Evaluating the performance of different classification methods on medical X-ray images

Amin Khatami¹ · Sahar Araghi² · Toktam Babaei¹

© Springer Nature Switzerland AG 2019

Abstract

X-ray images are broadly used for diagnosis and analysis in medical science. In a number of medical applications, such as disease detection, automatic classification systems are very beneficial. Therefore, finding an appropriate classification technique is helpful. In this paper, it is aimed to compare the performance of different classification methods on categorizing X-ray images. In this regard, we have done our experiments in two phases. In phase 1, the effect of wavelet transformation (WT) as a feature selection method is evaluated. Obtaining almost similar performance (about 78%) for probabilistic neural network (PNN) and WT-based PNN, while reducing computational cost, leads us to apply WT on datasets in phase 2. During phase 2, random forest (RF), decision tree (DT), support vector machine (SVM) and Naive Bayes (NB) are applied on obtained images from WT. The results reveal better performance of RF compared to other methods by 82% accuracy. With very close accuracy, SVM got the next place (81%). DT and NB classifiers are the next ones by about 66% accuracy.

Keywords Medical X-ray images classification · Machine learning · Wavelet transformation

1 Introduction

X-ray images are a very common and useful detecting tool for several diseases including skeleton defections, lung, kidney and gall problems, and dental cavities. Classification of these medical images is an essential step in analyzing them and will assist medical experts in making decision. However, manual and conventional keyword-based classification methods of medical images are very demanding in terms of time and labor requirements. To address this problem, many researches have been conducted on using computer vision and machine learning techniques to retrieve medical images from large databases. In this regard, the content-based image retrieval (CBIR) systems have played a major role in clinical image analysis context [1–3]. The content in CBIR may refer to colors, shapes, texture or any other information which

can be extracted from image itself rather than annotated indices. Designing a robust CBIR system needs to design a reliable classifier. Researchers have embedded different classification schemes.

El-Naqa et al. [3] proposed a CBIR system which uses a learning scheme based on human observer perception samples. They evaluated their method by retrieving images from a database containing clinical mammograms with clustered micro-calcifications.

Qasem et al. [2] developed a radial basis function (RBF) network in which centers and weights of the network are evolved at the same time using a multi-objective optimization technique. They have used this model for disease diagnosis task and reported the results as well.

Rahman et al. [4] proposed a SVM-based classifier that is used to build an image retrieval system.

✉ Sahar Araghi, saraghi@swin.edu.au; Amin Khatami, skhatami@deakin.edu.au; Toktam Babaei, tbabaeij@deakin.edu.au | ¹Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Geelong, Australia. ²The Centre for Transformative Innovation (CTI), Swinburne University, Melbourne, Australia.

Bhattacharya et al. [5] proposed another framework by linearly combining multi-class SVM (MSVM) and fuzzy c-mean clustering method. For a query image, a fusion-based similarity matching function uses the obtained membership scores to retrieve the images with maximum similarity to the query image.

Ko et al. [6] also presented a CBIR system that uses the RF classifier for processing of features extracted from local wavelet binary patterns. They conducted their experiments on IRMA X-ray image data and showed that their model outperforms other classification methods.

In another research [1], a hybrid multiple classifier system is introduced to analyze breast thermograms with asymmetric features. It consists of individual classifiers, a neural fuser that combines individual classifier, along with a fuzzy measure which evaluates the diversity of the ensemble and removes redundant classifier from the ensemble.

Dealing with high-dimensional databases, feature extraction is an usual strategy that results in a reduced set of features and dimensionality reduction. Among different feature extraction techniques, WT is an important frequently used one. For classification task, information contained in WT explains discriminative features of raw data. Indeed, the wavelet coefficients will serve as inputs to a classifier for further analysis. In medical image classification problem, several researchers have utilized different levels of WT to extract texture features [7–9], and [10].

ImageCLEFmed 2009 is a popular database of medical images, which has been used by many researchers in this field to benchmark their proposed image retrieval models. The aim of this study is to compare the performance of different common classification models when they are combined with WT. For this purpose, we are conducting our experiments on ImageCLEFmed 2009 database. The acquired results are then compared with each other based on the F-measure, recall and precision criteria.

This paper is structured as follows: The technical aspects of developed model and competitor models are described in Sect. 2. Comparative analysis of the results is presented in Sect. 3. Finally, conclusion is given in Sect. 4.

2 Methods

In this section, the fundamentals and concepts of the classification methods utilized in this work are given.

2.1 Wavelet transformation (WT)

WT is an analytic dimensionality reduction technique. Haar wavelet conversion is a commonly used wavelet

transformation in applied mathematics for feature extraction. Let us denote the Haar mother function by $\phi(x)$.

$$\phi\left[\frac{x-b}{a}\right], \text{ with } (a, b) \in \mathbb{R}^+ \times \mathbb{R} \quad (1)$$

where $a = 2^{-j}$ and $b = 2^{-j}k$ for $j, k \in \mathbb{N}$. The dilations of ϕ will be calculated as follows:

$$\phi_{jk}(x) \propto \phi(2^j x - k) \quad (2)$$

the $\{\phi_{jk}\}$ s construct an orthogonal basis in $L^2(\mathbb{R})$ i.e:

$$\langle \phi_{jk} \phi_{lm} \rangle = \int_{-\infty}^{\infty} \phi_{jk} \phi_{lm} dx = \delta_{jl} \delta_{km} \quad (3)$$

where δ is the Kronecker delta function.

Hence, any function $\in L^2(\mathbb{R})$ can be represented by a linear combination of $\phi_{jk}(x)$ s.

For classification purpose, the Haar wavelet-based extracted features are fed as input to the classifier. Various researches, including the present work, have been conducted on classifying medical images using wavelet coefficients as features [11–13].

2.2 Probabilistic neural network (PNN)

PNN introduced by Specht [14] is a feed-forward neural network model that uses RBF as a nonlinearity element in its learning procedure. It applies the Euclidean distance of original inputs.

Considering a sample x presented to the input layer, the next layer (pattern layer) produces an output as follows:

$$\phi_{ij} = \frac{1}{(2\pi)^{d/2} \sigma^d} \exp \left[-\frac{(x - x_{ij})^T (x - x_{ij})}{2\sigma^2} \right] \quad (4)$$

where d denotes the dimension of the sample vector, σ is called smoothing parameter and x_{ij} denotes neuron vector.

The next layer (summation layer) then computes the maximum likelihood of the sample being classified as c_i as below:

$$p_i(x) = \frac{1}{(2\pi)^{d/2} \sigma^d} \frac{1}{N_i} \sum_{j=1}^{N_i} \exp \left[-\frac{(x - x_{ij})^T (x - x_{ij})}{2\sigma^2} \right] \quad (5)$$

where N_i is the number of samples in class c_i .

Finally, the output (decision) layer decides on sample's class according to the following:

$$\hat{c}(x) = \operatorname{argmax}\{p_i(x)\}, \quad i = 1, 2, \dots, m \quad (6)$$

2.3 Decision tree (DT)

DT learning is a construction of a decision tree from training data. A decision tree is a flowchart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node.

DT classifiers usually include two phases of building and pruning. Building the tree includes a proper selection of features which maximizes dissimilarity between classes. Various approximation techniques were proposed in this regard. Among them, Gain ratio [15], Gini index [16] and Chi-square [17] are more popular. The Gini index measures the impurity of a given case with respect to the other classes. For a given training set, T , the Gini index is calculated as follows:

$$\sum_{i \neq j} \sum (f(c_i, T)/|T|)(f(c_j, T)/|T|) \quad (7)$$

where $f(c_i, T)/|T|$ is the probability that a selected case has the class label c_i . As such, given a combination of features, a decision tree is built up with no pruning.

Gini index is usually used by RF algorithm (next classifier) as a criterion for best split selection. In pruning phase, the leaves and branches, which include one or a few data, will be removed. Different algorithms have been proposed to develop such classifier including *ID3* [18], *C4* [19], *CHAID* [20], *CART* [21].

2.4 Random forest (RF)

Over the recent decades, ensemble learning algorithms have received increasing interest from the machine learning research community. This is because of their higher accuracy and their robustness to the noises compared to single classifiers.

RF is an ensemble learning framework suggested by Breiman [22]. RF classifier operates by building a pool of decision trees during the training. Each decision tree is developed by some random selection of features. The class label is estimated based on the most voted class label (mode) by all single base classifiers (trees) for the input vector (X).

$$\hat{C}_{RF} = \text{Mode}\{c_b(\hat{X})\}_1^B \quad (8)$$

where $c_b(X)$ is the class label estimated by b th RF tree. The various subsets of training data are generated by bootstrap or bagging methods, and then different decisions are built based on them [23].

RF is shown to be a powerful classifier. The key advantages of it include:

- It performs efficiently on both two-class and multi-class problems.
- It has high scalability, is able to process huge amount of data and has been shown to be efficient in high-dimensional problems.
- It has a good performance results even when data are noisy.
- It gives measures of feature's importance.

2.5 Naive Bayes (NB)

NB, which is also known as independence Bayes, is a probabilistic method that utilizes the Bayes theorem. It assumes that the features are independent given the class label. NB-based classifiers find a class by calculating the maximum posterior probability conditioned on the new sample [24]. For example, in a classification problem with K classes labeled as $c_1, c_2, \dots, c_K$, if $X = (x_1, x_2, \dots, x_n)$ denotes a sample coming from a dataset with n independent features, the following probabilities will be assigned to it according to Bayes theorem:

$$p(c_k | x_1, x_2, \dots, x_n) = \frac{p(c_k)p(X|c_k)}{p(X)} \quad (9)$$

Now, because of the dependency of features, the conditional distribution of the class variable c will be:

$$\begin{aligned} p(c_k | x_1, x_2, \dots, x_n) &\propto p(c_k, x_1, x_2, \dots, x_n) \\ &\propto p(c_k)p(x_1|c_k)p(x_2|c_k)p(x_3|c_k) \dots \\ &\propto p(c_k) \prod_{i=1}^n p(x_i|c_k) \end{aligned} \quad (10)$$

or:

$$p(c_k | x_1, x_2, \dots, x_n) = \frac{1}{Z} p(c_k) \prod_{i=1}^n p(x_i|c_k) \quad (11)$$

where $Z = p(X)$ is a scaling factor dependent on $x_1, x_2, \dots, x_n$. Combining this probability model with a decision rule, for example, picking the most probable class (maximum posteriori), will result in NB classifier as follows:

$$\hat{y} = \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} p(c_k) \prod_{i=1}^n p(x_i|c_k) \quad (12)$$

where $\hat{y}$ is the estimated output.

This method has been used for different classification problems, including text categorization and automated medical diagnosis. Despite their simplicity and

the strong assumption of independency, their performance is reasonable even in the case of complex classification problems [25]. Moreover, small amount of data is needed for training.

2.6 Support vector machine (SVM)

In machine learning, SVM is one of the state-of-the-art models for supervised learning schemes including classification. To perform nonlinear classification, SVM uses a kernel function to transform the input data into higher dimensional space and then classify them linearly. Training a SVM involves solving a quadratic programming (QP) problem. For instance, consider a two-class classification problem by denoting the input–output pairs as $(x_i, y_i), i = 1, 2, \dots, n$ where $y_i \in \{-1, +1\}$ is class labels. To train the corresponding SVM, the following QP problem needs to be solved:

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j K(x_i x_j) \alpha_i \alpha_j \quad (13)$$

subject to:

$$0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^n y_i \alpha_i = 0 \quad (14)$$

where α_i s are Lagrange multipliers. C is SVM's hyperparameter and $K(\cdot)$ is the kernel function.

In the present work, we built our SVM using sequential minimal optimization (SMO) algorithm, a kernel function suggested by El-Naqa et al. [3]. SMO is a numerical optimization method for dealing with the QP problem. It splits down the problem into subproblems, which are then solved analytically [26, 27]. Since we are dealing with huge matrices, it is appropriate to use SMO.

3 Classification results and discussions

In this section, the performance of different classification methods on body X-ray images is compared. A sample of images in each class is shown in Fig. 1. IRMA database [28] is used for experiments. Features are extracted from the deep structured architecture for a large database consisting of 35 different parts of body X-ray images. We used ImageCLEFmed2009 IRMA database, and our experiment includes 11,211 MRI images. Experimental data are randomly split into training and testing sets, which contain 80% and 20% of data, respectively.

MRI images are usually available in different shapes and sizes. In this paper, to reduce the computational time of experiments, grayscale models with two dimensions are considered. All images are represented in two dimensions, re-sized to 56×56 and reshaped into row images with 3136 features. Images are categorized by reference coding using a mono-hierarchical coding scheme [28]. A reference code is a string of characters consisting of four parts, for example: TTTT, DDD, AAA and BBB, which is an instance of reference code describing technique, direction, anatomy and bio-system of an image, respectively. Figure 2 presents the detailed properties of the images per category, which are analyzed by experts [28, 29].

The experiments in this paper are divided to two phases. First, we study the benefit of using WT method and PNN is the classification method applied in this part. During the next phase, we compare the accuracy of RF, DT, SMO and NB on classifying X-ray images.

As it is mentioned earlier, WT is a beneficial method to extract features from high-dimensional database. In this study, we have done fivefold experiments to measure the accuracy of PNN and WT-based PNN.

The applied PNN contains 400 hidden neurons, and this number is obtained by trial and error in the range of 50–500 neurones. In the first round of experiments, PNN is applied to the data while for the next experiment, first WT is applied on the row data for reducing the dimension and then PNN is applied on obtained images (WT-PNN).

Fig. 1 A sample of different categories of images in the IRMA database

Class Name	Anatomic	Direction
1	Cranium	Coronal
2	Cervical	Coronal
3	Lumbar spine	Coronal
4	Hand	Coronal
5	Radio carpal joint	Coronal
6	Elbow	Coronal
7	Shoulder	Coronal
8	Chest	Coronal
9	Chest bones	Coronal
10	Abdomen	Gastrointestinal system
11	Thoracic spine	Coronal
12	Shoulder	Sagittal
13	Pelvis	Coronal
14	Foot	Coronal
15	Ankle joint	Coronal
16	Lower leg	Coronal
17	Knee	Coronal
18	Upper leg	Coronal
19	Hip	Coronal
20	Left breast	Other orientation
21	Cervical spine	Sagittal
22	Thoracic spine	Sagittal
23	Radio carpal joint	Sagittal
24	Forearm	Sagittal
25	Elbow	Sagittal
26	Chest	Sagittal
27	Ankle joint	Sagittal
28	Knee	Sagittal
29	Right breast	Axial
30	Left breast	Axial
31	Knee	Axial
32	Facial cranium	Other orientation
33	Right breast	Other orientation
34	Hand	Other orientation
35	Neuro cranium	Sagittal

Fig. 2 Properties of the X-ray images per category

The results, which are presented in Table 1), show PNN and WT-PNN have very close accuracy. WT has re-sized 56×56 images to 28×28 . The average accuracy of PNN is 78.07% and of WT-PNN is 77.95%. This result illustrates that WT-PNN still has high accuracy while using a scaled-down image. This leads us to use WT for our subsequent experiments to reduce the computational cost while preserving almost similar accuracy.

Next contribution of this paper is comparison between four classification techniques on 35 classes of images. RF, DT, SMO and NB are the methods applied for image classification. Here, Haar WT is applied on images to reduce their size from 56×56 to 28×28 , and then different classifications are applied on re-sized images. Haar WT has orthogonality property that allows the filter feature selection methods to be employed efficiently since the dependence among features is ruled out. Figure 3 shows the process of evaluating the accuracy of each classification method.

It should be noted that for X-ray images using low pass filter is more appropriate. Approximation coefficients matrix

Fig. 3 Process of evaluating accuracy of classification methods on X-ray data

derived by a low pass filter conveys features better than the detail information. Radiological images are textured, and by applying a high-pass filter, the texture information might be removed [30].

3.1 Performance measurements

To evaluate the performance of the proposed classification methods, the results are analyzed in terms of accuracy and F-measure. F-measure is a single combined metric of a classifier's efficiency. F-measure (F_β) is formulated as below:

$$F_\beta = \frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \quad (15)$$

where F_β measures the effectiveness of retrieval with respect to a user who attaches times as much importance to recall as precision.

Precision (Eq. 16) is the number of accurately labeled samples coming with the positive class divided by the number of labeled samples as belonging to the positive class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (16)$$

Recall (Eq. 17) is the number of accurately labeled samples as belonging to the positive class divided by the total number of samples belonging to the positive class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (17)$$

where TP is true positive, FP is false positive and FN is false negative.

3.2 Results and discussion

RF, DT, SMO and NB are applied to the images. Precision, recall and f-measure are calculated for five tests, and

Table 1 Average percentage accuracy of five tests for PNN and WT-PNN

Method	PNN	WT-PNN
Average	78.07	77.95
Median	78.39	78.06

Table 2 Performance results for different classification methods

Method	Precision	Recall	F-measure	Time (s)
RF	0.82	0.83	0.82	46
DT	0.66	0.66	0.66	60
SMO	0.82	0.81	0.81	231
NB	0.68	0.65	0.66	1.8

average of them is shown in Table 2. The time required to build each model is also measured and presented.

Results reveal that the best performance is for RF with about 82% accuracy and 83% recall. SMO with about 81% accuracy and 81% recall has a very close performance to RF; however, having a comparison between the times required to build the model shows superiority of the RF. RF needs almost 46 s to build the model, while 231 s is needed when SMO is used.

DT and NB have close performance and are able to do classification by about 66% accuracy and 66% and 65% recall, respectively. NB requires the shortest time to build the model which is about 1.8 s. DT with almost similar performance takes 60 s to build the model. Note that about 78% accuracy is obtained for PNN method as mentioned in the previous section. Therefore, to present classification superiority order for X-ray image, we have RF, SMO, PNN and then DT and NB, respectively.

4 Conclusion

Extracting and classifying the medical images to different categories is an important issue in medical image processing. In the case of working with large databases, this issue gets more significant.

In this research work, both of these challenges are studied. At first, a set of experiments are done to evaluate the benefits of using WT technique. First, we used pure PNN, and next time, we applied WT combined with PNN. Obtained results illustrate reducing the computational cost of the process when WT is used. For the next phase, we applied RF, DT, SMO and NB to evaluate their accuracy of performance in classifying X-ray images. RF with about 82% accuracy has the highest place among other classification techniques. SMO with almost close performance has the second place, and PNN, DT and NB are the next ones, respectively. In our future work, we will study better ways of feature selection and work on larger datasets.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

- Krawczyk B, Schaefer G (2014) A hybrid classifier committee for analysing asymmetry features in breast thermograms. *Appl Soft Comput* 20:112–118
- Qasem SN, Shamsuddin SM (2011) Radial basis function network based on time variant multi-objective particle swarm optimization for medical diseases diagnosis. *Appl Soft Comput* 11(1):1427–1438
- El-Naqa I, Yang Y, Galatsanos NP, Nishikawa RM, Wernick MN (2004) A similarity learning approach to content-based image retrieval: application to digital mammography. *IEEE Trans Med Imaging* 23(10):1233–1244
- Rahman MM, Antani SK, Thoma GR (2011) A learning-based similarity fusion and filtering approach for biomedical image retrieval using svm classification and relevance feedback. *IEEE Trans Inf Technol Biomed* 15(4):640–646
- Bhattacharya P, Rahman M, Desai B (2006) Image representation and retrieval using support vector machine and fuzzy c-means clustering based semantical spaces. In: 18th international conference on pattern recognition, 2006. *ICPR 2006*, vol 2, pp 1162–1168
- Ko BC, Kim SH, Nam J-Y (2011) X-ray image classification using random forests with local wavelet-based cs-local binary patterns. *J Dig Imaging* 24(6):1141–1151
- Tan Y, Li G, Duan H, Li C (2014) Enhancement of medical image details via wavelet homomorphic filtering transform. *J Intell Syst* 23(1):83–94
- Torbati N, Ayatollahi A, Kermani A (2014) An efficient neural network based method for medical image segmentation. *Comput Biol Med* 44:76–87
- Lei B, Tan E-L, Chen S, Ni D, Wang T, Lei H (2014) Reversible watermarking scheme for medical image based on differential evolution. *Expert Syst Appl* 41(7):3178–3188
- Nguyen T, Khosravi A, Creighton D, Nahavandi S (2015) Medical data classification using interval type-2 fuzzy logic system and wavelets. *Appl Soft Comput* 30:812–822
- Martis RJ, Acharya UR, Min LC (2013) Ecg beat classification using pca, lda, ica and discrete wavelet transform. *Biomed Signal Process Control* 8(5):437–448
- Nanthagopal AP, Rajamony RS (2013) Classification of benign and malignant brain tumor ct images using wavelet texture parameters and neural network classifier. *J Vis* 16(1):19–28
- Yan S, Jianming L, Yahagi T (2006) Texture classification for liver tissues from ultrasonic b-scan images using testified pnn. *IEICE Trans Inf Syst* 89(8):2420–2428
- Specht DF (1990) Probabilistic neural networks. *Neural Netw* 3(1):109–118
- Quinlan JR (2014) C4. 5: programs for machine learning. Elsevier, Amsterdam
- Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC Press, Boca Raton
- Mingers J (1989) An empirical comparison of selection measures for decision-tree induction. *Mach Learn* 3(4):319–342
- Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1(1):81–106

19. Salzberg SL (1994) C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. Mach Learn 16(3):235–240
20. Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. J Royal Stat Soc: Ser C Appl Stat 29(2):119–127
21. Lewis RJ (2000) An introduction to classification and regression tree (cart) analysis. In: Annual meeting of the society for academic emergency medicine in San Francisco, California, pp 1–14
22. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
23. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
24. Duda RO, Hart PE et al (1973) Pattern classification and scene analysis, vol 3. Wiley, New York
25. David KY, Hand J (2001) Idiot's bayes: not so stupid after all? Int Stat Rev 69(3):385–398
26. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the 5th annual workshop on computational learning theory, ser. COLT'92. ACM, New York, NY, USA, pp 144–152. <https://doi.org/10.1145/130385.130401>
27. Osuna E, Freund R, Girosi F (1997) An improved training algorithm for support vector machines. In: Neural networks for signal processing [1997] VII. Proceedings of the 1997 IEEE workshop, pp 276–285
28. Lehmann TM, Gold M, Thies C, Fischer B, Spitzer K, Keysers D, Ney H, Kohnen M, Schubert H, Wein BB (2004) Content-based image retrieval in medical applications. Methods Inf Med 43(4):354–361
29. Lehmann TM, Güld MO, Deselaers T, Keysers D, Schubert H, Spitzer K, Ney H, Wein BB (2005) Automatic categorization of medical images for content-based retrieval and data mining. Comput Med Imaging Graph 29(2):143–155
30. Acharya T, Ray AK (2005) Image processing: principles and applications. Wiley, New York

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.