Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Zhang et al., J Biomet Biostat 2012, S7
DOI: 10.4172/2155-6180.S7-008

Research Article Open Access

Automatic Glaucoma Diagnosis with mRMR-based Feature Selection

Zhuo Zhang[1,2*], Chee Keong Kwoh[2], Jiang Liu[1], Carol Yim Lui Cheung[3], Tin Aung[3] and Tien Yin Wong[3,4]

[1]Institute for Infocomm Research, Singapore
[2]School of Computer Engineering, Nanyang Technology University, Singapore
+Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
[4]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Abstract

Glaucoma's irreversibility, lacking of glaucoma specialists and patient unawareness demand for an economic and effective glaucoma diagnosis system for screening. In this study we explore feature selection (FS) technologies to identify the most essential parameters for automatic glaucoma diagnosis.

Methods: We compose feature space from heterogeneous data sources, i.e., retinal image and eye screening data. A feature selection framework is proposed by exploring multiple feature ranking schemes and a wide range of supervised learners. The optimal feature set is derived automatically from the performance curve smoothed by measurement score regression.

Results: Under the proposed framework, the optimal feature set obtained using mRMR (minimum Redundancy Maximum Relevance) scheme contains only 1/4 of the original features. The classifiers trained upon the optimal feature set are more efficient with better performance in terms of Accuracy and F-score. A detailed investigation on the features in the optimal set indicates that they can be the essential parameters for glaucoma mass screening and image segmentation.

Conclusions: An intelligent Computer-aid-diagnosis (CAD) model is constructed for automatic disease diagnosis. The effectiveness of the model is demonstrated in our glaucoma study based on heterogeneous data sets. The effort not only improves the discriminative power, but also facilitates a better understanding of CAD process and may ease the data collection in glaucoma mass screening.

Introduction

Glaucoma is a chronic and irreversible neurodegenerative eye condition in which the optic nerve fibers and astrocytes are progressively damaged [1,2]. It is the second leading cause of blindness worldwide with estimated 60 million glaucoma cases globally in 2010 [3]. However, many glaucoma patients are not aware of the disease until late stage due to lacking of an effective early screening system. In Singapore, the SiMES eye study [4] showed that 90% of the glaucoma patients are unaware of their conditions. As the lost capability of the optic nerve cannot be recovered, early diagnosis and subsequent early treatment [5] are important to preserve the vision of the affected patients.

In clinical practice, glaucoma is diagnosed based on the analysis of patients' medical history, measurement of the intraocular pressure, testing of visual field loss, the manual assessment of the optic nerve head (ONH) via ophthalmoscopy of fundus imaging [6] and etc. Due to the complexity and variety of the disease pathology, the diagnosis of glaucoma relies heavily on the experiences of the glaucoma specialist. Glaucoma's irreversibility, lacking of glaucoma specialists and patient unawareness demand for an economic, effective and automatic glaucoma mass screening system.

In a traditional eye screening program, patients' health records are documented in a text database. Recently, techniques in human retina imaging provide complementary and structured information. In today's visual assessment, multiple modalities are available, for example, digital fundus photographs show information similar to what ophthalmologists see from ophthalmoscopes; the Heidelberg Retina Tomograph (HRT) [7] produces reflectance and topographic images of the retinal surface using confocal laser scanning; Optical Coherence Tomography (OCT) [8] captures 3D information about the different cell/tissue layers of the retina.

Advancement in medical image processing have enabled the development of image-based Computer-aided diagnosis systems [9-11]. These systems focus mainly on estimating vertical optic cup-to-disc ratio (vCDR), which, is an important risk factor for detecting the presence of glaucoma [12]. Besides vCDR, numerous pathological signs are often referred by ophthalmologists for glaucoma diagnosis. For instances, the following signs usually suggest high possibility of glaucoma: thinning of neuroretinal rim (NRR) in different quadrants (Inferior, Superior, Nasal and Temporal) [13]; NRR thickness distribution not following the `ISNT Rule' [14]; Retinal Nerve Fiber Layer (RNFL) defect [15] and presence of Alpha and Beta Peripapillary Atrophy (PPA) [16] etc. Advanced image processing techniques are being developed to identify and measure such image cues like vCDR value [9,17], existence of PPA [18], conformation of ISNT rule [19] and detection of RNFL [20] automatically.

With the availability of abundant data, one would expect better disease prediction algorithms. Nevertheless, the parameters and image cues extracted from heterogeneous data sources are of different levels of importance and sometimes interrelated with each other. The interrelated dependency may be structural, probabilistic or even functional. Some features will be redundant when they are highly correlated or derived from the similar primary source or data; some may be contradict with

***Corresponding author:** Zhuo Zhang, Senior Research Officer, Institute for Infocomm Research, 1 Fusionopolis Way, Singapore, E-mail: zzhang@i2r.a-star.edu.sg

Received March 13, 2012; **Accepted** April 09, 2012; **Published** April 12, 2012

Citation: Zhang Z, Khow CK, Liu J, Cheung YLC, Aung T, et al. (2012) Automatic Glaucoma Diagnosis with mRMR-based Feature Selection. J Biomet Biostat S7:008. doi:10.4172/2155-6180.S7-008

each other. Learning algorithms may not work satisfactorily with the complete list of data. In this study, we explore feature selection techniques to identify the optimal subset of important and clinically relevant features thus improving the prediction. Experiments show that the optimal feature set often contains the necessary essential parameters for automatic glaucoma detection. Meanwhile using the optimal feature set reduces the dimensional space and thus computation efforts.

In machine leaning and data mining, feature selection (FS) aims to remove irrelevancy and/or redundancy from the feature space. The advantages of FS are manifold, it helps 1) to avoid overfitting and improve model performance; 2) to provide faster and more cost-effective model; 3) to gain a deeper insight into the underlying model; 4) to reduce data storage requirements and the cost of future measurements and 5) to establish simpler and clearer decision rules.

Based on the selection scheme and the learning algorithm, FS techniques can be classified mainly into two categories: filters and wrappers. Filters [21] remove irrelevant attributes using general characteristics of the training data and are independent of the learning algorithms. On the other hand, wrappers [22] use the learning algorithm to evaluate the given subset, searching for features better suited to the modeling technique. Exhaustive search strategies in wrapper are usually too costly to be deployed, given a large number of features. More efficient algorithms have been developed using heuristic approaches, such as sequential forward selection (SFS) and sequential backward selection (SBS). Research has shown that heuristic search is less prone to data overfitting as compare to exhaustive search [23]. Feature ranking as a filter method is often employed as a principal selection mechanism that to be combined with heuristic wrappers [24].

We propose a feature selection framework for automatic glaucoma diagnosis. The objective is to build a classifier that accurately predicts the classes (glaucoma or normal) of new unlabeled samples, using an optimal subset of features to improve the interpretability and benefit the data collection in mass screening.

The content of this paper is organized as follows. Section II describes our methodologies for classification optimization via exploring various feature selection methods and learning methods. Section III reports the experimental result. Section IV discusses the essential parameters identified, issues and future work.

Methods

Heterogeneous data Sets

The presented work (as illustrated in Figure 1) is based on two heterogeneous data sets, including screening data from SiMES (Singapore Malay Eye Study) [4] and image data from ORIGA [25] database. SiMES is a population-based study conducted from 2004 to 2007 to assess the causes and risk factors of blindness and visual impairment in Singapore Malay community, containing 3280 subjects. The eye screening record in SiMES contains parameters fall into the following categories: 1) demographical data such as age, gender, height; 2) medical histories data acquired via interview; 3) ocular examination data, e.g. intraocular pressure (IOP) and corneal thickness etc. Moreover, diagnostic information such as glaucoma and cataract were available and are used as class label in this study. ORIGA contains 650 retinal images randomly selected from SiMES image collection. The images were segmented semi-automatically and verified by a group of professionals from Singapore Eye Research Institute. The image cues obtained from image segmentation possess valuable information for glaucoma diagnosis. For example, one can use I-S-N-T values to check the compliance of ISNT rule: the normal optic disc usually demonstrates a configuration in which the inferior neuroretinal rim is the widest portion of the rim, followed by the superior rim, and then the nasal rim, with the temporal rim being the narrowest portion.

The heterogeneous data sets are cleaned and fused with following steps: 1) remove features with more than 5% missing values; 2) remove subjects with more than 5% missing values with the remaining features; 3) impute the missing value with mode; 4) digitalize categorical

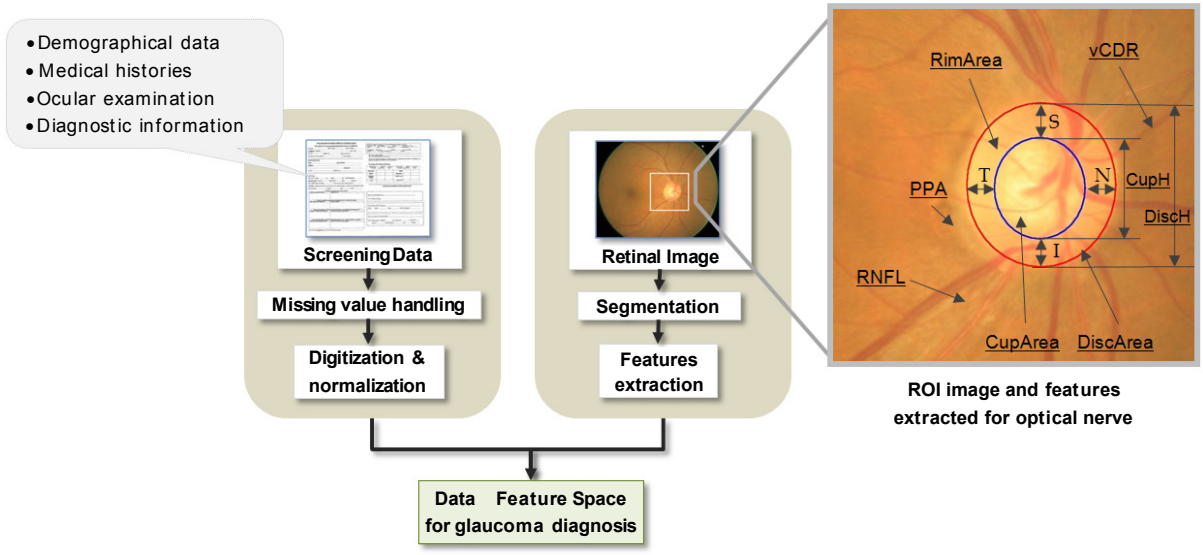

Figure 1: From heterogeneous data sets to feature space.
Retinal image segmentation info: RimArea - area of neuroretinal rim; CupArea - area of optic cup; DiscArea - area of optic disc; vCDR - vertical cup-disc ratio; cupH height of optic cup; DiscH height of optic disc RNFL - retinal nerve fiber layer defect; PPA - peripapillary atrophy; I - optic rim inferior thickness; S - rim superior thickness; N - rim nasal thickness; T - rim temporal thickness.

parameters; 5) merge the two data sets via subject matching. The fused feature space contains 104 features in total, from which 19 features are from retinal image and 85 features are from screening data.

Optimal feature set selection framework

We propose a framework to identify the optimal feature set for learning, as illustrated in Figure 2. In the feature ranking stage, multiple feature ranking criteria are explored. Subsequently, we generate candidate feature sets by employing increment selection method. The incrementally nested feature sets are then fed to a group of learning algorithms. For each classifier trained by different feature sets and learning methods, we conducted 10-fold cross validation to measure their Accuracy and F-score; followed by applying regression method to smooth the performance curve; finally, the optimal feature set is detected via first derivative test.

Feature ranking for incremental feature set selection

Feature ranking serves as a preprocessing step independent of the choice of the classifier, and is categorized as a filter method [22]. Many FS algorithms employ feature ranking as a principal selection mechanism because of its simplicity, scalability, and good empirical success.

In this study we explore and compare several ranking criterion. First let's introduce common notations used in this study. Consider a set of m subjects $\{D_k y_k | k = 1,\ldots m\}$, consisting of n input features $D_k = \{x_{k,i} | i = 1,\ldots n\}$ and one output variable y_k. We denote X_i as the feature vector corresponding to the i^{th} component of input D. Similarly, Y be the vector of which the outcome y is a realization.

Correlation criterion for feature ranking: The correlation criterion is based on Pearson correlation coefficient, defined as:

$$\Re(i) = \frac{cov(X_i, Y)}{\sqrt{var(X_i)var(Y)}} \quad (1)$$

where *cov* designates the covariance and *var* the variance. In discrete applications, the estimate of $R(i)$ is given by

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}} \quad (2)$$

The square of $R(i)$ is the coefficient of determination in linear regression, which represents the fraction of the total variance around the mean value $\bar{y}$ that is instanced by the linear relation between X_i and Y. Therefore, $R(i)^2$ detects linear dependencies between a feature and the target. Statistics inferred from $R(i)$ based on T-test yields p-values of features thus quantify the feature significances measured by correlation criteria. Such p-value is prone to type I errors when sample size is large, thus a Bonferroni correction is applied for a stricter cut-off point of statistics significance. The default cut-off p-value $\alpha_0 = 0.05$ is adjusted by sample number n and the new cut-off is $\alpha = \alpha_0/n$.

Information theoretic ranking criterion: The correlation based feature ranking has several limitations, e.g., it can't quantify the strength of a nonlinear relationship, and it is sensitive to extreme values (outliers). To measure the non-linear dependencies between a feature and the target, mutual information between each feature and the target is further investigated in information theoretic approach. The mutual information is defined by entropy I:

$$I(i) = \int_{x_i}\int_{y} p(x_i, y) log \frac{p(x_i, y)}{p(x_i)p(y)} dxdy \quad (3)$$

where $p(x_i)$ and $p(y)$ are the probability densities of x_i and y, and $p(x_i, y)$ is the joint density. The criterion $I(i)$ is a measure of dependency between the density of variable x_i and the density of the target y.

In the simple case of discrete or nominal features, the integral becomes a sum:

$$I(i) = \sum_{x_i}\sum_{y} P(X = x_i, Y = y) log \frac{P(X = x_i, Y = y)}{P(X = x_i)P(Y = y)} \quad (4)$$

The probabilities are then estimated from frequency counts. For case of continuous variables we can approximate those densities with a non-parametric method.

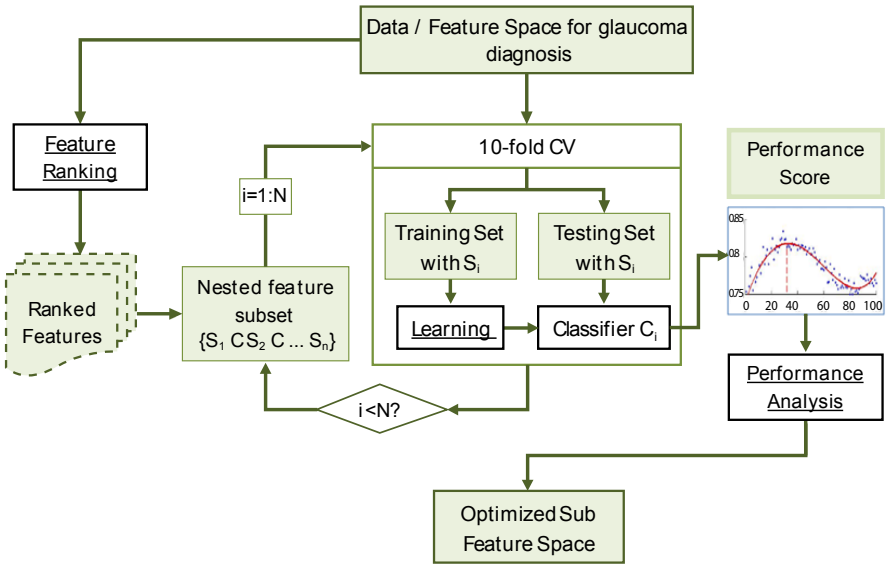

Figure 2: Framework of Optimal feature set searching for Glaucoma Classification.

Minimum Redundancy Maximum Relevance (mRMR) feature selection: Information theoretic ranking criteria takes into consideration of none-linear relationships between a feature and target, however, it assesses features independently and can not deal with feature redundancy problem. To address the issue, we explore minimum Redundancy Maximum Relevance (mRMR) [24] method, which aims at selecting optimal features for classification. For a feature set S with n_0 features $\{x_i\}$, (i = 1,.. n_0). maximum relavance is to search for features such that the mutual information values between individual feature and target should be maximized. Let $D(S, y)$ be the mean of the mutual information between individual features and target y. It is formally defined below,

$$\max D(S, y) = \frac{1}{|S|}\sum_{x_i \in S} I(x_i, y) \tag{5}$$

Although two features may have strong separability on the target class, it would be undesirable to include them if they are highly correlated. The idea of minimum redundancy is to select the features such that they are mutually maximally dissimilar. Let $R(S, y)$ be the mean of the mutual information between pairs of features in S. It is formally defined below,

$$\min R(S) = \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i, x_j) \tag{6}$$

The criterion combining the above two constraints is called minimal-redundancy-maximal-relevance (mRMR). The mRMR feature set is obtained by maximizing $D(S, y)$ and minimizing $R(S)$ simultaneously, which, requires combining the two measure into a single criterion function. Two simplest combination criteria are Mutual Information Difference criterion (MID), definded as $D(S, y)/R(S)$, and Mutual Information Quotient criterion (MIQ), defined as $D(S, y)/R(S)$.

In the following sessions, we denote the above four feature ranking criteria as *correlation-based, entropy-based, mRMR-MIQ and mRMR-MID respectively*.

Feature set selection from ranked features: Exhaustive search in the whole feature space is known to be NP-hard [26] and is prone to be computationally intractable. To compose candidate feature sets, We employ the incremental selection scheme [27] to select n sequential features from the input X, where ranked features are progressively incorporated into n nested subsets $S_1 \subset S_2 \subset \ldots \subset S_n$. These feature sets are sequentially fed to learning algorithms to build the n classifiers.

Machine learning and classifier evaluation

Learning machines: To conduct a comprehensive comparison on the different ranking schemes, a wide range of classification techniques are studied. K Nearest Neighbors (KNN) is an instance-based classifier which classifies objects based on closest training examples in the feature space. It is amongst the simplest of all machine learning algorithms. Linear Discriminant Analysis (LDA) is a commonly used technique for data classification and dimensionality reduction. It maximizes the ratio of between-class variance to the within-class variance to guarantee maximal separability. Logistic Regression (LR) is a generalized linear model used for binomial regression, and it is able to modelling the joint effects of multiple features. Naive Bayes (NB) is a simple probabilistic classifier based on applying Bayes' theorem (or Bayes's rule). NB classifier considers all of features independently contributing to the probability of target. C45 algorithm builds decision trees from a set of training data using the concept of information entropy. The decision tree can then be used for classification. Support Vector Machines (SVM) is a supervised machine learning algorithm that produces a linear boundary to achieve maximum separation between two classes of subjects (cases versus controls), by mathematical transformation (kernel function) of the input features for each subject. In this study we explore SVM with linear kernels (SVM-linear) and SVM with polynomial kernels (SVM-poly). Artificial Neural Networks (ANN) imitate the learning process of human brain and can process problems involving non-linear and complex data by identifying and learning correlated patterns between input data sets and the corresponding target.

In this study, we explore the effect of feature ranking on the above seven classification methods, ranging from lazy learning (KNN) to eager learning (SVM, ANN, C45); from linear to non-linear modelling and from probabilistic based to kernel based.

Evaluation criteria and performance analysis: To evaluate the performance of supervised learning classifiers, and to better reflect the natures of CAD and FS, we use two performance metrics: Accuracy and F-Score to compare different combinations of ranking schemes and learning techniques.

Accuracy is defined as

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{7}$$

where, fp - false positive, fn - false negative, tp - true positive and tn - true negative counts

F-score as

$$F - score = \frac{2 \times (precision \times recall)}{precision + recall} \tag{8}$$

where,

$$precision = \frac{tp}{tp + fp} \tag{9}$$

$$recall = \frac{tp}{tp + fn} = sensitivity \tag{10}$$

While Accuracy measures the overall discrimination power of the classifier, F-Score can better measure the identification of positive class (glaucoma). We will be more interested in improving F-score in a glaucoma CAD system as it is more balanced criterion.

Optimal feature set determination: Classifiers are built using each of the n sequential feature sets $S_1, S_2, \ldots, S_n$. The classification performance scores, denoted as $F_1, F_i, \ldots, F_n$, are obtained via 10-fold cross validation. Our experimental results show that, although in general, more features lead to a better performance score before reaching the optimal feature number, the increment of score might not be significant for each additional feature, there are fluctuation along the score distribution. Many factors count for these fluctuations. One cause is that additional features might be noisy. Another possible cause the cross-validation method might introduce some fluctuations. We use curve fitting to solve the problem.

As illustrated in Figure 3. An optimal solution can be detected on a regression line representing the trend of classification behavior. For $F_i = f(i)$ where i is the feature index and F_i is the performance score for classifier trained using feature set $S_i = \{x_1, x_2, \ldots, x_i\}$ We use 4th order polynomial curve fitting for regression as it best fits the original points in overall cases. The curve function is, $f(i) = \sum_{k=0}^{4} a_k \times i^k$,

where a_k is the polynomail parameter for the curve. The optimal solution can be found by first derivative and second derivative tests, with $f'(i) = 0$ and $f''(i) < 0$. As multiple solutions might exist over one curve, we choose the first turning point after size 10.

In Figure 3, blue `+' points are the raw measurement of performance

score, red line is the regression curve. Green lines are the 1st and 2nd derivative of $f(i)$. The turning point on the red regression line determine the optimal feature set for the classifier achieving the best prediction result. In this figure, the optimal feature set contains 30 features.

Experiment and Result

Performance of classifiers built on full feature set

To determine the effect of the feature set to the classification results of the classifiers, we first build a broad range of 8 machine learning classifiers that utilize full feature set. For every classifier, the input is the 104 normalized features, and the output is the likelihood of the whether the case is glaucomatous. These 8 supervised learning methods have been proven effective in various applications; and are able to learn complex patterns and trends in data as well as create a decision surface that fits the data.

To better utilize the samples, the classifiers are trained and tested using 10-fold cross validation. Following the normal practice, Accuracy and F-score are measured in the experiment. Parameters for each classifier are fine-tuned to obtain optimal performances.

From Table 1, it is shown that whether the kernel is linear or polynomial does not affect the result (in terms of Accuracy, F-score and elapsed time) much; and more complex machine learning ANN classifier (with the expense of elapsed time) outperforms the relative simple classifier like KNN and C4.5 in terms of Accuracy and F-Score. However, the Accuracy and F-Score performance differences among the 8 classifiers are not distinctive.

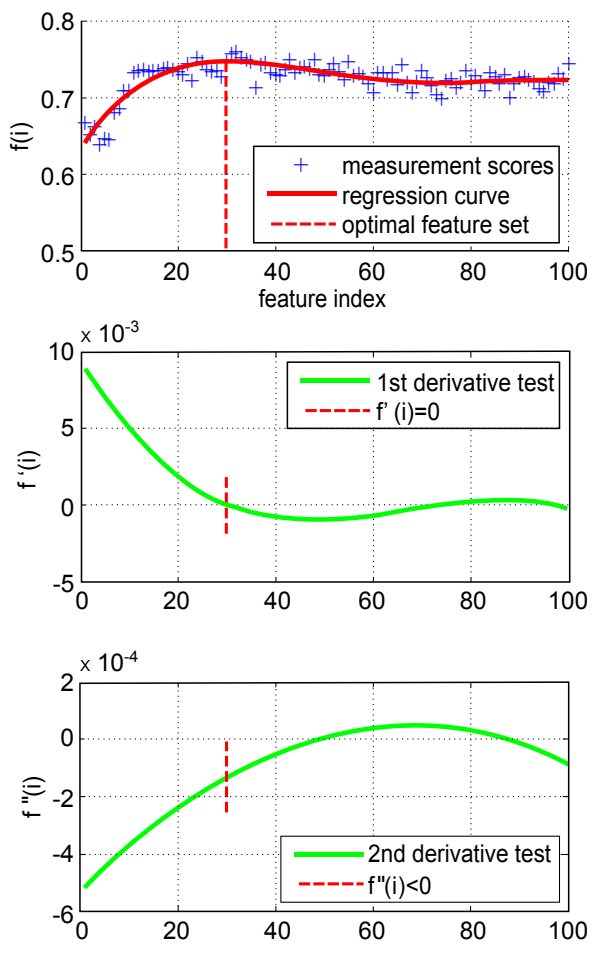

Figure 3: Detect Optimal feature set via regression and 1st derivative test.

Ranked features obtained via various ranking method

The data with full features are fed to four feature ranking methods, e.g., correlation-based, entropy-based, mRMR-MIQ and mRMR-MID based criteria, to obtain the ranked feature list. Table 2 lists the top 20 features picked up by different ranking methods.

It is observed that:

1. All four methods are able to pick up the most important parameter vCDR as the top feature, this is consistent with the clinical practice [12].
2. The common features among the 4 methods are related to the following parameters: vCDR and cup / disc area; age; ISNT rule; Intra-ocular pressure (IOP); peripapillary atrophy (PPA); which are also the important risk factors in glaucoma diagnosis.
3. Comparing the 2 mRMR based FS methods, it is noticed that

Method	Accuracy %	F-score %	Elapsed time for 10-fold cv
SVM-linear	0.809	0.726	9.70s
SVM-poly	0.808	0.737	10.61s
LR	0.821	0.769	6.86s
ANN	**0.825**	**0.777**	50.5s
LDA	0.811	0.755	0.11s
KNN	0.758	0.639	0.38s
C45	0.766	0.714	9.3s
NaiveBayes	0.805	0.739	8.7s

Table 1: Performance of Classifiers trained on full feature set. Results are obtained from 10-fold cross validation.

s/n	correlation-based	entropy-based	mRMR-MID	mRMR-MIQ
1	vCDR	vCDR	vCDR	vCDR
2	CupHMM	CupHMM	ms2	iopl
3	CupAreaMM	CupAreaMM	RNFL	ISNT
4	I	S	AlphaPPA	CupAreaMM
5	S	I	anisometropia	AlphaPPA
6	T	T	ISNT	glyn
7	N	ocular_htnl	smks_cat	age
8	RimAreaMM	ISNT	glyn	S
9	DiscAreaMM	iopl	DiscAreaMM	I
10	ocular_htnl	N	iopl	ocular_htnl
11	ISNT	RimAreaMM	T	RNFL
12	age	ocular_htn	age	CupHMM
13	iopl	agegp	CupAreaMM	T
14	AlphaPPA	DiscAreaMM	ocular_htnl	drlevel
15	ocular_htn	glyn	BestPPA	anisometropia
16	agegp	age	I	smks_cat
17	iopr	ocular_htnr	drlevel	ocular_htn
18	DiscHMM	AlphaPPA	S	BestPPA
19	ocular_htnr	RNFL	CupHMM	DiscAreaMM
20	glyn	DiscHMM	ocular_htnr	eyehist_re

Table 2: Top 20 features ranked via different feature ranking methods. Please refer to Appendix for explaination on feature symbols.

only 3 features are different (ms2 in mRMR-MID versus eyehist_re in mRMR-MIQ, age_Reg versus age, and ocular_htnr verses ocular_htn), and the last two pairs are very much related. Hight consistency is observed between the two methods.

4. Comparing mRMR-based FS methods with the none mRMR-based FS methods, mRMR-based methods are able to reduce feature redundancies: for example, a single age-related feature age (or age_rec) is picked up by mRMR-based methods, whereas both age (or age_rec) and agegp features are used in none mRMR-based methods. More importantly, mRMR-based methods are able to automatically calculate the relavance among the features as in the case of ISNT and vCDR. The ISNT relationship ($I \geq S \geq N \geq T$) is represented by 4 features (ISNT, S,I,T) in mRMR-based methods where as it is represented by 5 features (ISNT, S, I, T, N), which is redundant, in other classifiers. Similarly, the vCDR (vCDR = CupHMM/DiscHMM) is represented by 2 features (vCDR and CupHMM) in mRMR-based methods where as it is represented by 3 features (vCDR, DiscHMM and CupHMM), which is redundant.

5. Two important parameters related to glaucoma are ranked beyond the top 20 features by non-mRMR-based methods, which are smoking ('smks') and retinal fiber layer defect (`RNFL'). The reason can be that these features have less significant statistics comparing to those redundant features, and the effect on glaucoma of these features are co-factored with other features. For example, in the correlation-based test, p-value of `smks' feature is $p\text{-}value = 0.0067$, compare to the singificance level adjusted via Bonferroni correction, $\alpha = \alpha_0 / n = 0.05/650 = 7.7E\text{-}5$, the association with glaucoma is not statistically significant.

parameter group	feaure name	explanation
3*cup/disc related	vCDR	vertical Cup-Disc Ration
	CupHMM	cup hight in mm
	CupAreaMM	cup area in mm^2
4*ISNT rule related	I	rim inferior thickness
	S	rim superior thickness
	T	rim temporal thickness
	INST	following ISNT rule
2*Age related	agegp	age group
	age	age
IOP related	iopl / iopr	left eye IOP / right eye IOP
PPA related	AlphaPPA	Alpha PPA observed
Ocular hypertension related	ocular_htnr ocular_htnl	Ocular Hypertension right/ left eye

Table 3: Common Top Features identified by all 4 feature ranking methods.

A detail analysis of the common features identified by the 4 ranking methods are summarized in Table 3.

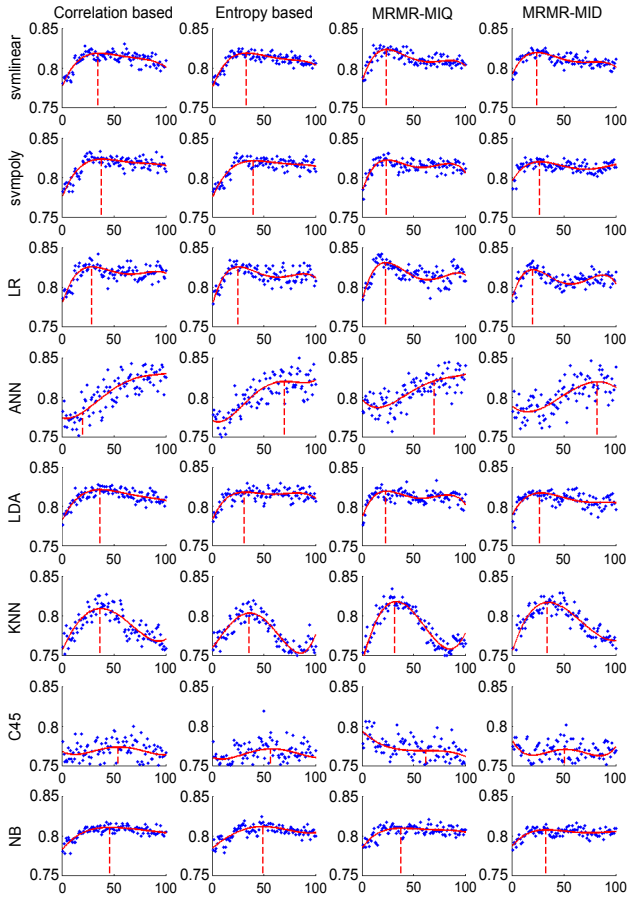

Figure 4: Accuracy curve with different classifier trained on incremental feature sets.

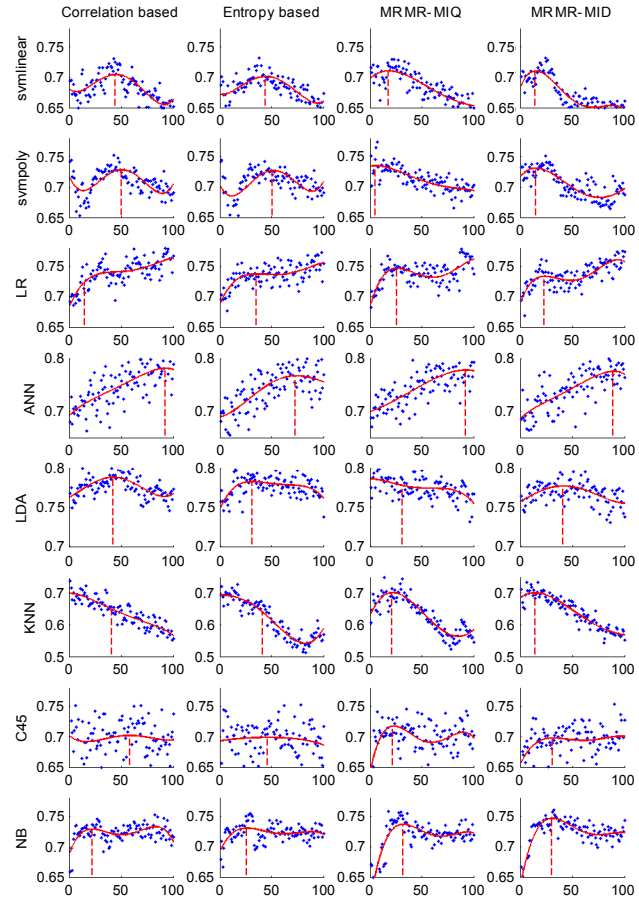

Figure 5: F-score curves with different classifier trained on incremental feature sets.

Search for optimal sub feature set

The optimal feature sets are obtained via curve fitting as illustrated above. Figure 4 and 5 illustrate the measurement scores of Accuracy and F-score respectively. From the raw scores, we employ 4th degree polynomial curve fitting for regression followed by first derivative and second derivative test to obtain their turning points, which are the optimal feature set sizes for the classifier. In the experiments, it is found that after excluding the very small numbers, the first turning point is the glaboal maximum. We use this observation to design classifier and without overfitting. Table 4 and 5 show the size of optimal feature sets with Accuracy and F-score measure on different classifiers.

We have the following observations:

1. mRMR-based ranking methods outperform non-mRMR based methods. For all cases listed in Table 4 and 5, mRMR based classifier find more compact feature set with better measurement scores.
2. The best classifier in terms of Accuracy measure is LR trained by top 23 features ranked by mRMR-MIQ, achiving an Accuracy of 0.83. Compared to the full-feature-ANN classifer with Accuracy of 0.825 (Table 1) and elapsed time 50.5 seconds, the elapsed time of the 23-feature-LR classifier is only 2.5 seconds.
3. The best classifier in terms of F-score measure is LR trained by top 23 features ranked mRMR-MIQ with an F-score of 0.774.
4. In Table 4, mRMR-MIQ outperforms other feature ranking methods in 4 outof 8 classifier, i.e., svm-linear, LR, LDA and KNN. Again in Table 5, mRMR-MIQ outperforms others in svm-linear, LR, LDA and KNN classifiers.
5. It is concluded that, with a compact feature set containing only about 1/4 of the original features, a simpler and faster machine learning method is able to achieve better performance.

Discussion and Conclusions

We explore feature selection and machine learning techniques to build a framework for automatic glaucoma diagnosis. We compose feature space from heterogeneous data sources, i.e., retinal image and eye screening data.

Under the proposed framework, we perform comprehensive studies on multiple feature ranking schemes and a wide range of supervised learners. The optimal feature set obtained using mRMR (minimum Redundancy Maximum Relevance) scheme contains only about 1/4 of the original features. The classifiers trained by the set are more efficient with overall better Accuracy and F-score. A detailed investigation on the features in the optimal set indicates that they can be the essential parameters for glaucoma mass screening and image segmentation.

Clinical Significance

In the ophthalmologic clinical practice, glaucoma diagnosis is based on evidences from multiple sources. Glaucoma specialists consider factors like patient's demographic data, medical history, vision measurement, IOP (Intra Ocular Pressure) as well as the assessment from various types of imaging equipment. Following the clinical decision making process, it is natural for us to design an automatic classifier being able to combine inputs from multiple heterogeneous data sources. However, the limitations of the black-box manner in the supervised learning classifiers offer little insight to the clinicians of how the classifier works, thus hinder the deployment of such systems.

The mRMR based feature ranking approach described in our paper, not only use an easily trained classification mechanism, but more importantly, present a clear list of ranked features to the clinicians. Comparing with Information theoritical based approach, features selected via mRMR provide a more balanced coverage of the feature space and capture broader characteristics of important information. This facilitates a better understanding of CAD process, and helps to explain what clinical features the classifier uses and how the system ranks the importance of the features in its prediction. With the confirmation from the clinicians, the features extracted may be used to guide the more efficient mass screening process in glaucoma early detection, leading to ease and reduced information to be collected. This

classifier	correlation-based		entropy-based		mRMR-MIQ		mRMR-MID		Accuracy on
	size	Accuracy	size	Accuracy	size	Accuracy	size	Accuracy	all features
svm-linear	35	0.818	33	0.818	**24**	**0.822**	24	0.819	0.809
svm-poly	**38**	**0.822**	40	0.820	24	0.821	27	0.819	0.808
LR	29	0.825	25	0.825	**23**	**0.831**	20	0.82162	0.821
ANN	29	0.789	70	0.819	23	0.788	82	0.819	**0.825**
LDA	37	0.821	31	0.816	**23**	**0.818**	27	0.816	0.811
KNN	37	0.808	36	0.803	**32**	**0.817**	34	0.816	0.758
C45	**54**	**0.773**	57	0.770	62	0.768	51	0.770	0.766
NB	46	0.810	**49**	**0.811**	38	0.809	33	0.807	0.805

Table 4: Optimal feature set and performance measured by Accuracy.

classifier	correlation-based		entropy-based		mRMR-MIQ		mRMR-MID		F-score on
	size	F-score	size	F-score	size	F-score	size	F-score	all features
SVM-linear	40	0.750	37	0.750	**22**	**0.757**	20	0.752	0.726
SVM-poly	**45**	**0.762**	47	0.759	22	0.761	23	0.758	0.737
LR	31	0.768	27	0.766	**23**	**0.774**	20	0.763	0.769
ANN	31	0.732	72	0.769	23	0.727	86	0.770	**0.777**
LDA	**37**	**0.772**	30	0.769	22	0.770	31	0.766	0.755
KNN	29	0.729	28	0.723	**29**	**0.747**	29	0.743	0.639
C45	55	0.707	56	0.703	**14**	**0.715**	49	0.702	0.714
NB	36	0.748	45	0.749	33	0.751	**31**	**0.753**	0.739

Table 5 : Optimal feature set and performance measured by F-score. Elapsed time is measured based on 10-fold cross validation.

can further reduce the cost and efforts for a mass screening program. The principle proposed in the paper can be applied to the automatic diagnosis of other eye diseases such as cataract and retinopathy, etc.

Authors Contributions

Z.Z contributes in experimental design and manuscript drafting; CK.K contributes in writing and method design; J.L contribute in analysis and manuscript writing; CYL.C, A.T and W.T.Y contribute in data source and manuscript writing.

Acknowledgements

This work was supported in part by the Singapore A*STAR SERC grand 092-148-0067 and Singapore MOE AcRF Grant MOE2008-T2-1-074. The authors thank SiMES group from Singapore Eye Research Institute, for the valuable dataset. The authors appreciate Liu Guimei's help and Adrianto Wirawan's help in usefull discussion.

References

1. Coleman AL (1999) Glaucoma. Lancet: 354: 1803-1810.
2. Weih LM, Nanjan M, McCarty CA, Taylor HR (2001) Prevalence and predictors of open-angle glaucoma: results from the visual impairment project. Ophthalmology 108: 1966-1972.
3. Quigley HA, Broman AT (2006) The number of people with glaucoma worldwide in 2010 and 2020. Br J Ophthalmol 90: 262-267.
4. Foong AW, Saw SM, Loo JL, Shen S, Loon SC, et al. (2007) Rationale and methodology for a population-based study of eye diseases in Malay people: The Singapore Malay eye study (SiMES). Ophthalmic Epidemiol 14: 25-35.
5. Michelson G, Warntges S, Hornegger J, Lausen B (2008) The papilla as screening parameter for early diagnosis of glaucoma. Dtsch Arztebl Int 105: 583-589.
6. Lin SC, Singh K, Jampel HD, Hodapp EA, Smith SD, et al. (2007) Optic nerve head and retinal nerve fiber layer analysis: a report by the American Academy of Ophthalmology. Ophthalmology 114: 1937-1949.
7. Malinovsky VE (1996) An overview of the Heidelberg Retina Tomograph. J Am Optom Assoc 67: 457-467.
8. Parikh RS, Parikh S, Sekhar GC, Kumar RS, Prabakaran S, et al. (2007) Diagnostic capability of optical coherence tomography (Stratus OCT 3) in early glaucoma. Ophthalmology 114: 2238-2243.
9. Liu J, Wong DWK, Lim JH, Li H, Tan NM, et al. (2009) ARGALI : An Automatic Cup-to-Disc Ratio Measurement System for Glaucoma Analysis Using Level-set Image Processing. 13th International Conference on Biomedical Engineering 23: 559-562.
10. Wong DW, Liu J, Lim JH, Tan NM, Zhang Z, et al. (2009) Intelligent fusion of cup-to-disc ratio determination methods for glaucoma detection in ARGALI. Conf Proc IEEE Eng Med Biol Soc 2009: 5777-5780.
11. Inoue N, Yanashima K, Magatani K, Kurihara T (2005) Development of a simple diagnostic method for the glaucoma using ocular Fundus pictures. Conf Proc IEEE Eng Med Biol Soc 4: 3355-3358.
12. Jonas JB, Bergua A, Schmitz-Valckenberg P, Papastathopoulos KI, Budde WM (2000) Ranking of optic disc variables for detection of glaucomatous optic nerve damage. Invest Ophthalmol Vis Sci 41: 1764-1773.
13. Jonas JB, Fernandez MC, Sturmer J (1993) Pattern of glaucomatous neuroretinal rim loss. Ophthalmology 100: 63-68.
14. Harizman N, Oliveira C, Chiang A, Tello C, Marmor M, et al. (2006) The ISNT rule and differentiation of normal from glaucomatous eyes. Arch Ophthalmol 124: 1579-1583.
15. Quigley HA, Katz J, Derick RJ, Gilbert D, Sommer A (1992) An evaluation of optic disc and nerve fiber layer examinations in monitoring progression of early glaucoma damage. Ophthalmology 99: 19-28.
16. Jonas JB (2005) Clinical implications of peripapillary atrophy in glaucoma. Curr Opin Ophthalmol 16: 84-88.
17. Joshi GD, Sivaswamy J, Krishnadas SR (2011) Optic Disk and Cup Segmentation From Monocular Color Retinal Images for Glaucoma Assessment. IEEE Trans Med Imaging 30: 1192-1205.
18. Wong D, Liu J, Tan N, Zhang Z, Yin F, et al. (2010) Automatic Detection of Peripapillary Atrophy in Digital Fundus Photographs. In Association for Vision Research and Ophthalmology Annual Meeting 2010, Fort Lauderdale, USA.
19. Tan NM, Liu J, Wong DWK, Zhang Z, Lu S, et al. (2010) Classification of Left and Right Eye Retinal Images. Medical Imaging : Computer - Aided Diagnosis 7624.
20. Jan J, Odstrcilik J, Gazarek J, Kolar R (2009) Retinal Image Analysis Aimed at Support of Early Neural-layer Deterioration Diagnosis. 9th International Conference on Information Technology and Applications in Biomedicine 506-509.
21. Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research 5: 1205-1224.
22. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artificial Intelligence 97: 273-324.
23. Liu H, Motoda H (2008) Computational Methods of Feature Selection. Data Mining and Knowledge Discovery Series, Chapman and Hall/CRC.
24. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27: 1226-1238.
25. Zhang Z, Yin FS, Liu J, Wong WK, Tan NM, et al. (2010) ORIGA(-light): an online retinal fundus image database for glaucoma analysis and research. Conf Proc IEEE Eng Med Biol Soc 2010: 3065-3068.
26. Amaldi E, Kann V (1998) On the approximation of minimizing non zero variables or unsatisifed relations in linear systems. Theoretical Computer Science 209: 237-260.
27. Hadley SW, Pelizzari C, Chen GTY (1996) Registration of Localization Images by Maximization of Mutual Information. In Proc Ann Meeting of the Am Assoc, Physicists in Medicine.

This article was originally published in a special issue, **Medical statistics: Clinical and experimental research** handled by Editor(s). Dr. Herbert Pang, Duke University, USA.