WEIGHTED CORRELATION MATRIX SIMILARITY: A NEW CLASSIFICATION ALGORITHM

Alexandre Serra Barreto
Ministry of Finance
Brasília-DF, Brazil, CEP 70.048-900

ABSTRACT

This paper presents a new classification algorithm: the Weighted Correlation Matrix Similarity (WCMS) classification algorithm. By means of weighted intra-class correlation matrices and a similarity measure between matrices, WCMS is able to assign a class to new unknown cases (samples). WCMS formulates no previous assumptions regarding the data and for better performance may be previously calibrated by means of a training set. Its process of classification is flexible and benefits both from the knowledge arisen from the training set and the information proportioned by the new sample to be classified itself. WCMS was applied for classification in practice and its performance was compared with other popular classification algorithms that are based on the data covariance structure using nine real datasets available in the UCI data repository. The results showed that the performance of the WCMS algorithm was at least as competitive as any of the other tested methods. Concerning the nine focused data sets, WCMS presented itself as superior in terms of classification accuracy in five of them. It was concluded that WCMS can be used as an effective classification tool in a wide range of data sets.

KEYWORDS

Classification, algorithm, data mining, prediction

1. INTRODUCTION

Mankind has performed classification since remote years, as a part of daily life and survival. With human evolution, our motivation to classify has become more complex and wide, comprehending classification in fields like engineering, management, banking, marketing, psychology etc. In the context of data mining, classification can be understood as the process of assigning to a new observation (sample) one amongst a set of previously known classes. In order to do that, a method or algorithm has to learn based on a training set (instances of the same data type and its actual classes) before classifying a new unknown case. The traditional classification techniques generally utilized in data mining are, among others, decision trees, discriminant analysis and its evolutions, Logistic Regression, and neural net (Han and Kamber, 2001).

There is a specific group of classifiers based on the data covariance structure, where the mean vector and the class covariance matrix are usually unknown and have to be estimated from a training set, but these estimates are optimal only asymptotically and can produce lower classification accuracy. Also, there are concerns involving covariance matrix inversion, and the assumption of data normality limits to some degree the use of these methods. Even though in most cases some level of violation of the theoretical assumptions of a classification method would not seriously compromise its practical application, strong deviations may pose serious problems and perhaps a major impediment in its applicability. In addition, the common known theoretical parametric forms do not fit the densities encountered in practice problems (Duda and Hart, 1973).

This article proposes a novel nonparametric classification method that is conceptually free from matrix inversion and also does not pose previous assumptions concerning the data, thus giving it a broad application: the Weighted Correlation Matrix Similarity (WCMS) classification algorithm. The algorithm is based on learning by analogy. By means of weighted intra-class correlation matrices calculated from data set attributes and a similarity measure between correlation matrices, WCMS is able to assign a class to new unknown cases (samples). WCMS should be previously calibrated by means of a training set for better performance. Its process of classification is of a flexible kind, given that the knowledge or generalization arisen from the

calibration and training set can be changed operationally by a weighting process in light of the information proportioned by the new sample to be classified. In summary, WCMS learns globally, as is usual, from the training set, as well as individually from the new sample itself.

2. CLASSIFIERS BASED ON THE DATA COVARIANCE STRUCTURE

In parametric classification techniques we learn from data under the assumption that the form for the underlying density function is known. The most common procedure is to consider the Normal distribution, as is the case of Gaussian Maximum Likelihood Classifier (GMLC). Suppose there are c distinct classes, given a sample vector $X^T = (x_1, x_2, ..., x_p)$ depicting p measurements made on the sample from p attributes, GMLC will assign to X the class h ($h=1,\ldots, c$) having the highest likelihood among the classes. GMLC assumes that the data follows the Normal density function:

$$(X \mid \mu_h, \Sigma_h) = \frac{1}{\sqrt{(2\pi)^p \mid \Sigma_h \mid}} \exp[-1/2(X-\mu_h)^T \Sigma_h^{-1}(X-\mu_h)].$$

In this equation, μ_h is the mean vector of class h, and Σ_h is the covariance matrix for class h. Usually these parameters are not known and must be estimated from training samples. The sample mean is typically the estimate for the density mean and the covariance matrix is usually estimated via the sample covariance matrix or the maximum likelihood covariance matrix estimate. The sample mean and the maximum likelihood covariance matrix estimates maximize the joint likelihood of training samples, which are assumed to be statistically independent (Hoffbeck and Landgrebe, 1996).

These GMLC features may pose some limitations to its applicability. The mean and covariance estimates are optimal only asymptotically and can produce lower classification accuracy when the training sample is small. Actually, Hoffbeck and Landgrebe (1996) stated that unless many more than $p+1$ samples are available the true covariance matrix is poorly estimated. Also, the assumption of the knowledge about the form for the underlying density function may be suspicious in most applications (Duda and Hart, 1973). Furthermore, the method involves the inversion of Σ_h estimate and in some cases this matrix can be ill-conditioned or even singular. Indeed, the problem of covariance matrix inversion is exacerbated in sample settings where there are less than $p+1$ training samples per class, implying that the resulting covariance matrix is singular, or where the training sample size per class is not considerably larger than p. Also, high dimensional data is more susceptible to presenting collinear data and then to result in ill-conditioned covariance matrices.

There is research proposing improvements, specifically concerning the covariance matrix estimation, like the use of the Leave-One-Out Covariance matrix estimate (LOOC) method to calculate an optimal covariance matrix mixture (Hoffbeck and Landgrebe, 1996) or Regularized Discriminant Analysis (RDA). Despite the significant improvements of maximum likelihood classification and its application in the case of limited training data, these approaches remain operating under the assumption that the form for the underlying density function is known. Additionally, they are also affected by calculus complexity. LOOC requires calculus approximations in the pursuit of an efficient implementation. Also, both for LOOC and RDA there is the requirement of an optimization of the mix of regularization parameters, so there are techniques for making this process more efficient.

Discriminant Analysis (DA) comprises general methods to produce rules to assign a predefined class to a sample characterized by certain attributes. Usually it also relies on the normal density and the previous computation of an estimate from the training set. Suppose that π_h is the prior probability of observing a class h member and $f_h(X)$ is a probability density, then the Bayes rule would assign to the object X the class h with maximal $f_h(X)\pi_h$, that is, [$\hat{f}_h(X)\hat{\pi}_h$ = Max $f_h(X)\pi_h$]. If the distribution for class h

is multivariate normal, the net Bayes rule minimizes (Venables and Ripley, 2002): $d_h(X) = (X - \mu_h)^T \Sigma_h^{-1} (X - \mu_h) + \ln |\Sigma_h| - 2 \ln \pi_h$.

This quantity is often called discriminant score for the h th class. Use of the classification rule [$\hat{d}_h(X)$ = Min $d_h(X)$] is known as Quadratic discriminant analysis (QDA). When all the classes h are assumed to have identical class covariance matrices, i.e. $\Sigma_h = \Sigma$, this rule is referred to as Linear Discriminant Analysis (LDA). QDA and LDA are expected to work well if the class conditional densities are approximately normal. But if the class sample sizes n_h are small with respect to p, the covariance matrix estimates become highly variable and biased, and will surely affect the process of classification (Friedman, 1989).

As already stated, one attempt to mitigate part of these problems is to try to obtain more reliable estimates from the sample covariance matrix by Regularized Discriminant Analysis (RDA) (Friedman, 1989), which involves an optimal mix of sample covariance matrix, global covariance matrix and the identity matrix, formalized as: $d_h(X) = (X - \mu_h)^T \Sigma_h^{-1}(\lambda, \gamma)(X - \mu_h) + \ln |\Sigma_h(\lambda, \gamma)| - 2 \ln \pi_h$; $0 \le \lambda \le 1$ and $0 \le \gamma \le 1$. See (Friedman, 1989) to access the formulas to calculate $\Sigma_h(\lambda, \gamma)$.

Concerning the regularization, if ($\lambda = 0$ and $\gamma = 0$) the last equation represents QDA, if ($\lambda = 1$ and $\gamma = 0$) it will represent LDA. One disadvantage of RDA is that there are often many possible tuning parameters (λ, γ) giving the same cross-validation error rate, but the test error rate based on them may vary significantly. Users are faced with "the dilemma of how to choose the best parameters" (Xu et al. 2009, p. 1675).

Finally, we should point out that both in LDA and QDA processing there still remains the concern with respect to matrix inversion and singularities. RDA overcomes it by modifying the sample covariance matrix, shrinking it towards a non-singular or well-estimated covariance matrix, for example, the identity matrix. A way to tackle this issue is to use the Moore–Penrose pseudo-inverse when the sample covariance matrix is singular. However, this method might have poor performance since the generalized inverse will be very unstable due to lack of observations (Guo et al., 2007, *apud* Xu et al., 2009, pp. 1675).

3. THE WEIGHTED CORRELATION MATRIX SIMILARITY CLASSIFICATION ALGORITHM

What is proposed is a new nonparametric classification method: the Weighted Correlation Matrix Similarity (WCMS) classification algorithm. WCMS theoretically does not make previous assumptions about data, for instance imposing normality of distribution and sample dimensionality independence. WCMS works with several empirical covariance matrices in its steps, and its applicability is insensitive to problems caused by small-sample set tings or data collinearity, since no matrix inversion is required. The new algorithm is formalized in the following subsections.

3.1 Description of the Algorithm

The idea behind the WCMS algorithm is that each subset of a population belonging to a specific class has its own correlation matrix pattern with regard to specific attributes (variables). Hence, the evaluation of the impact in this correlation matrix, caused by the addition of a new sample (case) to this population subset, can throw light on the actual class of the added unknown case. Note that the impact of the addition of a mean vector of a subset to this same subset is null in terms of matrix correlation structure.

Given c predefined classes, $h = 1,..., c$, and n sample vectors $X_i^T = (x_1, x_2, ..., x_p)$, $i = 1,..., n$, depicting p measurements made on the sample from p attributes, x_{ij} means the j th measurement,

$j = 1,..., p$, for the i th sample. Each sample vector X_i^T belongs to a known class h. So, let $\mathbf{X}$ be a data matrix of ty pe (n, p) w ith the m easurements of data (x_{ij}) as elem ents, $j = 1,..., p$ and $i = 1,..., n$. The WCMS algorithm functions as depicted below, considering the matrix $\mathbf{X}$ as a training set [with p attributes (variables), n instances (training samples) and c classes], x_{hij} as a n element of $\mathbf{X}$ belonging to a class h, $h = 1,..., c$, and t he ma trix $\mathbf{U}$ of type (s,p) as a ne w unknown set [with t he m easurements o f d ata (u_{sj}), $s = 1,..., v$, $j = 1,..., p$, as el ements, with p att ributes (variables), v ca ses (unknown sam ples) and c classes], the v new unknown samples having to be classified by the algorithm. Also, note below that the te rms *abs*( ) an d *round*( ) refer respectively to 'a bsolute val ues' and the 'rou nding proce ss' and that we must also introduce the auxiliary counters D_{hj} and D_{hs0}, $D_{hs1}, D_{hs2}, D_{hs3}$ and D_{hs4}. The Weighted Correlation Matrix Similarity (WCMS) Classification Algorithm Pseudocode is:

(0) begin algorithm (initialize variables and counters);
(1) for $h = 1$ to c do:
  (1.1) form original subsets from the training set containing all samples belonging to each class h (with measurements x_{hij}), each subset labeled as S_h (totalizing c subsets, each with n_h samples);
  (1.2) calculate the correlation matrix for the p attributes in each S_h subsets, each matrix labeled as A_h, (totalizing c correlation matrices);
  (1.3) calculate the preliminary number of replicas (rep_h): $rep_h = round(r_h \cdot n_h)$ (note: r_h is defined below);
end do (referring to step 1);
(2) for $s = 1$ to v do:
  (2.1) for $j = 1$ to p do:
    (2.1.1) for $h = 1$ to c do:
    access the deviation pattern of the sth new sample (with measurements u_{sj}) with respect to S_h, calculating:

$$D_{hj} = abs((u_{sj} - \bar{x}_{h.j}) / \sigma_{hj});$$

      (2.1.1.1) if $0 \le D_{hj} \le 1$ then do:
      $D_{hs0} = D_{hs0} + 1;$
      end if;
      (2.1.1.2) if $1 < D_{hj} \le 2$ then do:
      $D_{hs1} = D_{hs1} + 1;$
      end if;
      (2.1.1.3) if $2 < D_{hj} \le 3$ then do:
      $D_{hs2} = D_{hs2} + 1;$
      end if;
      (2.1.1.4) if $3 < D_{hj} \le 4$ then do:
      $D_{hs3} = D_{hs3} + 1;$

end if;
(2.1.1.5) if $D_{hj} > 4$ then do:
$D_{hs4} = D_{hs4} + 1$;
end if;
end do (referring to step 2.1.1);
end do (referring to step 2.1);
(2.2) weight the values of rep_h based on the sth new sample measurements and on the results from steps 1.3 and 2.1.1 in order to define the value for rep_{hs} (see detailed description of this step below);
(2.3) add rep_{hs} replicas of the sth new sample (with measurements u_{sj}) to each subset S_h, forming new S_{hs} subsets with x_{hsij} as elements (totalizing c new subsets, each with $n_{hs} = n_h + rep_{hs}$ samples);
(2.4) for $h = 1$ to c do:
(2.4.1) calculate the correlation matrix B_{hs} for the p attributes in each S_{hs} subsets;
(2.4.2) calculate a similarity measure $Sim_{(h,hs)}$ between the matrices A_h and B_{hs} for every A_h and B_{hs} with the same subscript h (see detailed description of this measure below);
end do (referring to step 2.4);
(2.5) assign to the sth new sample (with measurements u_{sj}) the class h referring to the lesser $Sim_{(h,hs)}$ similarity measure;
end do (referring to step 2);
(3) end of the algorithm.

Where in WCMS: $h = 1$ to c; $s = 1$ to v; n_h: number of instances in the subset S_h; n_{hs}: number of instances in the subset S_{hs}; r_h: percentage value to be multiplied by the number of instances n_h in the subsets S_h (see step 1.3), so as to define the preliminary number of replicas rep_h to be added to the S_h subsets (for better performance r_h should be defined by means of a process of calibration concerning the training set, or in light of previous knowledge about the data); rep_h: preliminary number of replicas to be added to the S_h subset; rep_{hs}: weighted number of replicas of the sth unknown case (sample) to be added to the subset S_h, as obtained in step 2.2; $\bar{x}_{h.j}$: mean for the jth attribute in the subset S_h; σ_{hj}: standard deviation for the jth measure in subset S_h.

Concerning step 2.2, the weighting phase of WCMS, we have to point out that the assessment of data variability and the presence of outliers are key points in the classification process. Also, note that infinite forms of calculations are possible, obviously yielding different results, including the use of weights and linear or exponential functions to weight data. Concerning functions, we could for instance use the linear function $w = 0.5 \cdot D_{hj} - 1$ to construct a weight (w) varying between 0 and 1 in function of the value for the deviation D_{hj} (see the step 2.1.1). Or we can also use an exponential function, maybe

($w = e^{D_{hj}-4}$), to weight this value in a dampened way. Consequently, this process itself is a scientific issue full of possibilities.

So, presented here is an individual solution for the weighting process in step 2.2, that gives more weight to the samples which are suspected to be an outlier concerning the training data set, as follows:

(2.2) weight the values of rep_h based on the sth new sample measurements and on the results from steps 1.3 and 2.1.1 in order to define the value for rep_{hs};

(2.2.1) for $h = 1$ to c do:

$w_{hs} = 1 - ((0.2 \cdot D_{hs2} / p) + (0.3 \cdot D_{hs3} / p) + (0.5 \cdot D_{hs4} / p))$;

$rep_{hs} = rep_h \cdot 1 / w_{hs}$;

$rep_{hs} = round(rep_{hs})$;

end do (referring to step 2.2.1);

The similarity measure in step (2.4.2) of WCMS uses the quadratic deviation between corresponding elements (below the diagonal) in matrices A_h and B_{hs}, and is formalized as follows: $Sim_{(h,hs)} = \sum_{i=2}^{p} \sum_{j=1}^{i-1} (b_{ij} - a_{ij})^2$. Where: p is the number of rows in matrices A_h and B_{hs}; a_{ij} is the element in the ith row and the jth column in A_h; b_{ij} is the element in the ith row and the jth column in B_{hs}. Note that the number of rows (or columns) in the correlation matrices A_h and B_{hs} equals the number of attributes in the data set. The lesser the measure $Sim_{(h,hs)}$ is, the more similar the matrices A_h and B_{hs} are. The analyst can use other similarity measures for the WCMS algorithm.

As mentioned in the algorithm depiction, for better performance WCMS should be calibrated by the submission of a training set to the WCMS steps (except for the steps 2.1 and 2.2), considering several combinations of values for r_h, $h = 1$ to c. The reason for not applying steps 2.1 and 2.2 is due to the fact that the objective of the calibration is to capture the overall pattern of data and to define the set of r_h in light of that. For this reason, step 2.3 in the calibration should be adjusted to calculate $rep_{hs} = rep_h$. We could say that in the calibration phase we would refer to a "CMS" algorithm, without the implementation of the weighting capabilities. After calibration, the best set of r_h with respect to some criteria (accuracy, sensitivity etc) should then be implemented in WCMS so as to classify the unknown set stored in $\mathbf{U}$.

It is important to mention that the WCMS algorithm is applicable in any situation concerning data dimensionality and sample sizes (p, n, n_h), including the cases where $n < p$ or $n_h < p$, provided that there exists intra-class variability for all considered attributes ($j = 1,..., p$). If it were not the case for a specific attribute, it should be discarded from the analysis.

Finally in this section, it should be added that recent literature involving the classifiers that are based on the data covariance structure makes no mention of an algorithm that works like WCMS. See, for instance, (Lu et al., 2009, Peng et al., 2008., Ji and Ye, 2008, Ye et al., 2006, Guo et al., 2008, Liu et al., 2008, Halbe and Aladjem, 2007, Lu et al., 2003., Xu et al., 2009).

3.2 Numerical Illustration with the Iris Data Set

Suppose that we change the location of our training data stored in $\mathbf{X}$. If $\bar{x}_{.j}$ is the global mean for the jth measurement in $\mathbf{X}$, a matrix $\mathbf{C}$ of type (n, p) can then be defined with the centered measurements

$(x_{ij} - \bar{x}_{.j})$ as elements. Each class h, $h = 1,..., c$, discriminates an h th subset with n_h samples. Then, theoretically WCMS works with correlation matrices given as: $\mathbf{R} = \mathbf{D}^{1/2}\mathbf{C}^T\mathbf{NCD}^{1/2}$. Where: $\mathbf{C}$ is a matrix of type (n, p) with the centered measurements of data $(x_{ij} - \bar{x}_{.j})$ as elements, $j = 1,..., p$ and $i = 1,..., n$; $\mathbf{N}$ is a diagonal matrix of type (n,p) with $(1/n)$ as elements; $\mathbf{D}$ is a diagonal matrix of type (n,p) with $(1/\sigma_j^2)$ as elements, σ_j^2 being the variance for the j th attribute in matrix $\mathbf{X}$.

So in step 1.2 WCMS calculates: $\mathbf{A}_h = \mathbf{D}_h^{1/2}\mathbf{C}_h^T\mathbf{N}_h\mathbf{C}_h\mathbf{D}_h^{1/2}$. Where: $\mathbf{A}_h$ is the correlation matrix for the p attributes in the subset S_h, $h = 1,..., c$; $\mathbf{C}_h$ is a matrix of type (n_h, p) with the centered measurements of data $(x_{hij} - \bar{x}_{h.j})$ as elements, $j = 1,..., p$ and $i = 1,..., n_h$, n_h being the number of samples enclosed in the subset S_h, x_{hij} being the j th measurement for the i th sample belonging to the subset S_h, and $\bar{x}_{h.j}$ being the mean for the j th attribute in the subset S_h; $\mathbf{N}_h$ is a diagonal matrix of type (n_h, p) with $(1/n_h)$ as elements; $\mathbf{D}_h$ is a diagonal matrix of type (n_h, p) with $(1/\sigma_{hj}^2)$ as elements, σ_{hj}^2 being the variance for the j th attribute present in the subset S_h.

In step 2.4.1 the following correlation matrices are obtained: $\mathbf{B}_{hs} = \mathbf{D}_{hs}^{1/2}\mathbf{C}_{hs}^T\mathbf{N}_{hs}\mathbf{C}_{hs}\mathbf{D}_{hs}^{1/2}$. Where: $\mathbf{B}_{hs}$ is the correlation matrix for the p attributes in the subset S_{hs}, $h = 1,..., c$, $s = 1,..., v$; $\mathbf{C}_{hs}$ is a matrix of type (n_{hs}, p) with the centered measurements of data $(x_{hsij} - \bar{x}_{hs.j})$ as elements, $j = 1,..., p$ and $i = 1,..., n_{hs}$, n_{hs} being the number of samples enclosed in the subset S_{hs}, x_{hsij} being the j th measurement for the i th sample belonging to the subset S_{hs}, and $\bar{x}_{hs.j}$ being the mean for the j th measurement in the subset S_{hs}; $\mathbf{N}_{hs}$ is a diagonal matrix of type (n_{hs}, p) with $(1/n_{hs})$ as elements; $\mathbf{D}_{hs}$ is a diagonal matrix of type (n_{hs}, p) with $(1/\sigma_{hsj}^2)$ as elements, σ_{hsj}^2 being the variance for the j th attribute present in the subset S_{hs}.

To illustrate numerically the WCMS algorithm, consider the Iris Data Set from the UCI Machine Learning Repository (Frank and Asuncion, 2010). A 10-fold cross validation process divided the data into ten equally populated blocs $(n_h = 15)$, 'bloc' is R program terminology used to describe data divisions (see section 4.1). Suppose that we are classifying the 6th validation bloc with the training set formed by the 135 leftover instances. Our 6th validation bloc is composed by the following Iris Data Set samples (V1, V2, V3, V4, V5=class): 4.6, 3.4, 1.4, 0.3, 1; 4.8, 3.4, 1.6, 0.2 , 1; 4.3, 3.0 , 1.1 , 0.1, 1; 5.7, 3.8, 1.7, 0.3, 1; 5.0 , 3.0 , 1.6, 0.2, 1; 5.0 , 3.2, 1.2 , 0.2 , 1; 5.5, 3.5, 1.3, 0.2, 1; 5.0 , 3.5, 1.6 , 0.6, 1; 4.8, 3.0, 1.4, 0.3, 1; 6.4 ,3.2, 4.5 ,1.5, 2; 7.1 ,3.0, 5.9 ,2.1, 3; 7.3, 2.9, 6.3, 1.8, 3; 6.4 , 2.7, 5.3, 1.9, 3; 7.7, 2.6, 6.9, 2.3, 3; 6.3 ,2.5, 5.0, 1.9, 3. Based on the respective training set, step 1.2 in WCMS produces the following correlation matrices for the four variables in the training subsets S_h related to each class h (in the matrices, only the elements below the diagonal are shown):

$$\mathbf{A}_1 = \begin{bmatrix} 1 & & & \\ 0.7765310 & 1 & & \\ 0.1921224 & 0.1120736 & 1 & \\ 0.3150621 & 0.2853089 & 0.2494696 & 1 \end{bmatrix}, \mathbf{A}_2 = \begin{bmatrix} 1 & & & \\ 0.5146851 & 1 & & \\ 0.7528799 & 0.5584643 & 1 & \\ 0.5388677 & 0.6570786 & 0.7857897 & 1 \end{bmatrix},$$

$$\mathbf{A}_3 = \begin{bmatrix} 1 & & & \\ 0.5443223 & 1 & & \\ 0.8514286 & 0.5050011 & 1 & \\ 0.2750327 & 0.5742875 & 0.3179154 & 1 \end{bmatrix}.$$

In order to classify bloc 6, our previous calibration process defined: ($r_1 = 0.15$, $r_2 = 0.15$, $r_3 = 0.11$). Let us now follow the next WCMS steps and the final classification for the 1st validation sample (V1=4.6, V2=3.4, V3=1.4, V4=0.3, $h = 1$). Step 1.3 calculates the preliminary number of replicas rep_h ($rep_1 = 6$, $rep_2 = 7$, $rep_3 = 5$). Step 2.2 weights the values for rep_h, ($w_{11} = 1$, $w_{21} = 0.65$, $w_{31} = 0.675$), and then calculates the values for rep_{hs} ($rep_{11} = 6$, $rep_{21} = 11$, $rep_{31} = 7$). After that, step 2.4.1 allows the calculation of the correlation matrices $\mathbf{B}_{11}$, $\mathbf{B}_{21}$ and $\mathbf{B}_{31}$, as follows (only the elements below the diagonal are shown):

$$\mathbf{B}_{11} = \begin{bmatrix} 1 & & & \\ 0.7249988 & 1 & & \\ 0.2346751 & 0.1163827 & 1 & \\ 0.1941009 & 0.2702554 & 0.2082107 & 1 \end{bmatrix}, \mathbf{B}_{21} = \begin{bmatrix} 1 & & & \\ -0.2361573 & 1 & & \\ 0.8740680 & -0.4710633 & 1 & \\ 0.8254459 & -0.4056114 & 0.9671269 & 1 \end{bmatrix},$$

$$\mathbf{B}_{31} = \begin{bmatrix} 1 & & & \\ 0.01282225 & 1 & & \\ 0.89329484 & -0.24545891 & 1 & \\ 0.76132273 & -0.16996754 & 0.9070960 & 1 \end{bmatrix}.$$

In step 2.4.2 the similarity measure $Sim_{(h,hs)}$ (defined in the previous section) applied to matrices $\mathbf{A}_1$, $\mathbf{B}_{11}$, $\mathbf{A}_2$, $\mathbf{B}_{21}$ and $\mathbf{A}_3$, $\mathbf{B}_{31}$ results in $S_{(1,11)} = 0.04209079$, $S_{(2,21)} = 5.765397$ and $S_{(3,31)} = 3.969925$. Thus, step 2.5 assigns to the 1st validation sample of the 6th bloc the class h of lesser $S_{(h,h1)}$, that is, class 1 (Iris-setosa), which matches the actual class for this sample in the Iris Data Set.

4. COMPARING WCMS TO OTHER CLASSIFICATION METHODS

4.1 Methodology

Some real data sets from the UCI repository (available from: http://archive.ics.uci.edu/ml/datasets/) are used to compare WCMS to alternative methods. These data sets allow the exploration of several unspecified empirical densities, high-dimensional situations, collinearity characteristics and also small-sample settings.

A k-fold cross validation is widely used in the related literature (Venables and Ripley, 2002; Licheng et al., 2006) to present a more stable estimate of the performance of a classification method. So to calibrate the WCMS algorithm and define the best set of r_h to be applied to each validation bloc, a previous 10-fold

cross-validation was performed for each of the 10 training blocs. Thereafter, WCMS was submitted to each validation bloc (once adjusted with the best r_h setting regarding the corresponding training bloc and its proper process of 10-fold cross-validation). In the calibration process, the chosen criteria was the greatest accuracy rate reached over the 10-fold cross-validation. If two r_h settings gave the same accuracy rate during the process of calibration, the chosen set of r_h was the set with the lower $\sum_{h=1}^{c} r_h$. In the calibration process we considered all possible sets of r_h ($0.01 \leq r_h's \leq 0.15$, with increments of 0.01). The overall results (r_h settings) and the processing times of the 10-fold cross-validation process are presented in the next subsection.

For comparison, RDA, LDA and GMLC algorithms were applied to exactly the same blocs generated by the depicted cross-validation process, all of them processed and programmed in R program (R Development Core Team, 2007). LDA was implemented in R package "MASS". RDA here refers to regularized discriminant analysis with the regularization parameter determined by cross-validation. The values of the regularization parameter tested are the default values given in the rda() function in R package "rda" (10 values equally spaced between 0 and 0.99 for "alpha" and between 0 and 3 for "delta"). We used the "Min–Min" rule (Guo et al., 2007, *apud* Xu et al., 2009, pp. 1675) for selecting the optimal parameters when multiple value-pairs gave the same misclassification error during cross-validation.

4.2 Presentation of Data Sets and Comparison of Classification Results

Pima Indians Diabetes Data Set comprises 768 entries (8 attributes and a class variable), 550 of them classified as 0 and 268 classified as category 1. Ten mutually exclusive folds were randomly sampled from the Pima data set (nine validation folds including 77 entries and the tenth fold comprising 75).

Breast Cancer Wisconsin Original Data Set comprises 699 entries (9 attributes and a class variable), 458 of them classified as category 2 and 241 classified as category 4 (recoded as 0 and 1, respectively). Ten mutually exclusive folds were randomly sampled from this data set (nine validation folds including 69 entries and the tenth fold comprising 62). 16 original entries with missing data were removed.

Haberman's Survival Data Set comprises 306 entries (3 attributes and a class variable), 81 of them classified as category 2 and the remaining 225 classified as category 1 (recoded as 1 and 0, respectively). Ten mutually exclusive folds were randomly sampled from this data set (nine validation folds including 31 entries and the tenth fold comprising 27).

Connectionist Bench Data Set comprises 208 entries (60 attributes and a class variable). The label associated with each record is a letter "R" or "M", with a total of 97 Rs and 111 Ms. We recoded the class variable (R=0 and M=1). Ten mutually exclusive folds were randomly sampled from this data set (nine validation folds including 21 entries and the tenth fold comprising 19).

SPECT Heart Data Set comprises 267 entries (22 attributes and a class variable). The label associated with each record is "0" or "1". Ten mutually exclusive folds were randomly sampled from this data set (nine validation folds including 36 entries and the tenth fold comprising 27). The 18th and the 19th variables were discarded since they do not have variability over class "0".

Mammographic Mass Data Set comprises 961 entries of data (5 attributes and a class variable). The class associated with each record is the field 'severity', 0 or 1. Ten mutually exclusive folds were randomly sampled from this data set (all of them with 83 entries). 131 original entries with missing data were removed.

Blood Transfusion Service Center Data Set (copyright Prof. I-Cheng Yeh) comprises 748 entries (4 attributes and a class variable). The class associated with records is a binary variable (0 or 1). Ten mutually exclusive folds were randomly sampled from this data set (9 of them with 75 and the tenth with 73 entries).

BUPA Liver Disorders Data Set comprises 345 entries (6 attributes and a class variable). The class associated with each record is a selector field (1 and 2, recoded, respectively, as 0 and 1). Ten mutually exclusive folds were randomly sampled from this data set (9 of them with 35 and the tenth with 30 entries).

Ionosphere Data Set comprises 351 instances of data (34 attributes and a class variable). The labels in the data are "g" or "b" (recoded as 0 and 1, respectively). Ten mutually exclusive folds were randomly sampled from this data set (nine validation folds including 36 entries and the tenth fold comprising 27 entries of data). The first and second attributes were discarded since they do not present variability over classes.

To illustrate the 10-fold cross-validation results for WCMS calibration, Table 1 summarizes the r_h settings that gave the greatest accuracy rate (%) for all 10 training blocs. The reader is reminded that there are two classes for all the datasets $(h = 0, h = 1)$. Thereafter, these r_h settings (in Table 1) were applied to step 1.3 in WCMS in order to classify the corresponding validation blocs.

Table 1. Summary of the 10-fold cross-validation process results. The (r_0, r_1) (%) settings giving the best accuracy rate concerning training blocs

	bloc 1	bloc 2	bloc 3	bloc 4	bloc 5	bloc 6	bloc 7	bloc 8	bloc 9	bloc 10
PI	7%, 10%	4%, 6%	5%, 7%	2%, 3%	2%, 3%	4%, 6%	2%, 3%	9%, 13%	8%, 12%	2%, 3%
BR	9%, 13%	5%, 8%	6%, 10%	5%, 9%	4%, 7%	4%, 7%	4%, 8%	5%, 8%	7%, 11%	9%, 14%
HB	2%, 5%	1%, 1%	4%, 8%	5%, 14%	4%, 9%	5%, 11%	2%, 11%	3%, 7%	5%, 7%	1%, 4%
CN	10%,12%	7%, 8%	11%,13%	12%,14%	9%, 11%	12%,14%	8%, 10%	4%, 4%	10%,12%	5%, 6%
IO	2%, 5%	1%, 3%	4%, 9%	5%, 12%	2%, 5%	4%, 8%	2%, 5%	2%, 5%	3%, 6%	2%, 5%
SP	4%, 4%	9%, 6%	10%, 8%	14%,12%	2%, 1%	6%, 1%	8%, 8%	8%, 7%	6%, 3%	15%, 9%
BU	13%,12%	4%, 5%	14%,13%	2%, 2%	8%, 8%	10%, 9%	15%,14%	5%, 5%	11%, 9%	7%, 7%
MA	9%, 9%	6%, 5%	9%, 8%	8%, 5%	13%,15%	4%, 3%	12%,14%	14%,11%	7%, 7%	10%, 7%
TR	1%, 3%	1%, 3%	4%, 12%	3%, 10%	1%, 4%	1%, 4%	1%, 3%	1%, 4%	4%, 13%	1%, 4%

DATA SETS: PI = PIMA; BR = BREAST; HB = HABERMAN'S; CN = CONNECTIONIST;
IO = IONOSPHERE; SP = SPECT; BU = BUPA; MA= MAMMOGRAPHIC; TR = TRANSFUSION.

With respect to the processing times in R program (version 2.14.0 – 64 bit) running on a notebook (processor: 2.13 GHz, 3 MB L3 cache; 3 GB DDR3 Memory; 320 GB HDD), Table 2 shows processing times for all compared algorithms concerning bloc 1 (running of calibration and classification). LDA and GMLC do not require a calibration phase. In Table 2 we see that all algorithms were very fast in running the classification itself. Also, under certain circumstances WCMS may be time-consuming in a 10-fold cross validation process, since it covers all possible settings for r_h ($0.01 \le r_h's \le 0.15$, with increments of 0.01). Undoubtedly, better and more recent machines or desktops can perform this calibration faster (even when encapsulated in a 10-fold cross validation process).

Table 2. Processing times for bloc 1 (for calibrating training bloc 1 and running the classification of validation bloc 1). Where: s (seconds) and m (minutes); * (to run the classification of validation bloc 1); ** (to calibrate training bloc 1)

	HB CN BR			PI	MA	TR	SP	IO	BU
LDA*	0.32 s	0.55 s	0.33 s	0.23 s	0.25 s	0.28 s	0.30 s	0.31 s	0.26 s
GMLC*	0.84 s	1.37 s	1.23 s	1.08 s	1.14 s	0.81 s	1.01 s	1.42 s	0.85 s
RDA*	0.25 s	0.22 s	1.30 s	1.90 s	2.38 s	1.81 s	0.25 s	0.32 s	0.29 s
RDA**	3.31 s	3.18 s	12.82 s	17.41 s	21.68 s	16.40 s	3.67 s	4.25 s	3.79 s
WCMS*	2.42 s	10.64 s	9.08 s	7.09 s	7.52 s	4.57 s	4.41 s	9.45 s	3.53 s
WCMS**	12.90 m	79.94 m	110.43 m	141.61 m	115.64 m	93.14 m	28.35 m	90.91 m	21.74 m

DATA SETS: PI = PIMA; BR = BREAST; HB = HABERMAN'S; CN = CONNECTIONIST;
IO = IONOSPHERE; SP = SPECT; BU = BUPA; MA= MAMMOGRAPHIC; TR = TRANSFUSION.

Finally, Table 3 shows synoptically the accuracy rate mean and standard error for all datasets and methods (the best results for each dataset are in bold). All methods were proficient in classifying data and obtained relatively similar results. As can be seen, WCMS attained the greatest accuracy rate in five of the data sets, RDA in three of them, and LDA in only one. Also note that the GMLC algorithm was incapable of classifying the Transfusion data set due to limitations involving matrix singularities. A positive aspect that should be highlighted here is that the results for the WCMS algorithm were obtained without posing previous assumptions regarding the data, as the other compared methods do.

Table 3. Classification Results for Real Data Sets - Accuracy Rate Mean % (SE)

WCMS		GMLC	LDA	RDA
PIMA	**76.57 (1.66)**	73.41 (2.25)	76.56 (1.52)	76.36 (1.53)
BREAST	**97.52 (0.43)**	95.00 (0.83)	96.07 (0.43)	96.51 (0.43)
HABERMAN'S	73.62 (2.52)	75.1 (2.42)	73.99 (2.39)	**75.28 (2.68)**
CONNECTIONIST	**77.79 (2.38)**	74.41 (2.71)	73.81 (3.57)	77.42 (2.44)
IONOSPHERE	**87.31 (1.65)**	86.85 (1.26)	85.19 (1.49)	84.91 (1.90)
SPECT	83.52 (2.00)	81.67 (2.77)	81.99 (3.13)	**83.94 (3.06)**
BUPA 61	.86 (2.07)	57.76 (2.66)	**67.29 (3.32)**	66.19 (3.97)
MAMMOGRAPHIC	**81.69 (0.84)**	80.00 (1.11)	81.33 (1.53)	80.48 (1.12)
TRANSFUSION	77.52 (1.73)	*	76.99 (1.35)	**78.04 (1.81)**

SE: Standard Error for Accuracy rate mean; *: algorithm does not run, (singular matrices).

5. CONCLUSION

Presented in this paper is a new classification algorithm: the Weighted Correlation Matrix Similarity (WCMS). By means of weighted intra-class correlation matrices and a similarity measure between matrices, WCMS assigns a class to new unknown cases (samples). Theoretically WCMS does not make previous assumptions of data, for instance imposing normality of distribution. Also, its applicability is insensitive to problems caused by small-sample settings or data collinearity since no matrix inversion is required. These characteristics give it a broad and more realistic practical application.

Differently from well known classification algorithms that deal with covariance structure, its process of classification benefits both from the training set and from information proportioned by the new sample to be classified. This last information has the capability of operationally changing, via a weighting step, the generalization arisen from the training set, thus impacting the classification results.

For better performance, the algorithm should be previously calibrated by means of a training set. We used a 10-fold cross validation process to calibrate the algorithm. Under certain circumstances, WCMS calibration through such a process may be time consuming and a proposed time-saving alternative would be to calibrate based only on two subsets (training and validation sets) or to augment the increments for the possible settings of r_h in the process of calibration.

WCMS was applied for data classification and its performance was compared with other widely used classification algorithms (RDA, LDA and GMLC, which are based on the data covariance structure) considering nine real datasets available in the UCI data repository. This data represents a range of different types of data dependence structure and dimensionality. The results showed that the performance of the WCMS algorithm was at least as competitive as any of the other tested methods. WCMS attained the greatest accuracy rate in five of the data sets. It was concluded that WCMS can be used as an effective classification tool in a wide range of data sets.

To conclude, today there is scope for algorithms that focus not only on the generalization from training data, but that also have a deep awareness of the information from the individual sample to be classified, as WCMS does in its weighting step, and which is its contribution towards that. Thus, the weighting phase of the algorithm is itself a scientific issue, with infinite possibilities of calculations and further research for better performance.

ACKNOWLEDGEMENT

The author would like to thank the UCI Machine Learning Repository and the data donors for putting real data sets at the disposal of the scientific community, and would also wish to thank the R foundation for Statistical Computing and its contributors for developing and making the R program available to the public. The Breast Cancer Wisconsin (Original) Data Set was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg, and the author would like to thank him.

REFERENCES

Duda, R. O. and Hart, P. E., 1973. *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, USA.

Elter, M. et al., 2007. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical Physics*, Vol. 34, No. 11, pp. 4164-4172.

Frank, A. and Asuncion, A., 2010. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Friedman, J. H., 1989. Regularized Discriminant Analysis. *In Journal of the American Statistical Association*, Vol. 84, No. 405, pp. 165-175.

Guo, P. et al., 2008. A study of regularized Gaussian classifier in high-dimension small sample set case based on MDL principle with application to spectrum recognition. *In Pattern Recognition*, Vol. 41, pp. 2842 – 2854.

Halbe, Z. and Aladjem, M., 2007. Regularized mixture discriminant analysis. *In Pattern Recognition Letters*, Vol. 28, pp. 2104–2115.

Han, J. and Kamber, M., 2001. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco.

Hoffbeck, J. P. and Landgrebe, D. A., 1996. Covariance Matrix Estimation and Classification with Limited Training Data. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 7, pp. 763-767.

Ji, S. and Ye, J., 2008. Kernel Uncorrelated and Regularized Discriminant Analysis: A Theoretical and Computational Study. *In IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 10, pp. 1311-1321.

Licheng, J. et al., 2006. An organizational coevolutionary algorithm for classification. *In IEEE Transactions on Evolutionary Computation*, Vol. 10, No. 1, pp. 67–80.

Liu, J. et al., 2008. A study on three linear discriminant analysis based methods in small sample size problem. *In Pattern Recognition*, Vol. 41 , pp. 102-116.

Lu, H. et al., 2009. Uncorrelated Multilinear Discriminant Analysis With Regularization and Aggregation for Tensor Object Recognition. *In IEEE Transactions on Neural Networks*, Vol. 20, No. 1, pp. 103-123.

Lu, J., et al., 2003. Regularized discriminant analysis for the small sample size problem in face recognition. *In Pattern Recognition Letters*, Vol. 24, pp. 3079–3087.

Mangasarian, O. L. and Wolberg, W. H., 1990. Cancer diagnosis via linear programming. *SIAM News*, Vol. 23, No. 5, pp. 1-18.

Peng, J. et al., 2008. Discriminant Learning Analysis. *In IEEE Transactions on Systems, Man, and Cybernetics—PART B: Cybernetics*, Vol. 38, No. 6, pp. 1614-1625.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

Venables, W. N. and Ripley, B. D., 2002. *Modern Applied Statistics with S*. Springer-Verlag, New York, USA.

Xu, P. et al., 2009. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics and Data Analysis*, Vol. 53 , pp. 1674–1687.

Ye, J. et al., 2006. Feature Reduction via Generalized Uncorrelated Linear Discriminant Analysis. *In IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 10, pp. 1312-1321.

Yeh, I-C. et al., 2009. Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, Vol. 36, Issue 3, pp. 5866–5871.