Toward Unsupervised Classification of Calcified Arterial Lesions

G. Brunner, U. Kurkure, D.R. Chittajallu,
R.P. Yalamanchili, and I.A. Kakadiaris

Computational Biomedicine Lab, Depts. of Computer Science, Elec. & Comp. Engineering, and Biomedical Engineering, Univ. of Houston, Houston, TX, USA
{gbrunner,ukurkure,drchittajallu,rpyalamanchili2,ioannisk}@uh.edu

Abstract. There is growing evidence that calcified arterial deposits play a crucial role in the pathogenesis of cardiovascular disease. This paper investigates the challenging problem of unsupervised calcified lesion classification. We propose an algorithm, *US-CALC* (UnSupervised Calcified Arterial Lesion Classification), that discriminates arterial lesions from non-arterial lesions. The proposed method first mines the characteristics of calcified lesions using a novel optimization criterion and then identifies a subset of lesion features which is optimal for classification. Second, a two stage clustering is deployed to discriminate between arterial and non-arterial lesions. A histogram intersection distance measure is incorporated to determine cluster proximity. The clustering hierarchies are carefully validated and the final clusters are determined by a new intra-cluster compactness measure. Experimental results indicate an average accuracy of approximately 80% on a database of electron beam CT heart scans.

1 Introduction

The National Heart Lung and Blood Institutes' 2007 disease statistics reported that in 2004, 872,000 deaths or 36% of all deaths in the United States were due to cardiovascular disease. The underlying causes of most heart attacks, and sudden cardiac deaths are inflamed, active, and growing atherosclerotic plaques believed to be vulnerable plaques. Several studies have established that the presence of calcified coronary plaque has a significant predictive value for coronary artery disease in both asymptomatic and symptomatic patients. One of the foci of modern cardiology is to identify high-risk individuals with no prior symptoms, where calcified arterial lesions could be used as biomarkers to assess a patients' risk for a cardiovascular event.

Arterial calcium can be measured by computed tomography (CT). To date, the default method is to assess the total calcium burden by various scores such as Agatston, volume or mass. These scores represent the total amount of calcified plaque in the coronary arteries [1].

The clinical standard for the initial selection of calcified arterial and non-arterial lesions is typically based on a simple threshold of contiguous pixels of at

D. Metaxas et al. (Eds.): MICCAI 2008, Part I, LNCS 5241, pp. 144–152, 2008.

least 1 mm^2 and a CT density of at least 130 Hounsfield units (HU). However, this thresholding has the effect of selecting calcified tissues, bone, metal stents or any other metallic objects that will be included in the data. Then, the calcified areas are manually classified as arterial. From a medical point of view most of these regions are not *real* lesions. However, for ease of terminology, we refer to all selected candidate regions as lesions throughout this paper[1]. The goal of our method is to automatically separate arterial vs. non-arterial lesions.

The gold standard for clinical calcium scoring is to manually identify coronary calcifications from a thresholded scan. Although there is extensive medical literature about manual calcium scoring, surprisingly little work has been conducted from a computer science point of view. To the best of our knowledge there is no published research on unsupervised arterial lesion classification in CT heart scans. However, clustering and other unsupervised classification techniques have been widely applied for related biomedical problems. A hierarchical clustering method was proposed by Carmo *et al.* [2] to compute topological flow patterns for haemodynamics of the cardiovascular system. Also a hierarchical unsupervised clustering technique was introduced by Tiwari *et al.* [3] for the identification of cancerous areas in the prostate in Magnetic Resonance Spectroscopy data. A novel clustering method was proposed by O'Donnell and Westin [4] and was deployed to a high dimensional feature space for the purpose of finding white matter fiber correspondences across a population of five brains. However, there are certain design choices for clustering algorithms that are often not sufficiently discussed in the literature such as the initial feature space representation, clustering tendency, and validation of the clustering.

In this paper, we present an unsupervised calcified arterial lesion classification method that mines calcified lesion characteristics to cluster candidate lesions as arterial or as non-arterial. By the term non-arterial, we refer to both calcified lesions which are not located in the coronaries, and artifacts.

Our contributions are the following. First, we present a novel optimization criterion for the determination of an *optimal* lesion feature subset based on: 1) feature correlations, 2) statistical properties, and 3) clustering tendency using Hopkins statistical hypotheses testing. Second, we present a novel two stage integrated clustering schema which identifies the optimal partitioning of the lesion feature space with respect to the cophenetic correlation coefficient and a histogram intersection metric. Third, the final arterial lesion clusters are identified based on a novel intra-cluster compactness measure.

2 Methods

One of the challenges of unsupervised classification is to discriminate between object classes without actually knowing the class labels [5]. *US-CALC* proposes a holistic approach to unsupervised classification of calcified arterial lesions and explores the high dimensional vector space containing a multitude of features. The method automatically derives an *optimal* subset of features with respect

[1] We only distinguish between arterial and non-arterial lesions.

to an optimization criterion which is based on clustering tendency, inter-feature correlations and individual feature statistics. *US-CALC* agglomerates a nested hierarchy of calcified arterial lesions and non-lesions using two sets of orthogonal features. The detailed steps of the method are outlined below.

Algorithm. *US-CALC*

1. **Feature Extraction**
   - Compute shape, texture and statistical lesion properties
2. **Cluster Tendency and Dimensionality Reduction**
   - Perform Hopkins statistical testing
   - Compute feature correlation and statistical properties
   - Determine *optimal* feature subset
3. **First Clustering Iteration**
   - Cluster using n-dim. feature vector per lesion
4. **Second Clustering Iteration**
   - Cluster using mean and std of distances between members for every cluster
5. **Identify arterial and non-arterial lesion clusters**

Feature Extraction: Each lesion region represents a blob like structure and can be described by a set of features. Based on the appearance of lesions we decided to use a set of features that captures the shape and, the morphology of the lesions. For instance, the density across a lesion can be captured by the respective HUs. The complete set of features for each lesion consists of: *area*, *compactness*, *first and second eigenvector*, *eccentricity*, *moment features* [6], *3D coordinates of the pixel with the peak intensity in the region*, *mean radial lesion length*, *standard deviation of radial lesion length*, *min. and max. HU*, *moments 1-4 of HU values for all lesion pixels*, *HU range*, *entropy of HUs* and *15 texture features* [7]. The extracted features describe shape, texture and statistical properties of each lesion.

Clustering Tendency and Dimensionality Reduction: So far, we have created a high dimensional feature space which serves as starting point for the clustering step of *US-CALC*. The quality of the clustering is largely determined by the properties of the feature space. That raises the question: *Is there a natural structure in our feature space at all?* In order to answer that question we investigate the feature space for *clustering tendency* by testing points for randomness. In detail, we carry out an analysis that is based on checking for the *Null Hypothesis* H_0, which is to be rejected for non-random points. As a test of merit *US-CALC* incorporates the Hopkins test [8], as it has been shown to be able to robustly determine the clustering tendency in an unknown dataset [9]. The Hopkins test is based on the nearest neighbor distance, that are the distances between randomly sampled points and points from the actual distribution. In detail, let $\mathbf{X} = \{\mathbf{x}_i,\ i = 1, ..., N\}$, be the dataset of lesions, where N is the total number of lesions. Further, let $\mathbf{X}_s$ be a subset of $\mathbf{X}$ with M randomly selected vectors from $\mathbf{X}$, defined as $\mathbf{X}_s = \{\mathbf{x}_s,\ s = 1, ..., M\}, M \approx \frac{N}{10}$. Also, let $\mathbf{X}_r = \{\mathbf{x}_r,\ r = 1, ..., M\}$, be a set of vectors randomly distributed according to the uniform distribution. Now, we compute d_{rs} as the distance from $\mathbf{x}_r \in \mathbf{X}_r$

to its nearest vector in $\mathbf{x}_s \in \mathbf{X}_s$. Moreover, let d_{sn} be the distance from $\mathbf{x}_s$, to its nearest neighbors. We compute the actual Hopkins statistics with the l^{th} powers of d_{rs} and d_{sn} as: $h = \frac{\sum_{r=1}^{M}(d_{rs})^l}{\sum_{r=1}^{M}(d_{rs})^l + \sum_{s=1}^{M}(d_{sn})^l}$, with $0 \leq h \leq 1$. The closer h is to one the stronger is the clustering tendency. The actual values of h are reported in Sec. 3, and indicate that the lesion feature space exhibits a strong clustering tendency. In addition to clustering tendency, feature redundancy is an important issue for any classification algorithm in general and especially for unsupervised classification methods. To that end, *US-CALC* performs a feature redundancy test resulting in an actual dimensionality reduction of the given lesion feature space. Ideally, we wish for non-overlapping and non-correlated features, as clustering techniques are more likely to generate *correct* results for that kind of vector space [5]. Hence, we introduce a criterion that identifies an *optimal* feature subset with respect to clustering tendency, feature correlation and statistical properties of each feature. Specifically, we compute the clustering tendency separately for each feature using the Hopkins test resulting in $\mathbf{h}_i$, $i = \{1, 2, ..., K\}$, with K as the number of features. Similarly, we compute the skewness and kurtosis for each feature resulting in SK_i and KU_i, for the i^{th} feature, respectively. The choice of skewness and kurtosis is motivated by the assumption that the number of arterial calcified lesions is much smaller than the number of non-arterial lesions - which usually holds. A large skewness and/or kurtosis is an indicator of an asymmetric distribution opposed to Gaussian distributions (zero skewness). Real world datasets often tend to be non-Gaussian. Next, we compute the correlation coefficient between all features CC_{ij}, with i and j each in the range of $\{1, 2, ..., K\}$. Highly correlated features carry redundant information that can be disregarded, since those features do not provide additional discriminative power. The actual optimization criterion is to identify features that exhibit a strong clustering tendency (i.e., h close to one), a low correlation with each other, and possess a large skewness and kurtosis at the same time. Note that many feature selection algorithms are specifically developed for supervised classification tasks. In our case, we work with unlabeled data and the proposed criterion is optimal with respect to the given features and with respect to the assumptions of our proposed clustering algorithm. A detailed analysis of our findings will be addressed in a forthcoming paper. The output of this step is a feature subset $\mathbf{F}$, that reveals a strong clustering tendency while it contains as little as possible redundant information. In the next step, *US-CALC* deploys a clustering algorithm to the optimal feature subset $\mathbf{F}$.

The clustering of the lesions is performed in two steps. *US-CALC* first forms a hierarchy (see Algorithm *US-CALC* Step 3) of the lesion feature space $\mathbf{F}$, and incorporates in the second step an orthogonal set of features that is computed on the fly in step one (see Algorithm *US-CALC* Step 4). These newly computed features can be interpreted as compactness measures of the clusters that were obtained in the first clustering process. Both clustering steps undergo a strict validation check.

First Clustering Iteration: Initially, *US-CALC* partitions the lesion feature space $\mathbf{F}$ into $\{C_n;\ n = 1, ..., N\}$ clusters, with N denoting the number of lesions,

where each cluster contains a vector that describes a single lesion. In the next step, *US-CALC* iteratively merges the initial partitions based on their feature space proximity into larger and larger clusters resulting in a nested hierarchy. This formation of a cluster hierarchy can be seen as the iterative application of a dissimilarity function to a set of possible pairs of clusters of the data matrix $\mathbf{X}$. In detail, we introduce the symmetric square $N \times N$ dissimilarity matrix $\mathcal{D}_b$, with elements d_{ij}^b. Each element represents a measure of distinction between the i^{th} and j^{th} lesion feature vector[2]. The actual dissimilarity measure $\mathcal{H}$ can be interpreted as *closeness* or similarity of two clusters or two groups of lesion clusters. Formally, it can be written as a function: $\mathcal{H} : \mathbf{X} \times \mathbf{X} \rightarrow r \in R^+$, with r as the set of real numbers. Hence, we introduce the histogram intersection measure:

$$\mathcal{H}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i<j} min\{x_i, x_j\}. \tag{1}$$

It determines which two clusters of lesion features are most similar to each other. At this point we want to stress that the lesion feature vector is *n-dim.* Thus, we cluster high dimensional features together rather than the physical coordinates (i.e., spatial location of lesions). Now we have a way to find any two closest clusters of lesion features. In the subsequent step, we identify the next closest cluster and next closest to that. Eventually *US-CALC* obtains a hierarchy of nested clusters, where each contains a different number of lesions. In detail, *US-CALC* performs $N-1$ subsequent clustering steps resulting in a hierarchy of lesion clusters. So far we have obtained a specific partitioning of our feature space. Next, we need to ensure that we actually compute the best possible hierarchy. To that end, *US-CALC* incorporates a validation step. Let us consider the cluster hierarchy $\mathcal{T}_{ij}$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ are for the first time merged in the same cluster. Further, $\mathcal{L}_{ij}$ is the proximity level where the clustering $\mathcal{T}_{ij}$ has been formed. Then, we can compute the distances between the proximity levels resulting in the cophenetic matrix $\mathcal{D}_a$[3]. Now, we can compute the cophenetic correlation coefficient CCC that measures the degree of similarity between the cophenetic matrix $\mathcal{D}_a$ and the dissimilarity matrix $\mathcal{D}_b$ obtained from the lesion features $\mathbf{X}$.

$$CCC = \frac{\frac{1}{O}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^a d_{ij}^b - \mu_a \mu_b}{\sqrt{\left(\frac{1}{O}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(d_{ij}^a)^2 - (\mu_a)^2\right)\left(\frac{1}{O}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(d_{ij}^b)^2 - (\mu_b)^2\right)}}, \tag{2}$$

where O is a normalization factor, and, μ_a and μ_b are the respective means, $\left(\mu_a = \frac{1}{O}\sum_{i=1}^{N-1} d_{ij}^a,\ \mu_b = \frac{1}{O}\sum_{i=1}^{N-1} d_{ij}^b\right)$. The value of CCC is in the range of $[-1, 1]$, where values closer to 1 indicate a better agreement between the cophenetic and the proximity matrix. Hence, the CCC is a measure of how

[2] $\mathcal{D}_b$ fulfills all requirements of a metric.

[3] Note that $\mathcal{D}_a$ is a metric and even the ultrametric inequality holds [10].

accurately the hierarchical tree represents the dissimilarities of the original input data. The actual values of CCC are reported in Sec. 3.

Second Clustering Iteration: The second clustering step is necessary as we want to identify clusters that only contain arterial lesions. The first clustering results in an *a prior* unspecified number of clusters. However, we know that there are typically much fewer arterial lesions than non-arterial ones and we expect two categories of clusters: 1) arterial lesion clusters that are very compact with fewer members and, 2) non-arterial lesion clusters that are more scattered and typically contain many members. At this point the advantage of a two stage clustering becomes apparent. The arterial lesion clusters are more compact within than the non-arterial ones, but are not necessarily *close* to each other in terms of a vector space distance measure. Therefore, a simple one stage clustering that searches for only two clusters in the feature space would fail[4]. In the second clustering step *US-CALC* identifies dense and compact clusters of arterial lesions, given the clustering from step one. Specifically, we compute distance matrices for all members of each cluster. Note that compact clusters tend to have shorter distances between it's members on the average. The clustering procedure in step two is identical to the first one with just one exception. Instead of using the n-dim. feature vector describing each lesion, *US-CALC* only uses the mean and the standard deviation for each distance matrix. Hence the input feature space for the second clustering step is of dimension $2\times P$, where P is the number of clusters obtained in the first clustering step. The second stage clustering results in a partitioning, too. Note, that the final clustering is also validated with the cophenetic correlation coefficient (Eq.2). In the final step *US-CALC* identifies which clusters contain arterial lesions. The most compact clusters - that are believed to be arterial lesion clusters - consists of *members* that are more *similar* to each other than to any other subgroup; such clusters show a *high mutual similarity*. As measure of compactness (intra-cluster distance) *US-CALC* incorporates:

$$\sigma_j = \left(\frac{1}{n_c} \sum_{\mathbf{x} \in C_j} ||\mathbf{x} - \mathbf{w}_j||^2 \right)^{\frac{1}{2}}, \tag{3}$$

where n_c refers to the number of clusters, and C_j denotes an individual lesion cluster with lesion members $\mathbf{x}$, and, $\mathbf{w}_j$ is the cluster representative for C_j. The final assignment of whether a lesion is arterial or not is based on the final clustering. Members of compact clusters according to Eq. 3 are labeled as arterial lesions and members of non-compact cluster are considered to be non-arterial lesions.

3 Results

Data: The heart scans have been obtained by EBCT imaging with a slice thickness of 3 mm and an x-y pixel spacing of 0.508 - 0.586 mm. We have created

[4] We have performed preliminary tests with k-means clustering and one stage clustering.

Table 1. Average parameters computed by *US-CALC* as described in Sec. 2

Dataset	*CC*	*KU*	*SK*	*h*	*CCC* Step 3	*CCC* Step 4
1	0.02	13.42	0.25	0.94	0.86	0.89
2	0.06	2.77	0.41	0.94	0.88	0.89
3	0.02	5.34	0.07	0.93	0.89	0.90
4	0.06	42.47	2.22	0.95	0.90	0.89

Table 2. Performance evaluation of *US-CALC* for the four datasets described in Sec. 3

Subsets	Performance measures per dataset [%]							
	DS-1		DS-2		DS-3		DS-4	
	Accuracy	f	Accuracy	f	Accuracy	f	Accuracy	f
1	80.00	80.14	83.03	84.04	80.11	80.21	80.14	80.13
2	80.00	80.00	82.35	82.05	79.64	79.67	79.73	79.69
3	77.13	77.98	84.17	84.54	75.62	75.20	81.09	81.65
4	76.61	76.97	83.07	82.27	70.55	70.42	80.13	80.06
Average	78.43	78.77	83.16	83.23	76.48	76.38	80.27	80.38

four mutually exclusive sets of arterial and non-arterial lesions. Each dataset contains between 92,296 to 102,600 lesions. In total there were more than 200 patient scans with approximately 20-35 CT image slices per scan. The number of arterial lesions per dataset was more than 1,300 (i.e., the number of non-arterial lesions is much larger). For each of the four datasets we have extracted four random subsets of non-arterial lesions that equal the number of arterial lesions.

Experiments: Upon extraction of the complete feature set, we computed the clustering tendency, the feature correlations, and the statistical properties. The optimal feature subset was determined with the values listed in Table 1. Typical selected features are the skewness and kurtosis of the HUs and the texture features described in Sec. 2. Interestingly, none of the shape features has been selected. This might be due to the similarity of shapes of the arterial and non-arterial lesions. The cophenetic correlation coefficient for **F** was approximately 0.9 for all four datasets. The value actually verifies that the clustering partitions obtained are matching *well* with the original input data. In the first clustering step the typical number of clusters was between 150 and 200.

For the actual validation of our results, manual annotations of arterial lesions were performed. Then, we compared the manually assigned labels with the actual *US-CALC* results for each of the four subsets resulting in 16 datasets. Table 2 summarizes the results for four random subsets for each of the four datasets. We report two widely used performance measures, accuracy and f-measure [11]. For the best subset *US-CALC* exhibited an accuracy of 84%.

In a second experiment, we evaluated *US-CALC* on a per patient basis. To that end, we randomly selected 10 patient scans from the four datasets. This time we included all lesions within a heart bounding box that was semi-automatically

computed and visually verified. Note, that the number of arterial lesions is the same as in the first experiment. However, the number of non-arterial lesions is reduced by approximately a factor of 4. The typical total number of lesions per patient is about 300-600 per patient. Averaged over ten randomly selected patient scans *US-CALC* exhibited an accuracy of 83.15% and an f-measure of 81.61%.

4 Discussion and Concluding Remarks

We have presented a novel method for the unsupervised classification of calcified arterial lesions. The method incorporates a novel optimization criterion that reduces the feature space dimensionality and identifies an *optimal* feature subset. *US-CALC* also performs a clustering tendency step before a two stage clustering method is deployed. The second clustering step discriminates between the arterial and non-arterial lesions using a novel intra-cluster compactness measure. The results obtained are promising and provide insight into which features are *more* suitable for the characterization of the arterial lesions. The results show that the texture and statistical properties of lesion intensities were always selected over shape features. Encouraged by the performance of *US-CALC* we plan to further investigate the unsupervised feature selection method.

Acknowledgments. This work was supported in part by the Biomedical Discovery Training Program of the W.M. Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia (NIH Grant No. 1 T90 DA022885 and 1 R90 DA023418), and in part by NSF Grant IIS-0431144.

References

1. Rumberger, J., Kaufman, L.: A rosetta stone for coronary calcium risk stratification: agatston, volume, and mass scores in 11,490 individuals. AJR Am. J. Roentgenol. 181(3), 743–748 (2003)
2. Carmo, B., Ng, Y., Prügel-Bennett, A., Yang, G.Z.: A data clustering and streamline reduction method for 3D MR flow vector field simplification. In: Barillot, C., Haynor, D.R., Hellier, P. (eds.) MICCAI 2004. LNCS, vol. 3216, pp. 451–458. Springer, Heidelberg (2004)
3. Tiwari, P., Madabhushi, A., Rosen, M.: A hierarchical unsupervised spectral clustering scheme for detection of prostate cancer from magnetic resonance spectroscopy MRS. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part II. LNCS, vol. 4792, pp. 278–286. Springer, Heidelberg (2007)
4. O'Donnell, L., Westin, C.F.: White matter tract clustering and correspondence in populations. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 140–147. Springer, Heidelberg (2005)
5. Jain, A., Duin, R., Mao, J.: Statistical pattern recognition: A review. IEEE Trans. Pattern Anal. Mach. Intell. 22(1), 4–37 (2000)
6. Shen, L., Rangayyan, R., Desautels, J.: Application of shape analysis to mammographic calcifications. IEEE Trans. on Medical Imaging 13(2), 263–274 (1994)

7. Laws, K.: Rapid texture identification. In: Proc. SPIE Conference on Missile Guidance, vol. 238, pp. 376–380 (1980)
8. Hopkins, B.: A new method for determining the type of distribution of plan-individuals. Annals of Botany 18, 213–226 (1954)
9. Peres, M., de Andrade-Netto, L.: A fractal fuzzy approach to clustering tendency analysis. In: Bazzan, A.L.C., Labidi, S. (eds.) SBIA 2004. LNCS (LNAI), vol. 3171, pp. 395–404. Springer, Heidelberg (2004)
10. Hartigan, J.: Representation of similarity matrices by trees. Journal of the American Statistical Association 62, 1140–1158 (1967)
11. Lewis, D., Gale, W.: A sequential algorithm for training text classifiers. In: SIGIR 1994: Proc. 17th ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 3–12. Springer, New York (1994)