Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 126 (2018) 576–585

Procedia Computer Science

www.elsevier.com/locate/procedia

22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Partially Lazy Classification of Cardiovascular Risk via Multi-way Graph Cut Optimization

Karma M.Fathalla[a,b], Anikó Ekárt[b], Doina Gherghel[c]

[a]Computer Engineering, Arab Academy for Science and Tehnology. Alexandria, Egypt
[b]Aston Lab for Intelligent Collective Engineering (ALICE), Aston University. Birmingham, UK
[c]Vascular Research Laboratory, Aston University. Birmingham, UK.

Abstract

Cardiovascular disease (CVD) is considered a leading cause of human mortality with rising trends worldwide. Therefore, early identification of seemingly healthy subjects at risk is a priority. For this purpose, we propose a novel classification algorithm that provides a sound individual risk prediction, based on a non-invasive assessment of retinal vascular function. So-called lazy classification methods offer reduced time complexity by saving model construction time and better adapting to newly available instances, when compared to well-known eager methods. Lazy methods are widely used due to their simplicity and competitive performance. However, traditional lazy approaches are more vulnerable to noise and outliers, due to their full reliance on the instances' local neighbourhood for classification. In this work, a learning method based on Graph Cut Optimization called *GCO_mine* is proposed, which considers both the local arrangements and the global structure of the data, resulting in improved performance relative to traditional lazy methods. We compare *GCO_mine* coupled with genetic algorithms (*hGCO_mine*) with established lazy and eager algorithms to predict cardiovascular risk based on Retinal Vessel Analysis (RVA) data. The highest accuracy of 99.52% is achieved by *hGCO_mine*. The performance of *GCO_mine* is additionally demonstrated on 12 benchmark medical datasets from the *UCI repository*. In 8 out of 12 datasets, *GCO_mine* outperforms its counterparts. *GCO_mine* is recommended for studies where new instances are expected to be acquired over time, as it saves model creation time and allows for better generalization compared to state of the art methods.

Keywords: Cardiovascular disease, Retinal Vessel Analysis, lazy classification, graph cut optimization, genetic algorithm.

1. Introduction

Cardiovascular disease (CVD) is usually manifested as a sudden life-threatening event [34], therefore early detection of asymptomatic subjects at risk is vital [17, 40]. Several risk scores have been established for identification of CardioVascular Risk (CVR), including Framingham Risk Score (FRisk) [12] and QRisk [23], which are appropriate

E-mail address: fathakmg@aston.ac.uk*

1877-0509
10.1016/j.procs.2018.07.292

for population based risk stratification [3, 9, 11, 24]. Identifying novel early risk markers is a developing area expected to lead to more reliable individual risk estimation and consequently better disease control and higher survival rate.

Retinal Vessel Analysis (RVA) [37] is a noninvasive method for identifying vessels' reactivity to stress. The association of variability in retinal vessel diameters' changes with CVD has been recently studied and established [26, 30]. We pioneered the use of RVA coupled with machine learning methods for early CVR prediction [18]. We applied established standard classification methods on a subset of features, including a combination of measures generated from RVA data and others known to contribute to the calculation of FRisk and QRisk scores. The highest achieved accuracy was 96.22% using Random Forest. Given the specifics of the problem, i.e., imbalanced data, with very limited number of patients in high risk category, expectation to recruit new patients, overlapping measurements ranges and high cost of misclassification, further improvement of accuracy is essential.

The main objective of this study is to propose a classification method that can correctly identify high cardiovascular risk subjects based solely on RVA data, while minimizing classification error overall. The novel classification algorithm *GCO_mine* produces reliable risk prediction and can handle the specifics of the data. RVA data have two main characteristics, namely (1) the boundaries between risk groups are not crisply defined and (2) the various interactions between the features are not clear. For our study, the data are continuously being collected, therefore a method that can adapt to newly arriving data is needed. Thus, instead of a global abstraction classification model using eager methods, an instance-based approach is proposed here. Our *GCO_mine* method is particularly suitable, because it respects the individual variation within one risk group and manages the overlapping pre-morbid and normal ranges of features.

Existing lazy learning solutions are outlined in the next section, together with examples of the use of graph cut optimization (*GCO*) for classification. In section 3, the proposed classification algorithm is described in detail, followed by the experimental evaluation in section 4. Conclusions are drawn in section 5.

2. Related Work

Here we focus on lazy learning methods specifically, as they can handle expanding datasets and overlapping ranges, which are essential for our dataset. The application of graph cut optimization for clustering and classification is discussed to illustrate its potential for tackling the limitation of pure locality in the existing lazy methods.

2.1. Lazy Learning

In lazy classification, local neighbourhood arrangements are constructed using similar training instances at testing time. This is done to induce a classification decision for unseen instances. A major drawback with lazy approaches is the construction of entirely local patterns, which disregard the overall structure of the data, leading to worse noise tolerance compared to eager methods.

The k-Nearest Neighbours algorithm (kNN) is an example of a well established lazy classifier. The k nearest neighbours of a test instance are first determined [2]. The nearest neighbours are located within the training set using Euclidean distance. Majority vote from the k-neighbours is used as the classification decision of the test instance. Extensions of kNN include introducing decision rules other than majority vote, using more efficient search strategies for neighbour search, determining the best value of k and investigating the distance function effect on performance.

Shang et al. combine fuzzy set theory with classical kNN [39]. The effect of the neighbouring samples is weighted by their distance to the test instance. A fuzzy membership to every class is computed based on neighbours' membership weighted by their distance. The test sample receives the class label of the highest membership. Kaveh-Yazdy et al. propose another attempt for decision rule improvement [29]. The distance of the neighbours is weighted by the size and dispersion of their class, where neighbours belonging to larger and more dispersed classes are allocated larger weight. This is applied to determine the impact of the neighbours more accurately. They also address the issue of search space reduction using linear discriminant analysis. Wang and Li introduce an approach for efficient space search using particle swarm intelligence to determine k-nearest neighbours and eliminate outliers quickly [41].

For determining the best value of k, the most common strategy is cross validation (brute force) [42]. However, while Wang et al. determine k locally using statistical confidence [42], Hassanat et al. employ ensemble classification to reduce the influence of a single k selection: weak kNN classifiers of different k values are applied first and then a weighted sum rule is used to combine the classifications of the weak classifiers [21].

The effect of the distance function is studied by Hu et al., showing that the performance is dependent on the feature data types of the dataset [25]. Moreover, Bao et al [5] combine several distance functions, such as heterogeneous Euclidean-Overlap metric and discretized value difference metric to determine different k-nearest neighbours groups first, then apply majority vote. The K^* algorithm [10] uses the entropic transformation function to determine the samples similarity. K^* has been shown to handle categorical data better than kNN due to its similarity function.

Another line of study is merging the nearest neighbours decision with the Naive Bayes (NB) classifier, where a NB model is either constructed locally based on test samples k-nearest neighbours [19] or multiple NB models are locally constructed with different k and then the most accurate model is selected to classify a test instance [43].

Despite the dedicated efforts to enhance the existing lazy learning methods, purely local approaches, which are vulnerable to noise and outliers, remain the most prevalent and the global resultant structure is overlooked.

2.2. Graph Cut Optimization in Classification Problems

Graph cut optimization (*GCO*) has been used as a solution to clustering problems [28, 16, 15]. Dhillon et al. [15] employed kernel k-means to optimize weighted graph cuts and overcome the equal-sized cluster restriction of Karypis and Kumar [28]. Despite the success in clustering problems, its use for classification problems in the data mining domain remains limited. Extensive application of *GCO* can be found in image segmentation and object classification: *GCO* has been widely used with hyperspectral images [4, 13], retinal images, where *GCO* produces bi-labels (artery/vein) from segmented images [14, 27, 36] and flower segmentation from colour images [44]. Even though *GCO* has been included within a broad set of learning models, not all the presented models use *GCO* as a stand alone classifier. The described approaches either utilize pre-classifiers (e.g. SVM and K-means) [4, 13] or apply *GCO* on presegmented images [14, 27, 36] or apply significant image-specific preprocessing before *GCO* [44].

We consider that further development of *GCO*-based generic classification methods is a promising avenue as (1) *GCO* is expected to lead to high accuracy classification for a range of problems and (2) it is a low cost method with guaranteed optimality bounds. Also, utilizing its formulation to modify the typical instance-based learning approach would achieve better generalization of the learning decision.

We propose a partially lazy mining (classification) algorithm based on Graph Cut Optimization *GCO_mine* that aggregates local connectivities into a globally connected graph, on which a global classification decision is taken. In addition, *GCO_mine* introduces the concept of a sample's direct membership (distance) to the given classes, which does not exist in traditional lazy approaches such as kNN. The *GCO_mine* approach strikes a favourable balance between merely local instance-based lazy methods and eager techniques, which build global latent models of the training data in a separate phase.

3. Proposed Partially Lazy Classification Method

In this section, we introduce *GCO_mine*, a partially lazy classification method that employs multi-label graph cut optimization (*GCO*). *GCO_mine* formulates the classification model as a graph cut minimization problem. *GCO_mine* reaches a solution by incorporating a smoothness prior, which employs similarity information from both training and test instances. The graph formulation introduces connectivity among the local neighbourhoods, thus allowing for the study of the global structure of the classes. We consider *GCO_mine* a partially lazy classifier as it includes caching intermediate results for parameter settings [1].

3.1. Graph Cut Formulation of GCO_mine

Using graph cut optimization [8, 31], we formulate our classification problem as an undirected graph. In an undirected graph, there exists a set of vertices v and edges e that connect these vertices. Each edge e_i is assigned a non-negative cost c_i denoting the penalty of cutting the edge e_i between two vertices. There is a special type of vertices called the terminals. Each terminal of the graph represents a label creating l-vertices for l class problems. The other vertices correspond to the data points (records). Given a dataset, its data points are represented by d-vertices. In our formulation, there exist d_l-vertices which belong to the set labeled data points D_l and d_u-vertices for unlabeled data points D_u. Each d-vertex is connected to all the terminals (l-vertices) through t-links of different costs. Also, the

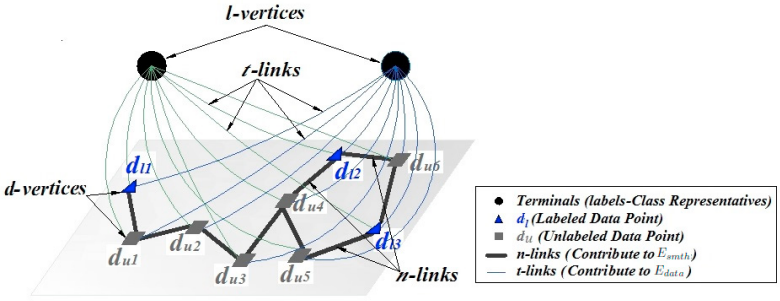


Fig. 1: Graph Formulation of Classification Problem

neighbouring data points have weighted links called n-links. The described graph cut formulation structure is outlined in Fig. 1. The energy of a cut per iteration depends on the costs of the severed t-links and n-links in a single α-expansion move (see [8] for details). Only the expansion move is utilised for optimization due to its guaranteed, proven optimality properties [8]. The objective is to find the cut with minimum energy E_{Total}, that partitions the data points d-vertices such that each d-vertex is associated with a single l-vertex corresponding to its label. When the designated cut is reached, a suboptimal labeling $Class_{labels}$ is realized.

3.2. Proposed Energy Function Definition for GCO_mine

The energy of a cut E_{Total} is defined as the sum of the costs of the edges it severs as it is formulated. E_{Total} is based on the data energy E_{data} and smoothness energy E_{smth}:

$$E_{Total} = E_{data} + \beta E_{smth} \tag{1}$$

where E_{data} measures the conformity of the data points with each label, β determines the contribution weight of E_{smth}, and E_{smth} quantifies the interaction penalty of the neighbouring data points with each other. For each candidate cut, E_{data} and E_{smth} are computed corresponding to the costs of the t-links and n-links, respectively. Both energy components are calculated using the standardized Euclidean distance ξ between data points (equations 3 and 7). A standardized distance metric is chosen to balance the contribution of each feature to the cost, as all features are converted to the same scale.

The data energy E_{data} (equation 2) comprises two cost functions: C_u and C_l, where C_u computes the cost of assigning an unlabeled data point d_i to a class l_i, where $d_i \in D_u$ (equation 3) and C_l sets the cost of classifying a labeled training data point d_i to class l_i, where $d_i \in \mathcal{T}$ and $\mathcal{T} \subset D_l$ (equation 4). In order to reduce computational cost and avoid the use of noise or outliers in the classification and energy calculations, only a subset $\mathcal{T}$ of the labeled samples D_l is used for training. E_{data}, C_u and C_l are defined as:

$$E_{data} = \sum_{d_i \in D_u} C_u(l_i|\, d_i) + \sum_{d_i \in \mathcal{T}} C_l(l_i|\, d_i) \tag{2}$$

$$C_u(l_i|\, d_i) = \xi(d_i, \eta) \tag{3}$$

$$C_l(l_i|\, d_i) = \begin{cases} 0, & \text{if } l_i = \mathfrak{C} \\ \infty, & \text{otherwise.} \end{cases} \tag{4}$$

C_u is measured as the distance between d_i and the representative η of class l_i. η is selected from the training subset $\mathcal{T}$. For a labeled data point d_i, the cost C_l is set to zero when l_i is the 'ground truth' target class $\mathfrak{C}$ and to ∞ (practically the largest integer) in all other cases, to direct the chosen cut and guide the classification process. This extremely large distance acts as a high penalty imposed for misclassification.

E_{smth} is calculated as the sum of the normalized costs of assigning two neighbours d_i and d_j to different classes (cutting their n-link), $\omega(d_i, d_j)$ (equation 5). This cost is calculated as the difference between the local maximum distance and the pairwise distance normalized to the local maximum distance (equation 7). The local maximum

distance $\max_{p\in\aleph}(\xi(d_i, d_p))$ is the largest pairwise distance among the calculated distances between d_i and a number m of its nearest neighbours $\aleph$, where $\aleph \subset \{D_u \cup \mathcal{T}\}$. Traditional lazy learning methods include only training samples for establishing neighbourhoods. Unlike these methods, in *GCO_mine* both labeled and unlabeled data can contribute to the classification decision of a point d_i, to achieve better structured classes.

$$E_{smth} = \sum_{d_i} \sum_{d_j \in \aleph} V(l_i, l_j |\, d_i, d_j).\ \omega(d_i, d_j) \tag{5}$$

$$V(l_i, l_j |\, d_i, d_j) = \begin{cases} 0, & \text{if } l_i = l_j \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

$$\omega(d_i, d_j) = \frac{\max_{p\in\aleph}(\xi(d_i, d_p)) - \xi(d_i, d_j)}{\max_{p\in\aleph}(\xi(d_i, d_p))} \tag{7}$$

Given these energy definitions, two aspects affect the classification process: the choice of the class representative η and the selection of the training samples subset $\mathcal{T}$. Firstly, the choice of a class representative significantly influences the C_u contributing to E_{data}, therefore it is important to investigate the impact of the representative choice. Two class representatives are studied, namely:

- Centroid (*Cent*): η is the data point with minimum overall (sum) distance to all $\mathfrak{C}$ members
- Closest-point (*Close*): η is a data point from $\mathfrak{C}$ with minimum distance to point d_i.

Secondly, the selection method for the labeled subset ($\mathcal{T}$) and the number of selected instances need to be considered. The commonly used selection strategies are random and guided. In random selection ($_r$), the training samples are drawn arbitrarily from the training set. This sampling process is simple and easy to apply, but may lead to the selection of clustered samples or outliers. A guided selection ($_g$) method is proposed to address this issue. The aim of the proposed guided method is to sample labeled data points of each class that are uniformly distributed in the feature space, while avoiding outliers. For this purpose, the centroid of each class is determined and the angle space around the centroid is partitioned into n_s slices, where n_s is the number of samples to be drawn. The angle between the class training samples and the centroid is calculated to locate each training instance within an angle partition. Then for each angle partition, the training instance closest to the median is selected to ensure that it is an appropriate representative (i.e., not an outlier). The number of selected instances n_s depends on a predefined sampling rate S_r and the number of training samples per class n_{tc} such that $n_s = S_r \times n_{tc}$.

Algorithm 1 Non-Parametric Classification via GCO

1: **procedure** $GCO\_mine(D_l, D_u, \beta, m, S_r)$
2: $\quad\mathcal{T} = Sample\_Training(D_l, S_r)$
3: $\quad E_{data} = Compute\_EData(D_u, \mathcal{T})$
4: $\quad E_{smth} = Compute\_ESmooth(D_u, \mathcal{T}, m)$
5: $\quad Class_{labels} = GCO(D_u, E_{data}, E_{smth}, \beta)$
6: **end procedure**

The main steps of the *GCO_mine* method are outlined in Algorithm 1. The differences in subset selection of the training samples and the class representative choices lead to variations in $Sample\_Training(D_l, S_r)$ (line 2) and $Compute\_EData(D_u, \mathcal{T})$ (line 3), respectively. Thus, there are four variants of *GCO_mine*: $GCO\_mine_{r,Cent}$, $GCO\_mine_{g,Cent}$, $GCO\_mine_{r,close}$ and $GCO\_mine_{g,close}$.

3.3. Parameter Setting for GCO_mine variants

The performance of *GCO_mine* is controlled by hyper parameters β, m and S_r. The optimization of these parameters can be done by search methods such as grid or heuristic search. Despite its simplicity, blind grid search becomes inefficient as the number of parameters increases and is not practical for searching continuous spaces. In contrast, genetic algorithms (GA) scale well with the increase in parameter numbers. Since the parameter space of *GCO_mine* variants is limited, both blind search and genetic algorithms can be applied in conjunction with the proposed variants to reach a near optimal setting. We employ both search techniques and compare them. The performance with each setting is evaluated using cross validation. In blind search, the parameter space is uniformly partitioned and the accuracy of the variants of *GCO_mine* with each setting is considered. With GAs, each chromosome encodes a possible set of values for β, m and S_r and the fitness of the solution is defined as the overall accuracy.

Table 1: Medical Dataset Characteristics

Dataset	#C	*S*	*DB*	*CH*
RVA_data	3	0.15	5.30	21.81
Pima Diabetes	2	0.16	4.42	24.29
Ecoli	8	0.35	1.57	81.17
Parkinson	2	0.25	2.85	13.05
Wisc.Breast Diagnostic	2	0.61	0.72	633.63
Breast Tissue	6	-0.36	3.69	6.73
VertebColumn 2C	2	0.09	1.56	55.65
VertebColumn 3C	3	0.08	2.09	97.71
Colic#1 (Surg_lesion)	2	-0.10	5.1	4.78
Colic#2 (Outcome)	3	0.04	2.94	26.13
Lymph	4	0.15	1.86	11.50
Dermatology	6	-0.13	5.82	2.07
HeartCleveland 2C	2	-0.05	6.30	2.77

4. Experimental Evaluation

We first introduce our experimental study and then discuss the results on both our RVA data and *UCI Machine Learning Repository* datasets.

4.1. Experimental Study

An extensive set of experiments are conducted to validate the effectiveness and suitability of *GCO_mine* on our RVA dataset. Additionally, the general applicability of the proposed method is demonstrated using 12 benchmark datasets, where its performance is shown to be competitive with well established approaches. The experiments employ the Weka [20] and MATLAB environments.

Experimental data: The proposed algorithm is first applied on the RVA data. Asymptomatic volunteers were recruited and investigated similarly to [35, 38]. Our study is based on 236 participants, eliminating subjects who had a positive diagnosis of severe cardio- or cerebro-vascular disease. Retinal vessel reactivity was measured using the Dynamic Vessel Analyser (DVA); IMEDOS GmbH, Jena, Germany[35]. For each subject, both artery and vein responses were measured over a period of 350 seconds, including three cycles $F1, F2$ and $F3$ of flicker light stimulation. Retinal vessel diameters were recorded at a frequency of 25 readings/sec. After applying polynomial regression on the recorded response, for each vessel and each flicker cycle and also for the averaged flicker per vessel [35, 38], a set of features are calculated per subject creating a total of 104 features [18].

For labeling the data, we adopt a scheme based on the Framingham Risk Score (FRisk) [12]. The FRisk provides a validated means of predicting cardiovascular risk in asymptomatic patients. It presents a 10-year risk score for each subject given physical examination findings and laboratory evaluations. The applied risk score thresholds are the de facto standard widely used in the literature [6]. Three groups are defined and the subjects are labeled accordingly:

- **Low Risk (LR):** Subjects with FRisk <10% (211 participants).
- **Medium Risk (MR):** Subjects with 10%≤ FRisk <20% (15 participants).
- **High Risk (HR):** Subjects with FRisk ≥ 20% and subjects with unknown FRS but have one or more risk factors (smoker, Family History of CVD, Diabetes Prone) (10 participants).

For successful cardiovascular risk prediction RVA data need to be oversampled [18]. ADAptive SYNthetic (ADASYN) oversampling [22] is applied on the data and the learning algorithms use the oversampled (real+synthesized) data for prediction. The resultant POST-ADASYN dataset include 211 low risk, 211 medium risk and 208 high risk samples.

The benchmark medical datasets are selected from *UCI Machine Learning Repository* [32] to have similar characteristics to our RVA dataset in terms of size and/or number of target classes. Imbalanced datasets are selected to determine the effectiveness of the proposed algorithm in case of skewed datasets. The datasets are outlined in Table 1, they depict various medical conditions and include numeric and categorical features. To indicate the cohesion and separation properties of the classes in each data set, three recognized evaluation measures are reported: Silhouette index (S), Davies Bouldin index (DB) and Calinski-Harabasz criterion (CH) [33]. Silhouette index measures the consistency within classes, while DB and CH assess the scatter within the classes relative to the separation in between them. DB uses the ratio between the intra- and inter- class distances and CH compares inter- to intra- class variances.

Table 2: *GCO_mine* Variants Performance on RVA_data

Variant	Blind Search (b)		Genetic Algorithm (i)	
	OA	$t_c[s]$	OA	$t_c[s]$
$GCO\_mine_{r,Cent}$	86.67	812.34	88.57	1020.26
$GCO\_mine_{g,Cent}$	88.89	1895.04	88.89	1998.94
$GCO\_mine_{r,close}$	93.70	1367.36	*98.89*	1125.85
$GCO\_mine_{g,close}$	*98.89*	2114.35	**99.52**	2056.42

Table 3: Classifiers Performance on RVA_data

	OA	AUC	TP_H	F_m	t_{sr}
RF	98.42	0.99	*1*	0.98	4.98
MLP	93.84	0.97	*1*	0.94	235.76
NB	85.80	0.96	*1*	0.87	1.26
kNN	83.28	0.89	0.84	0.83	**0.98**
hGCO_mine	**99.52**	**0.996**	*1*	**0.992**	2.13

Methods Implementation and Settings: Variants of *GCO_mine* are implemented using MATLAB R2016b. They rely on the MATLAB implementation of GCO Toolbox v3.0 [8, 7, 31]. For setting the parameters of *GCO_mine* variants, the allowed ranges are: [0.1, 20] for β, $\{1, 2, \ldots, 10\}$ for m and [0.1, 0.9] for S_r. These ranges are selected to offer a balance between performance and computation complexity.

Experimental Algorithms: We compare the results of the proposed algorithm with those of four well established classifiers: Random Forest (RF), Multilayer Perceptron (MLP), Naive Bayes (NB) and K-nearest neighbours (kNN). In our earlier study, we conducted a smaller set of experiments using a wide range of classifiers [18], where we found RF, MLP and NB performing the best. Hence, these were selected for the current experiments. Also, RF and MLP are known for their effectiveness and robustness, while NB offers low computation complexity. kNN is a popular lazy classifier, to which the proposed method is similar, because it bases the classification decision on the neighbourhood of the instance. The default Weka implementation of RF, MLP, NB and kNN is utilised in the experiments.

Experimental Procedure: First, experiments are conducted to evaluate the success of the proposed variants on the RVA data. Then, the variant with the best accuracy is applied on the benchmark datasets and its performance is compared to the selected well-established classifiers. The reported results are the average of five 10-fold cross validation runs on the available datasets.

Evaluation metrics: The utilised performance indicators are Overall Accuracy (OA), Area Under the ROC Curve (AUC), F-measure (F_m), and True Positives for High risk group when the class is clearly marked as High risk (TP_H). While OA is a measure commonly used in the literature, AUC and F_m are more suitable for imbalanced data. TP_H is reported in this study because a misclassification in the high risk group could be detrimental (i.e. not capturing the level of risk may lead to missing out on treatment and subsequently deteriorating health). For RVA_data, the total computation time t_c needed for parameter space search and training subset selection and *GCO_mine* variants evaluation is recorded. Also, the execution times of a single run t_{sr} for all classifiers on RVA_data are compared.

4.2. Results and Discussion

The performance of *GCO_mine* is evaluated on RVA data and other medical benchmark datasets using blind search and genetic algorithms for parameter setting.

4.2.1. RVA Data

The *GCO_mine* variants are applied on ADASYN oversampled RVA data [18]. The overall classification accuracy (OA) and computation time (t_c) needed for parameter setting by each method are recorded in Table 2. An accuracy improvement (of at least 0.63%) is offered by genetic algorithms with three variants. In terms of computation time t_c, there is no clear winner, as blind search presents lower t_c with *cent* variants and genetic algorithms show lower t_c with *close* variants for RVA data. Consequently both search methods will be applied on the benchmark datasets.

From these results it can be observed that *close* variants outperform *cent* variants with an improvement of OA of at least 4.81%. This can be explained by the fact that *cent* fails to represent scattered and partially coinciding classes, characteristic to our RVA dataset due to the presence of subjects at the borderline between two risk groups. Also, guided sampling contributes to better classification accuracy, at the expense of computation time increase of almost 100%. Thus, the selection of the sampling method becomes a design decision between high accuracy ($_g$) and low computation time ($_r$). For CVR prediction, high accuracy is essential to avoid consequences of misclassification, which can be detrimental for missed high risk patients or costly for allocation of low risk patients onto unnecessary treatment plans. Therefore, the highest accuracy $GCO\_mine_{g,close}$ with GA parameter setting is chosen for further investigation and comparison to RF, MLP, NB and kNN.

The results of $GCO\_mine_{g,close}$ with genetic algorithms (denoted as *hGCO_mine* in Table 3) and the set of established classifiers (RF, MLP, NB and kNN) are shown in Table 3. The results portray the superiority of $GCO\_mine_{g,close}$

Table 4: Classifiers Performance on Continuous Datasets

		OA	*AUC*	TP_H	F_m
Pima Diabetes	RF	75.26	0.81	0.60	0.75
	MLP	75.13	0.79	0.61	0.75
	NB	76.30	0.81	0.61	0.76
	kNN	72.52	0.79	0.58	0.72
	bGCO_mine	76.32	0.80	0.59	0.76
	hGCO_mine	**77.24**	**0.83**	**0.65**	**0.77**
Ecoli	RF	86.09	0.96		0.86
	MLP	85.71	0.95		0.86
	NB	85.41	0.96		0.86
	kNN	86.90	0.95		0.86
	bGCO_mine	88.48	0.97		0.87
	hGCO_mine	**89.39**	**0.98**		**0.89**
Parkinson	RF	91.28	0.96	0.75	0.91
	MLP	91.28	0.96	**0.83**	0.91
	NB	69.23	0.86	0.61	0.75
	kNN	93.84	*0.98*	0.81	*0.94*
	bGCO_mine	93.16	*0.98*	0.77	0.93
	hGCO_mine	**94.21**	*0.98*	0.81	*0.94*
Wisconsin Breast Diagnostic	RF	95.78	*0.99*	*0.98*	0.96
	MLP	96.66	*0.99*	0.95	*0.97*
	NB	92.97	0.98	0.94	0.93
	kNN	95.95	0.95	0.97	0.96
	bGCO_mine	96.25	*0.99*	0.96	*0.97*
	hGCO_mine	**96.79**	*0.99*	*0.98*	*0.97*
Breast Tissue	RF	71.69	*0.93*		0.72
	MLP	64.15	0.88		0.65
	NB	70.75	*0.93*		0.71
	kNN	71.69	0.83		0.72
	bGCO_mine	69.00	0.83		0.70
	hGCO_mine	**72.00**	*0.93*		**0.73**
Verteb Column 2C	RF	84.19	0.93	0.76	0.84
	MLP	84.51	0.93	0.70	0.85
	NB	77.74	0.88	0.87	0.80
	kNN	80.00	0.86	0.68	0.80
	bGCO_mine	**87.00**	0.93	*0.78*	*0.86*
	hGCO_mine	86.67	**0.94**	*0.78*	*0.86*
Verteb Column 3C	RF	83.54	0.96		0.84
	MLP	85.48	0.96		*0.86*
	NB	83.23	0.95		0.83
	kNN	77.42	0.91		0.78
	bGCO_mine	86.33	*0.98*		*0.86*
	hGCO_mine	**87.00**	*0.98*		*0.86*

Table 5: Classifiers Performance on Categorical Datasets

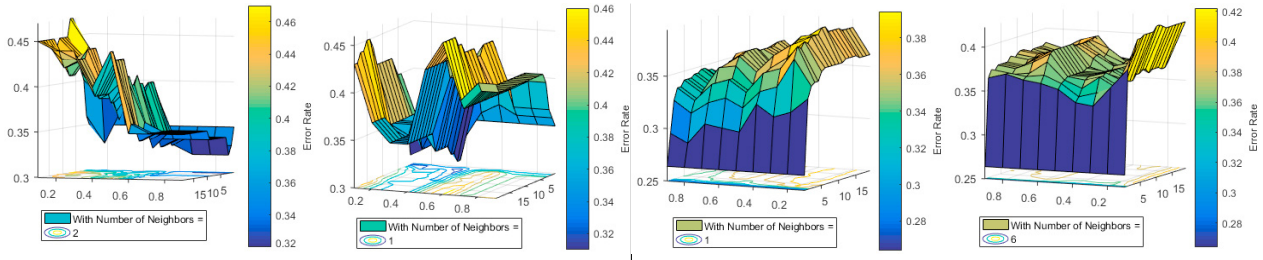
		OA	*AUC*	TP_H	F_m
Colic #1 (Surg_Lesion)	RF	**85.32**	**0.89**	**0.77**	**0.85**
	MLP	81.25	0.87	0.74	0.82
	NB	77.1	0.82	0.74	0.78
	kNN	84.23	0.88	0.73	0.84
	bGCO_mine	73.61	0.81	0.72	0.74
	hGCO_mine	73.61	0.81	0.72	0.74
Colic #2 (outcome)	RF	69.94	*0.83*		0.67
	MLP	69.39	0.77		*0.69*
	NB	68.30	*0.83*		*0.69*
	kNN	69.12	0.82		0.66
	bGCO_mine	**72.50**	0.79		0.66
	hGCO_mine	71.39	0.78		0.65
Lymph	RF	83.11	*0.93*		0.83
	MLP	**89.96**	*0.93*		**0.90**
	NB	85.13	0.89		0.85
	kNN	69.59	0.84		0.69
	bGCO_mine	85.71	0.88		0.84
	hGCO_mine	85.71	0.88		0.84
Dermatology	RF	96.44	*1.00*		0.97
	MLP	*97.54*	*1.00*		*0.98*
	NB	*97.54*	*1.00*		*0.98*
	kNN	95.62	0.99		0.95
	bGCO_mine	91.39	0.94		0.90
	hGCO_mine	91.67	0.94		0.91
Heart Cleveland 2C	RF	77.62	0.78	0.38	0.77
	MLP	76.71	*0.80*	**0.56**	0.77
	NB	78.89	*0.80*	0.51	0.79
	kNN	**81.74**	0.77	0.44	*0.80*
	bGCO_mine	77.14	0.77	0.35	0.79
	hGCO_mine	80.48	0.79	0.44	*0.80*

over its counterparts considering the *OA*, *AUC* and F_m evaluation metrics, while a TP_H of 1 is achieved by all algorithms except kNN. On the other hand, the least execution time t_{sr} is offered by kNN, while $GCO\_mine_{g,close}$ presents an average *OA* improvement of 14% over the accuracies of the faster kNN and NB alternatives. Compared to *RF*, the second best alternative, $GCO\_mine_{g,close}$ improves *OA*, *AUC*, F_m and reduces t_{sr} to 43%.

4.2.2. Benchmark Medical Data

$GCO\_mine_{g,close}$, RF, MLP, NB and kNN are applied on the benchmark datasets previously outlined in Table 1. Blind search and heuristic search using genetic algorithms denoted by $_b$ and $_h$ respectively are employed for parameter setting with $GCO\_mine_{g,close}$. Table 4 and Table 5 depict the performance evaluation results. Table 4 illustrates the results of the datasets with features having real continuous values, while Table 5 present results on the datasets that include categorical variables. As shown, $GCO\_mine_{g,close}$ is particularly effective on datasets of continuous real features and it presents a competitive performance on categorical datasets. This can be attributed to the utilization of standardized Euclidean distance metric in data cost calculation, since this metric is designed for real valued samples. A similar improvement could be achieved for categorical data through the future adoption of a specially designed distance function.

Overall, $GCO\_mine_{g,close}$ has higher *OA* on 8 out of 12 datasets with differences ranging from 0.13 % to 2.56 % to the second highest accuracy value. In some cases, $GCO\_mine_{g,close}$ shows a remarkable accuracy increase such as when applied on the Parkinson dataset, a 25 % increase is recorded when compared to NB. In comparison to kNN (a nonparametric classifier of similar principle), $GCO\_mine_{g,close}$ is superior in 9 out of 12 datasets; the improvement reaches 16.12 %. However, the performance of $GCO\_mine_{g,close}$ is lower for datasets of low compactness and sepa-



(a) Breast Tissue dataset **(b)** Colic #1 dataset

Fig. 2: Error surface plots for number of neighbours given by blind search (left) and GA (right)

rability: Colic# 1, Dermatology and HeartCleveland 2C. Their low compactness is indicated by negative silhouette index values, relatively high values of Davies Bouldin index and low Calinski-Harabasz criterion values. Even though *GCO_mine* and kNN rely on a similar concept for data cost determination, the inclusion of smoothing energy E_{smth} and the adoption of graph cut optimization lead to better classification for homogeneous classes. However, in cases of inconsistent classes, the filtering over-smoothing effect of *GCO_mine* leads to performance deterioration. For continuous variables (Table 4) GA performs better than blind search on all datasets with the exception of Verteb Column 2C, while in the case of categorical variables (Table 5) the difference between the results of the two methods is not consistent, i.e., equal in two cases, in one case blind search is better, in two cases GA is better. In order to understand the reasons for these, in Fig. 2 we plot the error against the two parameters $\beta \in [0.1, 20]$ and $S_r \in [0.1, 0.9]$, for the number of neighbours m chosen by the two methods, blind search (left) and GA (right), respectively. The two datasets chosen for comparison are Breast Tissue, as a representative with difference in accuracy and Colic#1, as representative for equal performance. The resulting phenotype fitness landscape in Fig. 2 (a) is more rugged and more complex, with several local optima, whereas the landscape in Fig. 2 (b) is smoother and simpler, with less local optima. Thus, it can be seen that the GA performs better than blind search on the more complex landscape and both methods perform similarly on the simpler landscape.

5. Conclusion

In this paper, an effective partially lazy classification method has been proposed to predict cardiovascular risk level from RVA data. *GCO_mine* has been created to accommodate continuous data collection and produce accurate instance based classification, as needed in this context. *GCO_mine* merges the concepts of decision locality and global optimization; thus, it handles the presence of outliers and noise better than the traditional lazy alternative kNN. Indeed, compared to the kNN lazy classifier and the RF, MLP and NB well-established eager classifiers, *GCO_mine* together with *heuristic* parameter setting (*hGCO_mine*) presents the highest accuracy on RVA data, namely 99.52%. Furthermore, *GCO_mine*'s general utility is demonstrated on 12 benchmark medical datasets from the *UCI Machine Learning Repository* [32]. *GCO_mine* manifests superior performance relative to NB on 9 out of 12 datasets, while showing similar results to MLP and RF. In conclusion, *GCO_mine* not only accurately predicts cardiovascular risk level based on RVA data, it also offers a competitive solution to a broad range of medical classification problems, with the additional intrinsic advantage of applicability to newly collected samples.

References

[1] Aha, D.W., 1997. Lazy Learning, Kluwer Academic Publishers, pp. 7–10.
[2] Aha, D.W., Kibler, D., Albert, M., 1991. Instance-based learning algorithms. Machine Learning 6, 37–66.
[3] Arts, E.E.A., Popa, C., Den Broeder, A.A., Semb, A.G., Toms, T., Kitas, G.D., van Riel, P.L., Fransen, J., 2015. Performance of four current risk algorithms in predicting cardiovascular events in patients with early rheumatoid arthritis. Ann Rheum Dis 74, 668–674.
[4] Bai, J., Xiang, S., Pan, C., 2013. A Graph-Based Classification Method for Hyperspectral Images. IEEE Trans. GRS. 51, 803–817.
[5] Bao, Y., Ishii, N., Du, X., 2004. Combining Multiple k-Nearest Neighbor Classifiers Using Different Distance Functions, in: Proc. 5th Int. Conf. IDEAL 2004., Springer Berlin Heidelberg. pp. 634–641.
[6] Bosomworth, N.J., 2011. Practical use of the Framingham risk score in primary prevention. Can Fam Physician 57, 417–423.

[7] Boykov, Y., Kolmogorov, V., 2004. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. IEEE Trans. PAMI 26, 1124–1137.
[8] Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. PAMI 23, 1222–1239.
[9] Brindle, P.M., McConnachie, A., Upton, M.N., Hart, C.L., Smith, G.D., Watt, G.C.M., 2005. The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study.
[10] Cleary, J.G., Trigg, L.E., 1995. K*: An instance-based learner using an entropic distance measure, in: Proc. 12th ICML, pp. 108–114.
[11] Cooney, M.T., Dudina, A.L., Graham, I.M., 2009. Value and Limitations of Existing Scores for the Assessment of Cardiovascular Risk: A Review for Clinicians. J. Am. Coll. Cardiol 54, 1209–1227.
[12] D'Agostino, R.B., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M., Kannel, W.B., 2008. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. Circulation 117, 743–753.
[13] Damodaran, B.B., Nidamanuri, R.R., Tarabalka, Y., 2015. Dynamic Ensemble Selection Approach for Hyperspectral Image Classification With Joint Spectral and Spatial Information. IEEE J. STARS 8, 2405–2417.
[14] Dashtbozorg, B., Mendonça, A.M., Campilho, A., 2014. An Automatic Graph-Based Approach for Artery/Vein Classification in Retinal Images. IEEE Trans. IP 23, 1073–1083.
[15] Dhillon, I.S., Guan, Y., Kulis, B., 2007. Weighted Graph Cuts without Eigenvectors A Multilevel Approach. IEEE Trans. PAMI 29, 1944–1957.
[16] Ding, C.H.Q., He, X., Zha, H., Gu, M., Simon, H.D., 2001. A min-max cut algorithm for graph partitioning and data clustering, in: Proc. 2001 IEEE Int. Conf. on Dat Min, pp. 107–114.
[17] Duprez, D.A., Cohn, J.N., 2007. Detection of Early Cardiovascular Disease. Springer London. pp. 1615–1622.
[18] Fathalla, K.M., Ekárt, A., Seshadri, S., Gherghel, D., 2016. Cardiovascular risk prediction based on Retinal Vessel Analysis using machine learning, in: 2016 IEEE Int.Conf. SMC, pp. 880–885.
[19] Frank, E., Hall, M., Pfahringer, B., 2003. Locally Weighted Naive Bayes, in: Proc. 19th Conf. UAI, pp. 249–256.
[20] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations 11, 10–18.
[21] Hassanat, A.B., Abbadi, M.A., Alhasanat, A.A., 2014. Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. IJCSIS 12, 33–39.
[22] He, H., Bai, Y., Garcia, E.A., Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proc.Int.Jt. Conf. Neural Netw. , 1322–1328.
[23] Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A., Brindle, P., 2008. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 336, 1475–1482.
[24] Hobbs, F.D.R., Jukema, J.W., Da Silva, P.M., McCormack, T., Catapano, A.L., 2010. Barriers to cardiovascular disease risk scoring and primary prevention in Europe. QJM: An International Journal of Medicine 103, 727–739.
[25] Hu, L., Huang, M., Ke, S., Tsai, C., 2016. The distance function effect on k-nearest neighbor classification for medical datasets.
[26] Ikram, M.K., De Jong, F.J., Bos, M.J., Vingerling, J.R., Hofman, A., Koudstaal, P.J., De Jong, P.T.V.M., Breteler, M.M.B., 2006. Retinal vessel diameters and risk of stroke: The Rotterdam Study. Neurology 66, 1339–1343.
[27] Joshi, V., Reinhardt, J., Garvin, M., Abramoff, M., 2014. Automated Method for Identification and Artery-Venous Classification of Vessel Trees in Retinal Vessel Networks. PloS one 9, e88061.
[28] Karypis, G., Kumar, V., 1998. Multilevelk-way Partitioning Scheme for Irregular Graphs. J. Parallel Distrib. Comput. 48, 96–129.
[29] Kaveh-Yazdy, F., Zare-Mirakabad, M., Xia, F., 2012. A Novel Neighbor Selection Approach for KNN: A Physiological Status Prediction Case Study, in: Proc. 1st Int. Workshop ContextDD, ACM. pp. 2:1–2:7.
[30] Kawasaki, R., Xie, J., Cheung, N., Lamoureux, E., Klein, R., Klein, B.E.K., Cotch, M.F., Sharrett, R., Shea, S., Wong, T., 2012. Retinal microvascular signs and risk of stroke: the Multi-Ethnic Study of Atherosclerosis (MESA). Stroke; J.Am. Heart Assoc. 43, 1984–1992.
[31] Kolmogorov, V., Zabin, R., 2004. What energy functions can be minimized via graph cuts? IEEE Trans. PAMI 26, 147–159.
[32] Lichman, M., 2013. UCI Machine Learning Repository. URL: http://archive.ics.uci.edu/m.
[33] Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., . Understanding of Internal Clustering Validation Measures, in: 2010 IEEE ICDM, pp. 911–916.
[34] Mendis, S., Puska, P., Norrving, B., . Global atlas on cardiovascular disease prevention and control. Technical Report. Geneva 2011.
[35] Mroczkowska, S., Ekart, A., Sung, V., Negi, A., Qin, L., Patel, S.R., Jacob, S., Atkins, C., Benavente-Perez, A., Gherghel, D., 2012. Coexistence of macro- and micro-vascular abnormalities in newly diagnosed normal tension glaucoma patients. Acta ophthalmol 90, 553–559.
[36] Rothaus, K., Jiang, X., Rhiem, P., 2009. Separation of the retinal vascular graph in arteries and veins based upon structural knowledge. Image and Vision Computing 27, 864–875.
[37] Seifertl, B.U., Vilser, W., 2002. Retinal Vessel Analyzer (RVA)–design and function. Biomed Tech (Berl) 47, 678– 681.
[38] Seshadri, S., Ekart, A., Gherghel, D., 2016. Ageing effect on flicker-induced diameter changes in retinal microvessels of healthy individuals. Acta ophthalmol 94, 35–42.
[39] Shang, W., Huang, H., Zhu, H., Lin, Y., Wang, Z., Qu, Y., 2005. An Improved kNN Algorithm – Fuzzy kNN, in: Proc. Int. Conf. CIS 2005 Part I, Springer Berlin Heidelberg. pp. 741–746.
[40] Thomas, M.R., Lip, G.Y.H., 2017. Novel Risk Markers and Risk Assessments for Cardiovascular Disease. Circulation Research 120, 133–149.
[41] Wang, J., Li, X., 2010. An improved KNN algorithm for text classification, in: Proc. 2010 ICINA, pp. V2.436–V2.439.
[42] Wang, J., Neskovic, P., Cooper, L.N., 2006. Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence. Pattern Recognition 39, 417–423.
[43] Xie, Z., Hsu, W., Liu, Z., Lee, M.L., 2002. SNNB: A Selective Neighborhood Based Naive Bayes for Lazy Learning, in: Proc. PAKDD, Springer Berlin Heidelberg. pp. 104–114.
[44] Zagrouba, E., Gamra, S.B., Najjar, A., 2014. Model-based graph-cut method for automatic flower segmentation with spatial constraints. Image Vis. Comput. 32, 1007–1020.