Computer Methods and Programs in Biomedicine 145 (2017) 85–93

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Pareto-optimal multi-objective dimensionality reduction deep auto-encoder for mammography classification

Saeid Asgari Taghanaki[*], Jeremy Kawahara, Brandon Miles, Ghassan Hamarneh

Medical Image Analysis Lab, Simon Fraser University, Canada

ARTICLE INFO

Article history:
Received 2 September 2016
Revised 21 March 2017
Accepted 12 April 2017

Keywords:
Breast cancer
Computer aided diagnosis
Feature reduction
Auto-encoder
Multi-objective optimization

ABSTRACT

Background and objective: Feature reduction is an essential stage in computer aided breast cancer diagnosis systems. Multilayer neural networks can be trained to extract relevant features by encoding high-dimensional data into low-dimensional codes. Optimizing traditional auto-encoders works well only if the initial weights are close to a proper solution. They are also trained to only reduce the mean squared reconstruction error (MRE) between the encoder inputs and the decoder outputs, but do not address the classification error. The goal of the current work is to test the hypothesis that extending traditional auto-encoders (which only minimize reconstruction error) to multi-objective optimization for finding Pareto-optimal solutions provides more discriminative features that will improve classification performance when compared to single-objective and other multi-objective approaches (i.e. scalarized and sequential).

Methods: In this paper, we introduce a novel multi-objective optimization of deep auto-encoder networks, in which the auto-encoder optimizes two objectives: MRE and mean classification error (MCE) for Pareto-optimal solutions, rather than just MRE. These two objectives are optimized simultaneously by a non-dominated sorting genetic algorithm.

Results: We tested our method on 949 X-ray mammograms categorized into 12 classes. The results show that the features identified by the proposed algorithm allow a classification accuracy of up to 98.45%, demonstrating favourable accuracy over the results of state-of-the-art methods reported in the literature.

Conclusions: We conclude that adding the classification objective to the traditional auto-encoder objective and optimizing for finding Pareto-optimal solutions, using evolutionary multi-objective optimization, results in producing more discriminative features.

1. Introduction

Although mammography is an effective modality for early breast cancer detection and diagnosis, on mammographic examinations, 10–30% of cancerous/noncancerous lesions may be misinterpreted [1]. To overcome this, computer aided diagnosis (CADx) systems have been developed. The accuracy of CADx for x-ray breast mammography still requires improvements to be useable as a flawless guide (an alert system flagging potential misclassification to the human operator) for radiologists or an independent clinical interpreter [2,3]. Recently, CADx systems have been developed to help radiologists classify suspicious lesions, e.g., labeling the lesion according to the Breast Imaging Reporting And Data System (BI-RADS) assessment categories [4]. In order to build a classification system, a large number of features can be calculated from mammograms, but using high-dimensional features with relatively few training samples can lead to the classifier over-fitting to the training data. This can degrade the predictive model performance as well as having a high computational cost. Since features in mammograms can be noisy and/or highly correlated with each other, feature transformation and reduction is often used to extract relevant features with high discriminatory power from a large number of potential candidate features [5,6].

Related mammography classification works. Mohanty et al. [7] designed an association rule mining based mammogram classification procedure to classify the extracted and hypothetically selected gray level co-occurrence matrix (GLCM) features. This method requires an accurate set of association rules between the features and labels to be defined. The high number of features required for breast cancer diagnosis makes defining these rules a difficult task, which may results in a large number of irrelevant associations. In one of the most recent works, Bria et al. [8] proposed a classification system based on a cascade boosting classifier. The authors defined 145 features to describe the micro-calcifications

[*] Corresponding author.
E-mail addresses: sasagarit@sfu.ca (S.A. Taghanaki), jkawahar@sfu.ca (J. Kawahara), bmiles@sfu.ca (B. Miles), hamarneh@sfu.ca (G. Hamarneh).

http://dx.doi.org/10.1016/j.cmpb.2017.04.012
0169-2607/© 2017 Elsevier B.V. All rights reserved.

but only discriminated between 2 classes, which are insufficient to address the variety of BI-RADS classes. Oliver et al. [9] designed a CAD system using PCA and Bayesian combination of kNN and C4.5 classifiers. This was tested on 184 selected views (all with at least one mass) from a private digital mammographic dataset. They proved that considering density information influences the performance of CAD systems for the detection of breast masses. Without considering breast density information they obtained a 92% accuracy, while by taking density information into account, they achieved an accuracy of 94%. Subshini et al. [10] selected 43 mammograms from the MIAS database and preprocessed them to remove the pectoral muscle and radiopaque artifacts. Next, they extracted statistical features e.g., entropy, uniformity, standard deviation and others from the filtered images for breast characterization. Then, using a SVM classifier they classified the data into three classes of breast density achieving an accuracy of 95.44%. Verma et al. [11] selected 200 ROIs from the DDSM dataset and extracted several features like mass margin, density, patient age, mass shape, subtlety value, and abnormality assessment rank. They proposed to classify the ROIs into two classes of benign and malignant with a soft-clustered direct learning algorithm. An accuracy of 97% was obtained by their proposed method. CAD was applied to standard mammograms from 127 cases in Sadaf et al. [12]. The authors analyzed the CAD sensitivity under 10 classes based on mode of presentation, breast density, lesion size, lesion type, and histopathology. Their overall CAD sensitivity was 91% (115 of 127 cases). Deserno et al. [13] used 2796 patches and defined 12 classes based on BI-RADS assessment categories, BI-RADS tissue density classes, and type of lesion. For feature extraction they applied PCA, 2DPCA, and SVM. Finally, they tested a SVM with three different kernels as a classifier. The best result observed was 80% using 2DPCA feature extraction and a SVM with a Gaussian kernel.

1.1. From shallow to deep learning dimensionality reduction

A significant amount of research has focused on shallow learning approaches such as support vector machines (SVM) [14], principal component analysis (PCA) [15], and linear discriminant analysis (LDA) [16]. Although SVMs are relatively easy to optimize and have good performance for feature transformation/reduction on continuous balanced data, even with advanced kernels, they do not perform well on imbalanced data which results in producing sub-optimal solutions [17]. Similar to PCA, the linearity and the underlying Gaussian assumption of LDA renders the LDA projections incapable of discriminating complex nonlinear data with non-Gaussian distributions. In 2006, the situation was changed by Hinton et al.'s revolutionary research on deep belief networks [18] along with work by Bengio et al. [19] and Poultney et al. [20]. This sparked a significant research effort into deep learning focused on solving the problems of training multiple layers in deep networks and improving initialization. To address this, several optimizations were proposed e.g., unsupervised greedy layer-wise pre-training of each layer [21], stochastic gradient descent methods, limited memory BFGS (L-BFGS) and conjugate gradient [22]. In recent years, deep learning strategies have been significantly improved. For a more detailed review on deep learning the reader is referred to [23–25].

1.1.1. Auto-encoders

Auto-encoders (AEs) encode high-dimensional input data into low-dimensional output codes and then recover the original data from the codes. Bengio et al., motivated the use of restricted Boltzmann machines (RBMs) as pre-training for AEs to build a deep structure [23]. To improve reconstruction fidelity, regularization of AEs was proposed. This can be divided into three models: sparse auto-encoders (SAEs), denoising auto-encoders (DAEs), and contractive auto-encoders (CAEs). SAEs were introduced by Ranzato et al. [26] and inspired by Bengio et al.s stacked AEs [27]. Sparsity of the representation could be obtained either by penalizing the hidden unit biases or by direct penalization of the hidden unit outputs. However, this penalty bias can potentially cause the weights to compensate for the bias, which weakens numerical optimization [24]. Vincent et al. [28] proposed DAE to modify the learning procedure from only reconstructing the raw data to reconstructing the corrupted (noisy) version of the data. These auto-encoders are optimized to, first, encode the noisy input data and second, recover the original input. A stacked denoising AE (SdAE) is constructed by stacking layers of DAEs. They utilize an additional layer to minimize the classification error, however, this is done sequentially not simultaneously [29]. CAEs [30] are an extension of DAEs, as they add a contractive penalty to the reconstruction error function, which penalizes attributes sensitivity to input variations. The fundamental weakness of the CAEs penalty is that it only considers the minuscule variations of input [24]. This was partially improved in [31], but not fully addressed.

Related multi-objective (semi-supervised) autoencoders. In the following paragraphs we focus on detailing the most relevant works. AEs have traditionally been used to perform unsupervised (i.e. without considering the classification task at hand) dimensionality transformation and reduction [24]. This process requires a large amount of unlabeled data samples to produce good feature encodings for reconstruction. However, this process may fail to capture the relevant class information in the data [26]. To reduce the requirement for input data and to find a more meaningful link between the unlabeled data and a classification problem, semi-supervised variants of AE were proposed [29,32,33], i.e. techniques that minimize both reconstruction error and classification error (either in sequential steps or by combining the multi-objective function into a scalarizing function).

Socher et al. [32] introduced a semi-supervised greedy recursive AE, in which a scalar cost function, summing up both reconstruction error and cross-entropy-based classification error, was used. They applied the L-BFGS algorithm for optimization. However, the L-BFGS is highly dependent on the pre-conditioner to avoid degenerating to the steepest descent method [34]. Furthermore, their method required careful tuning of a user-defined parameter that weighs the contributions of the reconstruction and the cross-entropy error terms. Similarly, other researchers [26,35–37] translated the multi-objective problem into a single-objective scalar function. More specifically, their scalarization (weighted-sum) method minimizes a positively weighted convex sum of the objectives (reconstruction error and discriminative error). However, as mentioned by Goldberg and Holland, [38]: “there are times when several criteria are present simultaneously and it is not possible (or wise) to combine these into a single number”. Moreover, it is difficult to define a set of appropriate weights to control the scalarized function to produce *Pareto-optimal solutions* (a solution is considered as a *pareto optimal* if it is not dominated by other solutions) [39]. Although several weight optimizations have been previously introduced e.g. [40], it is a complicated task to find relevance between the weights. Moreover, conducting several weight optimizations is computationally expensive. Additionally, the scalarization method suffers from two technical drawbacks. First, the relation between the *Pareto curve* and the objective function weights is a monotone spread of weight parameters, however, it does not generally produce uniformly distributed points on the *Pareto curve*. Second, minimizing the convex combinations of the objective functions does not necessarily result in reaching the non-convex portions of the *Pareto set* [41]. More recently, Almousli et al. [33] changed the cost function of the DAE in order to produce more accurate results for a supervised task. They included