ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

On the analysis of feature selection techniques in a conjunctival hyperemia grading framework

L. S. Brea[1], N. Barreira[1], N. Sánchez[1], A. Mosquera[2], C. García-Resúa[3] and E. Yebra-Pimentel[3] *

1- Dept. Computer Science, Univ. of A Coruna, Spain
{luisa.brea,nbarreira,nsanchez}@udc.es

2- Dept. Electronics and Comp. Science, Univ. of Santiago de Compostela, Spain

3- Dept. Applied Physics, Univ. of Santiago de Compostela, Spain

Abstract. Hyperemia is a parameter that describes the degree of redness in a tissue. When it affects the bulbar conjunctiva, it can serve as an early indicator for pathologies such as dry eye syndrome. Hyperemia is measured using scales, which are collections of images that show different severity levels. Features computed from the images can be used to develop an automatic grading system with the help of machine learning algorithms. In this work, we present a methodology that analyses the influence of each feature when determining the hyperemia level.

1 Introduction

Hyperemia is the occurrence of redness in a tissue. It appears when blood vessels are engorged, and it is an early symptom of several pathologies. When the affected tissue is the conjunctiva, it is frequently related with allergies, conjunctivitis, or dry eye syndrome. Hyperemia is measured as a degree in a scale. There are several scales available for specialists, such as Efron and CCLRU grading scales. Both scales consist in a set of photos or drawings that represent levels of severity. The specialist compare these images to the patient's eyes and assigns a value in the scale with respect to the most similar picture.

This time-consuming process presents a high level of intra and inter expert subjectivity. Those drawbacks can be solved if we develop an automatic methodology, yet there are few approaches proposed on the subject, and they are not completely automatic [1, 2]. There are several works regarding some of the steps involved in the measurement, specially the creation of grading scales and the automatic computation of image features [3, 4]. However, there are few attempts of one of the most important steps: perform comparisons of features and define which ones are the most relevant. In this work, we propose a methodology to solve this issue, determining which features are the most significant in relation to hyperemia grading.

*This work has been supported by the Secretaría de Estado de Investigación of the Spanish Government (Grant TIN2015-65069-C2-1-R).

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

The paper is structured as follows: Section 2 will depict the feature selection approach. Section 3 will explain performed experiments and will show the obtained results and, lastly, Section 4 will present the conclusions.

2 Feature selection

Hyperemia grading depends on several features of the image. Specialists take into account the general hue of the conjunctiva, and also the vessel tonality. A red or yellow colouration can both be hints of hyperemia, while the whiter the conjunctiva, the lower the level. The disposition of the vessels and its width is also relevant, as hyperemia is produced by vessel engorgement. Several width vessels, or a large number of thinner ones, usually imply a higher hyperemia level. The features we consider were calculated using different colourspaces, in order to find which one reflects better the expert perception. We employ 25 features, some of them proposed by earlier works [5] and some suggested by optometrists. Three of the features are related with vessel quantity, one measures vessel width, four study the vessels colour, nine compute different hues in the conjunctiva, and eight measure the colouration of the full image. Once the features are computed, we need to transform them to the grading scale ranges by means of machine learning algorithms [6]. In this work, we analyse different feature combinations in order to select the most influential ones.

Once several features have been computed for each image, it is expected that some of them will be related, as they refer to similar characteristics (for example, red level in different colourspaces). This arises the need to determine which features provide the most useful information. Feature selection methods examine the original set of features in order to obtain a smaller subset that preserves most of the information. To that end, they use a certain criteria (information gain, correlation) to decide if each factor is worth including or not. We can distinguish three groups of feature selection techniques: filters, wrappers and embedded methods. Wrappers evaluate each feature subset by building a predictive model and looking at the accuracy using these features while filters can directly compute a statistic on the given feature subset. Hence, wrappers are slower than filters, but also to provide better results [7]. Finally, embedded methods blend the feature selection with the training process of the prediction model, offering more accurate results than filters but at a higher computational cost.

In this work, the three approaches were tested. On one hand, we used Correlation based Feature Selection (CFS) [8], which is a filter method and, therefore, independent from the learning method. It was originally designed to be used in classification problems, so a first discretisation stage is needed in order to transform the data. To that end, the used algorithm is MDL [9]. CFS returns as output a subset containing the relevant features. Moreover, we tested Relief

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

[10], another filter method. Instead of constructing a subset of features, it orders them in a ranking. It returns the whole list of features, but sorted by relevance. If a cross-validation technique is used, the filter could return a different sorted list for each fold, and it will be necessary to define a threshold from which create the different subsets. On the other hand, we employed a wrapper that uses M5 algorithm [11]. It generates a decision list for regression problems using the separate-and-conquer technique. In each iteration it builds a model tree using M5 and makes the best leaf into a rule. Search strategy is best first. Also, we implemented a wrapper that uses the Support Vector Machine for regression. The algorithm uses the improvements proposed in [12] for the Sequential Minimal Optimisation (SMO) method. Finally, we implemented an embedded method that uses Recursive Feature Elimination with Support Vector Regression [13]. It starts with the full set of features and assigns them weights. The features whose absolute weights are the smallest are removed and the process continues iterating until the minimum number of features, previously established, is reached.

3 Results

We performed experiments with 105 frames extracted from hyperemia videos provided by the Optometry Group of the Department of Applied Physics (University of Santiago de Compostela). These frames were labelled by two experts and the average gradings were used as outputs. The features were computed for every frame, each one of them providing values in different continuous ranges. Then, these values were used to train a system using a 10-fold cross validation. The feature selection experiments were performed using the data mining software Weka [14]. Then, the classifiers were trained and tested again using only the selected features.

We used a 10-fold cross validation when computing the features. Other cross-validation techniques, such as leave-one-out, were also tested with similar results. We maintain the same subsets for all the tests in order to compare the results. We can observe in Figures 1 and 2 the feature selection by folds, where F_n represents each of the 25 features, and the radius represents the number of folds where the feature is selected. The normalised values were computed by applying the feature selection technique after transforming the values within the $[-1, 1]$ range.

Filter methods are the only ones affected by normalisation. Even so, the features that appear in a higher number of folds ended up being almost the same for the two options. There are some features that are commonly chosen by all methods in most folds, such as F_{14} or F_{21}. However, others such as F_{10} and F_{15} appear frequently only in one case (CCLRU scale with Relief).

Filter approaches provide similar results in all ten folds, which leads to more confidence when picking the subset of features for the whole image set. Wrappers

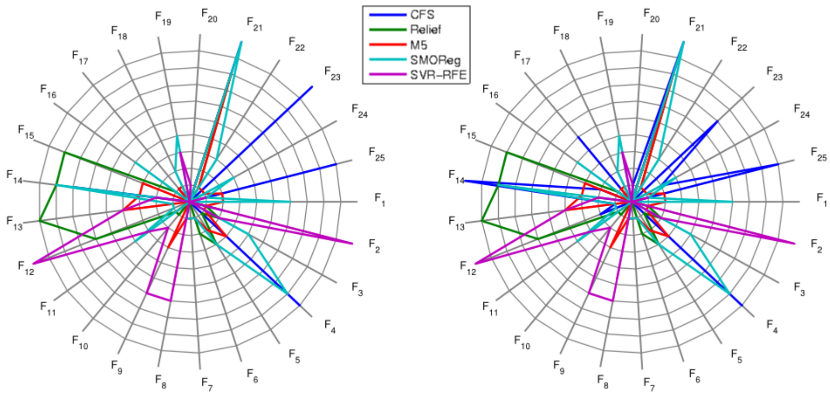

Fig. 1: Efron data. Left: raw values. Right: normalised values.

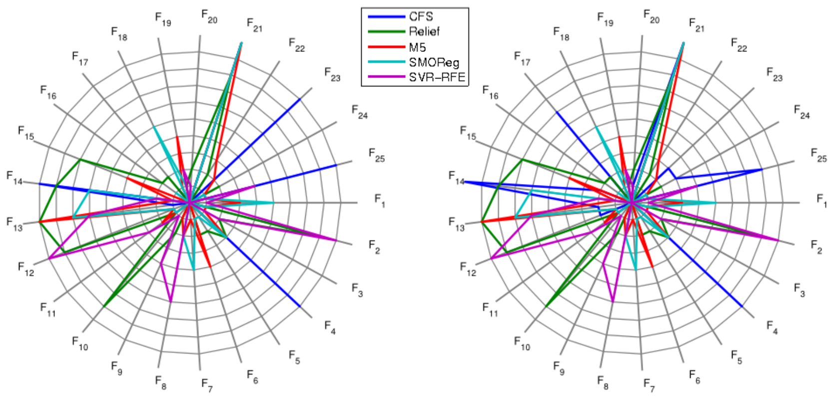

Fig. 2: CCLRU data. Left: raw values. Right: normalised values.

results present a higher variability between folds because they perform an additional 5-fold cross validation when determining the features. Also, they usually obtain smaller sets because of the search strategy, which starts with an empty set and then adds features one by one until the accuracy is not improved any further. The aimed number of features for the embedded method was set to 5. Tests were performed with higher subsets, but they presented the same high variability.

Since we obtain a set of best features for each fold, we need to create a set combining these results. There are several possibilities depending on how many features are selected on average and how much information is lost by removing features. For the ranker method, we established a minimum value averaging the maximum relevance of the ten folds and considered only the features which ranked higher than half that value. In view of the results, we decided to choose

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

those features that have been selected in at least 7 out of 10 folds (Table 1). This ensures that selected features are relevant enough, as they are chosen in most of the folds, and provides better results than choosing those that appear in a higher number of folds, as the condition will be too restrictive and subsets will be too small.

CFS is the only method that provides a different subset for normalised and not normalised values. There is only one different feature (17 instead of 23 in CCLRU), and both features measure a similar concept. We chose to keep them both in the final tests, as the resulting subset is still small.

Table 1: Features that appear in at least 7 out of 10 folds.

Method	Efron		CCLRU	
	#	selected features	#	selected features
CFS	5	4, 14, 21, 23, 25	5	4, 14, 21, 23, 25
CFS (norm.)	5	4, 14, 21, 23, 25	5	4, 14, 17, 21, 25
Relief	4	13, 14, 15, 21	7	2, 10, 12, 13, 14, 15, 21
Relief (norm.)	4	13, 14, 15, 21	7	2, 10, 12, 13, 14, 15, 21
M5	1	21	3	8, 13, 21
SMOReg	3	4, 14, 21	2	14, 21
SVR-RFE	2	2, 12	2	2, 12

Three classifiers were trained with the features, selected to cover different approaches: Multi-Layer Perceptron (MLP), Decision Trees (DT), and Naive-Bayes (NB). Other approaches, such as SVM, were tested but they did not improve the results. Results are depicted in Table 2. We can observe how, in most cases, the mean squared error remains in a similar range as the one from the complete feature set. In some cases, this value improves, so that some of the features were adding noise to the system instead of providing useful information. There are exceptions, the more relevant one is M5 method for Efron, that worsens the results in MLP and DT. This happens because only one feature was chosen for this case, and it does not provide enough information on its own.

Table 2: Comparison of mean square error values.

Method	MLP		DT		NB	
	Efron	CCLRU	Efron	CCLRU	Efron	CCLRU
All	0.21787	0.13740	0.18621	0.12718	0.50636	0.34273
Relief	0.21802	0.13747	0.18762	0.13754	0.44364	0.21000
CFS	0.10770	0.13716	0.23290	0.11966	0.48364	0.39727
M5	0.86141	0.06083	0.19468	0.12129	0.45909	0.24364
SMOReg	0.11443	0.07507	0.19271	0.11590	0.48364	0.17091
SVR-RFE	0.12839	0.08192	0.16621	0.11648	0.45000	0.31364

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

4 Conclusions

Hyperemia is an early indicator of several pathologies, making it necessary to perform a prompt and accurate evaluation of the patient. However, the current manual process is tedious and subjective, hence the need of implementing an automatic grading methodology. In this work, we tackle one of the steps of such methodology: the evaluation of the different combinations of features of the image, determining which are the most relevant by means of feature selection techniques. We are able to reduce our feature set from 25 to 2-3 values. In the Efron scale, the MLP classifier achieves the lowest error when using the features selected with the CFS approach and on the CCLRU scale, the MLP achieves the lowest error when using the features selected with the M5 wrapper.

Our next objective is to integrate these results in a framework for the automatisation of hyperemia grading.

References

[1] JD Rodriguez, PR Johnston, GW Ousler III, LM Smith, and MB Abelson. Automated grading system for evaluation of ocular redness associated with dry eye. *Clinical ophthalmology (Auckland, NZ)*, 7:1197, 2013.

[2] T Yoneda, T Sumi, A Takahashi, Y Hoshikawa, M Kobayashi, and A Fukushima. Automated hyperemia analysis software: reliability and reproducibility in healthy subjects. *Japanese journal of ophthalmology*, 56(1):1–7, 2012.

[3] PJ Murphy, JSC Lau, MML Sim, and RL Woods. How red is a white eye? clinical grading of normal conjunctival hyperaemia. *Eye*, 21(5):633–638, 2007.

[4] JS Wolffsohn. Incremental nature of anterior eye grading scales determined by objective image analysis. *British journal of ophthalmology*, 88(11):1434–1438, 2004.

[5] EB Papas. Key factors in the subjective and objective assessment of conjunctival erythema. *Investigative ophthalmology & visual science*, 41(3):687–691, 2000.

[6] L Sánchez, N Barreira, H Pena-Verdeal, and E Yebra-Pimentel. A novel framework for hyperemia grading based on artificial neural networks. *LNCS: Advances in Computational Intelligence (IWANN 2015)*, 2015.

[7] S Das. Filters, wrappers and a boosting-based hybrid for feature selection. In *ICML*, volume 1, pages 74–81, 2001.

[8] MA Hall and LA Smith. Practical feature subset selection for machine learning. In *21st Australian Computer Science Conference*, pages 181–191, 1998.

[9] KB Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *13th International Joint Conference on AI*, pages 1022–1029, 1993.

[10] M Robnik-Šikonja and I Kononenko. An adaptation of relief for attribute estimation in regression. In *Machine Learning: Proc. 14th Int. Conf.*, pages 296–304, 1997.

[11] JR Quinlan et al. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348, 1992.

[12] SK Shevade, SS Keerthi, C Bhattacharyya, and KRK Murthy. Improvements to the smo algorithm for svm regression. *Neural Networks, IEEE Trans. on*, 11(5):1188–1193, 2000.

[13] I Guyon, J Weston, S Barnhill, and V Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[14] M Hall, E Frank, G Holmes, B Pfahringer, P Reutemann, and IH Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.