided system to offer real time detection It improves the detection sensitivity.	The speckle noises present in images affects the pixel classification Low scan speed.
5	Decision Tree	Thermograph images was projected for the feature extraction	Credibility and sensitivity are not accurate.

Table. 2. Comparison Table

IV. DISCUSSIONS

1. SVM

In this algorithm the local image features can be easily classified for analysing the data. Support vector machine is formally said to be discriminative classifier. So it scales the high dimensional data. For the best performance of searching the algorithm, SVM kernels are used. SVM also uses the non -traditional data like trees and strings which are used as input instead of feature factors. So both the small and large datasets can be applicable in SVM algorithm. To classify the MR image classification on the basis of local image features the SVM algorithm is best suitable to that. Also the noisy features are identified while analysing the data and it is clearly removed by some other classification process.

2. CART

The classification and regression tree (CART) algorithm is mainly used for the classification of different tissues in image mining, which is on the basis of several derived parameters. The recursive partitioning method used in the CART algorithm to introduce the tree based modelling which is later converted to the statistical mainstream. To select the optimal tree value the algorithm involves the cross validation scheme from some rigorous approaches. Based on the technique called surrogate splits the algorithm automatically handles the missing values. For example the variable (x=t1) is selected then the greatest separation is produced so (x=t1) is said to be split. If this variable X it sends to which is less than t1 then the data is send to left or else it sends to right. The process is repeated for all the nodes. So that it is easy to conclude that CART algorithm uses only the binary splits.

3. K-Means

K-Means algorithm is said to be an unsupervised clustering algorithm. It works well for numerical data alone. The pixel-by-pixel image classification is possible by defining single and multiple thresholds. So that histogram statistics is used in this algorithm for the pixel based classification. The main work of this process is to check whether the histogram is bimodal or not. If it is then the gray value will be appeared otherwise the images get partitioned into several regions. The threshold of gray value can be determined using the peak values. However it converges only the local minimum values. So the algorithm involves number of clusters for the optimization.

4. Naive Bayes

The Naive bayes algorithm is the most powerful technique. It does the testing process easily and the classification problems can be solved. It can be able to build a model fastly and giving better predictions. To find the missing data the naïve bayes algorithm plays a major role. The unseen data can be easily predicted by characterizing the problem in naïve bayes method. During the construction time and prediction time this algorithm separates the attributes value. The probability of each attributes in isolation process needs only the enough data. So, there is no need of more data collection in this algorithm. Finally, if the data has high correlated features the performance will be degraded.

5. Decision Tree

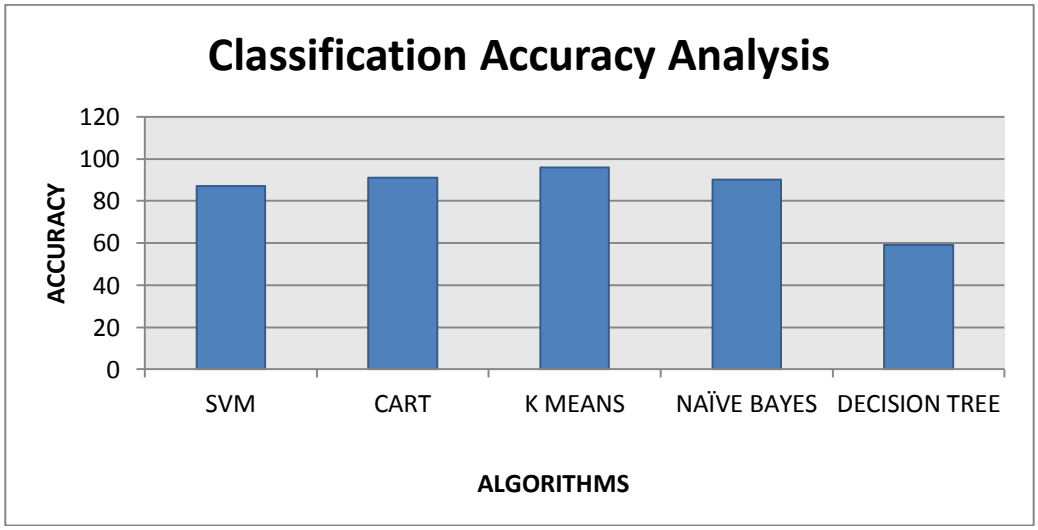
Decision tree algorithm is one of the classifier technique which is in the form of tree structure. For classification and prediction, the powerful tools are available in this algorithm. It has four divisions such as Decision node, leaf node, edge and path. A single attribute is represented in the decision node. Leaf node defines the target attribute. Splitting of one attribute is edge and the path is a final decision. For continuous attribute this algorithm is not applicable.

6. Performance Analysis

The overall performance of given algorithms such as SVM, CART, K- Means, Naive Bayes and Decision Tree are taken for the analysis in data mining.

S. No	METHOD	CLASSIFICATION ACCURACY
1	SVM	87%
2	CART	91%
3	K-Means	96%
4	Naive Bayes	90%
5	Decision Tree	59%

Table 3. Performance Analysis



From this study, the classification accuracy is analysed from various data mining algorithms. The accuracy level has both the positive and negative values.

V. CONCLUSION

In this study, the overall performance of various algorithms present in this paper was analyzed based on the classification accuracy. Accuracy of such algorithms found to be SVM gives 87%, CART is 91%, k- means shows 96%, Naïve bayes 90% and finally Decision tree shows 59%.The above accuracy in image classification is the main idea of evaluating the performance in data mining algorithms. The overall result shown in this paper is step into further development in future technology. To evaluate the best indications clinical studies are more essential. This paper projected several parameters which is used by data mining methods. By comparing these algorithms, k-means results shows better performance among the other methods presented in this work and it is also gives the best classification accuracy as well as saves the computing time. In future, we have planned to enhance decision tree method to improve the classification accuracy, sensitivity, specificity and as well as processing time.

REFERENCES

[1] C. Lakshmi Devasena, T.Sumathi, Dr. M. Hemalatha "An Experiential Survey on Image Mining Tools, Techniques and Applications" International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 3, pp. 5061-5067, 2011.

[2] E. Arnoldi, M. Gebregziabher, U.J. Schoepf, R. Goldenberg, L. Ramos-Duran, P.L. Zwerner, K. Nikolaou, M.F. Reiser, P. Costello, C. Thilo "Automated computeraided stenosis detection at coronary CT angiography: initial experience", Elsevier, Vol. 20, No.5, pp. 1160-1167, 2010.

[3] Lefkovits Sz, Leftkovits L. "Enhanced Gabor Filter Based Facial Feature Detector, The proceeding of the European Integration- Between Tradion and modernity", Elsevier, computer science section, vol. 1, pp. 78738-8863, 2013.

[4] R.Venkata Ramana Chary, Dr.D.Rajya Lakshmi and Dr. K.V.N Sunitha "Feature extraction methods for color image similarity", Advanced Computing: An International Journal (ACIJ), Vol.3, No.2, pp. 147-157, 2012.

[5] Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, Alok Choudhary, "A Lung Cancer Outcome Calculator Using Ensemble Data Mining on SEER Data", BIOKDD 2011, No. 5, 2011.

[6] S.Balan and T.Devi, "Design and Development of an Algorithm for Image Clustering In Textile Image Retrieval Using Color Descriptors", International Journal of Computer Science, Vol.2, No.3, pp. 199-211, 2012.

[7] Vamsidhar Enireddy, Kiran Kumar Reddi, "A Data Mining Approach for Compressed Medical Image Retrieval", International Journal of Computer Applications (0975 – 887) Vol. 52, No.5, 2012.

[8] D. Deshpande, "Association Rule Mining Based on Image Content", International Journal of Information Technology and Knowledge Managem ent, Vol. 4, No. 1, pp. 143-146, 2011.

[9] Li-Hong Juang , Ming-Ni Wu "MRI brain lesion image detection based on color-converted K-means clustering segmentation", elsevier, computer science section, vol.43 ,No.7 , pp. 941-949, 2010.

[10] S.L.A. Leea, A.Z. Kouzania, E.J. Hub "Random forest based lung nodule classification aided by clustering", elsevier, computer science section, vol.34, No.7, pp. 535-542, 2010.

[11] Mahnaz EtehadTavakol , Vinod Chandran , E.Y.K. Ng , Raheleh Kafieh, "Breast cancer detection from thermal images using bispectral invariant features", elsevier, International journal of thermal sciences, vol.69, pp. 21-36, 2013.

[12] Ming-Yih Leea, Chi-Shih Yang, "Entropy-based feature extraction and decision tree induction for breast cancer diagnosis with standardized thermograph images", elsevier, computer science methods and programs in biomedicine, vol.100, No.3, pp. 269-282, 2010.

[13] Ye Chena, Judd Storrs, Lirong Tana, Lawrence J. Mazlackc, Jing-Huei Leeb, Long J. Lua, "Detecting brain structural changes as biomarker from magnetic resonance images using a local feature based SVM approach", elsevier, Journal of Neuroscience methods, vol. 221 , No. 15 , pp. 22-31, 2014.

[14] Wen-Jie Wua, Shih-Wei Lina, Woo Kyung Moonb, "Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images", elsevier, computer science section, vol. 36 , No. 8 , pp. 627-633, 2012.

[15] Jarich W. Spliethoff , Daniel J. Ever, Houke M. Klomp, Johanna W. van Sandick, Michel W. Wouters, Rami Nachabe, Gerald W. Lucassen, Benno H.W. Hendriks, Jelle Wesseling, Theo J.M. Ruers," Improved identification of peripheral lung tumors by using diffuse reflectance and fluorescence spectroscopy", elsevier, computer science section, vol.80 , No. 2 , pp. 165-171, 2013.

[16] Min-Chun Yang, Chiun-Sheng Huang, "Whole breast lesion detection using naive bayes classifier for portable ultrasound", elsevier, computer science section, vol. 38, No.11, pp. 1870-1880, 2012.

[17] Shengjun Zhou, Yuanzhi Cheng, Shinichi Tamura, " Automated lung segmentation and smoothing techniques for inclusion of juxtapleural nodules and pulmonary vessels on chest CT images", elsevier, computer science section, vol. 13, pp. 62-70, 2014.

[18] U Ravi Babu, Y. Venkaswarlu, Aneel Kumar Chintha. "Handwritten Digit Recognition Using K-Nearest Neighbour Classifier" IEEE World Congress on Computing and Communication Technologies ISBN: 978-1-4799-2876-7 pp 60-65, 2014.

[19] U Ravi Babu, Y. Venkaswarlu, Aneel Kumar Chintha. "Handwritten Digit Recognition Using K-Nearest Neighbour Classifier" IEEE World Congress on Computing and Communication Technologies ISBN: 978-1-4799-2876-7 pp 60-65, 2014.

[20] Umut Konur, Fikret S. Gurgen, " Computer aided detection of spina bifida using nearest neighbor classification with curvature scale space features of fetal skulls extracted from ultrasound images, vol. 85, pp. 80-95, 2015.

