World Applied Sciences Journal 35 (10): 2124-2128, 2017
ISSN 1818-4952
© IDOSI Publications, 2017
DOI: 10.5829/idosi.wasj.2017.2124.2128

Architecture Model for Semantic Classification of Chest X-Ray Images Using Machine Learning Algorithm

[1]*Norkhairani Abdul Rawi,* [1]*Mohd Nordin Abdul Rahman,*
[1]*Mokhairi Makhtar,* [2]*Norhasiza Mat Jusoh,* [1]*Abd Rasid Mamat*

[1]Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Malaysia
[2]Faculty of Medicines, Universiti Sultan Zainal Abidin, Malaysia

Abstract: The increment of digital images captured through multiple computerized medical facilities has been an attraction for the researcher in the field of image processing. Even though there are number of system for medical images, it is still lack in term of supporting image analysis at low cost. Currently digital medical images, specifically chest x-ray (CXR) are classified manually by radiologists where it is labor intensive, repetitive and requires a highly trained experts. Therefore misclassifications can be occurred because it can be influenced by human subjectivity or tiredness and other reasons such as to experimental condition and variable image quality. Thus this research aims to have a reference database that represents the meaning or semantic of digital images so the classification is done by computerized program. This paper will discuss the related work in related field and proposing a approach to carry out the whole research. The model will be used in the next phase of research.

Key words: Chext X Ray Images · Machine learning algorithm · classifications

INTRODUCTION

Medical imaging occupied 30% of world storage capacity in 2010 and mammography data in US in 2009 amounts 2.5 petabytes. In has been projected that it will increase up to 140 petabytes space need for digital images in picture archiving and communication systems (PACS) in 2016. Some have similarities and differential among them that can be classified to various categories of diagnose. Unfortunately, assessing chest radiographs (CXRs) requires specialized personnel that are not always available, which is a major obstacle toward their use, especially in the aforementioned resource-constrained regions [1]. Considering these circumstances, computer-aided detection (CAD) systems can prove valuable, as they can produce the desired assessment with limited or no human intervention at all as example shown by Maduskar *et Al*. [2] and Jaeger *et al*. [3]. Therefore a simple tool to detect basic class for chest x ray images which is opacity and lucencies will help early detection for any abnormalities regarding chest x ray images.

The structure of this paper is as follows. Introduction section briefs about the current situation in medical imaging and problems raised in the domain. The next section, related matters in semantic of images and categories of machine learning. The Architecture Model section portrays the architecture model to be deployed in the research. It also outlines the steps need to be taken and the formula to calculate the expected result. Finally, the Future Works section outlined the next task to be completed.

Related Work: The vast amount of multimedia data available today need to be managed in a proper way since the amount of data keep on increasing from time to time. A fast grow of multimedia data in various modalities (video, audio, 3D objects, etc) lead to a harder management, distribution and access to multimedia material, either for lay and professional users. Organizing the semantic element of multimedia data is even tougher. It is still need a further explored how technologies in different areas can be utilized to increase the value of

Corresponding Author: Norkhairani Abdul Rawi, Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Besut Campus, 22200, Terengganu, Malaysia.

multimedia data, e.g. for representation, analysis, annotation and clustering with semantic and knowledge based methods.

In the last decade, substantial progress has been made in the creation, transmission, presentation and analysis of multimedia data to facilitate the development of large scale multimedia information system. Research on how semantics technologies can be applied to the acquisition, generation, transmission, storage, processing and retrieval of large scale multimedia information system, still on a long way to go [4].

The fundamental issues of concern in managing multimedia data is the notorious semantic gap between what stored in the database and what the end-user understands and expect [5]. It leads to an ambiguous, subjective and transient (AST) semantic problem of multimedia data. Therefore, Li, Q., Zhao, J., & Zhu, X. [5] suggested necessary context information needs to be made available and provided in order to conduct meaningful and effective query processing of different paradigms and at various levels of abstraction.

In medical area, the accuracy in extracting information from medical imaging such as x-rays is a vital issue since it will determine the action and treatment to be taken. Diagnosing the right medical problem normally need to be done few time involving several procedure and sometimes different kind of experts. Therefore it is much easier if the images can be analyzed by a system that can depict the early signs of disease through the classification of images.

Soft computing can better exploit visual information in medical imaging for decision support where it able to find similar cases, use these including outcomes for diagnosis support. There is also a need to develop scalable solutions that allow treating the volumes produced in hospitals. It helps in detecting small regions of interest in medical images. It is done by mapping images to semantics, store only regions of interest. It then link information in reports with image data. Consequently, it make work of radiologists more efficient. It has been applied by few researchers (Qi, D., Denton, E. R., & Zwiggelaar, R., [6], Ogiela, L., Tadeusiewicz, R., & Ogiela, M. R., [7], Maduskar *et al*. [2] and Jaeger *et al*. [3]).

Machine learning is part of soft computing studies on computer algorithms for learning to do stuff. We might, for instance, be interested in learning to complete a task, or to make accurate predictions, or to behave intelligently. The learning that is being done is always based on some sort of observations or data, such as examples, direct experience, or instruction [8]. So in general, machine learning is about learning to do better in the future based on what was experienced in the past. The emphasis of machine learning is on automatic methods. In other words, the goal is to devise learning algorithms that do the learning automatically without human intervention or assistance. The machine learning paradigm can be viewed as "programming by example." Often we have a specific task in mind, such as spam filtering. But rather than program the computer to solve the task directly, in machine learning, we seek methods by which the computer will come up with its own program based on examples that we provide.

Machine learning is a core subarea of artificial intelligence. It is very unlikely that we will be able to build any kind of intelligent system capable of any of the facilities that we associate with intelligence, such as language or vision, without using learning to get there. These tasks are otherwise simply too difficult to solve. Further, we would not consider a system to be truly intelligent if it were incapable of learning since learning is at the core of intelligence. Although a subarea of AI, machine learning also intersects broadly with other fields, especially statistics, but also mathematics, physics, theoretical computer science and more. Machine learning tasks can be of several forms. Among approaches that available in machine learning are:

Supervised Learning: Algorithms are trained on labelled examples, i.e., input where the desired output is known. The supervised learning algorithm attempts to generalize a function or mapping from inputs to outputs which can then be used speculatively to generate an output for previously unseen inputs.

Unsupervised Learning: Algorithms operate on unlabeled examples, i.e., input where the desired output is unknown. Here the objective is to discover structure in the data (e.g. through a cluster analysis), not to generalize a mapping from inputs to outputs.

Semi-Supervised learning: Combines both labeled and unlabeled examples to generate an appropriate function or classifier.

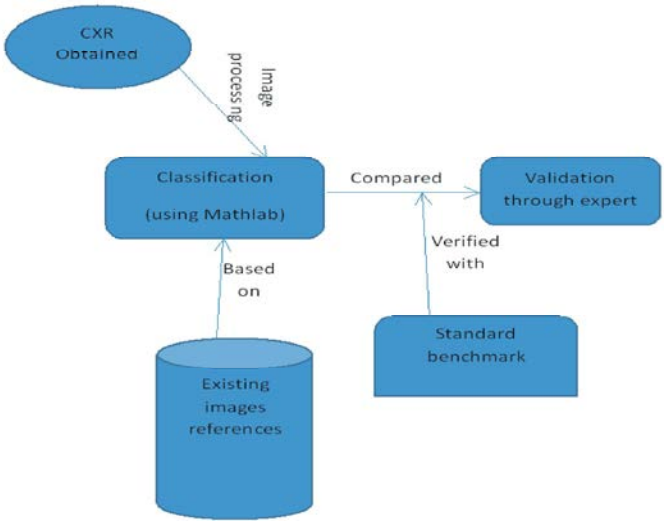
Transduction: Or transduction inference, tries to predict new outputs on specific and fixed (test) cases from observed, specific (training) cases.

Reinforcement Learning: Is concerned with how intelligent agents ought to act in an environment to maximize some notion of reward. The agent executes actions which cause the observable state of the environment to change. Through a sequence of actions, the agent attempts to gather knowledge about how the environment responds to its actions and attempts to synthesize a sequence of actions that maximizes a cumulative reward.

Learning to Learn: Learns its own inductive bias based on previous experience.

Developmental Learning: Elaborated for robot learning, generates its own sequences (also called curriculum) of learning situations to cumulatively acquire repertoires of novel skills through autonomous self-exploration and social interaction with human teachers and using guidance mechanisms such as active learning, maturation, motor synergies and imitation.

Architecture Model: Based on related work discussed in previous sections, an architecture model to execute the research in modelling a semantic classification for CXR images has been developed to serve as supporting tools for radiologists. It can be represented by the diagram below.



The researchers will carefully examine the existing CXR and processed it using established techniques and methods in image processing. Haralick features extraction which had fourteen (14) features will be used to extract data from the CXR images. Haralick features extraction is one of established techniques in image processing and has been widely used by many researchers (Wibmer *et al*. [9], Mohanaiah, P., Sathyanarayana, P., & GuruKumar, L. [10], Nagarajan, M. B. *et al*. [11], Samanta, S. [12] & Mamat, A. R. *et al*. [13]). Standard value for each of features will be identified to be the base of comparison using MathLab. Feature selection also will be done to select the most appropriate and significance features. For this section, existing images reference will be used to identify the ground value for each features. The value will be verified by the experts in order to ensure that ground value is accurate. As a result a standard benchmark will be produced.

Next, using appropriate Machine Learning Algorithm, where each algorithm from supervised learning category will be tested towards images to obtain the classification. Supervised learning is focused since the result to be achieved is known, however the accuracy according to experts need to be tested. The preprocessed data then will be tested using WEKA to classify the images. It will lead to choosing any appropriate algorithm to be combined to produce better result. To perform supervised learning, there is a need for a few steps as listed bellows:

Determine the Type of Training Examples: Before doing anything else, the researcher should decide what kind of data is to be used as a training set. In the case of handwriting analysis, for example, this might be a single handwritten character, an entire handwritten word, or an entire line of handwriting. For this research, the images of the whole lungs will be used.

Gather a Training Set: The training set needs to be representative of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.

Determine the Input Feature Representation of the Learned Function: The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object. The number of features should not be too large, because of the curse of dimensionality; but should contain enough information to accurately predict the output.

Determine the Structure of the Learned Function and Corresponding Learning Algorithm: For example, the engineer may choose to use support vector machines or decision trees.

Complete the Design: Run the learning algorithm on the gathered training set. Some supervised learning algorithms require the researcher to determine certain control parameters. These parameters may be adjusted by optimizing performance on a subset (called a validation set) of the training set, or via cross-validation.

Evaluate the Accuracy of the Learned Function: After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set.

Later, the specific class for CXR can be obtained and develop the semantics of the images can be interpreted. This collection of class will be used as gold standard database for further reference. Datasets available at GH and HU will be collected. The actual CXR obtained will be processed MathLab and classified according to selected Machine Learning Techniques using WEKA based on gold standard database as reference. The result of classification will compared to expert evaluation and verified against standard benchmark. Result of this process will be used to improve the model until it reach the acceptable standard benchmark.

Last part, the evaluation and user feedback. The developed model will be run in experiment session together with the experts particularly the radiologist from GH and HU chosen. The experiment session will evaluate the degree of accuracy (ACC), sensitivity (TPR) and specificity (SPC) of developed model using Machine Learning Techniques in helping diagnosing the disease related to CXR. The formula to compute are as follows:
(number of) positive samples (P)
(number of) negative samples (N)
(number of) true positive (TP)
(number of) true negative (TN)
(number of)false positive (FP)
(number of)false negative (FN)
1.0 Accuracy (ACC)

$$ACC = (TP + TN) / (TP + FP + FN + TN)$$

2.0 Sensitivity or true positive rate (TPR)

$$TPR = TP/P = TP / (TP + FN)$$

3.0 Specificity (SPC) or true negative rate

$$SPC = TN/N = TN / (TN + FP)$$
$$SPC = TN/N = TN / (TN + FP)SPC = TN/N = TN / (TN + FP)$$

Future Works: The proposed architecture will be used in the next part of research where processing of the images from the reference data will be preprocessed using MathLab and classified using WEKA. Results of classification will be verified by the experts in radiology and any further improvements will be carried out. The main focus of the research to improve the performance of classifications algorithm either by using assemble method or hybrid approach.

ACKNOWLEDGMENTS

This research is funded by the university under the code of UniSZA/2017/DPU/23 for twenty four (24) months. Authors like to thank the Research Management, Innovation and Commercialization Centre (RMIC) and Faculty of Informatic & Computing for all the supports given.

REFERENCES

1. Melendez, J., B. Van Ginneken, P. Maduskar, R.H. Philipsen, K. Reither, M. Breuninger and C.I. Sánchez, 2015. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays. IEEE Transactions on Medical Imaging, 34(1): 179-192.

2. Maduskar, P., M. Muyoyeta, H. Ayles, L. Hogeweg, L. Peters-Bax and B. van Ginneken, 2013. Detection of tuberculosis using digital chest radiography: automated reading vs. interpretation by clinical officers. The International Journal of Tuberculosis and Lung Disease, 17(12): 1613-1620.
3. Jaeger, S., A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan and G. Thoma, 2014. Automatic tuberculosis screening using chest radiographs. IEEE Transactions on Medical Imaging, 33(2): 233-245.
4. Troncy, R., B. Prabhakaran and Y. Cao, 2012. Multimedia data semantics: guest editors introduction. Multimedia Tools and Applications, 59(2): 579-583.
5. Li, Q., J. Zhao and X. Zhu, 2008. Multimedia data modeling through a semantic view mechanism. World Wide Web, 11(2): 287-311.
6. Qi, D., E.R. Denton and R. Zwiggelaar, 2006. Semantic analysis on medical images: a case study. In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on (Vol. 1, pp: 1260-1263). IEEE.
7. Scranton Gillette Communications. 2012. Storage of Medical Images to Become More Centralized. Available and retrieved on 28th March 2016 at http://www.itnonline.com/content/storage-medical-images-become-more-centralized.
7. Ogiela, L., R. Tadeusiewicz and M.R. Ogiela, 2008. Cognitive Approach to Medical Image Semantics Description and Interpretation. In proc. INFOS2008, 6th Int. Conf. on Informatics and Systems.
8. Syed, U., M. Bowling and R.E. Schapire, 2008. Apprenticeship learning using linear programming. In Proceedings of the 25th international conference on Machine learning (pp: 1032-1039). ACM.
9. Wibmer, A., H. Hricak, T. Gondo, K. Matsumoto, H. Veeraraghavan, D. Fehr and V.E. Reuter, 2015. Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. European Radiology, 25(10): 2840-2850.
10. Mohanaiah, P., P. Sathyanarayana and L. Guru Kumar, 2013. Image texture feature extraction using GLCM approach. International Journal of Scientific and Research Publications, 3(5): 1.
11. Nagarajan, M.B., M.B. Huber, T. Schlossbauer, G. Leinsinger, A. Krol and A. Wismüller, 2013. Classification of small lesions in breast MRI: evaluating the role of dynamically extracted texture features through feature selection. Journal of Medical and Biological Engineering, 33(1).
12. Samanta, S., S.S. Ahmed, M.A.M.M. Salem, S.S. Nath, N. Dey and S.S. Chowdhury, 2015. Haralick features based automated glaucoma classification using back propagation neural network. In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014 (pp. 351-358). Springer, Cham.
13. Mamat, A.R., M.K. Awang, N.A. Rawi, M.I. Awang and M.F.A. Kadir, 2016. Average Analysis Method in Selecting Haralick's Texture Features on Color Co-occurrence Matrix for Texture Based Image Retrieval. International Journal of Multimedia and Ubiquitous Engineering, 11(2): 79-88.
00. Bhatt, C.A., and M.S. Kankanhalli, 2011. Multimedia data mining: state of the art and challenges. Multimedia Tools and Applications, 51(1): 35-76.
00. Clinchant, S., J. Ah-Pine and G. Csurka, 2011. Semantic combination of textual and visual information in multimedia retrieval. In Proceedings of the 1st acm international conference on multimedia retrieval (pp: 44). ACM.
00. Depeursinge, A., A. Vargas, A. Platon, A. Geissbuhler, P.A. Poletti and H. Müller, 2012. Building a reference multimedia database for interstitial lung diseases. Computerized Medical Imaging and Graphics, 36(3): 227-238.
00. Han, J., M. Kamber and J. Pei, 2011. Data mining: concepts and techniques. Elsevier.
00. Versi, E., 1992. Gold standard" is an appropriate term.". BMJ 305(6846): 187. doi:10.1136/bmj.305.6846.187-b. PMC 1883235.PMID 1515860
00. Ko, B.C., S.H. Kim and J.Y. Nam, 2011. X-ray image classification using random forests with local wavelet-based CS-local binary patterns. Journal of Digital Imaging, 24(6): 1141-1151.
00. Xie, L. and R. Yan, 2008. Extracting semantics from multimedia content: challenges and solutions. In Multimedia Content Analysis (pp: 1-31). Springer US.
00. Peckham, J. and F. Maryanski, 1988. Semantic data models. ACM Computing Surveys (CSUR), 20(3): 153-189.