Multi-criteria Evaluation of Class Binarization and Feature Selection in Tear Film Lipid Layer Classification

Rebeca Méndez, Beatriz Remeseiro, Diego Peteiro-Barral and Manuel G. Penedo
Departamento de Computación, Universidade da Coruña, Campus de Elviña s/n, 15071, A Coruña, Spain

Keywords: Tear Film Lipid Layer, Class Binarization Techniques, Feature Selection, Filters, Multiple Criteria Decision Making, Multilayer Perceptron.

Abstract: Dry eye is an increasingly popular syndrome in modern society which can be diagnosed through an automatic technique for tear film lipid layer classification. Previous studies related to this multi-class problem lack of analysis focus on class binarization techniques, feature selection and artificial neural networks. Also, all of them just use the accuracy of the machine learning algorithms as performance measure. This paper presents a methodology to evaluate different performance measures over these unexplored areas using the multiple criteria decision making method called TOPSIS. The results obtained demonstrate the effectiveness of the methodology proposed in this research.

1 INTRODUCTION

The tear film is a complex layer of liquid covering the anterior surface of the eye. It was classically defined by Wolff (E.Wolff, 1954) as a three-layered structure which consists of an anterior lipid layer, an intermediate aqueous layer and a deep mucin layer. The tear film is a essential component of the eye which plays some important functions (Korb, 2002), such as visual and cleaning functions. Also, it plays an essential role in the maintenance of ocular integrity by removing foreign bodies from the front surface of the eye.

The lipid layer is the outermost and thinnest layer of the tear film and it is mainly secreted by the meibomian glands (Nichols et al., 2004). It is a crucial component of the tear film because it provides a smooth optical surface for the cornea and retards evaporation of the eye during the inter-blink period (Bron et al., 2004). Other functions of the lipid layer are establishing the tear film or sealing the lid margins during sleep.

Quantitative or qualitative changes in the normal lipid layer have a negative effect on the evaporation of tears from the ocular surface and on the quality of vision (Rolando et al., 1998). In fact, these changes are associated with the *evaporative dry eye* (EDE), since it refers to disorders of the tear film caused by poor tear quality, reduced tear production or excessive evaporation (Rolando et al., 1983). The international committee of Dry Eye Workshop (DEWS) defined the EDE as follows (Lemp et al., 2007):

> Dry Eye is a multifactorial disease of the tears and the ocular surface that results in symptoms of discomfort, visual disturbance, and tear film instability with potential damage to the ocular surface. It is accompanied by increases osmolarity of the tear film inflammation of the ocular surface

This disease affects a wide sector of the population, specially among contact lens users, and worsens with age. The proportion of people with EDE has increased due to the current work conditions (Lemp et al., 2007), such as computer use.

EDE diagnosis is very difficult to accomplish, basically because of its multifactorial nature. There are several clinical tests which measures the tear quality and the quantity of tears. One of these test is called *lipid layer pattern assessment* and consists on evaluating tear film quality and lipid layer thickness by non-invasively imaging the superficial lipid layer by interferometry. This test is based on a standard classification defined by Guillon (Guillon, 1998), who established various categories of lipid layer patterns: open meshwork, closed meshwork, wave and color fringe. Note that EDE is associated with the lipid layer thickness since a thinner lipid layer speeds up water evaporation, which means a reduction in tear film stability. Many eye care professionals have abandoned this test because it is very difficult to inter-

Méndez R., Remeseiro B., Peteiro-Barral D. and G. Penedo M..
Multi-criteria Evaluation of Class Binarization and Feature Selection in Tear Film Lipid Layer Classification.
DOI: 10.5220/0004224300620070
In *Proceedings of the 5th International Conference on Agents and Artificial Intelligence* (ICAART-2013), pages 62-70
ISBN: 978-989-8565-39-6

pret the lipid layer patterns, specially the thinner ones which lack color and/or morphological features. Nevertheless, there is no doubt that this technique is a valuable test which provides relevant information by using noninvasive techniques. For this reason, the tear film lipid layer automatic classification could become a key step to diagnose EDE.

Some techniques have been designed to objectively calculate the lipid layer thickness by analyzing the interference color with an interference camera (Goto et al., 2003) or by using a sophisticated optic system (King-Smith et al., 1999). However, first attempts to automatize the *lipid layer pattern assessment* test can be found in (Calvo et al., 2010; Ramos et al., 2011; García-Resúa et al., 2012) which demonstrate how the interference phenomena can be characterized as a color texture pattern. Therefore, the automatic test can save time for experts and eliminate the subjectivity of the process. Further investigation was carried out in (Remeseiro et al., 2011) where a set of color texture analysis techniques was applied to tear film lipid layer classification and the previous results were improved. Regarding machine learning techniques, the behavior of five different algorithms was studied over this set of color texture analysis methods in (Remeseiro et al., 2012). A statistical comparison of them was performed using only the accuracy of the classifiers.

To the best knowledge of the authors, there are no attempts in the literature to study this multi-class problem using class binarization techniques. Class binarization techniques may improve performance on multi-class problems of learners which could directly handle multi-class classification (Furnkranz, 2002a; Rifkin and Klautau, 2004). Furthermore, all previous researches analyses the color texture characterization based on the accuracy of the classifiers, no other performance measures were studied. In relation to machine learning techniques, there is no deep study about the performance of *artificial neural networks* (ANNs). Finally, the number of features which define the color texture pattern used to characterize the interference phenomena is large enough to consider the use of feature selection techniques.

In this sense, there are a lot of unexplored areas of study in tear film lipid layer automatic classification. Thus, a research methodology is proposed in this work to analyze the performance of class binarization techniques and feature selection methods applied to tear film classification using ANNs. For this purpose, the obtained results will be analyzed in terms of a wide set of performance measures and a multiple criteria decision making method will be used in order to validate the different approaches.

This paper is organized as follows: section 2 describes the steps of the research methodology, section 3 explains the experimental study performed, section 4 shows the results and discussion, and section 5 includes the conclusions and future lines of research.

2 RESEARCH METHODOLOGY

The methodology proposed in this search aims to evaluate tear film lipid layer classification in terms of several criteria when using class binarization techniques and feature selection methods.

2.1 Class Binarization Techniques

Methods can be roughly divided between two different approaches—the "single machine" approaches, which construct a multi-class classifier by solving a single optimization problem, and the "error correcting" approaches, which use the ideas from error correcting coding theory to combine a set of binary classifiers (Rifkin and Klautau, 2004). There exist several techniques for turning multi-class problems into a set of binary problems (Dietterich and Bakiri, 1995; Crammer and Singer, 2002; Furnkranz, 2002b; Hsu and Lin, 2002). A class binarization is a mapping of a multi-class learning problem to several two-class learning problems in a way that allows a sensible decoding of the prediction (Furnkranz, 2002b).

- The "*one-vs-all*" strategy consists in constructing one classifier per class, which is trained to distinguish the samples of one class from the samples of all remaining classes. These two-class problems are constructed by using the examples of class i as the positive examples and the examples of the rest of the classes as the negative examples.
- The "*one-vs-one*" strategy consists in training one classifier for each pair of classes. Thus, for a problem with c classes, $\frac{c(c-1)}{2}$ subproblems are constructed to distinguish the samples of one class from the samples of another class. The binary classifier for a problem is trained with examples of its corresponding classes i, j, whereas examples of the rest of classes are ignored for this problem.

2.1.1 Decoding Methods

If the classifiers are soft, as is the case of ANNs, they compute the "likelihood" of classes for a given input, that is they obtain a confidence p for the *positive* class and a confidence of $1-p$ for the *negative* class. The decoding method in the *one-vs-all* technique, if we assume the *one*-part as the positive class and the *all*-part

as the negative class, is simply done according to the maximum probability p among classes. However, this method is not appropriate for *one-vs-one* binarization techniques. Therefore, several decoding methods for *one-vs-one* binarization techniques are described as follows,

- *Hamming Decoding*. Dietterich and Bakiri (Dietterich and Bakiri, 1995) suggested the use of a matrix $M \in \{-1,1\}^{N\times F}$, where N is the number of classes and F is the number of binary classifiers. The i-th row of the matrix induces a partition of the classes into two "metaclasses", where a sample x_i is placed in the positive metaclass for the j-th classifier if and only if $M_{y_i j} = 1$ (Rifkin and Klautau, 2004), where y_i stands for the desired class of sample x_i. If a new sample appears for classification, the Hamming distance between the sign of the output of every binary classifier $f_1(x),\ldots,f_F(x)$ and each row of the matrix M is then compared as follows, choosing the minimizer,

$$f(x) = arg \min_{r=1\ldots N} \sum_{i=1}^{F} \left(\frac{1 - sign(M_{ri} f_i(x))}{2} \right)$$

  where $sign(z) = +1$ if $z > 0$, $sign(z) = -1$ if $z < 0$, and $sign(z) = 0$ if $z = 0$. In (Allwein et al., 2001), Allwein, Schapire and Singer extended the earlier work of Dietterich and Bakiri. They chose the matrix $M \in \{-1,0,1\}^{N\times F}$, rather than only allowing -1 and 1 as entries in the matrix. If $M_{y_i j} = 0$, then example x_i is not used when the j-th classifier is trained.

- *Loss-based Decoding*. The major disadvantage of Hamming decoding is that it ignores the significance of the predictions, which can be interpreted as a measure of confidence. If the classifiers are soft, in (Allwein et al., 2001) the authors suggest using the loss function L instead of the Hamming distance. They proposed that the prediction for a sample x should be the class n that minimizes the total loss under the assumptions that the label for sample x in the f-th binary classifier is M_{nf}:

$$f(x) = arg \min_{r=1\ldots N} \sum_{i=1}^{F} L(M_{ri} f_i(x))$$

  The loss function depends on the learning algorithm. In this research, the most appropriate loss function is the logistic regression $L(z) = log(1 + e^{-2z})$ (Allwein et al., 2001).

- *Accumulative probability with threshold*. If the classifiers obtain a confidence p for the *positive* class and a confidence of $1-p$ for the *negative* class, the accumulative probability for every class is computed as the sum of their corresponding probabilities p. The prediction for a sample should be the class that maximizes the accumulative sum. The accumulative probability with threshold takes into consideration binary classifiers that will be ignored if the difference between p and $1-p$ is under a threshold ε. It is assumed that ignored classifiers will correspond with class samples not used for their training procedure. In other words, only *significant* positive or negative probabilities will be considered.

2.2 Feature Selection

Feature selection is a dimensionality reduction technique aimed at detecting relevant features and discarding irrelevant ones, with the goal of obtaining a subset of features that describes properly the given problem with minimum degradation of performance (Guyon et al., 2006). Thus, feature selection is helpful in reducing the computational effort, allocated memory and training time.

There exists three different models for feature selection: filter, wrapper and embedded methods. Wrappers use a prediction method to score subsets of features. Filters rely on the general characteristics of the training data to select features with independence of the classifier. Halfway these two models, embedded methods perform feature selection as part of the training process of the classifier. It is well-known that wrappers and embedded methods have the risk of overfitting when having more features than samples (Loughrey and Cunningham, 2005), as it is the case in this research. Therefore, filters were chosen because they prevent the risk of overfitting and also allow for reducing the dimensionality of the data without compromising time and memory requirements of learning algorithms.

The three filters used in this work will be described as follows. They were selected based on previous researches (Bolón-Canedo et al., 2011; Bolón-Canedo et al., 2011).

- *Correlation-based Feature selection* (CFS) is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function (Hall, 1999). The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class. Redundant features should be screened out as they will be highly correlated

with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features. CFS's feature subset evaluation function is defined as,

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}$$

where M_S is the heuristic "merit" of a feature subset S containing k features, $\overline{r_{cf}}$ is the mean feature-class correlation ($f \in S$) and $\overline{r_{ff}}$ is the average feature-feature intercorrelation. The numerator of this equation can be thought of as providing an indication of how predictive of the class a set of features is; and the denominator of how much redundancy there is among the features.

- *Consistency-based filter* (Dash and Liu, 2003) evaluates the worth of a subset of features by the level of consistency in the class values when the training instances are projected onto the subset of attributes. The algorithm generates a random subset S from the number of features in every round. If the number of features of S is less than the current best,the data with the features prescribed in S is checked against the inconsistency criterion. If its inconsistency rate is below a pre-specified one, S becomes the new current best. The inconsistency criterion, which is the key to the success of this algorithm, specifies to what extent the dimensionally reduced data can be accepted. If the inconsistency rate of the data described by the selected features is smaller than a pre-specified rate, it means the dimensionally reduced data is acceptable.
- *INTERACT* (Zhao and Liu, 2007) is a subset filter based on symmetrical uncertainty (SU) (Press et al., 1986), which is defined as the ratio between the information gain (IG) and the entropy (H) of two features, x and y:

$$SU(x,y) = \frac{2IG(x|y)}{H(x)+H(y)}$$

where the information gain is defined as:

$$IG(x|y) = H(y)+H(x)-H(x,y)$$

being $H(x)$ the entropy and $H(x,y)$ the joint entropy. Besides SU, INTERACT also includes the consistency contribution (c-contribution). C-contribution of a feature is an indicator about how significantly the elimination of that feature will affect consistency. The algorithm consists of two major parts. In the first part, the features are ranked in descending order based on their SU values. In the second part, features are evaluated one by one starting from the end of the ranked feature list. If c-contribution of a feature is less than an established threshold, the feature is removed, otherwise it is selected. The authors stated in (Zhao and Liu, 2007) that INTERACT can thus handle feature interaction, and efficiently selects relevant features.

2.3 Multiple-criteria Decision-making

Classification algorithms are normally evaluated in terms of multiple criteria such as accuracy, precision or training time. Thus, algorithm selection can be modeled as a multiple-criteria decision-making (MCDM) problem. MCDM methods evaluate classifiers from different aspects and produce rankings of classifiers (Kou et al., 2012). Among many MCDM methods that have been developed up to now, *technique for order of preference by similarity to ideal solution* (TOPSIS) (Hwang and Yoon, 1981) is a well-known method that will be used in this research.

2.3.1 TOPSIS

TOPSIS is a MCDM method proposed by Hwang and Yoon in 1981 (Hwang and Yoon, 1981). It finds the best algorithms by minimizing the distance to the ideal solution whilst maximizing the distance to the anti-ideal one. The extension of TOPSIS proposed by Opricovic and Tzeng (Opricovic and Tzeng, 2004) and Olson (Olson, 2004) is used in this research,

1. Compute the decision matrix consisting of m alternatives and n criteria. For alternative A_i, $i = 1,\ldots,m$, the performance measure of the j-th criterion C_j, $j = 1,\ldots,n$, is represented by x_{ij}.
2. Compute the normalized decision matrix. The normalized value r_{ij} is calculated as,

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}$$

3. Develop a set of weights w, where w_j is the weight of the j-th criterion and $\sum_{j=1}^{n} w_j = 1$, and compute the weighted normalized decision matrix. The weighted normalized value v_{ij} is computed as,

$$v_{ij} = x_{ij}w_j$$

4. Find the ideal alternative solution S^+ and the anti-ideal alternative solution S^-, which are computed as,

$$S^+ = \{v_1^+, \ldots, v_n^+\} = \\ = \left\{ \left(\max_i v_{ij} | i \in I' \right), \left(\min_i v_{ij} | i \in I'' \right) \right\}$$

and

$$S^- = \{v_1^-, \ldots, v_n^-\} = \\ = \left\{ \left(\min_i v_{ij} | i \in I' \right), \left(\max_i v_{ij} | i \in I'' \right) \right\}$$

respectively, where I' is associated with benefit criteria and I'' is associated with cost criteria.

5. Compute the distance of each alternative from the ideal solution and from the anti-ideal solution, using the Euclidean distance,

$$D_i^+ = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^+)^2}$$

and

$$D_i^- = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^-)^2}$$

respectively.

6. Compute the ratio R_i^+ equal to the relative closeness to the ideal solution,

$$R_i^+ = \frac{D_i^-}{D_i^+ + D_i^-}$$

7. Rank alternatives by maximizing the ratio R_i^+.

3 EXPERIMENTAL STUDY

The aim of this research is to evaluate the influence of binarization and feature selection in tear film lipid layer classification. The multilayer perceptron (MLP) was selected as base learning algorithm.

3.1 Data Source

The methodology proposed in this research has been tested on the VOPTICAL-I1 dataset (Remeseiro, 2012). This set includes 105 images categorized by optometrists from the School of Optics and Optometry of the University of Santiago de Compostela (Spain). All these images were acquired from healthy subjects aged from 19 to 33 years. The dataset includes 29 open meshwork, 29 closed meshwork, 25 wave and 22 color fringe images. In (Remeseiro et al., 2011), it was demonstrated that the interference phenomena can be characterized as a color texture pattern and the automatic classification into Guillon categories is feasible. The results presented by Remeseiro et al. (Remeseiro et al., 2011) show how co-occurrence features (Haralick et al., 1973), as a feature extraction method, and the Lab color space (McLaren, 1976) provide the highest discriminative power from a wide range of methods analyzed. From a single image, a quantitative vector composed of 588 features is obtained to categorize it.

3.2 Performance Measures

Most performance measures in machine learning are defined to be used in two-class problems. Since a multi-class problem is studied in this research, all these measures will be calculated for each class individually. As tear film lipid layer classification is a 4-class problem, the total number of measures would be four times the number of binary measures. In order to reduce the total amount of measures, each multi-class measure will be obtained as the minimum of its four binary measures according to (Fernandez Caballero et al., 2010). Thus, the performance of the learning algorithms are computed as a lower bound, or pessimistic, estimation.

The binary performance measures considered are: accuracy, true positive rate (TPR), true negative rate (TNR), precision, F-measure and area under the curve (AUC).

Finally, the training time of the learning algorithms is also considered. It comprises the time elapsed for training a learning model. Notice that this comprises training a set of classifiers when class binarization techniques are used. Note also that the testing time, that is the time elapsed for outputting a new classification, is negligible thus it will not be considered as a selection criterion.

3.3 Experimental Procedure

A leave-one-out cross-validation was used, which consists in using a single sample from the dataset as the test set and the remaining samples are retained as the training set. This process is repeated such that each sample is used once as the test set. The experimental research was carried out as follows,

1. Apply the three feature selection methods (CFS, consistency-based and INTERACT) to the VOPTICAL-I1 dataset, to provide the subset of features that properly describes the given problem. Note that the binarization techniques modify the output of the dataset thus the feature selection

methods have to be applied on *each* "dataset", that is,

- In the *one-vs-all* technique, four subsets of features are obtained corresponding with *1-vs-all*, *2-vs-all*, *3-vs-all* and *4-vs-all* datasets.
- In the *one-vs-one* technique, six subsets of features are obtained corresponding with *1-vs-2*, *1-vs-3*, *1-vs-4*, *2-vs-3*, *2-vs-4* and *3-vs-4* datasets.

2. Train a MLP for each combination of binarization technique, feature selection method, and number of hidden units. In (Hecht-Nielsen, 1990), it was demonstrated that a MLP that contains a single hidden layer with sufficient number of hidden units is able to approximate any function. Thus, only the number of hidden units will vary in this research ranging from 2 to 64. In particular, 2, 4, 8, 16, 32, and 64 hidden units were tested. Empirical results showed risk of overfitting for a larger number of hidden units. Finally, the mean square error was used as error function and the hyperbolic tangent sigmoid was used as transfer function in the processing units.
3. Compute the performance measures, that is, accuracy, TPR, TNR, precision, F-measure, AUC and training time.
4. Apply TOPSIS in order to evaluate the different binarization techniques, feature selection methods and number of hidden units proposed in this research. The values of the weights (see Section 2.3.1) are assigned equally, except for the training time that is reduced to 0.01. Notice that the training step is executed off-line, making its value not as relevant as the other performance measures. Note also that the training time is a cost criteria while the other measures are benefit criteria.

Experimentation was performed on an Intel© Core™ i5-650 CPU @ 4M Cache, 3.20 GHz with RAM 6 GB DDR3. Matlab was the software used to train the MLP networks.

4 RESULTS

Table 1 shows the number of features selected by the three feature selection filters (CFS, consistency-based, and INTERACT) in single machine, one-vs-all, and one-vs-one approaches. The median percentage of features selected (out of 588 features) is in parenthesis.

Broadly speaking, consistency-based filter performed the most aggressive selection retaining only the 1.02%, 0.59%, and 0.43% of the features in single machine, one-vs-all, and one-vs-one approaches, respectively. CFS retained from four to ten times more features (4.59%, 3.74%, and 4.34%) than the former. Halfway, INTERACT selected in average 3.57%, 2.38%, and 2.38% of the features, respectively. As expected, in average the percentage of features selected in the single machine approach is larger than the percentage in binarization. Notice that binarization may reduce the complexity of the problem.

Table 1: Number of features selected by the three filters in single machine, one-vs-all, and one-vs-one approaches (median percentage in parentheses).

Technique		Feature selection		
		CFS	Cons	INT
Single		27	6	21
	Median(%)	(4.59%)	(1.02%)	(3.57%)
One-vs-all	1-vs-all	17	2	14
	2-vs-all	27	6	17
	3-vs-all	11	3	14
	4-vs-all	33	4	14
	Median(%)	(3.74%)	(0.59%)	(2.38%)
One-vs-one	1-vs-2	20	2	12
	1-vs-3	53	1	53
	1-vs-4	23	1	23
	2-vs-3	27	3	14
	2-vs-4	24	3	14
	3-vs-4	27	4	13
	Median(%)	(4.34%)	(0.43%)	(2.38%)

The set of techniques, methods and topologies used in this research lead to 120 alternatives in total. Thus, for purposes of simplicity only the most significant results are shown. Table 2 shows the top 20 results ranked by TOPSIS in terms of the binarization method, feature selection filter, number of hidden units (H), ratio R^+ (see TOPSIS, Section 2.3.1), accuracy, TPR, TNR, precision, F-measure, AUC and training time (in seconds). Note that *single* stands for the single machine, multi-class, approach.

In general, the techniques and methods proposed in this research outperform the single machine approach (see Table 2). In the top 20, 15 out of 20 classifiers use binarization, and 9 out of 20 classifiers apply feature selection. Moreover, binarization leads to smaller topologies in the MLP. In the top 20, the average number of hidden units is 29.47 in binarization against 51.20 in the single machine approach. In a similar way, the average number of hidden units is 35.77 when feature selection is applied whilst it is 42.18 when no feature selection selection is applied. These are logical results because feature selection and binarization reduce the size of the input and output

Table 2: Top 20 measure results obtained by TOPSIS.

#	Method	Filter	H	R^+	Acc.	TPR	TNR	Prec.	F	AUC	Time(s)
1	1-vs-all	None	16	0.9876	0.96	0.92	0.97	0.92	0.93	0.95	118.12
2	1-vs-all	None	8	0.9698	0.96	0.90	0.97	0.93	0.92	0.94	95.60
3	Single	None	64	0.9560	0.96	0.91	0.96	0.90	0.92	0.94	125.83
4	Single	CFS	64	0.9498	0.95	0.91	0.96	0.90	0.91	0.94	116.18
5	1-vs-1†	CFS	64	0.9427	0.95	0.90	0.97	0.91	0.90	0.93	223.71
6	Single	None	32	0.9404	0.95	0.91	0.96	0.89	0.91	0.94	90.18
7	1-vs-1†	None	16	0.9386	0.95	0.90	0.96	0.90	0.90	0.94	214.07
8	1-vs-1¶	None	16	0.9354	0.95	0.90	0.96	0.90	0.90	0.93	195.33
9	1-vs-1¶	None	32	0.9353	0.95	0.91	0.97	0.89	0.90	0.94	221.36
10	1-vs-1†	None	64	0.9352	0.95	0.91	0.97	0.89	0.90	0.94	298.69
11	1-vs-1†	None	32	0.9294	0.95	0.89	0.96	0.90	0.90	0.93	231.25
12	1-vs-1¶	None	64	0.9293	0.95	0.89	0.96	0.90	0.90	0.93	287.23
13	1-vs-1†	CFS	8	0.9159	0.94	0.89	0.96	0.90	0.89	0.92	185.72
14	1-vs-all	None	32	0.9158	0.94	0.89	0.96	0.90	0.89	0.92	200.46
15	Single	CFS	32	0.9155	0.95	0.88	0.96	0.89	0.90	0.93	95.65
16	1-vs-1¶	CFS	16	0.9142	0.95	0.89	0.96	0.88	0.90	0.93	187.08
17	1-vs-1¶	CFS	8	0.9058	0.94	0.88	0.96	0.88	0.90	0.93	183.44
18	Single	INT	64	0.9040	0.94	0.88	0.95	0.88	0.90	0.93	130.61
19	1-vs-1†	CFS	2	0.9037	0.94	0.88	0.96	0.89	0.89	0.92	205.10
20	1-vs-1¶	CFS	64	0.9020	0.95	0.89	0.96	0.87	0.89	0.94	221.06

Decoding methods in 1-vs-1 binarization: Hamming decoding¶ Loss-based decoding†.

space, respectively. Notice that the low number of samples in the dataset, which is composed of 105 images, does not favor the use of the *one-vs-one* technique since the training datasets are reduced to the samples of two classes.

5 CONCLUSIONS AND FUTURE RESEARCH

Three binarization techniques and three feature selection methods have been used in this research for tear film lipid layer classification. The evaluation of the techniques and methods was based on several criteria: accuracy, TPR, TNR, precision, F-measure, AUC and training time. TOPSIS method was used as a tool for selecting classification algorithm when algorithm selection involves more than one criterion. In general terms, binarization and feature selection outperform the single machine, multi-class, approach. To the best knowledge of the authors, the use of binarization techniques, features selection filters, and MCDM methods was not attempt so far in the literature for improving classification performance in the assessment of the tear film lipid layer. These results demonstrate the soundness of the methods presented in this research.

For future work, the authors plan to extend this research to different learning algorithms (e.g. naive Bayes classifier or decision trees) and different MCDM methods. Since different MCDM methods will evaluate different learning classifiers from different criteria, they may produce divergent rankings. Thus, the authors plan to implement an approach to resolve disagreeing rankings.

ACKNOWLEDGEMENTS

This research has been partially funded by the Secretaría de Estado de Investigación of the Spanish Government and FEDER funds of the European Union through the research projects PI10/00578, TIN2009-10748 and TIN2011-25476; and by the Consellería de Industria of the Xunta de Galicia through the research project CN2011/007. Beatriz Remeseiro and Diego Peteiro-Barral acknowledge the support of Xunta de Galicia under *Plan I2C* Grant Program.

We would also like to thank the Escuela de Óptica y Optometría of the Universidade de Santiago de Compostela for providing us with the annotated image datasets.

REFERENCES

Allwein, E., Schapire, R., and Singer, Y. (2001). Reducing multiclass to binary: A unifying approach for mar-

gin classifiers. *The Journal of Machine Learning Research*, 1:113–141.

Bolón-Canedo, V., Peteiro-Barral, D., Alonso-Betanzos, A., Guijarro-Berdiñas, B., and Sánchez-Maroño, N. (2011). Scalability analysis of ANN training algorithms with feature selection. *Advances in Artificial Intelligence*, pages 84–93.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2011). On the behavior of feature selection methods dealing with noise and relevance over synthetic scenarios. In *The 2011 International Joint Conference on Neural Networks (IJCNN)*, pages 1530–1537. IEEE.

Bron, A., Tiffany, J., Gouveia, S., Yokoi, N., and Voon, L. (2004). Functional aspects of the tear film lipid layer. *Experimental Eye Research*, 78(3):347–360.

Calvo, D., Mosquera, A., Penas, M., García-Resúa, C., and Remeseiro, B. (2010). Color Texture Analysis for Tear Film Classification: A Preliminary Study. In *Lecture Notes in Computer Science: International Conference on Image Analysis and Recognition (ICIAR)*, volume 6112, pages 388–397.

Crammer, K. and Singer, Y. (2002). On the learnability and design of output codes for multiclass problems. *Machine Learning*, 47(2):201–233.

Dash, M. and Liu, H. (2003). Consistency-based search in feature selection. *Artificial intelligence*, 151(1-2):155–176.

Dietterich, T. and Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286.

E.Wolff (1954). *Anatomy of the eye and orbit (4th edition)*. H. K. Lewis and Co., London.

Fernandez Caballero, J., Martínez, F., Hervás, C., and Gutiérrez, P. (2010). Sensitivity versus accuracy in multiclass problems using memetic pareto evolutionary neural networks. *Neural Networks, IEEE Transactions on*, 21(5):750–770.

Furnkranz, J. (2002a). Pairwise classification as an ensemble technique. *Machine Learning: ECML 2002*, pages 9–38.

Furnkranz, J. (2002b). Round robin classification. *The Journal of Machine Learning Research*, 2:721–747.

García-Resúa, C., Giráldez-Fernández, M., Penedo, M., Calvo, D., Penas, M., and Yebra-Pimentel, E. (2012). New software application for clarifying tear film lipid layer patterns. *Cornea*.

Goto, E., Yagi, Y., Kaido, M., Matsumoto, Y., Konomi, K., and Tsubota, K. (2003). Improved functional visual acuity after punctual occlusion in dry eye patients. *Am J Ophthalmol*, 135(5):704–705.

Guillon, J. (1998). Non-invasive tearscope plus routine for contact lens fitting. *Cont Lens Anterior Eye*, 21 Suppl 1.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. (2006). *Feature Extraction: Foundations and Applications*. Springer Verlag.

Hall, M. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato.

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621.

Hecht-Nielsen, R. (1990). *Neurocomputing*. Addison-Wesley.

Hsu, C. and Lin, C. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425.

Hwang, C. and Yoon, K. (1981). *Multiple attribute decision making: methods and applications: a state-of-the-art survey*, volume 13. Springer-Verlag New York.

King-Smith, P., Fink, B., and Fogt, N. (1999). Three interferometric methods for measuring the thickness of layers of the tear film. *Optom Vis Sci*, 76:19–32.

Korb, D. (2002). *The Tear Film: Structure, Function and Clinical Examination*. Butterworth-Heinemann.

Kou, G., Lu, Y., Peng, Y., and Shi, Y. (2012). Evaluation of Classification Algorithms using MCDM and Rank Correlation. *International Journal of Information Technology & Decision Making (IJITDM)*, 11(01):197–225.

Lemp, M., Baudouin, C., Baum, J., Dogru, M., Foulks, G., Kinoshita, S., Laibson, P., McCulley, J., Murube, J., Pfugfelder, S., Rolando, M., and Toda, I. (2007). The definition and classification of dry eye disease: Report of the definition and classification subcommittee of the internation dry eye workshop (2007). *Ocular Surface*, 5(2):75–92.

Loughrey, J. and Cunningham, P. (2005). Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets. *Research and Development in Intelligent Systems XXI*, pages 33–43.

McLaren, K. (1976). The development of the CIE 1976 (L*a*b) uniform colour-space and colour-difference formula. *Journal of the Society of Dyers and Colourists*, 92(9):338–341.

Nichols, K., Nichols, J., and Mitchell, G. (2004). The lack of association between signs and symptons in patients with dry eye disease. *Cornea*, 23(8):762–770.

Olson, D. (2004). Comparison of weights in TOPSIS models. *Mathematical and Computer Modelling*, 40(7-8):721–727.

Opricovic, S. and Tzeng, G. (2004). Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research*, 156(2):445–455.

Press, W., Flannery, B., Teukolsky, S., Vetterling, W., et al. (1986). *Numerical recipes*, volume 547. Cambridge Univ Press.

Ramos, L., Penas, M., Remeseiro, B., Mosquera, A., Barreira, N., and Yebra-Pimentel, E. (2011). Texture and color analysis for the automatic classification of the eye lipid layer. In *LNCS: Advances in Computational Intelligence (International Work Conference on Artificial Neural Networks-IWANN 2011)*, volume 6692, pages 66–73.

Remeseiro, B. (2012). *VOPTICAL_I1, VARPA optical dataset annotated by optometrists from the Faculty of Optics and Optometry, University of Santiago de Compostela (Spain)*.

http://www.varpa.es/voptical_I1.html, last access: october 2012.

Remeseiro, B., Penas, M., Mosquera, A., Novo, J., Penedo, M., and Yebra-Pimentel, E. (2012). Statistical comparison of classifiers applied to the interferential tear film lipid layer automatic classification. *Computational and Mathematical Methods in Medicine*, 2012.

Remeseiro, B., Ramos, L., Penas, M., Martínez, E., Penedo, M., and Mosquera, A. (2011). Colour texture analysis for classifying the tear film lipid layer: a comparative study. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 268–273, Noosa, Australia.

Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141.

Rolando, M., Iester, M., Marcrí, A., and Calabria, G. (1998). Low spatial-contrast sensitivity in dry eyes. *Cornea*, 17(4):376–379.

Rolando, M., Refojo, M., and Kenyon, K. (1983). Increased tear evaporation in eyes with keratoconjunctivitis sicca. *Arch Ophthalmol*, 101(4):557–558.

Zhao, Z. and Liu, H. (2007). Searching for interacting features. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1156–1161. Morgan Kaufmann Publishers Inc.