Unsupervised breast masses classification through optimum-path forest

International Symposium on Computer-Based Medical Systems, 28th, 2015, São Carlos e Ribeirão Preto.
http://www.producao.usp.br.handle/BDPI/49187

Downloaded from: Biblioteca Digital da Produção Intelectual - BDPI, Universidade de São Paulo
Unsupervised Breast Masses Classification Through Optimum-Path Forest

São Paulo State University
Department of Computing
Bauru, São Paulo - Brazil
Email: patriciabellin@yahoo.com.br, leandropassosjr@gmail.com, luis@iontec.com.br, kelton, papa@fc.unesp.br

Abstract—Computer-Aided Diagnosis (CAD) can be divided into two main categories: CADe (Computer-Aided Detection), which is focused on the detection of structures of interest, as well as to assist radiologists to find out signals of interest that might be hidden to human vision; and the CADx (Computer-Aided Diagnosis), which works as a second observer, being responsible to give an opinion on a specific lesion. In CADe - based systems, the identification of mammograms with and without masses is highly needed to reduce the false positive rates regarding the automatic selection of regions of interest. The main contribution of this study is to introduce the unsupervised classifier Optimum-Path Forest to identify breast masses, and to evaluate its performance against with two other unsupervised techniques (Gaussian Mixture Model and k-Means) using texture features from images obtained from a private dataset composed by 120 images with and without the presence of masses.

Keywords—Optimum-Path Forest, Breast masses, Mammography

I. INTRODUCTION

Breast cancer early detection is one of the most important factors that affects the possibility of recovering, thereby increasing the chances of survival [2]. Advances in early diagnosis of breast cancer have been obtained through the development of technologies such as mammography, ultrasound and magnetic resonance imaging, among others [3], [4].

The mammography exam can diagnose malignant or benign tumours early, when we have the best individual’s chance of survival. The opposite is faced by the self-examination, when the detected nodules may present a significant size [3], [5]. Therefore, mammography is identified as a major diagnostic method for breast cancer detection. The sensitivity of this procedure varies according to the patient age, breast density, size, location and tumor aspect [6], [7]. The most common lesions found on mammography exams are the micro-calcifications and breast masses [3].

The authors are grateful to FAPESP grants #2013/20387-7 and #2014/16250-9, and CNPq grants #303182/2011-3, #470571/2013-6 and #306166/2014-3.

Aiming to aid the diagnosis of mammographic findings, technologies based on Computer-Aided Diagnosis (CAD) have been the subject of extensive research [8], [9]. CAD schemes use digital images - obtained through digitized mammography films with high resolution equipment or by means of computed and digital radiography - together with the patient information. Such schemes can work as a second opinion to radiologists, since they can refer to the location and nature of these structures [10], [11]. According to Freitas et al. [12], the use of CAD schemes has resulted in a reduced number of false-positive (FP) cases.

Unsupervised classifiers, such as k-Means clustering, Gaussian Mixture Model (GMM), Principal Component Analysis (PCA), Self-Organizing Maps, Fuzzy c-Means and others can be found in the literature to diagnose breast cancer. Kim et al. [13], Al-Olfe et al. [14] and Choi et al. [15] designed a new type of classification combining unsupervised and supervised models to classify malignant and benign masses on mammograms. Abdel-Qader and Abu-Amara [16] presented a hybrid Computer-Aided Detection and Diagnosis (CADD) system for breast cancer identification based on PCA, Independent Component Analysis and a fuzzy classifier to identify and label suspicious regions. Sarfraz et al. [17] proposed a framework that integrates PCA, Fisher Linear Discriminant, and Nearest Neighbour Classifier for the detection of abnormalities in mammograms as well.

A few years ago, Papa et al. [18], [19], [20] proposed a pattern classifier called Optimum-Path Forest (OPF), which reduces the pattern recognition problem by partitioning a graph into optimum-path trees rooted at prototype samples. In this approach, any element belonging to a given tree is more strongly connected to its root than to any other. This connectivity is established by a path-cost function defined in some way according to an specific class of functions [21]. Currently, two supervised variants have been proposed: OPF with complete graph (supervised learning) [19], [20] and OPF with a k-Nearest Neighbors-based graph [18], as well as one unsupervised OPF for data clustering [22].

The purpose of this work is to introduce the unsupervised...
OPF for the classification of mammography images aiming to identify the presence of breast masses, as well as to compare OPF against with Gaussian Mixture Model [23] and k-Means [24]. As far as we know, unsupervised OPF has never been applied to this context up to date. The remainder of this paper is organized as follows: Sections II and III introduce the OPF classifier and the methodology employed in this paper, respectively. Section IV presents the experimental results, and Section V states conclusions and future works.

II. Optimum-Path Forest Clustering

The design of classifiers based on Optimum-Path Forest has been proposed as a graph-based methodology to exploit connectivity relations between data samples in a given feature space. The methodology interprets a training set as a graph, whose nodes are the samples and the arcs connect pairs of samples that satisfy a given adjacency relation. For a suitable path-value (connectivity) function, the optimum-path forest algorithm [21] partitions the graph into optimum-path trees rooted at some key samples, named prototypes. The prototypes compete among themselves for the most closely connected samples in the training set, such that each sample is assigned to the tree whose prototype offers it an optimum path. Classification of a new sample is done by finding its most closely connected root in an incremental way through the evaluation of the optimum-path values of the training samples.

Let \( Z \) be a dataset such that for every sample \( s \in Z \) there exists a feature vector \( \vec{v}(s) \). Let \( d(s,t) \) be the distance between \( s \) and \( t \) in the feature space. For instance, \( d(s,t) = \| \vec{v}(t) - \vec{v}(s) \| \) — the Euclidean distance between \( \vec{v}(t) \) and \( \vec{v}(s) \). A graph \( (Z,A_k) \) can be defined such that the arcs \( (s,t) \in A \) connect \( k \)-nearest neighbors in the feature space. The arcs are weighted by \( d(s,t) \) and the nodes \( z \in Z \) are weighted by a probability density value \( \rho(s) \):

\[
\rho(s) = \frac{1}{\sqrt{2\pi}\sigma^2|A_k(s)|} \sum_{\forall t \in A_k(s)} \exp\left(-\frac{d^2(s,t)}{2\sigma^2}\right)
\]

where \( |A_k(s)| = k, \sigma = \frac{d_f}{3}, \) and \( d_f \) is the maximum arc weight in \((Z,A_k)\). This parameter choice considers all adjacent nodes for density computation, since a Gaussian function covers most samples within \( d(s,t) \in [0,3\sigma] \). Moreover, since \( A_k \) is asymmetric, symmetric arcs must be added to it on the plateaus of the probability density function (pdf) in order to guarantee a single root per maximum.

The traditional method to estimate a pdf is by Parzen-window. Equation (1) can provide the Parzen-window estimation based on an isotropic Gaussian kernel when we define the arcs by \((s,t) \in A_k \) if \( d(s,t) \leq d_f \). However, this choice presents problems with the differences in scale and sample concentration. Solutions for this problem lead to adaptive choices of \( d_f \) depending on the region of the feature space [25]. By taking into account the \( k \)-nearest neighbors, the method handles different concentrations and reduces the scale problem to the one of finding the best value of \( k \), say \( k^* \) within \([k_{\text{min}},k_{\text{max}}]\), for \( 1 \leq k_{\text{min}} < k_{\text{max}} \leq |Z| \).

The solution proposed by Rocha et al. [22] to find \( k^* \) considers the minimum graph cut among all clustering results for \( k \in [1,k_{\text{max}}] \) \((k_{\text{min}} = 1)\), according to the normalized measure \( GC(A_k,L,d) \) suggested by Shi and Malik [26]:

\[
GC(A_k,L,d) = \sum_{i=1}^{c} \frac{W_i'}{W_i + W_i'},
\]

\[
W_i = \sum_{\forall(s,t) \in A_k(L(s) = L(t) = i)} \frac{1}{d(s,t)},
\]

\[
W_i' = \sum_{\forall(s,t) \in A_k(L(s) = i, L(t) \neq i)} \frac{1}{d(s,t)},
\]

where \( L(t) \) is the label of sample \( t \), \( W_i' \) uses all arc weights between cluster \( i \) and other clusters, and \( W_i \) uses all arc weights within cluster \( i \). For \( i = 1,2,\ldots,c \).

The method defines a path \( \pi_t \) as a sequence of adjacent samples starting from a root \( R(t) \) and ending at a sample \( t \), being \( \pi_t(1) \) a trivial path and \( \pi_t(s,t) \) the concatenation of \( \pi_s \) and arc \((s,t)\). It assigns to each path \( \pi_t \) a value \( f(\pi_t) \) given by a connectivity function \( f \). A path \( \pi_t \) is considered optimum if \( f(\pi_t) \geq f(\tau_t) \) for any other path \( \tau_t \).

Among all possible paths \( \pi_t \) from the maxima of the pdf, the method assigns to \( t \) a path whose minimum density value along it is maximum. That is, the method finds \( V(t) = \min_{\forall \pi_t(1)} \{f(\pi_t)\} \) for \( f(\pi_t) \) defined by:

\[
f(\langle t \rangle) = \begin{cases} 
\rho(t) & \text{if } t \in R, \\
\rho(t) - \delta & \text{otherwise,}
\end{cases}
\]

\[
f(\pi_s \cdot (s,t)) = \min \{f(\pi_s), \rho(t)\},
\]

for \( \delta = \min_{(s,t) \in A_k, \rho(t) \neq \rho(s)} |\rho(t) - \rho(s)| \) and \( R \) being a root set, discovered on-the-fly, with one element per each maximum of the pdf. It should be noted that higher values of \( \delta \) reduce the number of maxima. We are setting \( \delta = 1.0 \) and scaling real numbers \( \rho(t) \in [1,1000] \) in this work. The OPF algorithm maximizes the connectivity map \( \bar{V}(t) \) by computing an optimum-path forest — a predecessor map \( P \) with no cycles that assigns to each sample \( t \notin R \) its predecessor \( P(t) \) in the optimum path from \( R \) or a marker \( \text{nil} \) when \( t \in R \). Algorithm 1 implements this procedure.

Algorithm 1: – OPF Algorithm

**Input:** Graph \((Z,A_k)\) and distance function \(d\).

**Output:** Optimum-path forest \(P\), connectivity map \(V\) and label map \(L\).

**Auxiliary:** Priority-path forest \(P\), connectivity map \(V\) and label map \(L\).

1. For each \( s \in Z \), do
2. \[ \text{Compute } \rho(s) \text{ using Equation (1).} \]
3. \[ \text{Set } P(s) \leftarrow \text{nil}, V(s) \leftarrow \rho(s) - \delta, \text{ and insert } s \text{ in } Q. \]
In Algorithm 1, Lines 1-3 initialize the variables, and also inserts all samples in the priority queue \( Q \). The main loop in Lines 4-12 is responsible to run the OPF clustering algorithm. It first removes a sample \( s \) from \( Q \) with maximum connectivity value \( V(s) \). If \( s \) has not been conquered by any other sample, then \( P(s) = \text{nil} \) (Line 6) and \( s \) is a root of the connectivity map (a maximum of the pdf). Since \( s \in \mathcal{R}_l \), by Equation (5), its connectivity value is reset to \( \rho(s) \) (Line 7), which in addition to the fact that \( \mathcal{A}_k \) is symmetric on plateaus of the pdf will make root \( s \) to conquer the remaining samples of its plateau. It is also assigned to it a new distinct label (cluster) for optimum-path propagation to the rest of its dome. The inner loop in Lines 8-12 evaluates all neighbours \( t \) to which \( s \) can offer a better connectivity value (i.e., \( V(t) < V(s) \)). If the path \( \pi_s \cdot (s, t) \) offers a higher connectivity value to \( t \) (Lines 9-10), then the current path \( \pi_t \) is substituted by the new path \( \pi_s \cdot (s, t) \), being the maps \( V(t), L(t), \) and \( P(t) \) updated accordingly (Lines 11-12).

III. MATERIALS AND METHODS

In this section, we describe the experimental methodology employed in this work, as well as the dataset and feature extraction approach.

A. Dataset

A dataset of 120 images containing regions of interest [28] of several sizes were selected in agreement with medical reports supplied for each mammography image, being 60 of them containing suspect masses and 60 without masses. In this work, we are interested in classifying each image in two classes: with (positive) and without breast mass (negative), being the data of the former class composed by images with benign and malign samples. The original digital mammograms were obtained from films digitized by a Lumiscan (Lumisys, Inc.) scanner, with 12 bits of contrast and spatial resolution of 0.15mm and 0.075 mm per pixel.

B. Feature Extraction

In the context of breast masses identification, the texture plays an important role [29], [30], [9], thus providing good measures that describe an image based on the variation of intensity or subtle changes between the object and the image’s background. Therefore, it is possible to obtain important information such as the softness, roughness and regularity of a certain object.

In this work, we used texture features provided by the well-known approach proposed by Haralick [27], which aims at representing the co-occurrence matrix of grey-levels for each image’s pixel. The main idea is to compute the probability of the combined occurrence between grey levels in different angles, being the distance between pairs of pixels with similar intensity values based on this matrix of Haralick. Therefore, for each image (specifically the region of interest, i.e., masses) we extracted the Haralick texture features, being the best ones selected using a Gaussian distribution [9]. Thus, the selected features were used as the input to the unsupervised methods employed in this work, i.e., OPF, GMM and \( k \)-Means.

C. Statistical Evaluation

In this work, we performed a multiple hold out procedure over 20 runnings, being 50% of the dataset used for training, 20% for validation and the remaining 30% employed for classification purposes (such percentages have been empirically chosen). The dataset was also normalized within the range \([0, 1]\).

The purpose of using a validation set concerns with the fine-tuning of techniques addressed in this work, i.e., \( k_{\text{max}} \) for OPF, and the number of input clusters for \( k \)-means and GMM. Notice all parameters have been optimized within the range \([1, 100]\) with steps of 5. Additionally, the results have been analyzed by means of the Wilcoxon signed-rank test [32].

D. Quality Measure

We adopted the purity as the clustering quality measure, since it is simple and transparent [33], [34]. A perfect clustering has a purity of 1, whereas a bad clustering has a purity of 0. The calculation of the purity is performed as follows:

\[
Purity = \frac{1}{N} \sum_{k=1}^{C} \psi_k, \tag{6}
\]

where \( N \) and \( C \) stand for the number of elements and the clusters of the dataset, respectively, and \( \psi_k \) denotes the number of elements that belong to the label of highest occurrence in cluster \( k \).

IV. EXPERIMENTAL RESULTS

In this section, we present the experimental results concerning the unsupervised breast masses classification by means of OPF, GMM and \( k \)-means. As aforementioned, prior to the application of the clustering techniques to the testing set, we applied a fine-tuning procedure in order to choose the parameters that maximized the purity over the validation set. Figure 1 displays this experiment concerning
OPF classifier, in which $k_{\text{max}} \in [1, 60]$ has been the one that obeyed such restriction.

![Figure 1](image1.png)

Figure 1. Purity of OPF clustering considering different values of $k_{\text{max}}$.

Figure 2 illustrates the purity with respect to GMM and $k$-Means. A similar behavior for both techniques can be observed, being the highest purity obtained with 100 clusters. An interesting point can be drawn here: although OPF has one parameter to be fine-tuned, it requires much less knowledge about the problem when compared to GMM and $k$-means. Notice the purity value is much less affected by different values of $k_{\text{max}}$ than by the number of classes required by GMM and $k$-means, since the curve displayed in Figure 2 oscillates more than the one presented in Figure 1.

![Figure 2](image2.png)

Figure 2. Purity of GMM and $k$-Means considering different values of clusters.

Table I presents the purity results of OPF, GMM and $k$-Means, being the technique in bold the most accurate according to the Wilcoxon signed-rank test. We also presented the parameters that led to the best purity values: $k_{\text{max}} = 10$ considering OPF, and $k = 91$ and $k = 93$ for GMM and $k$-means, respectively. Additionally, the statistical test has pointed out a similarity between GMM and $k$-means. Therefore, OPF has been the most accurate and stable technique, since it has obtained the smaller standard deviation.

![Table 1](table1.png)

Table I

PURITY VALUES OBTAINED BY EACH CLASSIFICATION ALGORITHM CONSIDERING EACH HOLD OUT RUNNING.

Table II displays the recognition rate (hit rate) for each class, being the class #1 representing the presence of masses, and class #2 the opposite (i.e., the absence of mass). Once again, OPF has obtained the best recognition rates for both classes, as well as the overall recognition rate, which is basically the average of the recognition rate of each class. Additionally, the Wilcoxon statistical test has pointed out OPF as the most accurate technique.

Since the dataset is originally labeled, we have also computed the hit rate curves during the fine-tuning process, as displayed in Figures 3 and 4. A similar behaviour can be observed when compared to the purity curves, showing the purity is an interesting metric to evaluate the quality of clustering-based techniques.

![Table 2](table2.png)

Table II

PURITY VALUES OBTAINED BY EACH CLASSIFICATION ALGORITHM CONSIDERING EACH HOLD OUT RUNNING.

V. CONCLUSIONS

In this work, we introduced the unsupervised OPF algorithm in the context of breast masses identification in mammography images, being its effectiveness compared against with the well-known GMM and $k$-means classifiers.

Prior to the classification phase, we conducted a fine-tuning in order to estimate the best set of parameters that minimized a clustering quality measure called purity over a validating set. After that, the fine-tuned techniques were then applied for classification purposes. We have observed OPF was more accurate than both GMM and $k$-means considering
the purity and the recognition rates, since we have computed the accuracy over each class, as well as a global one. Therefore, OPF has shown to be a very suitable tool for unsupervised breast masses identification.

In regard to the future works, we intend to compare OPF against with other techniques out there, such as Self Organizing Map, Principal Components Analysis and Hierarchical Clusters Analysis.

ACKNOWLEDGMENTS

The authors would like to thank FAPESP grants #2013/20387-7 and #2014/16250-9, Capes and CNPq grants #306166/2014-3, #470571/2013-6 and #303182/2011-3.

<table>
<thead>
<tr>
<th>Running Number</th>
<th>% Hit rate OPF</th>
<th>% Hit rate GMM</th>
<th>% Hit rate k-Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>0.983</td>
<td>0.950</td>
</tr>
<tr>
<td>2</td>
<td>0.997</td>
<td>0.939</td>
<td>0.961</td>
</tr>
<tr>
<td>3</td>
<td>1.000</td>
<td>0.939</td>
<td>0.961</td>
</tr>
<tr>
<td>4</td>
<td>0.997</td>
<td>0.989</td>
<td>0.972</td>
</tr>
<tr>
<td>5</td>
<td>1.000</td>
<td>0.967</td>
<td>0.970</td>
</tr>
<tr>
<td>6</td>
<td>1.000</td>
<td>0.978</td>
<td>0.981</td>
</tr>
<tr>
<td>7</td>
<td>0.997</td>
<td>0.972</td>
<td>0.953</td>
</tr>
<tr>
<td>8</td>
<td>1.000</td>
<td>0.989</td>
<td>0.948</td>
</tr>
<tr>
<td>9</td>
<td>1.000</td>
<td>0.983</td>
<td>0.983</td>
</tr>
<tr>
<td>10</td>
<td>0.995</td>
<td>0.986</td>
<td>0.959</td>
</tr>
<tr>
<td>11</td>
<td>0.997</td>
<td>0.956</td>
<td>0.975</td>
</tr>
<tr>
<td>12</td>
<td>1.000</td>
<td>0.939</td>
<td>0.961</td>
</tr>
<tr>
<td>13</td>
<td>1.000</td>
<td>0.975</td>
<td>0.959</td>
</tr>
<tr>
<td>14</td>
<td>1.000</td>
<td>0.975</td>
<td>0.961</td>
</tr>
<tr>
<td>15</td>
<td>0.997</td>
<td>0.983</td>
<td>0.967</td>
</tr>
<tr>
<td>16</td>
<td>1.000</td>
<td>0.981</td>
<td>0.956</td>
</tr>
<tr>
<td>17</td>
<td>1.000</td>
<td>0.967</td>
<td>0.959</td>
</tr>
<tr>
<td>18</td>
<td>1.000</td>
<td>0.970</td>
<td>0.945</td>
</tr>
<tr>
<td>19</td>
<td>1.000</td>
<td>0.986</td>
<td>0.939</td>
</tr>
<tr>
<td>20</td>
<td>1.000</td>
<td>0.961</td>
<td>0.936</td>
</tr>
</tbody>
</table>

Average 0.999 0.971 0.960
Average class 1 0.999 0.963 0.953
Average class 2 0.999 0.979 0.967
Standard deviation 0.002 0.016 0.012

Table II GLOBAL AND CLASS - ORIENTED AVERAGE HIT RATE.

REFERENCES


