Classifier Selection based on the Correlation of Diversity Measures: When Fewer is More

Fabio A. Faria, Jefersson A. dos Santos
Institute of Computing
University of Campinas (UNICAMP)
Campinas, São Paulo, Brazil
{ffaria,jsantos}@ic.unicamp.br

Sudeep Sarkar
Dept. of Computer Science and Engineering
University of South Florida (USF)
Tampa, FL, USA
sarkar@usf.edu

Anderson Rocha, Ricardo da S. Torres
Institute of Computing
University of Campinas (UNICAMP)
Campinas, São Paulo, Brazil
{anderson.rocha,rtorres}@ic.unicamp.br

Abstract—**The ever-growing access to high-resolution images has prompted the development of region-based classification methods for remote sensing images. However, in agricultural applications, the recognition of specific regions is still a challenge as there could be many different spectral patterns in a same studied area. In this context, depending on the features used, different learning methods can be used to create complementary classifiers. Many researchers have developed solutions based on the use of machine learning techniques to address these problems. Examples of successful initiatives are those dedicated to the development of learning techniques for data fusion or Multiple Classifier Systems (MCS). In MCS, diversity becomes an essential factor for their success. Different works have been using diversity measures to select appropriate high-performance classifiers, but the challenge of finding the optimal number of classifiers for a target task has not been properly addressed yet. In general, the proposed solutions rely on the *a priori* use of *ad hoc* strategies for selecting classifiers, followed by the evaluation of their effectiveness results during training. Searching by the optimal number of classifiers, however, makes the selection process more expensive. In this paper, we address this issue by proposing a novel strategy for selecting classifiers to be combined based on the correlation of different diversity measures. Diversity measures are used to rank pairs of classifiers and the agreement among ranked lists guides the classifier selection process. A fusion framework has been used in our experiments targeted to the classification of coffee crops in remote sensing images. Experiment results demonstrate that the novel strategy is able to yield comparable effectiveness results when contrasted to several baselines, but using much fewer classifiers.**

Keywords-**multiple classifier system; ensemble of classifiers; diversity measures; coffee crop recognition;**

I. Introduction

Remote sensing images (RSI) are widely used as data source in agricultural studies. The recognition of regions in the images, given a pattern of interest, is a major application. It is typically addressed by using supervised classification techniques [1], [2], [3], [4].

With the increase in sensor technology, large collections of RSIs have become available for research and new techniques have been proposed to process such data. Concerning the high-resolution images, region-based classification methods have become a trend in the literature [5].

Despite advances in sensors and computational techniques, the classification of agricultural regions is still a challenging task. A typical problem is the presence of multiple patterns in the same area of study. Coffee crops, for example, are usually cultivated in mountainous regions (for example, in Brazil) [2]. In these areas, shadows and distortions usually impact the quality of available spectral information. Moreover, the growing of coffee is not a seasonal activity, and, therefore, in the same region, there may be coffee plantations of different ages, which also affects the observed spectral patterns.

With all these particularities, it is essential to use several image descriptors to properly encode the different existing patterns [6]. Furthermore, the use of several classifiers is a suitable alternative, since their results may be complementary in the sense they may differ depending on the application and the used features [7].

Many ensemble techniques for remote sensing are reported in [8], [9]. More recently, Faria et al. [10], for example, proposed an innovative and effective ensemble system by using meta-learning based on support vector machines.

The use of diversity measures [11] has been observed in the literature as an important tool for identifying potential classifiers for fusion. These measures assess the degree of agreement or disagreement between classifiers. According to Kuncheva et al. [11], [12], the good performance of an ensemble system is related to the diversity between the learning methods involved. The authors studied different diversity measures as well as discussed their impacts on the final accuracy of ensemble systems.

In [13], many diversity measures are employed in an ensemble strategy for produce recognition. In [14], the authors used a single measure of correlation to assess the combination of descriptors at different scales of segmentation. They showed that it is possible to reduce the number of selected classifiers in an ensemble by selecting only the pairs with less correlation. In [15], the authors evaluated several Multiple Classifier Systems (MCS) in remote sensing and used different strategies to improve those systems (e.g., classifier ensemble approaches, pairwise and non-pairwise diversity measures, and some modified algorithms).

In this paper, we address the problem of selecting the most important classifiers for fusion by proposing a novel strategy for selecting classifiers to be combined based on the correlation of different diversity measures. Diversity measures are used to rank pairs of classifiers and the agreement among ranked lists guides the classifier selection process. We evaluate the use of the proposed classifier selection and fusion targeted to support RSI classification tasks. Experiment results show that we can improve the classifier selection process of the investigated framework towards developing an effective and lightweight approach for classifying RSIs.

The remainder of this paper is organized as follows. Section II presents related concepts necessary for understanding this paper. Section III describes a new strategy for selecting classifiers that has been implemented in a recently proposed framework for classifier fusion. Section IV shows the experimental protocol we devised to validate our work, while Section V discusses the results. Finally, Section VI presents the conclusions and future research directions.

II. Related Concepts

The following subsections describe related concepts necessary for the understanding of this paper.

A. Support Vector Machine (SVM)

Support Vector Machine is a machine learning method introduced in [16]. The goal is to construct an optimum separation hyperplane or set of hyperplanes, which can be used to separate an n-dimensional feature space. The hyperplane is calculated such that it maximizes the margin between two classes (the standard SVM is a two-class classifier). The margin can be seen as the minimum distance of one point of one class to the other. It can be interpreted as a separation measure between two classes and represents the separability degree between them (quality measure of classification). The points on borders between the classes are called support vectors. When it is not possible to find a linear separator for the classes, the data are mapped on-the-fly onto higher dimensional spaces through a non-linear mapping using the kernel trick [17]. The reason for choosing SVM in this work is that by using the kernel, SVMs gain flexibility in the choice of the form of the threshold separating the classes of interest, which do not need to be linear and even do not need to have the same functional form for all data. Also, SVMs deliver a unique solution, since the optimality problem is convex.

B. Pairwise Diversity Measures

Let $\mathcal{M}$ be a matrix 2×2 containing the relationship between a pair of classifiers with percentage of agreement. This relationship matrix $\mathcal{M}$ has the percentage of hit and miss for each classifier c_i and c_j. The value a is the percentage of images that both classifiers c_i and c_j classified correctly in the validation set. Values b and c are the percentage of images that c_j classified correctly but c_i missed and vice-versa. The value d is the percentage of images that both classifiers missed.

In [11], Kuncheva et al. presented several measures to assess diversity, considering pairs of classifiers. Following their work, in our experiments, we have used *Correlation Coefficient p* (COR), *Double-Fault Measure* (DFM), *Disagreement Measure* (DM), *Interrater Agreement k* (IA), and *Q-Statistic* ($QSTAT$). Those measures are defined as follows:

$$COR(c_i, c_j) = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \quad (1)$$

$$DFM(c_i, c_j) = d, \quad (2)$$

$$DM(c_i, c_j) = \frac{b+c}{a+b+c+d}. \quad (3)$$

$$QSTAT(c_i, c_j) = \frac{ad - bc}{ad + bc}, \quad (4)$$

$$IA(c_i, c_j) = \frac{2(ac - bd)}{(a+b)(c+d)+(a+c)(b+d)}, \quad (5)$$

The diversity is greater if the measures *Double-Fault Measure*, *Q-Statistic*, *Interrater Agreement* k, and *Correlation Coefficient* p are lower among pairs of classifiers c_i and c_j. In the case of the *Disagreement Measure*, the greater the measure, the greater the diversity [11].

C. The Classifier Fusion Framework

This section presents a framework for classifier selection and fusion, as devised in [13].

1) Overview: Given a visual classification problem, one has a set of characterization or description techniques (descriptors) and a set of learning methods that will be used to learn patterns from available instances for training in order to classify new and unseen instances.

Once one trains all necessary classifiers along with different image descriptors, the learned knowledge undergoes a selection process of the most relevant learning methods and descriptors to be combined by another learning method (meta-learning approach) aiming at selecting the most discriminative methods as well as boosting the classification performance at test time by selecting less, but more effective, classifiers.

The classifiers (one classifier is a tuple learning/descriptor) are selected in a selection process that uses diversity measures calculated at training time to show the degree of agreement/disagreement between involved classifiers pointing out the most interesting ones to be further used in a combination scheme.

Sections II-C2 and II-C3 present a formal description of a framework for classifier selection and fusion along with examples when necessary.

2) Formalization: Let $\mathcal{L}$ be any set of learning methods (e.g., Decision Tree, Naïve Bayes, kNN, etc.) and $\mathcal{F}$ be a set of image descriptors (e.g., Color Histogram). Suppose that classifiers are created by combining each available learning method with each image descriptor. For example, three classifiers could be created by combining the learning methods Decision Tree, Naïve Bayes, and kNN with the Color Histogram descriptor. Let $\mathcal{C}$ be the set of classifiers created by that combination, where $|\mathcal{C}| = |\mathcal{L}| \times |\mathcal{F}|$.

Let $\mathcal{S}$ be a set of images, where the class of $s_i \in S$ $(1 < i \leq |\mathcal{S}|)$ is known. The set $\mathcal{S}$ is used to construct both the training (T) and validation (V) sets, where $T \cup V = \mathcal{S}$ and $T \cap V = \emptyset$. As the scenario of interest is a supervised learning scenario, the actual classes for training and validation data points are known *a priori*.

Initially, all classifiers $c_j \in \mathcal{C}$ $(1 < j \leq |\mathcal{C}|)$ are trained on set T. Next, the outcomes of each classifier on the validation set V is computed and stored into a matrix M_V, where $|M_V| = |V| \times |\mathcal{C}|$.

In the following, M_V is used as input to select a set $\mathcal{C}^* \subset \mathcal{C}$ of classifiers that are good candidates to be combined. In this framework, diversity measures are employed to determine $\mathcal{C}^*$ (see Section II-C3). Note that a new matrix $M_V^* \subset M_V$ is created by using the selected classifiers in $\mathcal{C}^*$.

Given a new image I, one can use each classifier $c_k \in \mathcal{C}^*$ $(1 < k \leq |\mathcal{C}^*|)$ to determine the class of I, producing k outcomes. The k outcomes are used as input of a fusion technique (e.g., majority voting, SVM, etc.) that takes the final decision regarding the definition of the class of I. In the case of a fusion technique that requires prior training (e.g., SVM), M_V^* is used.

Figure 1 illustrates a framework for combining classifiers.

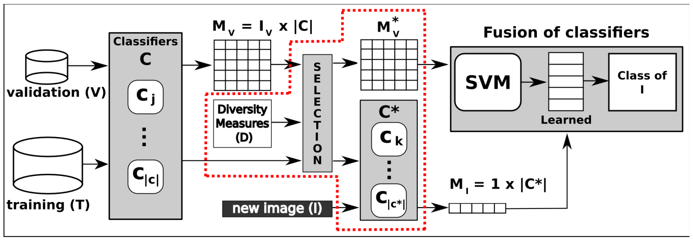


Figure 1. Framework for classifier selection and fusion [13]. Given a classification problem with training examples, one trains different classifiers and image descriptors and by means of diversity measures, he/she selects the most discriminative ones to be combined in a meta-level using any other classifier. Notice that, in this particular example, the SVM technique has been used for classifier fusion. The classifier selection process is delimited by dashed red line

3) Classifier Selection: Figure 2 illustrates the proposed five-step approach for selecting classifiers based on diversity measures [13].

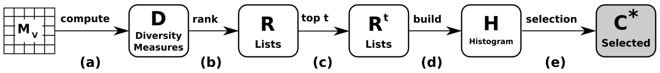


Figure 2. The five steps for classifier selection are: (a) Compute diversity measures from the validation matrix M_V; (b) $\mathcal{R}$ lists sorted by diversity measures scores; (c) $\mathcal{R}^t$ lists with top t; (d) counts the number of occurrences of each classifier that satisfy a defined threshold; (e) Selected classifiers $|\mathcal{C}^*|$.

Consider the previously defined $\mathcal{C}$ (set of classifiers) and M_V (a matrix such that $|M_V| = |V| \times |\mathcal{C}|$), containing the outcomes of classifiers $c_j \in \mathcal{C}$ from the validation set V.

Let $\mathcal{D}$ be a set of diversity measures. Each diversity measure $d_l \in \mathcal{D}$ is used to compute the agreement and disagreement between two classifiers $c_{j_n}, c_{j_m} \in \mathcal{C}$, considering all possible combinations of classifiers (arrow (a) in Figure 2).

Let $\mathcal{R}_{d_l} = \{(c_{j_n}, c_{j_m}), score_{d_l}(c_{j_n}, c_{j_m}), mean(c_{j_n}, c_{j_m})\}$, be a ranked list of pairs of classifiers defined by the score of the diversity measure d_l and the mean accuracy of pairs of classifiers (c_{j_n}, c_{j_m}).

Let $\mathcal{R} = \{\mathcal{R}_{d_1}, \mathcal{R}_{d_2} \dots \mathcal{R}_{d_{|\mathcal{D}|}}\}$ be the set of ranked lists defined for each available diversity measure. This process is illustrated by arrow (b). Let $\mathcal{R}^t$ be a set of ranked lists, where each ranked list contains the top t pairs of classifiers (t pairs of classifiers that are good candidates to be combined) – arrow (c), and $\mathcal{H}$ be a histogram that counts the number of occurrences of a classifier that satisfy the condition $mean(c_{j_n}, c_{j_m}) > T$ (where T is a threshold representing the average accuracy among all $mean(c_{j_n}, c_{j_m})$ of pairs of classifiers) in all ranked lists of $\mathcal{R}^t$ – arrow (d).

The set $\mathcal{C}^*$ of classifiers that are combined by the proposed fusion approach is the $h = |\mathcal{C}^*|$ most frequent classifiers in $\mathcal{H}$. This step is represented by arrow (e).

For more details about the classifier selection process, please refer to [13].

III. Classifier Selection Based on the Correlation of Multiple Diversity Measures

In this section, we expand upon previous work in the literature [13] and introduce a new strategy for guiding the selection of classifiers based on the *opinion* of multiple diversity measures.

A. Formalization

We propose to use multiple diversity measures to determine which classifiers should be combined. Our hypothesis is that by exploring complementary information provided by different diversity measures, more appropriate classifiers are selected to be combined.

Recall from Section II-B that a diversity measure indicates the agreement of pairs of classifiers. In that sense, different diversity measures would rank pairs of classifiers differently. Therefore, we propose to explore different strategies to select classifiers based on *correlation* scores among ranked lists of pairs of classifiers. Ranked lists are defined by different diversity measures.

1) Defining ranked lists of pairs of classifiers: As mentioned before, let $\mathcal{C}$ be the set of classifiers created by the combination of learning methods and image descriptors. Let $\mathcal{P} = \{p_1, p_2, \dots, p_{|\mathcal{C} \times \mathcal{C}|}\}$ be a set of all possible pairs of classifiers, i.e., $p_l = (c_i, c_j)$, where $(c_i, c_j) \in \mathcal{C} \times \mathcal{C}$.

Let $\mathcal{D} = \{d_1, d_2, \dots, d_{|\mathcal{D}|}\}$ be a set of diversity measures, such that each diversity measure $d_k \in \mathcal{D}$ defines a distance function $\rho : \mathcal{P} \to \mathbb{R}$, where $\mathbb{R}$ denotes real numbers. Eqs. 1–5 define different criteria for implementing the function ρ. Consider $\rho(p_l) \geq 0$ for all $p_l \in \mathcal{P}$ and $\rho(p_l) = 0$, with $p_l = (c_i, c_j)$, if $c_i = c_j$. The distance $\rho(p_l)$ among all pairs of classifiers $p_l = (c_i, c_j) \in \mathcal{C} \times \mathcal{C}$ can be computed to obtain a $|\mathcal{C}| \times |\mathcal{C}|$ distance matrix A.

Given a diversity measure $d_k \in \mathcal{D}$, we can compute a ranked list $\mathcal{R}_{d_l}$ by taking into account the distance matrix A. The ranked list $\mathcal{R}_{d_l}$=$\{p_1, p_2, \dots, p_{|\mathcal{C} \times \mathcal{C}|}\}$ (where $p_l = (c_i, c_j)$ is a pair of classifiers) can be defined as a permutation of the collection $\mathcal{P}$, such that, if p_l is ranked at lower positions than p_m, i.e., p_l is ranked before p_m, then $\rho(p_l) < \rho(p_m)$. In this way, pairs of classifiers are ranked according to their agreement score defined in terms of a diversity measure.

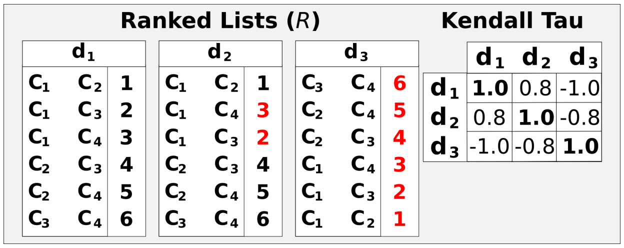
2) Measuring the correlation of ranked lists: We propose to exploit the correlation of ranked lists of pairs of classifiers to select the more appropriate ones to be combined. In this paper, we use the *Kendall* tau rank correlation coefficient (τ) to measure the degree of concordance between two different ranked lists of the same set of observed samples. We will use only the term '*Kendall*' for *Kendall* tau rank, thus avoiding possible confusion with 'tau' index (evaluation measure).

The *Kendall* correlation $\tau(\mathcal{R}_{d_i}, \mathcal{R}_{d_j})$ between two ranked lists $\mathcal{R}_{d_i}$ and $\mathcal{R}_{d_j}$ is defined in terms of the number of concordant pairs NC in $\mathcal{R}_{d_i}$ and $\mathcal{R}_{d_j}$, the number of discordant pairs ND, and the number of positions n in the ranked lists. Eq. 6 defines the *Kendall* correlation:

$$\tau(\mathcal{R}_{d_i}, \mathcal{R}_{d_j}) = \frac{NC - ND}{\frac{1}{2}n(n-1)}, \tag{6}$$

Figure 3 shows an example to illustrate the use of the *Kendall* correlation. In this example, we consider four classifiers c_1, c_2, c_3, and c_4 whose agreement is measured by means of three diversity measures (d_1, d_2, and d_3). Each diversity measure defines three ranked lists ($\mathcal{R}_{d_1}$, $\mathcal{R}_{d_2}$, and $\mathcal{R}_{d_3}$). We highlight in red the differences of $\mathcal{R}_{d_2}$, and $\mathcal{R}_{d_3}$ when compared to $\mathcal{R}_{d_1}$. Note that, in $\mathcal{R}_{d_2}$, just two pairs of classifiers are inverted. Pairs of classifiers in $\mathcal{R}_{d_3}$, in turn, are ranked in the inverse order, when compared to $\mathcal{R}_{d_1}$.

Figure 3 also shows in the table on the right side, the τ correlation scores among the three ranked lists. The correlation coefficient value $\tau(\mathcal{R}_{d_1}, \mathcal{R}_{d_2})$, as expected, is high, which means that ranked lists $\mathcal{R}_{d_1}$ and $\mathcal{R}_{d_2}$ have high degree of concordance. However, the correlation between ranked lists $\mathcal{R}_{d_1}$ and $\mathcal{R}_{d_3}$ is low (-1.0 stands for the lowest possible correlation score).



Ranked Lists (*R*)

d₁			d₂			d₃		
C₁	C₂	1	C₁	C₂	1	C₃	C₄	6
C₁	C₃	2	C₁	C₄	3	C₂	C₄	5
C₁	C₄	3	C₁	C₃	2	C₂	C₃	4
C₂	C₃	4	C₂	C₃	4	C₁	C₄	3
C₂	C₄	5	C₂	C₄	5	C₁	C₃	2
C₃	C₄	6	C₃	C₄	6	C₁	C₂	1

Kendall Tau

	d₁	d₂	d₃
d₁	**1.0**	0.8	-1.0
d₂	0.8	**1.0**	-0.8
d₃	-1.0	-0.8	**1.0**

Figure 3. Example of three computed ranked lists ($\mathcal{R}_{d_1}$, $\mathcal{R}_{d_2}$, and $\mathcal{R}_{d_3}$) and *Kendall* scores between them. Both ranked lists ($\mathcal{R}$) and *Kendall* are computed by using the validation matrix M_V (see Section II-C3).

B. A Novel Strategy for Selecting Classifiers

We propose a novel strategy, named *Kendall* classifier selection (KCS), to define appropriate classifiers to be used in the classification framework presented in [13]. KCS makes use of the degree of agreement of different diversity measures. This agreement is measured in terms of the *Kendall* correlation among ranked lists of classifiers, as presented in Section III-A.

Let d_{H_1} and d_{H_2} be the diversity measures with the highest correlation scores, which are defined by the *Kendall* correlation. Let $\mathcal{R}_{d_{H_1}}$ and $\mathcal{R}_{d_{H_2}}$ be the ranked lists of pairs of classifies defined by d_{H_1} and d_{H_2}, respectively. KCS defines the top-ranked pairs of classifiers in $\mathcal{R}_{d_{H_1}}$ and $\mathcal{R}_{d_{H_2}}$ as the most appropriate ones to be used in the classification framework presented in [13].

We also tested in our experiments selected classifiers defined in terms of the lowest correlated diversity measures (d_{L_1} and d_{L_2}). In this case, we use classifiers defined in the top-ranked positions of $\mathcal{R}_{d_{L_1}}$ and $\mathcal{R}_{d_{L_2}}$.

Figure 4 summarizes in six steps the new approach for selecting classifiers based on *Kendall* correlation.

Finally, it is important to highlight that all steps regarding the selection of classifiers for fusion are performed during the training phase of the decision-making framework. Using a validation set separated during training allows us to evaluate different descriptors and learning techniques, assess their outcomes when classifying the validation examples, and properly selecting, by means of the proposed *Kendall*-based methodology, the most suitable classifiers for deployment during testing.

IV. Experimental Methodology

This section presents our experimental goals and evaluation criteria, the used image dataset, image descriptors, learning methods, evaluation measures, and the validation protocol.

A. Experimental Goals and Evaluation Criteria

In these experiments, we aim at showing the performance of the selection process on the framework of selection and fusion regarding the best baselines of the literature.

This performance evaluation considers both effectiveness and efficiency aspects. Effectiveness analysis is based on accuracy, kappa, and tau results in coffee crop classification tasks. Efficiency analysis, in turn, is based on the number of classifiers on the framework to achieve the same effectiveness results than the best baselines.

Finally, statistical tests are performed to assess which classifier selection strategy yields the best effectiveness results while boosting efficiency.

B. Baselines

Five different diversity measures are considered in our study (see Section II-B). Two different baselines, each one with different strategies to select classifiers based on the available diversity measures, are considered: the *1-Diversity Classifier Selection (Single)*, and the *5-Diversity Classifier Selection (ALL)*. The 1-Diversity Classifier Selection (Single) refers to the use of only one diversity measure to define appropriate classifiers to be selected. The 5-Diversity Classifier Selection (ALL) considers the opinion of all available diversity measures.

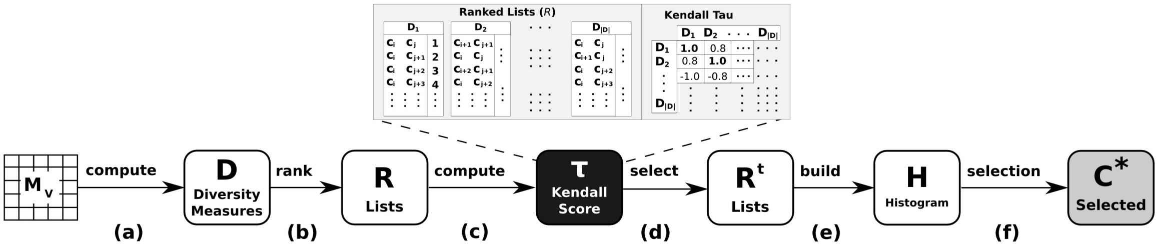


Figure 4. The six steps for new classifier selection are: (a) Compute diversity measures from the validation matrix M_V; (b) Sort $\mathcal{R}$ lists by diversity measures scores; (c) Compute *Kendall* correlation coefficients among all ranked lists of classifiers $\mathcal{R}$; (d) Select $\mathcal{R}_{d_{H_1}}$ and $\mathcal{R}_{d_{H_2}}$ or $\mathcal{R}_{d_{L_1}}$ and $\mathcal{R}_{d_{L_2}}$ ranked lists to be used in the next step; (e) $\mathcal{R}^t$ lists with top t; (f) Count the number of occurrences of each classifier that satisfy a defined threshold; (g) Select the most appropriate classifiers $|\mathcal{C}^*|$.

C. Dataset

In this paper, we consider $4,885$ images created via the method for multi-scale segmentation proposed by Guigues et al. [18], which separated into regions a SPOT satellite image of resolution $1,000 \times 1,000$ pixels. The SPOT satellite image corresponds to the Monte Santo de Minas county, in the State of Minas Gerais, Brazil, a traditional place of coffee cultivation. The region where this image was captured is mountainous. Therefore, the spectral patterns tend to be affected by the topographical differences and interference generated by the shadows. Another problem is that coffee is not a seasonal crop. Thus, in the same area, there may be crops of different ages. Concerning classification aspects, we have several completely different patterns representing the same class while some of these patterns are much closer to other classes.

To evaluate the accuracy, we use a ground truth that indicates all coffee regions in the image. As the experiments were performed with region level image and the ground truth is in pixel level, it was necessary to define a rule to label each region: if more than 80% of a region contains pixels of coffee, that region was labeled as “coffee”; otherwise it is a non-coffee region.

D. Image Descriptors

As we stated in Section I, there is no silver bullet to solve all image classification problems with just one machine learning classifier or even with just one image characterization technique. To choose the most appropriate descriptors is also a hard task.

Agricultural specialists usually perform analysis of agricultural targets by exploiting vegetation indices, such as NDVI [19]. With those indices, it is possible to estimate production and differentiate some objects in the surface.

Thus, in this work, the feature extraction algorithms are performed mainly on the bands corresponding to Red (R), Green (G) and Near-Infrared (NI). These bands are the most interesting for agricultural targets since are the basis for the computation of the main vegetation indices.

The used framework can consider a diverse set of classifiers and descriptors and point out the most interesting ones to solve a problem. In this sense, here we have used several image descriptors comprising color-, and texture-based methods. The used color descriptors include Border/Interior Pixel Classification (BIC) [20], Color Coherence Vector (CCV) [21], and Global Color Histogram (GCH) [22]. The used texture descriptors include Quantized Compound Change Histogram (QCCH) [23], Steerable Pyramid Decomposition (SID) [24], and Unser [25].

The criteria for choosing the image descriptors in this work are based on extensive experiments performed in [26], [6] pointing out to some of the most interesting image descriptors in the current computer vision literature.

E. Learning Methods

We used six learning methods in the framework of selection and fusion: Decision Tree (DT), Naïve Bayes (NB), Naïve Bayes Tree (NBT), and k-Nearest Neighbours (kNN) using $k = 1$, $k = 3$, and $k = 5$. Those methods are simple and fast, being suitable to be combined in a recognition system. The framework aims at automatically finding suitable combinations of classifiers formed by descriptors and learning methods. We have used the implementation of those learning methods available in the WEKA[1] data mining library. All learning methods were used with default parameters which means we did not optimize them whatsoever.

F. Evaluation Measures

In the experiments, we have calculated evaluation measures on the confusion matrix. The three evaluation measures are: accuracy, kappa [27], and tau [28] indices.

G. Cross-validation Protocol

We consider a k-fold cross-validation protocol for all experiments we perform. In this protocol, the original dataset is randomly separated into k non-overlapping subsets. A subset is chosen for testing set, and the $k - 1$ subsets are used for training a learning technique. The cross-validation process is repeated k times (rounds) and each subset is used only once as test set. The final result (the classification accuracy) from this process can be the arithmetic mean among all subsets. In the experiments, we have considered $k = 5$ fold cross-validation protocol. Each training set (consisting of four folds)

[1] http://www.cs.waikato.ac.nz/~ml/weka (As of May 2013).

can be further divided into validation and actual training (for instance, three folds can be used for training and the fourth for assessing the classifier being developed).

V. Results and Discussion

This section discusses the correlation analysis among diversity measures, behavior of each diversity measure into the selection process, and results regarding the effectiveness of the framework against different baselines from the literature.

A. Correlation Analysis of Diversity Measures

According to [13], the authors have used all five diversity measures together in the classifier selection process (Section II-C3). This experiment aims at showing that the use of these measures may provide different opinions about which classifiers can be better selected, thus potentially improving the quality of results in classification tasks.

Table I shows obtained results when using the well-known *Kendall* correlation scores [29] among ranked lists defined by different diversity measures.

Table I
Kendall score between five diversity measures.

Diversity Measures	COR	DFM	DM	IA	QSTAT
COR	**1.00**	0.14	0.87	-0.15	0.95
DFM	0.14	**1.00**	0.14	-0.09	0.15
DM	0.87	0.14	**1.00**	-0.15	0.88
IA	-0.15	-0.09	-0.15	**1.00**	-0.16
QSTAT	0.95	0.15	0.88	-0.16	**1.00**

As we can observe, the measures COR, DM, and $QSTAT$ have high correlation coefficients between them. DFM shows to be more correlated with COR, DM, and $QSTAT$ and less correlated with IA. In the case of IA, it has the lower correlation coefficients for all other measures analyzed which means it can be a very good candidate to consider when selecting diversity measures. Notice that none of the used measure is highly non-correlated with each other. This means that, although they are different diversity measures, all of them have an agreement degree about which classifiers should be combined.

As we can observe in Table I, $COR \times QSTAT$ has the highest correlation coefficient (in blue), while $IA \times QSTAT$ has the lowest (in red). Although it is not expensive to consider more diversity measures (their calculation requires simple operations with lists), ruling out some of them might be interesting.

Next section shows a study on combinations of diversity measures and impacts of these combinations in the selection process. In addition, we show how the methodology we propose in this paper can be used for selecting the most appropriate classifiers in a given problem.

B. Behavioral Analysis

This section shows three different analysis on the behavior of diversity measures in the classifier selection process in the framework proposed in [13]. First, we show an analysis of the diversity measures in isolation, followed by the combination of all diversity measures envisaged in Section II-B. Finally, we complete this study with the combination taking into account our novel approach (see Section III-B) which relies upon the *Kendall*'s scores defined in Table I.

Table II
Kappa index computed for all diversity measures using 5-folds cross-validation protocol for different number of classifiers ($|\mathcal{C}^*|$). Similar behaviors can be observed for other evaluation measures (accuracy and Tau index).

Type	Diversity Measures	Number of Classifiers $|\mathcal{C}^*|$					
		5	**10**	**15**	**20**	**25**	**30**
Single	**COR**	0.515	0.562	0.597	0.610	0.618	0.624
	DFM	0.490	0.540	**0.610**	**0.620**	0.620	**0.630**
	DM	0.507	0.549	0.577	0.611	0.622	0.622
	IA	0.557	0.579	0.601	0.606	0.607	0.613
	QSTAT	0.515	0.562	0.597	0.610	0.618	0.624
Baseline	**ALL**	0.472	0.553	0.592	0.610	**0.628**	0.623
Kendall	**COR+QSTAT**	0.515	0.562	0.597	0.610	0.618	0.624
	IA+QSTAT	**0.560**	**0.590**	0.594	0.615	0.618	0.617

For better understanding of this analysis, consider Figure 5. The x-axis denotes the number of classifiers $|C^*|$ than have been selected in the selection process. Notice that the values ranging from 5 to 36, where 5 is the lowest number of classifiers selected and 36 is the total amount of possible classifiers that can be selected (six descriptors and six learning methods result in 36 different classifiers). The y-axis denotes the classification effectiveness measured in terms of the kappa index.

In this figure, 'ALL' represents the baseline curve published in [13] that uses all of the five diversity measures into the selection process.

Figure 5-(a) shows a comparison between the SINGLE curves which use only one out of five diversity measures in the selection process against ALL curve. In this figure, it can be seen that for $|C^*| = \{5, 10\}$, the IA starts with the best kappa scores, but the more classifiers are available, the lower the effectiveness of this diversity measure. In fact, IA has the worst performance after 20 classifiers are considered. COR and $QSTAT$ curves have similar behavior. This experiment shows results in accordance with those in Table I. These measures show to have similar opinions and, therefore, are highly correlated. DFM is generally the best curve, showing to have the highest growth in the range $|C^*| \in \{5, \ldots, 15\}$, achieving the best kappa index when 30 classifiers are considered, i.e., $|C^*| = 30$. ALL starts with the worst kappa for $|C^*| = 5$, but has the highest growth for $|C^*| \in \{5, \ldots, 10\}$, and the best kappa for $|C^*| = 25$. This fact shows that the combination of different measures of diversity can be promising in order to obtain better results than using the measures in isolation.

Figure 5-(b) shows a comparison between the fusion of diversity measures with the lowest correlation coefficient (IA and $QSTAT$ — red scores in Table I), the fusion of the diversity measures with the highest coefficient (COR and $QSTAT$ — blue scores in the Table I), and ALL. ALL has achieved the worst kappa indices for $|C^*| = \{5, 20\}$, but it has achieved the best kappa for $|C^*| = 25$. $QSTAT$ combined with IA (IA+$QSTAT$) has achieved the best results for $|C^*| = \{5, 10\}$ and the worst for $|C^*| = \{25, 30\}$. This fact shows that $IA{+}QSTAT$ might be good for combining fewer classifiers, but not recommended when $|C^*|$ is large.

Table II shows kappa indices for all performed experiments. In bold, the best kappa index for each number of classifiers ($|C^*|$) defined in the selection process.

In Table II, we can see that there is no selection approach that achieves the best results for any number of classifiers. This illustrates the difficulty of automatically finding the optimal combination among diversity measures. The investigation of optimal combinations of diversity measures has showed to be a promising research venue and will be targeted in future work.

C. *Effectiveness Analysis*

In these experiments, seven fusion techniques are compared: the approach using SVM (FSVM-KERNEL-$|C|$) considering $|C| \in \{36, 49\}$, the approach considering fewer classifiers (FSVM-KERNEL-10) which effectively selects more promising learning and description methods using the *Kendall* method (combination of diversity measures $IA{+}QSTAT$, as showed in red, in Table II), Adaboost (BOOST-36), Bagging (BAGG-36), and Majority Voting (MV-36 and MV-49). Recall that FSVM-RBF-49 and MV-49 have been presented in [10].

Table III presents the results obtained for each fusion technique, considering three different evaluation measures (Accuracy, Kappa, and Tau). Notice that BOOST and BAGG techniques show up with the suffix 'ALL', which means the concatenation of the feature vectors produced by the six different image descriptors considered. Thus BAGG-ALL-36, BOOST-ALL-36 techniques are 36 iterations using six image descriptors.

Table III
CLASSIFICATION EFFECTIVENESS OF THE PROPOSED SELECTION FRAMEWORK AND BASELINES, WITH THEIR RESPECTIVE STANDARD DEVIATIONS.

Techniques	Accuracy	kappa	Tau
FSVM-RBF-49	89.09%±1.09	0.62±0.02	0.70±0.02
FSVM-NORM-36	89.09%±1.16	0.63±0.02	0.70±0.02
FSVM-NORM-10	87.88%±1.34	0.59±0.03	0.67±0.02
MV-49	88.50%±1.34	0.59±0.04	0.68±0.03
MV-36	89.13%±0.77	0.63±0.03	0.70±0.02
BAGG-*ALL*-36	88.33%±0.87	0.60±0.04	0.68±0.02
BOOST-*ALL*-36	89.46%±0.97	0.64±0.03	071±0.02

D. *Statistical Test of Significance (t-test)*

T-tests have been performed to verify the statistical significance of the results. In these tests, if the p-value is less than 0.05 (confidence of 95%) there is a significant difference between a pair of classifiers. If the p-value is greater than 0.05, there is no statistical difference.

Table IV shows a comparison among our approach (FSVM-NORM-10) with fewer classifiers selected using the methodology proposed in Section III (the result of the best selection process from Table II – $IA{+}QSTAT$, in red) against each one of the baselines.

Table IV
SIGNIFICANCE TESTS FOR OUR APPROACH AGAINST ALL BASELINES USED IN THE EXPERIMENTS.

Pair of Techniques	t-test	Significant
FSVM-NORM-10 × FSVM-RBF-49	0.0679	No
FSVM-NORM-10 × FSVM-NORM-36	0.0585	No
FSVM-NORM-10 × MV-49	0.3619	No
FSVM-NORM-10 × MV-36	0.1815	No
FSVM-NORM-10 × BAGG-ALL-36	0.5575	No
FSVM-NORM-10 × BOOST-ALL-36	0.0765	No

Notice that FSVM-NORM-10 has not statistical difference with any baseline. This means that our selection approach achieved similar results than all baselines but using far fewer classifiers. For example, FSVM-NORM-10 has used 10 classifiers against MV-36 that has used 36 to perform the same classification task. In addition, if all diversity measures are employed (ALL approach) as proposed in [13], 15 classifiers would be necessary to get the same result (kappa = 0.592, result in blue in Table II).

VI. FINAL REMARKS AND FUTURE WORK

This paper presented a novel strategy for selecting classifiers to be combined based on the *Kendall* correlation among different diversity measures. Those diversity measures were used to rank pairs of classifiers and the agreement of ranked lists was employed to guide the classifier selection process. In addition, we performed three different analysis with diversity measures in the classifier selection process of a classifier selection and fusion framework [13].

First, a correlation analysis using *Kendall* score has showed to be possible that different diversity measures have different opinions. In this experiment, we have showed that $COR \times QSTAT$ achieved the highest correlation coefficients, while $IA \times QSTAT$, the lowest . High correlation coefficients mean that both diversity measures have similar opinions about which classifiers might be selected. Low correlation coefficients in turn, mean that both diversity measure have a certain degree of divergence about which classifiers to select.

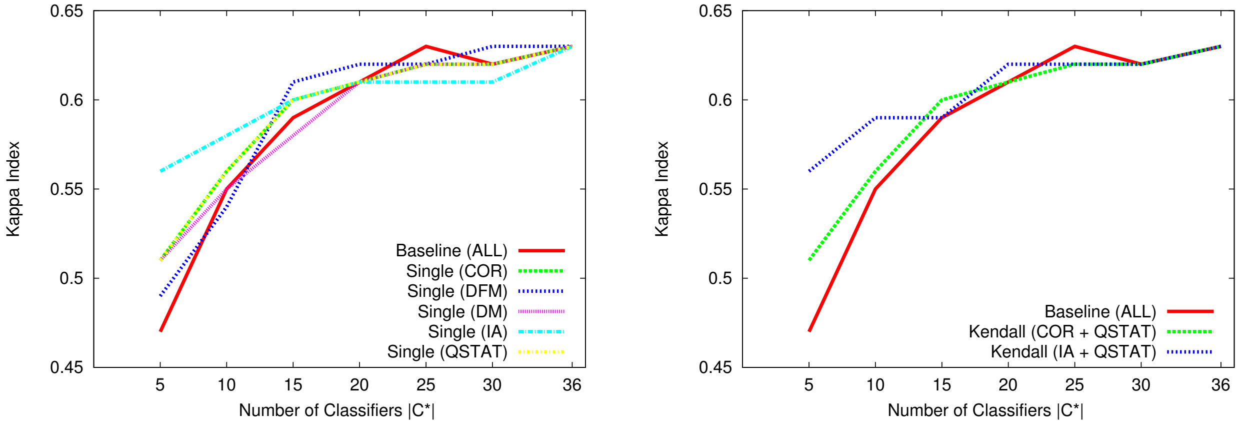
We also performed a behavioral analysis, based on which we showed two forms for selecting classifiers: (1) Single, which used only one diversity measure in the classifier selection process; (2) *Kendall*, which used two measures combined through *Kendall* correlation coefficients.

Finally, a comparison using the classifier accuracy has been performed using the best classifier selection approach that we could find using Table II by means of the use of diversity measures and the proposed methodology based on *Kendall*. The $IA + QSTAT$ approach has achieved the same results than all baselines using fewer classifiers than the original approach (ALL). Statistical tests have been performed to corroborate the claims.

As a future work, we plan to investigate additional strategies and metrics to automatically select the optimal diversity measures set for the classifier selection process, use non-pairwise diversity measures, and perform other experiments in other applications.

ACKNOWLEDGMENT

The authors are grateful to CAPES (grant 1260-12-0), CNPq (grants 306580/2012-8, 484254/2012-0, and 304352/2012-



(a) Single.

(b) *Kendall*.

Figure 5. Behavioral analysis of each diversity measures in the classifier selection process.

8), São Paulo Research Foundation – FAPESP (grants 2010/14910-0, 2010/05647-4, and 2012/18768-0), and Microsoft Research for the financial support.

References

[1] Castillejo-González, López-Granados, García-Ferrer, Peña-Barragán, Jurado-Expósito, de la Orden, and González-Audicana, "Object- and pixel-based analysis for mapping crops and their agro-environmental associated measures using quickbird imagery," *Elsevier Computer and Electronics in Agriculture*, 2009.

[2] J. A. dos Santos, F. A. Faria, R. Calumby, R. da S Torres, and R. Lamparelli, "A genetic programming approach for coffee crop recognition," in *IEEE Geoscience and Remote Sensing Symposium*, 2010.

[3] R. Pouteau and B. Stoll, "Svm selective fusion (self) for multi-source classification of structurally complex tropical rainforest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 4, pp. 1203 –1212, aug. 2012.

[4] D. C. Duro, S. E. Franklin, and M. G. Dubé, "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using spot-5 hrg imagery," *Remote Sensing of Environment*, vol. 118, no. 0, pp. 259 – 272, 2012.

[5] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2010.

[6] J. A. dos Santos, O. A. B. Penatti, and R. da S. Torres, "Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification," in *Intl. Conf. on Computer Vision Theory and Applications*, Angers, France, May 2010, pp. 203–208.

[7] A. Rocha, J. Papa, and L. Meira, "How far you can get using machine learning black-boxes," in *Conf. on Graphics, Patterns and Images*, 30 2010-sept. 3 2010, pp. 193 –200.

[8] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *Intl. Journal of Remote Sensing*, vol. 28, no. 5, pp. 823–870, 2007.

[9] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247 – 259, 2011.

[10] F. A. Faria, J. A. dos Santos, R. d. S. Torres, A. Rocha, and A. X. Falcão, "Automatic fusion of region-based classifiers for coffee crop recognition," in *IEEE Geoscience and Remote Sensing Symposium*, 2012.

[11] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, 2003.

[12] C. A. Shipp and L. I. Kuncheva, "An investigation into how adaboost affects classifier diversity," *IPMU*, pp. 203–208, 2009.

[13] F. A. Faria, J. A. Santos, A. Rocha, and R. d. S. Torres, "Automatic classifier fusion for produce recognition," in *Conf. on Graphics, Patterns and Images*, 2012.

[14] J. A. dos Santos, F. A. Faria, R. d. S. Torres, A. Rocha, P.-H. Gosselin, S. Philipp-Foliguet, and A. Falcao, "Descriptor correlation analysis for remote sensing image multi-scale classification," in *Intl. Conf. on Pattern Recognition*, 2012, pp. 3078–3081.

[15] P. Du, J. Xia, W. Zhang, K. Tan, Y. Liu, and S. Liu, "Multiple classifier system for remote sensing image classification: a review," *Sensors*, vol. 12, no. 4, p. 4764, 2012.

[16] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Workshop on Computational Learning Theory*, ser. COLT '92, 1992, pp. 144–152.

[17] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[18] L. Guigues, J. Cocquerez, and H. Le Men, "Scale-sets image analysis," *Intl. Journal of Computer Vision*, 2006.

[19] R. B. Myneni, F. G. Hall, P. J. Sellers, and A. L. Marshak, "The interpretation of spectral vegetation indexes," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 33, no. 2, pp. 481–486, 1995.

[20] R. Stehling, M. Nascimento, and A. Falcao, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *ACM Conf. on Information and Knowledge Management*, 2002, pp. 102–109.

[21] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *ACM Intl. Conf. on Multimedia*, 1996, pp. 65–73.

[22] M. Swain and D. Ballard, "Color indexing," *Intl. Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[23] C. Huang and Q. Liu, "An orientation independent texture descriptor for image retrieval," in *Intl. Conf. on Communications, Circuits and Systems*, 2007, pp. 772–776.

[24] J. Zegarra, N. Leite, and R. Torres, "Wavelet-based feature extraction for fingerprint image retrieval," *Journal of Computational and Applied Mathematics*, vol. 227, no. 2, pp. 294–307, 2008.

[25] M. Unser, "Sum and difference histograms for texture classification," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 8, no. 1, pp. 118–125, 1986.

[26] O. A. B. Penatti, E. Valle, and R. d. S. Torres, "Comparative study of global color and texture descriptors for web image retrieval," *Journal of Visual Communication and Image Representation*, 2012.

[27] R. L. Brennan and D. J. Prediger, "Coefficient kappa: Some uses, misuses, and alternatives," *Educational and psychological measurement*, 1981.

[28] Z. Ma and R. L. Redmond, "Tau coefficients for accuracy assessment of classification of remote sensing data." *Photogrametric Engineering and Remote Sensing*, 1995.

[29] M. G. Kendall, "A New Measure of Rank Correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, Jun. 1938.