The Computer Journal Advance Access published September 19, 2014

doi:10.1093/comjnl/bxu085

On Classification with Pairwise Comparisons, Support Vector Machines and Feature Domain Overlapping

Ryszard Janicki[1,*] and Mohammad Hadi Soudkhah[2]

[1]*Department of Computing and Software, McMaster University, Hamilton, ON, Canada L8S 4K1*
[2]*UXP Systems Inc., Toronto, ON, Canada M2N 7E9*
[*]*Corresponding author: janicki@mcmaster.ca*

Most existing classification algorithms either consider all features as equally important (equal weights), or do not analyze the consistency of weights assigned to features. When features are not equally important, assigning *consistent* weights is not an obvious task. In general, we have two cases. The first case assumes that a given sample of data *does not contain* any clues about the importance of features, so *the weights are provided by a pool of experts* and they are usually inconsistent. The second case assumes that the given sample *contains* some information about feature importance, hence we *can derive the weights directly from the sample*. In this paper, we deal with both cases. Pairwise comparisons and weighted support vector machines (SVMs) are used for the first case. For the second case, a new approach based on the observation that *the feature importance could be determined by the discrimination power of features* has been proposed. For the first case, we start with pairwise comparisons to rank the importance of features, then we use distance-based inconsistency reduction to refine the weight assessment and make the comparisons more precise. Next, we calculate the weights through the fully consistent or almost consistent pairwise comparison tables. For the second case, a *novel concept of feature domain overlappings* has been introduced. It can measure the feature discrimination power. This model is based on the assumption that less overlapping means more discriminatory ability, and this can be used to calculate weights characterizing the importance of particular features. For both cases, weighted SVMs are used to classify the data. Both methods have been tested using two benchmark datasets, Iris and Vertebral. The results were especially superior to those obtained without weights.

Keywords: classification; weighted features; pairwise comparisons; distance-based inconsistency; support vector machines; feature domain overlapping

Received 30 September 2013; revised 30 July 2014
Handling editor: Mohamed Jmaiel

1. INTRODUCTION

Classification of multifeature objects has many practical applications, and there are many well-known classification methods. Some methods assume that all features are equally important, other assign some weights to features.

Support vector machines (SVMs) were introduced by Vapnik in 1995 [1], and provide a framework for many efficient classification techniques. In principle, SVMs are supervised learning models [2] with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. They can work with small training sets, non-linear, high dimension learning problems and provide stable classifiers with high accuracy. When feature weights are added, we have *weighted SVMs* that have been proposed by Suykens *et al*. 2002 [3] and later used extensively [4–6]. The weights usually improve accuracy, especially for small training sets [5].

When features are not equally important, assigning *consistent* weights is not an obvious task. Weight assignment is a data preprocessing task for classification and its importance should not be underestimated [7]. In general, we have two cases. The first case assumes that a given sample of data *does not contain*

any clue about the importance of features, so *the weights are provided by a pool of experts* and they are usually inconsistent. The second case assumes that the given sample *contains* some information about feature importance, hence we *can derive the weights directly from the sample*.

In this paper, we deal with both cases. For the first case, we will use *pairwise comparisons paradigm*.

The *pairwise comparisons* method is based on the observation that it is much easier to rank the importance of *two* objects than it is to rank the importance of *several* objects. This very old idea goes back to Ramon Llull in the end of XIII century. Its modern version is due to the influential 1785 paper by Marquis de Condorcet, where he used this method in an election process where voters rank candidates based on their preferences, and the 1860 paper by Fechner. However, it was Thurstone in 1927 and Saaty in 1977 who provided the mathematical foundations that allowed this method to be effectively used in multicriteria decision making and analysis (see [8–10] for detailed references). Moreover, pairwise comparisons allow one to tackle the problem of inconsistent weights which has been so far neglected in SVM-based models.

In this paper, we start with pairwise comparisons to rank the importance of the features, then we use distance-based inconsistency reduction to refine the weight assessment and make the comparisons more precise. As the next step, we calculate the weights through the fully consistent or almost consistent pairwise comparison tables.

The proposed solution to the second case is based on the observation that *feature importance could be determined by the discriminatory power of features*.

This solution starts with the introduction of the *novel* concept of *feature domain overlappings*. This concept, also influenced by pairwise comparisons, allows one to measure the discriminatory power of a feature.

The basic idea is based on the observation that *less overlapping means more discriminatory power* and consequently bigger weights. The technical problem is how to measure this discriminatory power and how to derive the weight values from it. Our approach uses the pairwise comparisons paradigm and fundamental properties of geometric and arithmetic means [11–16].

When the weights have been calculated, either by pairwise comparisons or feature domain overlappings, weighted SVMs were used to classify the data. Both methods have been tested using two benchmark datasets, Iris [17] and Vertebral [18]. The results were especially superior to these obtained without weights.

This paper is a substantially extended and refined version of [19]. In many parts, it is based on the results of Soudkhah's Master Thesis [20], however, it also contains entirely new results that are neither in [20] nor [19], especially in Section 6.

The paper is structured as follows. In the next section, we briefly discuss the basic ideas and techniques of pairwise comparisons. SVMs and weighted SVMs are presented in Section 3. Section 4 contains the results of the weight assignment for the Iris and Vertebral datasets using pairwise comparisons, while Section 5 shows the results of classifications when the results from Section 4 were used. Section 6 contains the main contribution of this paper, a novel method of calculating weights by measuring the overlappings of feature values. The classification results with weights provided by overlappings are given and analyzed in Section 7. Section 8 contains final comments.

2. PAIRWISE COMPARISONS METHOD

Let $C_1, \ldots, C_n$ be a finite set of objects to be judged and/or analyzed. These objects are usually called *criteria*, *alternatives*, *attributes*, etc. In this paper, we will call them *features*. The first step in pairwise comparisons is to establish the relative preference or relationship of two features. This relationship may be *qualitative* (*relational*) [8, 13] or *quantitative* (*numerical*) [14, 21].

2.1. Quantitative model

A quantitative relationship between features C_i and C_j is represented by the number a_{ij}. We assume $a_{ij} > 0$ and $a_{ij} = 1/a_{ji}$, for $i, j = 1, \ldots, n$ (which implies $a_{ii} = 1$ for all i). If $a_{ij} > 1$, then C_i is more important (preferred, better, etc.) than C_j and a_{ij} is a measure of this relationship (bigger a_{ij} implies bigger difference), if $a_{ij} = 1$, then C_i and C_j are indifferent. The matrix of such relative comparison coefficients, $A = [a_{ij}]_{n\times n}$, is called a *pairwise comparison matrix*.

Since the features $C_1, \ldots, C_n$ are not random (on the contrary, they are usually carefully chosen and interrelated) the values of a_{ij} are not random, and they should be somehow *consistent*.

A pairwise comparison matrix $A = [a_{ij}]_{n\times n}$ is *consistent* [21] if and only if

$$a_{ij}a_{jk} = a_{ik}, \tag{1}$$

for $i, j, k = 1, \ldots, n$. Saaty's Theorem [21] states that a pairwise comparison matrix A is consistent if and only if there exist positive numbers $w_1, \ldots, w_n$ such that $a_{ij} = w_i/w_j$, $i, j = 1, \ldots, n$. The values w_i are unique up to a multiplicative constant. They are called *weights*, interpreted as a measure of importance and often scaled to $w_1 + \cdots + w_n = 1$ (or 100%).

In practice, the values a_{ij} are very seldom consistent, so some measurements of inconsistency are needed. Saaty [21] proposed an inconsistency index based on the value of the largest eigenvalue of A. However, this method does not give any clue where most inconsistent values of A are located [8, 14, 22], so we will not use it.

In [14], Koczkodaj provided an inconsistency index based on the analysis of all triads a_{ij}, a_{jk} and a_{ik} from $A = [a_{ij}]_{n\times n}$.

It is now called *distance-based inconsistency* and it is defined as follows:

$$cm_A = \max_{(i,j,k)} \left(\min \left(\left| 1 - \frac{a_{ij}}{a_{ik}a_{kj}} \right|, \left| 1 - \frac{a_{ik}a_{kj}}{a_{ij}} \right| \right) \right). \quad (2)$$

In this case, the most inconsistent triad is localized, which helps in the process of inconsistency reduction [22]. We will use this index for our purposes in this paper.

Acceptable levels of inconsistency depend on the inconsistency index definition and particular interpretation of $C_1, \ldots, C_n$. The acceptable values for the inconsistency index based on the value of the largest eigenvalue [21] are different from those for the distance-based consistency index cm_A used in this paper. The value of cm_A is based on triad analysis, so taking this into account and using heuristics similar to those used in statistics to justify P-values, it was argued in [9, 14] that a threshold of 0.3 is a reasonable assumption if no other factors are known. The results of the random experiments in [23] also support this threshold. It is highly doubtful that any threshold can ever be 'set in stone' as it represents our level of ignorance (or lack of the precise knowledge) and as such depends on the particular application [10, 14, 22].

Removing inconsistencies or lowering them to an acceptable level is a kind of art when the eigenvalue-based inconsistency index is used. When distance-based inconsistency index (Equation (2)) is used, since the biggest 'troublemakers' are localized, we can improve consistency step by step, by small changes of values of the triple that results in the maximal inconsistency index. It was proved in [9] that this process converges. In most cases, it converges very fast initially, and since there is no practical reason to continue decreasing the inconsistency indicator to zero, a matrix with acceptable level of inconsistency or even full consistency, can be found in a small number of steps by following common sense heuristics.

There are several approaches for deriving a suitable value w_i from an inconsistent, but with acceptable level of inconsistency, matrix A. Most often the weights $w_1, \ldots, w_n$ are either defined as the principal eigenvector of the matrix A (proposed in [21]), or as the geometric means of columns (or equivalently, rows) of the matrix A, i.e.

$$w_i = \sqrt[n]{\prod_{j=1}^{n} a_{ij}}, \quad (3)$$

for $i = 1, \ldots, n$ (proposed in [24]). Geometric means are used in this paper since they are easier to calculate and more intuitive in our opinion. For small values of cm_A, the differences between the two methods are negligible.

2.2. Qualitative pairwise comparisons

When mostly subjective judgment is involved, providing quantitative relationships between two entities is usually difficult and almost always controversial. It is not easy to justify statements like 'C_i is 1.5 times better than C_j'. It is much more convincing and trustworthy just to provide a *qualitative* assessment like 'C_i is much better than C_j' or 'C_i is only slightly better than C_j', etc.

TABLE 1. Non-linear comparison scale proposed in [8].

Value and range of a_{ij}		Relation	Definition of intensity or
Range	Default value	symbol R_{ij}	importance (C_i vs. C_j)
1.00–1.27	1	$C_i \approx C_j$	Indifferent
1.28–1.94	1.6	$C_i \sqsupset C_j$	Slightly in favor
1.95–3.17	2.6	$C_i \supset C_j$	In favor
3.18–6.14	4.7	$C_i > C_j$	Strongly better
6.15–	7.0	$C_i \succ C_j$	Extremely better

A comprehensive theory of qualitative pairwise comparisons has been proposed in [13, 25]. Instead of numerical coefficients a_{ij}, the binary relations R_{ij}, denoted as $\approx$, $\sqsubset$, $\subset$, $<$ and $\prec$ and are used. The interpretation of these relations is given in the two right columns of Table 1.

The number of relations has been limited to five because of the known restrictions of human mind[1] when it comes to subjective judgments [27, 28].

The relations $\approx, \sqsubset, \subset, <, \prec, \sqsupset, \supset, >, \succ$ are disjoint and cover the all cases, i.e. for every C_i, C_j we have $C_i \triangle C_j$ where $\triangle$ is one from $\approx, \sqsubset, \subset, <, \prec, \sqsupset, \supset, >, \succ$. The relation $\approx$ is symmetric and includes identity. The relations $\sqsupset, \supset, >, \succ$ are the inverses of $\sqsubset, \subset, <, \prec$.

The model of [13, 25] considers two sets of such relations $\mathcal{R}_d = \{\approx_d, \sqsubset_d, \subset_d, <_d, \prec_d\}$, called *ranking data*, and $\mathcal{R}_s = \{\approx_s, \sqsubset_s, \subset_s, <_s, \prec_s\}$, called *ranking system*.

The ranking data $\mathcal{R}_d$ is created from expert reports, so no reasonable consistency in any sense is expected, for example, the case $C_i < C_j < C_k < C_i$ *might* happen. For $\mathcal{R}_d$, the only requirements are covering the whole space, disjointedness and symmetry of $\approx$.

For the ranking system $\mathcal{R}_s$, it is additionally assumed that the relations[2] $\hat{\prec}_s = \prec_s$, $\hat{<}_s = \prec_s \cup <_s$, $\hat{\subset}_s = \prec_s \cup <_s \cup \subset_s$ and $\hat{\sqsubset}_s = \prec_s \cup <_s \cup \subset_s \cup \sqsubset_s$ are *partial orders*, i.e. irreflexive and transitive relations (cf. [29]). Quite often it is also required the relation $\approx_s$ to be an equivalence relation, i.e. $\hat{\sqsubset}_s$ is a *weak order* (see [13, 25, 29] for details).

Consistency in pairwise comparison-based models means that the relationships of C_i vs. C_j and C_j vs. C_k influence the relationship of C_i vs. C_k. For quantitative pairwise comparisons, it is given by the formula $a_{ij} \cdot a_{jk} = a_{ik}$.

For the qualitative ranking system $\mathcal{R}_s$, the *qualitative consistency* is defined by a set of axioms it must satisfy. The

[1] There are also some mathematical results supporting smaller scales [26].

[2] These relations are interpreted as follows: $\hat{\sqsubset}_s$ means *at least* slightly in favor, $\hat{\subset}_s$—*at least* in favor, etc.

number of axioms is substantial (45 in the version of [25]) as all combinations of all relational compositions must be taken care of. We will not present these axioms in this paper, the full set can easily be found in [8, 25], however, the idea on which all those axioms are constructed is very simple:

> composition of relations should be relatively continuous and must not change preferences in a drastic way.

Consider the following composition of preferences:

$a \approx b \wedge b \sqsubset c$. What relationships between a and c are consistent? Intuitively, $a \approx c$ and $a \sqsubset c$ are for sure consistent, $a \subset c$ is debatable, while $a \prec c$ and $a \prec\!\!\prec c$ are definitively inconsistent. This reasoning leads to [25, Axiom 2.1]:

2.1 $(a \approx b \wedge b \sqsubset c) \vee (a \sqsubset b \wedge b \approx c) \implies (a \approx c \vee a \sqsubset c \vee a \subset c)$.

There are two efficient algorithms that start with arbitrary ranking data $\mathcal{R}_d$ and derive consistent ranking systems $\mathcal{R}_s$ [25].

2.3. 'Mixed' pairwise comparisons

In reality, when subjective judgments are involved, all pairwise comparisons start with some qualitative judgments, however, in most cases, this is not considered formally as a separate step (cf. [10, 14, 21]). In this paper, we will explicitly consider qualitative comparisons as a separate steps.

The following natural seven steps process has been used:

(1) Experts provide *qualitative* judgments using relations from two right columns of Table 1 (see Table 2 as an example).
(2) The results are verified for *qualitative consistency*. If they are qualitatively inconsistent, they are corrected (using an algorithm from [25], or just common sense) and sent back to the experts.
(3) *If the results are qualitatively consistent, the qualitative judgments are transformed into quantitative values.*
(4) If the pairwise comparisons matrix constructed in Step (3) has an inconsistency index cm_A that is too large (Equation (2)), it is transformed into one with an acceptable inconsistency index.
(5) *The pairwise comparisons matrix is next transformed into the qualitative matrix.*
(6) Both the final qualitative and quantitative matrix are sent back to the experts for an analysis.
(7) If the experts accept both matrices from Step (6), the weights are calculated using geometric means (Equation (3), otherwise we return to Step (1)).

Steps (3) and (5) are the most problematic, and likely they will always be the most problematic. *Our idea of qualitative–quantitative relationship is based on the assumption that any transformation in either way should preserve consistency.* The mutual relationship between quantitative and qualitative pairwise comparisons has been analyzed in [8], which provides the following result.

TABLE 2. *Iris dataset*: initial *qualitative* judgments of mutual relationship provided by experts.

	Sepal length	Sepal width	Petal length	Petal width
Sepal length	$\approx$	$\sqsupset$	$\prec$	$\prec$
Sepal width	$\sqsubset$	$\approx$	$\prec$	$\prec\!\!\prec$
Petal length	$\succ$	$\succ$	$\approx$	$\sqsubset$
Petal width	$\succ$	$\succ\!\!\succ$	$\sqsupset$	$\approx$

The relational symbols from Table 1 were used.

PROPOSITION 2.1 [8]. *If the matrix $[a_{ij}]_{n\times}$ is consistent and each a_{ij} is transformed into R_{ij} by using ranges from far left column of Table 1, then the resulting set of relations is consistent with respect to qualitative consistency of* [25].

Proposition 2.1 can immediately be used in Step (5). A dual theoretical result, that could immediately be used in Step (3) does not exist yet, however, random tests have shown that when the numbers from the two left columns of Table 1 are used, a qualitatively consistent ranking system is transformed into a pairwise comparisons matrix with an often acceptable consistency index.

The default values in first four rows of Table 1 are middle points of appropriate ranges, while the default value in the fifth row is an educated guess. The default values are to be used when no other data or information or 'feelings' are available.

3. SVMs AND WEIGHTED FEATURE SVMs

In principle, the SVMs method [1] separates data points via hyperplanes for classification problems. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Formally, we start with a two-class training set $\mathbf{T} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k)\}$, such that each $\mathbf{x}_i \in \mathbb{R}^n$, where $\mathbb{R}$ denotes real numbers, and each $y_i \in \{-1, +1\}$. Each $\mathbf{x}_i = (x_{i1}, \ldots, x_{in}) \in \mathbb{R}^n$ is a vector of values of n different *features*. We want to find the maximum-margin hyperplane, either linear or non-linear, that divides the points having $y_i = -1$ from those having $y_i = +1$. For all mathematical details regarding these linear and non-linear optimization problems, we refer the reader to [1, 6].

In the standard SVM method, it is assumed all the features of training samples have equal contributions to construct the optimal separating hyperplane. However, for certain real-world data sets, some features may be more relevant than others. Such

relevance is often modeled by adding appropriate weights and assuming that bigger weight means bigger relevance. SVMs with weighted features, or weighted feature SVMs (WFSVM) have been proposed in [3] and analyzed, among others, in [5, 6]. In principle, given a dataset, the feature weighting methods assign real-valued numbers to each feature of the dataset. The bigger the number is, the more relevance its corresponding feature possesses.

Let $\mathbf{w} = (w_1, \ldots, w_n)$ be a vector of *feature weights*, each w_i is an indicator of the relative importance of the feature i for the classification, $i = 1, \ldots, n$.

The fundamental difference between SVMs and WFSVMs is that in the latter case we use the vectors

$$\mathbf{y}_i = (w_1 x_{i1}, \ldots, w_n x_{in}) = \mathrm{diag}(\mathbf{w}) \cdot \mathbf{x}_i,$$

instead of the vectors $\mathbf{x}_i$ alone, in all appropriate optimization formulas (up to some slightly different but equivalent formulations of optimization problem, see [5, 6, 30] for details).

There are several methods that can be used to calculate the weights vector $(w_1, \ldots, w_n)$. The most known are based on the concepts of: values of deviations [5], degrees of standard deviation [31] and Shannon's information theory [6].

The problem with these methods is that they assume some objective, unrelated to human judgment, relationships between features that can be objectively measured (but perhaps with errors). However, in many cases the importance of features is based mainly on subjective judgments. For such cases, the use of pairwise comparisons paradigm as described in the previous section is probably more justified. This will be the first approach discussed in this paper. The second one is based on the observation that not all features have the same discrimnatory power. We can measure this by analyzing feature domain overlappings, and propose weights on the basis of such analysis.

It is often assumed that all $w_i \geq 0$, and $\sum_{i=1}^{n} w_i = 1$. While scaling sum of weights to one (or 100%) displays the importance of features in a very intuitive manner, it is not always the best input for WFSVM algorithms. Due to computation errors, too small and too big weights may result in loss of accuracy. The optimal range of weights depends on the nature of a problem and is often found by experiments and heuristics (cf. [30]). Hence, two equivalent sets of weights are often used, one scaled for the results presentation and another, unscaled, for calculations.

Both SVMs and WFSVMs were originally designed for binary classification. How to effectively extend them for multiclass classification is still an on-going research issue. The most popular methods are called *one-against-all*, *one-against-one* and *DAGSVM* (cf. [32, 33]).

For all our calculations, we used the non-linear LibSVM algorithm from [30] which uses *one-against-one* approach for multiclass problem. In principle, the one-against-one approach consists in constructing one SVM for each pair of classes.[3] Thus, for a problem with n classes, $n(n-1)/2$ SVMs are trained to distinguish the samples of one class from the samples of another class. The classification is then done according to the maximum voting, where each SVM votes for one class (see [32, 33] for details).

4. ASSIGNING WEIGHTS TO THE IRIS AND VERTEBRAL DATASETS BY PAIRWISE COMPARISONS

Two datasets *Iris* [17] and *Vertebral* [18] are used in this paper.

The *Iris*[4] dataset consists of 150 samples belonging to three classes : *Setosa*, 50 samples, *Versicolor*, 50 samples and *Virginica*, 50 samples. The Setosa class is linearly separable from the other two; the Versicolor and Virginica classes are not linearly separable from each other.

Each sample has four features: *Sepal width* in cm, *Sepal length* in cm, *Petal width* in cm and *Petal length* in cm.

The *Vertebral*[5] is a biomedical dataset that consists of 310 samples. Each sample represents a patient by six biomechanical attributes that were derived from the shape and orientation of the pelvis, and lumbar spine: *pelvic incidence* (*PI*), *pelvic tilt* (*PT*), *lumbar lordosis angle* (*LLA*), *sacral slope* (*SS*), *pelvic radius* (*PR*) and *grade of spondylolisthesis* (*GOS*).

Each sample belongs to one of the following three classes: *Disk Hernia*, 60 samples, *Spondylolisthesis*, 150 samples and *Normal*, 100 samples.

Tables 2–5 contain the numerical judgments for the mutual relationship of Iris features and the weights derived from those judgments, while Tables 6 and 7 contain the numerical judgments for the mutual relationship of Vertebral features and the weights derived from those judgments. In both cases, we start with the initial judgments assigned by experts, who have expertise in botany (for the Iris dataset), and medicine and anatomy (for the Vertebral dataset). The experts followed the suggestions given in Table 1. They started with qualitative judgment (Table 2 for the Iris dataset and Table 8 for the Vertebral dataset), then translated it into numerical scale using the scale from Table 1 (Table 3 for Iris dataset and Table 6 for Vertebral dataset).

In both cases, the initial judgments resulted in inconsistency levels that were slightly too larger, larger than the traditional acceptance level which is 0.3 (see Section 2.1). Then we modified the elements of the appropriate pairwise comparisons matrices to lower the inconsistency indexes cm_A. In all cases,

[3]The one-against-one approach could also be seen as another instance of the pairwise comparison paradigm.

[4]The Iris dataset, created by Fisher in 1936 [17], is the most famous dataset to be found in the pattern recognition literature.

[5]The Vertebral dataset has been built by H. da Mota during a medical residence period in the Group of Applied Research in Orthopaedics of the Centre Médico-Chirurgical de Réadaptation des Massues, Lyon, France [18].

TABLE 3. *Iris dataset*: initial *quantitative* judgments of mutual relationship provided by experts when the initial values from Table 1, and qualitative judgments from Table 2 were used.

	Sepal length	Sepal width	Petal length	Petal width
Sepal length	1.0	1.6	1/4.7	1/4.7
Sepal width	1/1.6	1.0	1/4.7	1/7
Petal length	4.7	4.7	1.0	1/1.6
Petal width	4.7	7.0	1.6	1.0
Feature weights	0.5188	0.3713	1.9276	2.6936
Scaled weights	9.4%	6.7%	35.0%	48.9%
Inconsistency coefficient $cm_A = 0.38$, >0.3				

TABLE 4. *Iris dataset*: consistent matrix of pairwise comparisons derived from Table 3.

	Sepal length	Sepal width	Petal length	Petal width
Sepal length	1.0	1.61	1/2.9	1/4.65
Sepal width	1/1.61	1.0	1/4.65	1/7.5
Petal length	2.9	4.65	1.0	1/1.61
Petal width	4.65	7.5	1.61	1.0
Feature weights	0.5878	0.3653	1.7012	2.7174
Scaled weights	10.9%	6.8%	31.6%	50.8%
Inconsistency coefficient $cm_A = 0.0$, i.e. matrix is consistent				

TABLE 5. *Iris dataset*: final *qualitative* judgments of mutual relationship derived from Table 4 using ranges from Table 1.

	Sepal length	Sepal width	Petal length	Petal width
Sepal length	$\approx$	$\sqsupset$	$\subset$	$<$
Sepal width	$\sqsubset$	$\approx$	$<$	$\prec$
Petal length	$\supset$	$>$	$\approx$	$\sqsubset$
Petal width	$>$	$\succ$	$\sqsupset$	$\approx$

The gray color indicates the difference from Table 2.

we used Equation (2) to calculate cm_A and in all cases we ended with $cm_A = 0$, i.e. full consistency. The weights were then calculated using the geometric means.

For the Iris dataset, the final matrix of qualitative judgments (Table 5) differs from the initial one (Table 2). Our experts were contacted and they accepted the final qualitative judgments. In Table tab5, the ratio of *Petal length* to *Sepal length* is 2.9, close to the border between the relation $\supset$ and the relation $>$ according to Table 1, but the values in Table 1 are not set in stone, they should only be considered as reliable suggestions.

TABLE 6. *Vertebral dataset*: initial *quantitative* judgments of mutual relationship provided by experts when the initial values from Table 1, and qualitative judgments from Table 8 were used.

	PI	PT	LLA	SS	PR	GOS
PI	1	1/2.6	1/2.6	1/4.7	1/4.7	1/7
PT	2.6	1	1	1/1.6	1/1.6	1/4.7
LLA	2.6	1	1	1/1.6	1/1.6	1/4.7
SS	4.7	1.6	1.6	1	1	1/2.6
PR	4.7	1.6	1.6	1	1	1/2.6
GOS	7	4.7	4.7	2.6	2.6	1
Feature weights	0.3139	0.7747	0.7747	1.2909	1.2909	3.1857
Scaled weights	3.5%	9.9%	9.6%	16.2%	16.4%	44.4%
Inconsistency coefficient $cm_A = 0.43$, >0.3						

TABLE 7. *Vertebral dataset*: consistent matrix of pairwise comparisons derived from Table 6.

	PI	PT	LLA	SS	PR	GOS
PI	1	1/2.8	1/2.8	1/4.76	1/4.33	1/13.45
PT	2.8	1	1	1/1.7	1/1.55	1/4.8
LLA	2.8	1	1	1/1.7	1/1.55	1/4.8
SS	4.76	1.7	1.7	1	1	1.1/2.82
PR	4.33	1.55	1.55	1/1.1	1	1/3.1
GOS	13.45	4.8	4.8	2.82	3.1	1
Feature weights	0.2678	0.8304	0.7558	1.2857	1.2857	3.5991
Scaled weights	3.3%	10.3%	9.4%	16.0%	16.0%	44.9%
Inconsistency coefficient $cm_A = 0$, i.e. matrix is fully consistent						

The table of *qualitative* judgments derived from this table is exactly the same as Table 8.

5. CLASSIFICATION OF IRIS AND VERTEBRAL DATASETS WITH WEIGHTS OBTAINED BY PAIRWISE COMPARISONS

The results of applying both SVM and WFSVM procedures (non-linear LibSVM algorithm from [30]) for the Iris and Vertebral Datasets are presented in Tables 9 and 10. All tables have the same structure. We start with a small set of training samples to evaluate abilities of learning with small training sets. The division of samples into training and testing sets was always done randomly. All test samples were completely independent of training sets. Number of support vectors (SVs) indicates the classification complexity, as it is proportional to the number of SVs.

The accuracy is defined as n_{cc}/n_{ts} where n_{cc} is the number of correct classifications while n_{ts} is the total number of test samples.

For both the Iris and Vertebral datasets, weights significantly decreased the number of SVs, for training sets of all sizes.

For the Iris dataset, the weights also substantially increased the accuracy but only for small training sets, while for the Vertebral dataset the accuracy was increased for all sizes of training sets. For both datasets, there was not much difference between fully consistent weights and weights with an inconsistency index that was slightly too large. This could be due to the fact that the threshold for this index is not set in stone. We used 0.3, as recommended in [9, 14, 23] for cases where no other information is given, however, there are known cases where even 0.5 may give reasonable results, and in our case the initial inconsistency indexes were 0.38 for the Iris dataset and 0.43 for the Vertebral dataset, so the similarity of results was not entirely unexpected.

Summing up, using the weights provided by pairwise comparisons has improved accuracy, especially for small training sets, and decreased the number of SVs.

TABLE 8. *Vertebral dataset*: initial *qualitative* judgments of mutual relationship provided by experts.

	PI	PT	LLA	SS	PR	GOS
Pelvic incidence = PI	$\approx$	$\subset$	$\subset$	$<$	$<$	$\prec$
Pelvic tilt = PT	$\supset$	$\approx$	$\approx$	$\sqsubset$	$\sqsubset$	$<$
Lumbar lordosis angle = LLA	$\supset$	$\approx$	$\approx$	$\sqsubset$	$\sqsubset$	$<$
Sacral slope = SS	$>$	$\sqsupset$	$\sqsupset$	$\approx$	$\approx$	$\subset$
Pelvic radius = PR	$>$	$\sqsupset$	$\sqsupset$	$\approx$	$\approx$	$\subset$
Grade of spondylolisthesis = GOS	$\succ$	$>$	$>$	$\supset$	$\supset$	$\approx$

The relational symbols from Table 1 were used.

6. CALCULATING WEIGHTS WITH OVERLAPPING OF FEATURE VALUES

In this section, we will present a new method for assigning weights to features, by analyzing and measuring their discriminatory power. The method will employ the ideas

TABLE 9. Classification results for *Iris dataset*.

Samples		Without weights		Weights from Table 3, $cm_A = 0.38$		Weights from Table 4, $cm_A = 0$	
Train.	Test.	No. SV	Acc. %	No. SV	Acc. %	No. SV	Acc. %
15	135	14	94.1	9	96.3	9	95.6
30	120	17	94.2	16	96.7	16	97.5
45	105	24	96.2	19	97.1	18	97.1
60	90	26	94.4	21	95.6	23	95.6
75	75	31	93.3	25	96.0	24	96.0
90	60	34	98.3	26	96.7	26	96.7
105	45	39	97.8	27	97.8	28	97.8
120	30	42	96.7	30	96.7	31	96.7
135	15	44	100.0	31	100.0	36	100.0
Averages		30.1	96.1	22.7	97.0	23.4	97.0

Gray cells indicate significantly better results (weights vs. no weights).

TABLE 10. Classification results for *Vertebral dataset*.

Samples		Without weights		Weights from Table 3, $cm_A = 0.43$		Weights from Table 4, $cm_A = 0.0$	
Train.	Test.	No. SV	Acc. %	No. SV	Acc. %	No. SV	Acc. %
30	280	28	67.5	28	75.7	28	75.4
60	250	51	71.2	47	80.0	47	79.2
90	220	72	76.8	61	82.7	61	82.7
120	190	92	80.0	79	84.2	79	85.3
150	160	107	78.8	89	86.3	89	86.3
180	130	130	75.4	102	80.8	102	81.5
210	100	139	85.0	112	91.0	112	93.0
240	70	153	88.6	120	92.9	120	92.3
270	40	162	85.0	126	92.5	126	92.5
Averages		103.8	78.7	84.9	85.1	84.9	85.4

Gray cells indicate significantly better results (weights vs. no weights).

of pairwise comparisons [13] and fundamental properties of modeling with arithmetic and geometric means [11, 12, 15, 16].

In what follows, we will use the following notation. By $\mathbb{R}$, we will denote the set of all real numbers, by $\mathbb{N}$ the set of all natural numbers and by $|A|$ the cardinality of a set A. We will write $\langle i_1, \ldots, i_m\rangle$ to denote the set $\{i_1, \ldots, i_m\} \subseteq \mathbb{N}$, such that $i_s \leq i_r \iff s \leq r$ for all $i_s, i_r \in \{i_1, \ldots, i_m\}$.

We will also write

$$\langle i_1, \ldots, i_m\rangle \subseteq \langle j_1, \ldots, j_l\rangle$$

if $\{i_1, \ldots, i_m\} \subseteq \{j_1, \ldots, j_l\}$, $i_s \leq i_r \iff s \leq r$ for all $i_s, i_r \in \{i_1, \ldots, i_m\}$, and $j_s \leq j_r \iff s \leq r$ for all $j_s, i_r \in \{j_1, \ldots, j_m\}$.

For example, $\langle 2, 4, 6\rangle \subseteq \langle 1, 2, 3, 4, 5, 6\rangle$, $\langle 3, 5\rangle \subseteq \langle 2, 3, 5, 8\rangle$, etc.

For every $p \in \mathbb{N}$, the Cartesian product $\mathbb{R}^p$ can be written as $\mathbb{R}^p = \mathbb{R}_1 \times \cdots \times \mathbb{R}_p$, where $\mathbb{R}_i = \mathbb{R}$ for $i = 1, \ldots, p$. Such notation allows us to state explicitly that a value x belongs to the ith dimension of $\mathbb{R}^p$, just by writing $x \in \mathbb{R}_i$. It also allows us to describe precisely any particular subproduct of $\mathbb{R}^p$, namely for every $\langle i_1, \ldots, i_m\rangle \subseteq \langle 1, \ldots, p\rangle$ we will define $\mathbb{R}^{\langle i_1,\ldots,i_m\rangle}$ as

$$\mathbb{R}^{\langle i_1,\ldots,i_m\rangle} = \mathbb{R}_{i_1} \times \cdots \times \mathbb{R}_{i_m},$$

where $\mathbb{R}_{i_j} = \mathbb{R}$ for $j = 1, \ldots, m$.

We can now define the standard projection on the set of dimensions $\langle i_1, \ldots, i_m\rangle \subseteq \langle 1, \ldots, p\rangle$ as a mapping $\pi_{i_1 i_2,\ldots,i_m} : \mathbb{R}^p \to \mathbb{R}^{\langle i_1,\ldots,i_m\rangle}$ such that for any $\mathbf{z} = (z_1, \ldots, z_p) \in \mathbb{R}^p$:

$$\pi_{i_1,\ldots,i_m}((z_1, \ldots, z_p)) = (z_{i_1}, \ldots, z_{i_m}) \in \mathbb{R}^{\langle i_1,\ldots,i_m\rangle}.$$

Suppose that we have a sample of n vectors $\mathbf{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, p features $F_1, \ldots, F_p$ (so $\mathbf{x}_i \in \mathbb{R}^p$ for each $i = 1, \ldots, n$), and k groups $\mathbf{G}_1, \ldots, \mathbf{G}_k$ such that $\mathbf{S} = \mathbf{G}_1 \cup \cdots \cup \mathbf{G}_k$ and $\mathbf{G}_i \cap \mathbf{G}_j = \emptyset$ if $i \neq j$.

The main motivation for the approach presented in this section is the observation that not all the features play the same role in the partition of $\mathbf{S}$ into $\mathbf{G}_1, \ldots, \mathbf{G}_k$, some features have more discriminatory power than others. *The more discriminatory power a feature has the more important this feature is*. If, for example, there is a feature F_{i_0} such that $\pi_{i_0}(\mathbf{G}_r) \cap \pi_{i_0}(\mathbf{G}_t) = \emptyset$ if $r \neq t$, then the feature F_{i_0} alone is sufficient to construct the partition $\mathbf{G}_1, \ldots, \mathbf{G}_k$ of $\mathbf{S}$. For example, if $\mathbf{S} = \{(1, 2, 2), (1, 2, 3), (3, 2, 3), (2, 2, 3), (2, 2, 4)\} \in \mathbb{R}^3$, $\mathbf{G}_1 = \{(1, 2, 2), (1, 2, 3)\}$ and $\mathbf{G}_2 = \{(3, 2, 3), (2, 2, 3), (2, 2, 4)\}$, the feature F_1 provides enough information to divide $\mathbf{S}$ into $\mathbf{G}_1$ and $\mathbf{G}_2$. We have here $\pi_1(\mathbf{G}_1) = \{1\}$ and $\pi_1(\mathbf{G}_2) = \{2, 3\}$, so if $\pi_1(x) = 1$, then $x \in \mathbf{G}_1$, and if $\pi_1(x) \in \{2, 3\}$, then $x \in \mathbf{G}_2$. A less abstract and more realistic example is given in the bottom part of Fig. 1 where we can see very little overlapping for feature Petal in the Iris dataset.

Nevertheless having plenty of overlapping for each particular dimension does not necessarily imply huge overlapping when two (or more) dimensions are considered together. This is illustrated in Fig. 2 where each dimension has almost 100% overlapping, but the 2D overlapping is rather very small.

We will provide some measures of feature importance based on their discriminatory power.

One problem with projections is that often for different $\mathbf{x}$ and $\mathbf{y}$ we have $\pi_{i_1,\ldots,i_m}(\mathbf{x}) = \pi_{i_1,\ldots,i_m}(\mathbf{y})$, but we still want to remember that these identical projections came from different sources.

Let $F_{i_1}, \ldots, F_{i_m}$ be a subset of features $F_1, \ldots, F_p$. For every $x = (x_{i_1}, \ldots, x_{i_m}) \in \mathbb{R}^{\langle i_1,\ldots,i_m\rangle}$ and every $\mathbf{A} \subseteq \mathbb{R}^n$, define a set $\mathsf{C}_{i_1,\ldots,i_m}(x, \mathbf{A})$ and an index $\mathsf{c}_{i_1,\ldots,i_m}(x, \mathbf{A})$ as

$$\mathsf{C}_{i_1,\ldots,i_m}(x, \mathbf{A}) = \{\mathbf{y} \in \mathbf{A} \mid \pi_{i_k}(\mathbf{y}) = x_{i_k}, k = 1, \ldots, m\}, \quad (4)$$

$$\mathsf{c}_{i_1,\ldots,i_m}(x, \mathbf{A}) = |\mathsf{C}_{i_1,\ldots,i_m}(x, \mathbf{A})|. \quad (5)$$

The set $\mathsf{C}_{i_1,\ldots,i_m}(x, \mathbf{A})$ is the set of all $\mathbf{x}$ in $\mathbf{A}$ with their i_jth coordinate equal to x_{i_j}, while the index $\mathsf{c}_{i_1,\ldots,i_m}(x, \mathbf{A})$ simply states how many $\mathbf{x}$ in $\mathbf{A}$ has the i_jth coordinate equal to x_{i_j}. For example, for $\mathbf{A} = \{(1, 2, 2), (1, 2, 3), (3, 2, 3), (2, 2, 3), (2, 2, 4)\} \subset \mathbb{R}^3$, we have $\mathsf{C}_{1,2}((1, 2), \mathbf{A}) = \{(1, 2, 2), (1, 2, 3)\}$ so $\mathsf{c}_{1,2}((1, 2), \mathbf{A}) = 2$, while $\mathsf{C}_{2,3}((1, 2), \mathbf{A}) = \{(1, 2, 3), (3, 2, 3), (2, 2, 3)\}$ so $\mathsf{c}_{2,3}((2, 3), \mathbf{A}) = 3$.

The below result shows expected monotonicity of the index $\mathsf{c}_{i_1,\ldots,i_m}(x, \mathbf{A})$.

Fact 6.1. If $\langle i_1, \ldots, i_m\rangle \subseteq \langle j_1, \ldots, j_l\rangle$ and $x \in \mathbb{R}^p$, then

$$\mathsf{c}_{i_1,\ldots,i_m}(\pi_{i_1,\ldots,i_m}(x), \mathbf{A}) \geq \mathsf{c}_{j_1,\ldots,j_l}(\pi_{j_1,\ldots,j_l}(x), \mathbf{A}).$$

Proof. Since $\mathsf{C}_{j_1,\ldots,j_l}(x, \mathbf{A}) \subseteq \mathsf{C}_{i_1,\ldots,i_m}(x, \mathbf{A})$. □

We extend the index $\mathsf{c}_{i_1,\ldots,i_m}$ to sets as follows, for each $B \subseteq \mathbb{R}^{\langle i_1,\ldots,i_m\rangle}$:

$$\mathsf{c}_{i_1,\ldots,i_m}(B, \mathbf{A}) = \sum_{x \in B} \mathsf{c}_{i_1,\ldots,i_m}(x, \mathbf{A}).$$

For $\mathbf{A}$ as above and $B = \{(1, 2), (2, 2)\} \subseteq \mathbb{R}^{(1,2)}$, we have $\mathsf{c}_{1,2}(B, \mathbf{A}) = 4$, as $\mathsf{c}_{1,2}((1, 2), \mathbf{A}) = 2$ and $\mathsf{c}_{1,2}((2, 2), \mathbf{A}) = 2$.

For two different groups $\mathbf{G}_r$ and $\mathbf{G}_t$, we define

$$\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m} = \mathsf{c}_{i_1,\ldots,i_m}(\pi_{i_1,\ldots,i_m}(\mathbf{G}_r) \cap \pi_{i_1,\ldots,i_m}(\mathbf{G}_t), \mathbf{G}_r), \quad (6)$$

$$\mathsf{c}^{(rt)t}_{i_1,\ldots,i_m} = \mathsf{c}_{i_1,\ldots,i_m}(\pi_{i_1,\ldots,i_m}(\mathbf{G}_r) \cap \pi_{i_1,\ldots,i_m}(\mathbf{G}_t), \mathbf{G}_t). \quad (7)$$

The index $\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m}$ indicates how many elements of the group $\mathbf{G}_r$ have exactly the same values in $\mathbb{R}^{\langle i_1,\ldots,i_m\rangle}$ as the elements of the group $\mathbf{G}_t$. Similarly, for the index $\mathsf{c}^{(rt)t}_{i_1,\ldots,i_m}$. For example, if $\mathbf{G}_1 = \{(1, 2, 1), (1, 2, 2), (1, 2, 3), (2, 2, 4), (1, 3, 2), (2, 3, 4)\}$ and $\mathbf{G}_2 = \{(1, 2, 4), (1, 3, 5), (2, 2, 3), (2, 2, 4), (1, 4, 3)\}$, then we have $\pi_{(1,2)}(\mathbf{G}_r) \cap \pi_{(1,2)}(\mathbf{G}_t) = \{(1, 2), (2, 2)\}$ and $\mathsf{c}^{(1,2)1}_{(1,2)} = 4$, $\mathsf{c}^{(1,2)2}_{(1,2)} = 3$.

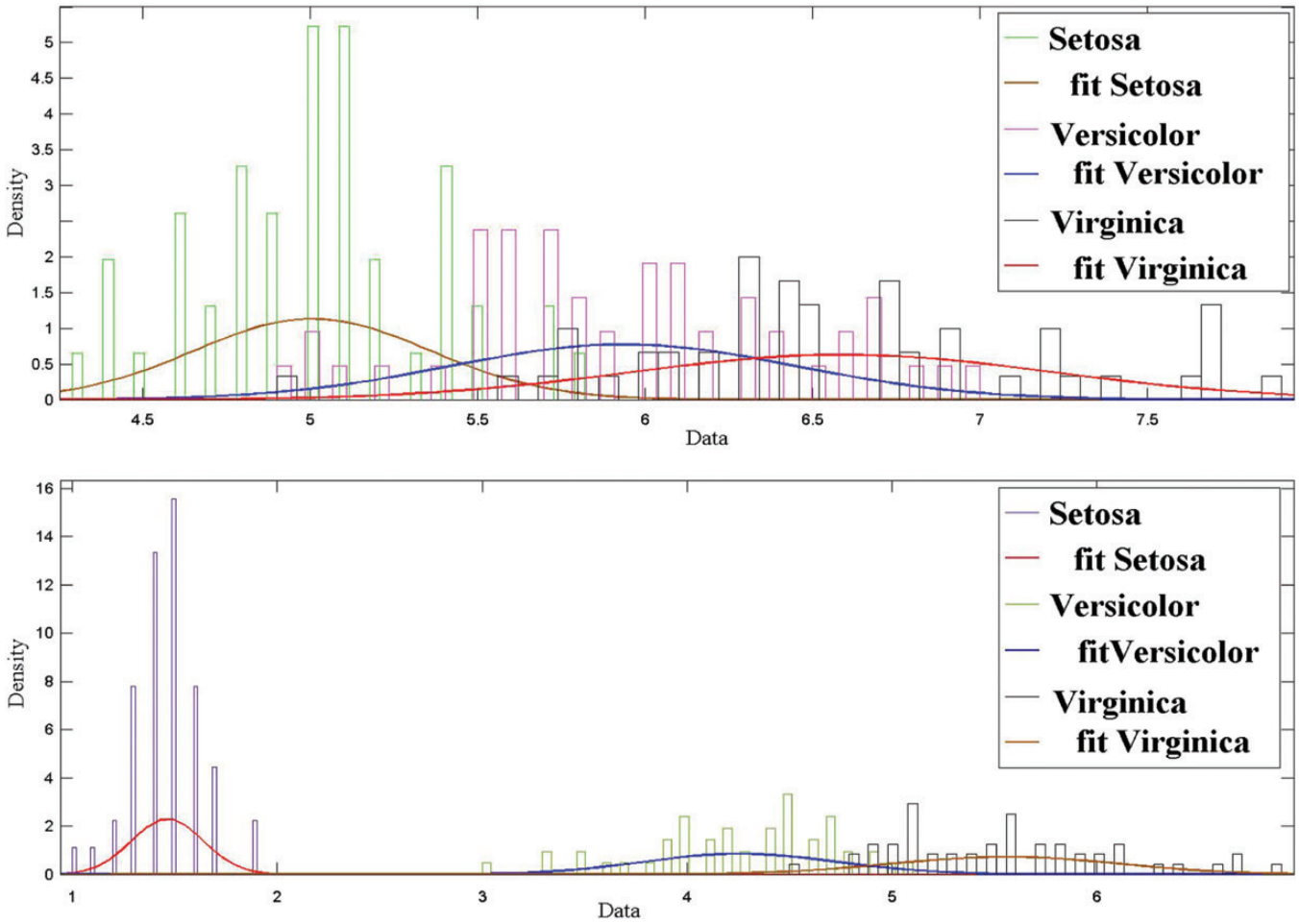

FIGURE 1. Distributions of Sepal width (top) and Petal width (bottom) datasets. Plenty of overlapping for Sepal and little for Petal. Distributions for Sepal and Petal lengths are structurally similar.

The basic properties of $\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m}$ are the following.

PROPOSITION 6.1. (1) *If* $\langle i_1,\ldots,i_m\rangle \subseteq \langle j_1,\ldots,j_l\rangle$, *then* $\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m} \geq \mathsf{c}^{(rt)r}_{j_1,\ldots,j_l}$.
(2) $\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m} = 0 \iff \mathsf{c}^{(rt)t}_{i_1,\ldots,i_m} = 0$.

Proof. (1) Directly from Fact 6.1.
(2) If $\pi_{i_1,\ldots,i_m}(\mathbf{G}_r) \cap \pi_{i_1,\ldots,i_m}(\mathbf{G}_t) = \emptyset$, then clearly $\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m} = \mathsf{c}^{(rt)t}_{i_1,\ldots,i_m} = 0$. Suppose that $x \in \pi_{i_1,\ldots,i_m}(\mathbf{G}_r) \cap \pi_{i_1,\ldots,i_m}(\mathbf{G}_t)$. From Equations (6) and (7), we immediately have $\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m} \geq 1$ and $\mathsf{c}^{(rt)t}_{i_1,\ldots,i_m} \geq 1$. □

If $\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m} \neq 0$, then there is no relationship between the values of $\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m}$ and $\mathsf{c}^{(rt)t}_{i_1,\ldots,i_m}$, and the difference between them can be arbitrary.

The *mutual overlapping* of $\mathbf{G}_r$ and $\mathbf{G}_t$ over the subdomain $\mathbb{R}^{\langle i_1,\ldots,i_m\rangle}$ is defined as

$$\mathsf{overlap}^{rt}_{i_1,\ldots,i_m} = \sqrt{\frac{\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m}}{|\mathbf{G}_r|} \cdot \frac{\mathsf{c}^{(rt)t}_{i_1,\ldots,i_m}}{|\mathbf{G}_t|}}. \tag{8}$$

The above equation, which is a geometric mean, is based on the pairwise comparisons paradigm. It measures the mutual overlapping of the groups $\mathbf{G}_r$ and $\mathbf{G}_t$, over the space $\mathbb{R}^{(i_1,\ldots,i_m)}$, and is used as a base for other calculations. For the $\mathbf{G}_1$ and $\mathbf{G}_2$ above, $\mathsf{overlap}^{1,2}_{(1,2)} = \sqrt{\frac{4}{6}\cdot\frac{3}{5}} = 0.6325$.

Why geometric mean? Mainly because we want the smaller of $\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m}/|\mathbf{G}_r|$ and $\mathsf{c}^{(rt)s}_{i_1,\ldots,i_m}/|\mathbf{G}_t|$ to be the dominant factor. In particular, we want the value of $\mathsf{overlap}^{rt}_{i_1,\ldots,i_m}$ to be close to zero if either of $\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m}/|\mathbf{G}_r|$ and $\mathsf{c}^{(rt)s}_{i_1,\ldots,i_m}/|\mathbf{G}_t|$ is close to zero. The below result shows that the geometric mean has this property.

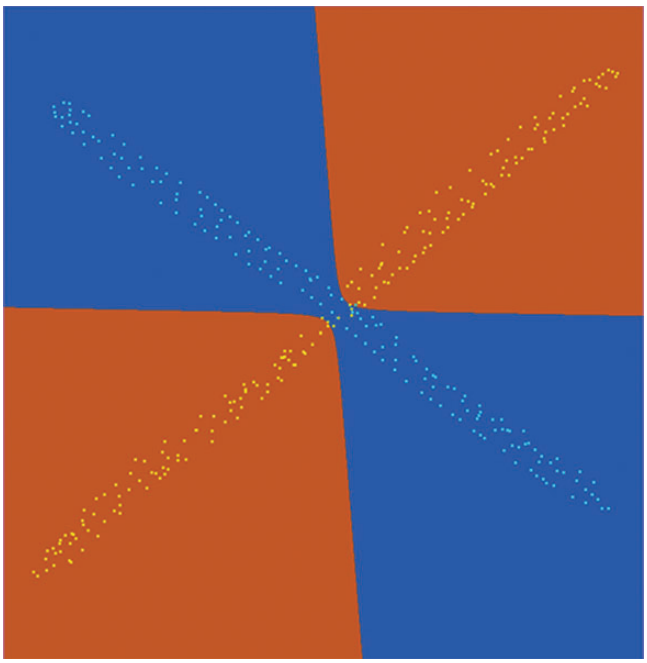

FIGURE 2. Example of overlapping in two dimensions. Both dimensions have almost 100% overlapping, but 2D overlapping is small.

PROPOSITION 6.2. *If* $0 < \varepsilon < a \leq b$, *then*

$$\sqrt{ab} - \sqrt{(a-\varepsilon)b} > \sqrt{a(b+\varepsilon)} - \sqrt{ab}.$$

Proof. First note that $ab + \sqrt{ab(a-\varepsilon)(b+\varepsilon)} > 0$ while $(a-b)\varepsilon \leq 0$. Hence,

$$\begin{aligned} &2ab + 2\sqrt{ab(a-\varepsilon)(b+\varepsilon)} > a\varepsilon - b\varepsilon \iff \\ &2\sqrt{ab} > \sqrt{a(b+\varepsilon)} - \sqrt{(a-\varepsilon)b} \iff \\ &\sqrt{ab} - \sqrt{(a-\varepsilon)b} > \sqrt{a(b+\varepsilon)} - \sqrt{ab}. \end{aligned}$$

□

Proposition 6.2 simply states that decrementing the smaller value changes the geometric mean more than the same increment to the bigger value (see also [11, 12, 16] for more arguments).

Note also that $0 \leq \mathsf{overlap}^{rt}_{i_1,\ldots,i_m} \leq 1$ and $\mathsf{overlap}^{rt}_{1,\ldots,p} = 0$ for all distinct $\mathbf{G}_r$ and $\mathbf{G_t}$.

The value of $\mathsf{overlap}^{rt}_{i_1,\ldots,i_m}$ is some measure of overlapping between *a pair* $\mathbf{G}_r$ and $\mathbf{G}_t$. In the spirit of the pairwise comparisons paradigm, we will derive a measure of the mutual overlappings for the features $F_{i_1}, \ldots, F_{i_m}$, denoted by $\mathsf{overlap}_{i_1,\ldots,i_m}$, from $\mathsf{overlap}^{rt}_{i_1,\ldots,i_m}$ for all pairs r and t. Since the value of $\mathsf{overlap}_{i_1,\ldots,i_m}$ is a *measure of central tendency* of all $\mathsf{overlap}^{rt}_{i_1,\ldots,i_m}$, we will use *the arithmetic mean* to measure it (cf. [11, 12] and especially [15]).

Formally, we define $\mathsf{overlap}_{i_1,\ldots,i_m}$ as

$$\mathsf{overlap}_{i_1,\ldots,i_m} = \frac{1}{\binom{p}{2}} \sum_{r>t}^{r,t=1,\ldots,p} \mathsf{overlap}^{rt}_{i_1,\ldots,i_m}. \tag{9}$$

We always have $0 \leq \mathsf{overlap}_{i_1,\ldots,i_m} \leq 1$ and the *smaller the value of* $\mathsf{overlap}_{i_1,\ldots,i_m}$ *is the more important* the subset $F_{i_1}, \ldots, F_{i_m}$ of features is.

The proposition below shows that $\mathsf{overlap}_{i_1,\ldots,i_m}$ is monotone with respect to $\langle i_1, \ldots, i_m \rangle$.

PROPOSITION 6.3. *If* $\langle i_1, \ldots, i_m \rangle \subseteq \langle j_1, \ldots, j_l \rangle$, *then*

$$\mathsf{overlap}_{i_1,\ldots,i_m} \geq \mathsf{overlap}_{j_1,\ldots,j_l}.$$

Proof. From Proposition 6.1(1), we have $\mathsf{c}^{(rt)r}_{i_1,\ldots,i_m} \geq \mathsf{c}^{(rt)r}_{j_1,\ldots,j_l}$ for all r, t, so consequently $\mathsf{overlap}^{rt}_{i_1,\ldots,i_m} \geq \mathsf{overlap}^{rt}_{j_1,\ldots,j_l}$, and $\mathsf{overlap}_{i_1,\ldots,i_m} \geq \mathsf{overlap}_{j_1,\ldots,j_l}$. □

From Proposition 6.3, it follows that if $\mathsf{overlap}_{i_1,\ldots,i_m} = 0$, then for each $\langle j_1, \ldots, j_l \rangle$, if $\langle i_1, \ldots, i_m \rangle \subseteq \langle j_1, \ldots, j_l \rangle$, then $\mathsf{overlap}_{j_1,\ldots,j_l} = 0$.

For a given feature F_i, the measure of its complete m-dimensional overlapping is again a measure of central tendency of all different $\mathsf{overlap}_{i_1,\ldots,i_m}$ such that $i \in \{i_1, \ldots, i_m\}$, so it is defined as the following arithmetic mean:

$$\mathsf{overlap}^{(m)}_i = \frac{1}{\binom{p-1}{m-1}} \sum_{i \in \{i_1,\ldots,i_m\}}^{\langle i_1,\ldots,i_m \rangle \subseteq \langle 1,\ldots,p \rangle} \mathsf{overlap}_{i_1,\ldots,i_m}. \tag{10}$$

Since $\mathsf{overlap}^{rt}_{1\ldots p} = 0$, if $r \neq t$, we have $\mathsf{overlap}^{(p)}_i = 0$ for all $i = 1, \ldots, p$.

Moreover, $\mathsf{overlap}^{(m)}_i$ is also monotone with respect to the number of dimensions m.

PROPOSITION 6.4. *For all* $i = 1, \ldots, p$ *and* $m, m' = 1, \ldots, n-1$, *if* $m' > m$, *then*:

$$\mathsf{overlap}^{(m)}_i \geq \mathsf{overlap}^{(m')}_i.$$

Proof. Due to the transitivity of the relation $\geq$ and the fact that $m, m' \in \mathbb{N}$, it suffices to show that $\mathsf{overlap}^{(m)}_i \geq \mathsf{overlap}^{(m+1)}_i$, for $m \geq 1$. First note that if $i \in \{i_1, \ldots, i_m\}$ and $\langle i_1, \ldots, i_m \rangle \subseteq \langle j_1, \ldots, j_m, j_{m+1} \rangle \subseteq \langle 1, \ldots, p \rangle$, then: $|\{\langle j_1, \ldots, j_{m+1} \rangle \mid \langle i_1, \ldots, i_m \rangle \subseteq \langle j_1, \ldots, j_{m+1} \rangle\}| = p - m$, and $|\{\langle i_1, \ldots, i_m \rangle \mid \langle i_1, \ldots, i_m \rangle \subseteq \langle j_1, \ldots, j_{m+1} \rangle\}| = m$.

Define

$$SUM^{(i)}_{\langle i_1,\ldots,i_m\rangle} = \sum_{i\in\{i_1,\ldots,i_m\}}^{\langle i_1,\ldots,i_m\rangle\subseteq\langle 1,\ldots,p\rangle} \mathsf{overlap}_{i_1,\ldots,i_m}$$

and

$$SUM^{(i)}_{\langle j_1,\ldots,j_{m+1}\rangle} = \sum_{i\in\{j_1,\ldots,j_{m+1}\}}^{\langle j_1,\ldots,j_{m+1}\rangle\subseteq\langle 1,\ldots,p\rangle} \mathsf{overlap}_{j_1,\ldots,j_{m+1}}.$$

By Proposition 6.4, $\mathsf{overlap}_{i_1,\ldots,i_m} \geq \mathsf{overlap}_{j_1,\ldots,j_{m+1}}$, so

$$(p-m)\mathsf{overlap}_{i_1,\ldots,i_m} \geq \sum_{seq_i\subseteq seq_j} \mathsf{overlap}_{j_1,\ldots,j_{m+1}},$$

where $seq_i = \langle i_1,\ldots,i_m\rangle$ and $seq_j = \langle j_1,\ldots,j_{m+1}\rangle$ and

$$(p-m)\cdot SUM^{(i)}_{\langle i_1,\ldots,i_m\rangle} \geq m\cdot SUM^{(i)}_{\langle j_1,\ldots,j_{m+1}\rangle}.$$

Hence,

$$\frac{1}{\binom{p-1}{m-1}} SUM^{(i)}_{\langle i_1,\ldots,i_m\rangle} \geq \frac{m}{\binom{p-1}{m-1}(p-m)} SUM^{(i)}_{\langle j_1,\ldots,j_{m+1}\rangle}.$$

But $\frac{m}{\binom{p-1}{m-1}(p-m)} = \frac{1}{\binom{p-1}{m}}$, so

$$\frac{1}{\binom{p-1}{m-1}} SUM^{(i)}_{\langle i_1,\ldots,i_m\rangle} \geq \frac{1}{\binom{p-1}{m}} SUM^{(i)}_{\langle j_1,\ldots,j_{m+1}\rangle},$$

which is equivalent to $\mathsf{overlap}^{(m)}_i \geq \mathsf{overlap}^{(m+1)}_i$. □

One consequence of Proposition 6.4 is that if $\mathsf{overlap}^{(m)}_i = 0$ for some m, then $\mathsf{overlap}^{(m')}_i = 0$ for all $m' > m$. Quite often $\mathsf{overlap}^{(1)}_i \approxeq 1.0$ for some i, and usually $\mathsf{overlap}^{(m)}_i \approxeq 0.0$ for bigger m. The latter is expected also due to Proposition 6.4.

The *total measure of the overlapping* of the feature F_i over all subdomains of $\mathbb{R}^p$ is given by the following formula:

$$\mathsf{OVERLAP}_i = \sum_{j=1}^{p-1} \frac{\mathsf{overlap}^{(j)}_i}{j}. \tag{11}$$

The value of $\mathsf{OVERLAP}_i$ is the *weighted sum* of all $\mathsf{overlap}^{(j)}_i$, and the jth weight is equal to $1/j$.

The motivation for this value is the following. Assume that $\langle i_1,\ldots,i_m\rangle \subseteq \langle j_1,\ldots,j_l\rangle$ and $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$. If for all $r = 1,\ldots,l$, $\pi_{j_r}(\mathbf{x}) = \pi_{j_r}(\mathbf{y})$, then clearly $\pi_{i_s}(\mathbf{x}) = \pi_{i_s}(\mathbf{y})$ for all $s = 1,\ldots,m$. Formally, this is expressed by the property $\mathsf{C}_{j_1,\ldots,j_l}(x,\mathbf{A}) \subseteq \mathsf{C}_{i_1,\ldots,i_m}(x,\mathbf{A})$ that was used in the proof of Fact 6.1. It also means that if $m < j$, the values of $\mathsf{overlap}^{(m)}_i$ and $\mathsf{overlap}^{(j)}_i$ *are not* independent, the elements that are used to calculate the value of $\mathsf{overlap}^{(m)}_i$ are also used to calculate the value of $\mathsf{overlap}^{(j)}_i$. Moreover, bigger values of $\mathsf{overlap}^{(m)}_i$ imply bigger values of $\mathsf{overlap}^{(j)}_i$, for all $j > m$. This phenomenon is illustrated in Fig. 2 where all points of 2D overlapping are also the points of both 1D overlappings. This kind of redundancy can be avoided, for example, by defining

$$\hat{\mathsf{C}}_{seq}(x,\mathbf{A}) = \mathsf{C}_{seq}(x,\mathbf{A}) \setminus \bigcup_{seq\subseteq seq'} \mathsf{C}_{seq'}(x,\mathbf{A}),$$

for every $seq = \langle i_1,\ldots,i_m\rangle$, and then using $\hat{\mathsf{C}}_{seq}(x,\mathbf{A})$ instead of $\mathsf{C}_{seq}(x,\mathbf{A})$ in the definition of $\mathsf{c}_{seq}(x,\mathbf{A})$ (Equation (5)). However, such an approach assumes that overlapping (or rather lack of it) over m dimensions has the same importance as overlapping over k, where $k < m$, dimensions; including $k = 1$ and $k = 2$. Our general assumptions are just the opposite, a low number of dimensions overlapping is more important as lack of overlap (or a small amount) makes considering bigger numbers unnecessary.

Summing up, allowing this kind of redundancy and then compensating for it seems to be a better approach. The weight $1/j$ is used to compensate for this redundancy. It takes into account Proposition 6.3 and an observation that in most cases the 1D overlappings are the most influential. The weights $1/j$ are an estimation based on heuristics and experiments. We believe a possible generic formula can be formulated as follows:

$$\mathsf{OVERLAP}_i = \sum_{j=1}^{p-1} \phi(j)\cdot \mathsf{overlap}^{(j)}_i, \tag{12}$$

where $\phi(j)\in\langle 0,1\rangle$, $\phi(1) = 1$ and $\phi(j)$ is decreasing, however, the exact formulas for $\phi(j)$ are unknown and it

TABLE 11. Weights of features for *Iris dataset* calculated with overlapping method.

	Sepal length	Sepal width	Sepal L. + W.	Petal length	Petal width	Petal L. + W.
Experts (initial)	9.4%	6.7%	16.1%	35%	48.9%	83.9%
Pairwise comparisons	10.9%	6.8%	17.7%	31.6%	50.8%	82.4%
Overlapping (1D)	9.1%	6.5%	15.6%	42.7%	41.7%	84.4%
Overlapping (1D and 2D)	9.7%	7.1%	16.8%	41.4%	41.7%	83.1%
Overlapping (1D, 2D and 3D)	10.0%	7.4%	17.4%	40.5%	42.1%	82.6%
Overlapping unscaled	1.5802	1.1636	2.7438	6.4129	6.6590	13.0719

TABLE 12. *Iris dataset*: final *qualitative* judgments of mutual relationship derived from the bottom raw of Table 11 using ranges from Table 1.

	Sepal length	Sepal width	Petal length	Petal width
Sepal length	$\approx$	$\sqsupset$	$<$ ($\subset$)	$\prec$
Sepal width	$\sqsubset$	$\approx$	$\prec$	$\prec$
Petal length	$>$ ($\supset$)	$\succ$	$\approx$	$\approx$ ($\sqsubset$)
Petal width	$\succ$	$\succ$	$\approx$ ($\sqsupset$)	$\approx$

The gray color indicates a difference from Table 5, and values from Table 5 are in parenthesis.

is not obvious how they could be derived. Since usually $\mathsf{overlap}_i^{(j)} \cong 0.0$ for bigger j, only the first few values of α_j are important and the values of $\phi(j)$ are irrelevant for bigger j. One very simple heuristic is that the elements used to calculate $\mathsf{overlap}_i^{(m)}$ have also been used to calculate $\mathsf{overlap}_i^{(1)}, \ldots, \mathsf{overlap}_i^{(m-1)}$. We do not have any formal proof that this relationship can be approximated by a linear formula, but an analysis of more than one hundred random examples has shown that $\phi(j) = 1/j$ is a reasonably good estimation of overlapping redundancy.[6]

Note also that if $\mathsf{overlap}_i^{(1)} = 0$, then $\mathsf{OVERLAP_i} = 0$ too, for all $i = 1, \ldots, p$.

The smaller $\mathsf{OVERLAP_i}$ *the more discrimination ability the feature* F_i *has. Moreover, if* $\mathsf{OVERLAP_i} = 0$, *then the feature* F_i *alone defines the partition* $\mathbf{G}_1, \ldots, \mathbf{G}_k$ *of* $\mathbf{S}$.

Hence, the formula for weights should emphasize the importance of small values of $\mathsf{OVERLAP_i}$ and it also should treat separately the case of $\mathsf{OVERLAP_i} = 0$.

Taking all the above arguments into account, we define the weight w_i, the measure the importance of F_i as

$$w_i = \frac{1}{\mathsf{OVERLAP_i}} \tag{13}$$

if $\mathsf{overlap_i}^{(1)} > \varepsilon$, for all $i = 1, \ldots, p-1$, where $\varepsilon > 0$ is some approximation of zero.

If there is i_0 such that $\mathsf{overlap_{i_0}}^{(1)} \leq \varepsilon$, which is interpreted as $\mathsf{overlap_{i_0}}^{(1)} = 0$, then the feature F_{i_0} alone defines the partition $\mathbf{G}_1, \ldots, \mathbf{G}_k$ of $\mathbf{S}$. Hence, $w_{i_0} = 1$ and $w_i = 0$ otherwise. If there is more than one such i_0, we pick one randomly, as it does not matter which one we choose for separation calculations.

We may not always want to use all subdomains of $\mathbb{R}^p$, so it is useful to define

$$\mathsf{OVERLAP_i}^{(l)} = \sum_{j=1}^{l} \frac{\mathsf{overlap}_i^{(j)}}{j}, \tag{14}$$

[6]The fact that $\sum_{i=1}^{\infty} 1/i = \infty$ is not a problem as the number of features, i.e. p, is relatively small.

TABLE 13. Classification results for the *Iris dataset* using the overlapping method for weight assignment.

Samples		Overlappings						Pairwise comparison			
		Without weights		Iteration up to one dimension		Iteration up to two dimensions		Iteration up to three dimensions		$cm_A = 0.0$	
Train.	Test.	No. SV	Acc. %	No. SV	Acc. %	No. SV	Acc. %	No. SV	Acc. %	No. SV	Acc. %
15	135	14	94.1	11	95.6	11	95.6	11	95.6	9	95.6
30	120	17	94.2	16	96.7	16	96.7	16	96.7	16	97.5
45	105	24	96.2	16	96.2	17	96.2	16	96.2	18	97.1
60	90	26	94.4	19	97.8	20	97.8	20	97.8	23	95.6
75	75	31	93.3	26	96.0	27	97.3	27	97.3	24	96.0
90	60	43	98.3	30	98.3	29	98.3	28	98.3	26	96.7
105	45	39	97.8	33	100.0	33	100.0	33	100.0	28	97.8
120	30	42	96.7	34	100.0	35	100.0	35	100.0	31	96.7
135	15	44	100.0	38	100.0	37	100.0	37	100.0	36	100.0
Averages		30.1	96.1	24.8	97.8	25.0	98.0	24.8	98.0	23.4	97.0

Classification results with pairwise comparisons also included for comparison.

$$w_i^{(l)} = \frac{1}{\mathsf{OVERLAP_i}^{(l)}}, \tag{15}$$

for all $l = 1, \ldots, p-1$. In most cases, we expect $\mathsf{OVERLAP_i}^{(l)} \approxeq \mathsf{OVERLAP_i} = \mathsf{OVERLAP_i}^{(p-1)}$ for bigger l, so in such cases we can stop calculations earlier.

Most of the SVM procedures, including LibSVM (cf. [30]), cannot separate the points that are very close, so we have to include this factor in our weight calculation. To tackle this problem, for each dimension i, $i = 1, \ldots, p$ we introduce a *discriminator* d_i, and for each $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$, we round each x_i to the nearest value of $l \cdot d_i$ for some l. Formally, we calculate $\hat{x}_i = l_0^i \cdot d_i$, where l_0^i is such that for each l, $|l_0^i \cdot d_i| \leq |l \cdot d_i|$. Then we just replace $\mathbf{x}$ by $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_p)$. For the calculations made for this paper, d_i was set as half of the mode of the distances of the values of samples of the set $\mathbf{S}$ belonging to the dimension $\mathbb{R}^{(i)}$. We did this only for calculating weights, not for segregation using LibSVM procedures.

7. ANALYSIS OF IRIS AND VERTEBRAL DATASETS WITH OVERLAPPINGS

The method described above was applied to both the Iris and Vertebral datasets. The results for the Iris dataset are presented in Tables 11–13. Table 11 gives weights calculated by the three methods discussed in this paper, pure expert judgments, expert judgments refined by pairwise comparisons and measurements of overlapping—using one, two and finally all three dimensions (as in this case $p-1=3$). Table 12 shows qualitative judgments provided by overlappings.

While for Sepal the weights, i.e. external importance, are close for all three methods, for Petal, even though its total for both length and width is also close for all three methods, the distribution between length and width is different for overlapping than for pairwise comparisons. The experts recommended *Petal Length* $\sqsubset$ *Petal Width*, the pairwise comparison refinement preserved this relationship (slightly enhancing the importance of width), but the overlapping method produced the relationship *Petal Length* $\approx$ *Petal Width*. Moreover, the experts initially recommended *Petal Length* $>$ *Sepal Length*, consistency analysis for pairwise comparisons changed it to *Petal Length* $\supset$ *Sepal Length*, but overlappings suggest again *Petal Length* $>$ *Sepal Length*.

Figure 1 illustrates overlapping of Sepal width and Petal width. One can see plenty of overlapping for Sepal width and very little for Petal width, which is reflected in the values of weights calculated by overlapping (3D): 7.4% for Sepal width and 42.1% for Petal width.

Table 13 gives the classification results. Assigning weights by overlapping resulted in significantly better accuracy than the remaining methods, while the number of SVs was similar to that for pairwise comparisons. In this case, the values of $\mathsf{overlap}_i^{(1)}$, $i = 1, 2, 3, 4$, are dominant. This could just be a property of the Iris dataset, however, we expect the values of $\mathsf{overlap}_i^{(j)}$ to be often (but not always!) negligible for bigger j (see Proposition 6.4).

The results for the Vertebral dataset are presented in Tables 14–16. Table 14 gives weights calculated by the three methods discussed in this paper, pure expert judgments, expert judgments refined by pairwise comparisons and measurements of overlapping—using one, two and finally all three dimensions (again in this case $p-1=3$). Table 15 shows qualitative judgments provided by overlappings.

As opposed to the Iris dataset case, where 1D overlapping dominates the total overlapping, here using 2D and 3D overlappings is necessary as 1D overlappings differ substantially from

TABLE 15. *Vertebral dataset*: initial *qualitative* judgments of mutual relationship derived from overlappings.

	PI	PT	LLA	SS	PR	GOS
PI	$\approx$	$\subset$	$<$ ($\subset$)	$<$	$<$	$\prec$
PT	$\supset$	$\approx$	$\approx$	$\sqsubset$	$\sqsubset$	$<$
LLA	$>$ ($\supset$)	$\approx$	$\approx$	$\sqsubset$	$\sqsubset$	$<$
SS	$>$	$\sqsupset$	$\sqsupset$	$\approx$	$\approx$	$<$ ($\subset$)
PR	$>$	$\sqsupset$	$\sqsupset$	$\approx$	$\approx$	$<$ ($\subset$)
GOS	$\succ$	$>$	$>$	$>$ ($\supset$)	$>$ ($\supset$)	$\approx$

The relational symbols from Table 1 were used. The gray color indicates a difference from consistent pairwise comparisons (Table 8), and values from Table 8 are in parenthesis.

TABLE 14. Weights of features for *Vertebral dataset* using overlap method.

	PI	PT	LLA	SS	PR	GOS
Experts (initial)	4.1%	10.2%	10.2%	16.9%	16.9%	41.7%
Pairwise comparison	3.43%	9.64%	9.64%	16.3%	14.88%	46.13%
Overlapping (1D)	16.76%	13.94%	14.76%	14.86%	14.34%	25.33%
Overlapping (1D and 2D)	5.48%	10.55%	11.21%	13.57%	14.33%	46.86%
Overlapping (1D, 2D and 3D)	2.93%	8.62%	9.66%	13.55%	13.99%	51.26%
Overlapping unscaled	0.7304	2.15	2.41	3.38	3.49	12.79

TABLE 16. Classification results for the *Vertebral dataset* using the overlapping method for weight assignment.

Samples		Without weights		Overlappings						Pairwise comparison	
				Iteration up to 1 dimension		Iteration up to 2 dimensions		Iteration up to 3 dimensions		$cm_A = 0.0$	
Train.	Test.	No. SV	Acc. %	No. SV	Acc. %	No. SV	Acc. %	No. SV	Acc. %	No. SV	Acc. %
30	280	28	67.5	29	68.2	28	72.9	28	77.1	28	75.4
60	250	51	71.2	51	74.0	50	78.4	49	81.2	47	79.2
90	220	72	76.8	71	79.5	65	82.7	60	84.1	61	82.7
120	190	92	80.0	90	81.6	81	84.7	78	86.8	79	85.3
150	160	107	78.8	99	80.6	92	84.4	90	86.3	89	86.3
180	130	130	75.4	117	78.5	111	80.0	106	82.3	102	81.5
210	100	139	85.0	129	88.0	117	90.0	110	91.0	112	93.0
240	70	153	88.6	144	90.0	126	91.4	120	92.9	120	92.3
270	40	162	85.0	151	87.5	133	92.5	124	92.5	126	92.5
Averages		103.8	78.7	97.9	80.9	89.2	84.1	85.0	86.0	84.9	85.4

Classification results with pairwise comparisons also included for comparison.

total overlappings. One-dimensional overlappings are roughly identical for PI, PT, LLA, SS and PR, appropriate total overlappings (i.e. when all three dimensions are used) are very different, the total overlapping for PI ($1/0.7304 = 1.3691$) is almost five times bigger that for PR ($1/3.49 = 0.2865$).

While in general the weights calculated by overlappings are closed to the weights derived by pairwise comparisons there are some important differences. The experts recommended $LLA \supset PI$, $GOS \supset SS$ and $GOS \supset PR$, and pairwise comparison refinement preserved this relationship, but the overlapping method produced the relationship $LLA > PI$, $GOS > SS$ and $GOS > PR$.

The differences between the weights produced by overlappings and the weights produced by pairwise comparisons, for both the Iris dataset and Vertebral dataset, may mean that either experts made misjudgments, or the dataset is incomplete, or the border values in Table 1 do no properly fit those cases (*which is most likely true*), or all combinations of the above, however, any analysis of this is beyond the scope of this paper.

Table 16 gives the classification results. While accuracy of classification without weights is definitely the worst, accuracy of overlapping restricted to one dimension is not much better. This is because it provides almost identical weights for PI, PT, LLA, SS and PR. Overlapping when one dimension and two dimensions are used give significantly better accuracy, but still worse than pairwise comparisons. Overlapping with all three dimensions used provides the best accuracy. The results for number of SVs are similar with one exception, the number of SVs for overlapping with all three dimensions is similar to that for pairwise comparisons.

8. FINAL COMMENT

It appears that adding proper weights improves classification when SVM techniques are used. It both decreases the number of SVM used and increases the accuracy, especially for smaller training sets. Two methods of providing weights have been proposed.

The first one assumes that a given sample of data *does not contain any clue* about the importance of features. In this case, domain experts and the pairwise comparison paradigm are used. The experts provide measures of mutual relationship between features, and the distance-based consistency is then used to correct the value provided by experts. Then the weights are calculated by techniques provided by pairwise comparisons (geometric means in this paper). When the experts provided measures with relatively small values of inconsistency index (as it was in this case), the weights calculated on the basis of their initial judgment are not much different than the weights after decreasing inconsistency. Nevertheless, the possibility of calculating and correcting inconsistency is a very useful and important tool that increases the trustfulness of the approach.

The second method *assumes that the given sample contains information about feature importance*, and that the feature importance is determined by the discriminatory power of features. The method has been tested on the Iris dataset and better accuracy was achieved than for other methods. The results were especially superior to these obtained without weights. The method also uses the pairwise comparisons ideas.

Summing up, the contributions can be divided into two categories: theory and application. The idea of feature domain overlapping and the proposed method of weights calculation based on this idea is a contribution to the theory of objects classification. Merging pairwise comparisons methods with weighted SVMs techniques, and testing both approaches on Iris and Vertebral data sets belongs to applications category.

ACKNOWLEDGEMENTS

The authors would like to thank the Natural Science and Engineering Council of Canada (NSERC) and McMaster Centre for Software Certification for their partial support of this work. Maria Janicka, Samareh Khamseh, Amir Soudkhah are thanked for providing expertise in botany and anatomy. The authors are also grateful to the anonymous referees for useful comments and suggestions.

FUNDING

Natural Science and Engineering Council of Canada, Discovery Grant No. 36573-2010.

REFERENCES

[1] Vapnik, V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

[2] Kotsiantis, S.B., Zaharakis, I.D. and Pintelas, P.E. (2007) Machine learning: a review of classification and combining techniques. *Artif. Intell. Rev.*, **26**, 159–190.

[3] Suykens, J.A.K., De Brabanter, J., Lukas, L. and Vandewalle, J. (2002) Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, **48**, 85–105.

[4] Amutha, A.L. and Kavitha, S. (2011) Features based classification of images using weighted feature support vector machines. *Int. J. Comput. Appl.*, **26**, 24–29.

[5] Sun, B., Song, S.-J. and Wu, C. (2009) A New Algorithm of Support Vector Machine Based on Weighted Feature. *Proc. Int. Conf. Machine Learning and Cybernetics*, Baoding, China, pp. 1616–1620.

[6] Xing, H.J., Hu, M.G. and Tian, D.Z. (2009) Linear feature-weighted support vector machine. *Fuzzy Inf. Eng.*, **1**, 289–305.

[7] Zhang, S., Zhang, C. and Yang, Q. (2003) Data preparation for data mining. *Appl. Artif. Intell.*, **17**, 375–381.

[8] Janicki, R. and Zhai, Y. (2011) Remarks on Pairwise Comparison Numerical and Non-Numerical Rankings. *Proc. RSKT 2011*, Banff, Canada, October 2011, Lecture Notes in Artificial Intelligence 6954, pp. 290–300. Springer.

[9] Koczkodaj, W.W. and Szarek, S.J. (2009) On distance-based inconsistency reduction algorithms for pairwise comparisons. *Logic J. IGPL*, **18**, 859–869.

[10] Saaty, T.L. (2005) *Theory and Applications of the Analytic Network Process*. RWS Publications, Pittsburgh, PA, USA.

[11] Fleming, P.J. and Wallace, J.J. (1986) How not to lie with statistics: the correct way to summarize benchmark results. *Commun. ACM*, **29**, 218–221.

[12] Huff, D. (1954) *How to Lie with Statistics*. W. W. Norton, New York.

[13] Janicki, R. (2009) Pairwise comparisons based non-numerical ranking. *Fundam. Inform.*, **94**, 197–217.

[14] Koczkodaj, W.W. (1993) A new definition of consistency of pairwise comparisons. *Math. Comput. Model.*, **18**, 79–84.

[15] Marnich, M.A. (2008) A knowledge structure for the arithmetic mean: relationship between statistical conceptualizations and mathematical concepts. PhD Thesis, University of Pittsburgh, Pittsburgh, PA, USA.

[16] Spizman, L. and Weinsten, M.A. (2008) A note on utilizing the geometric mean: when, why and how the forensic economist should employ the geometric mean. *J. Leg. Econ.*, **15**, 43–55.

[17] Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–188 (also http://archive.ics.uci.edu/ml).

[18] da Rocha Neto, A.R., Sousa, R., de A. Barreto, G. and Cardoso, J.S. (2011) Diagnostic of Pathology on the Vertebral Column with Embedded Reject Option. *Proc. IbPRIA 2011*, Las Palmas de Gran Canaria, Spain, June 2011, Lecture Notes in Computer Science 6669, pp. 588–595. Springer (see also http://archive.ics.uci.edu/ml).

[19] Soudkhah, M.H. and Janicki, R. (2013) Weighted Features Classification with Pairwise Comparisons, Support Vector Machines and Feature Domain Overlapping. *Proc. 22nd IEEE WETICE Conf.*, Hammamet, Tunisia, pp. 172–177. IEEE Publ., doi:10.1109/WETICE.2013.70.

[20] Soudkhah, M.H. (2013) Weighted features classification with pairwise comparisons, support vector machines and feature domain overlapping. MSc Thesis, Department of Computing and Software, Mcmaster University, Hamilton, ON, Canada.

[21] Saaty, T.L. (1977) A Scaling methods for priorities in hierarchical structure. *J. Math. Psychol.*, **15**, 234–281.

[22] Bozóki, S. and Rapcsák, T. (2008) On Saaty's and Koczkodaj's inconsistencies of pairwise comparison matrices. *J. Glob. Optim.*, **42**, 157–175.

[23] Koczkodaj, W.W. (1998) Testing the accuracy enhancement of pairwise comparisons by a Monte Carlo experiment. *J. Stat. Plan. Inference*, **69**, 21–32.

[24] Barzilai, J. (1997) Deriving weights for pairwise comparison matrices. *J. Oper. Res. Soc.*, **48**, 1226–1232.

[25] Janicki, R. and Zhai, Y. (2012) On a pairwise comparison based consistent non-numerical ranking. *Logic J. IGPL*, **20**, 667–676.

[26] Fülöp, J., Koczkodaj, W.W. and Szarek, S.J. (2010) A Different Perspective on a Scale for Pairwise Comparisons. In Nguyen, N.T. and Kowalczyk, R. (eds), *Transactions on Computational Intelligence I*, Lecture Notes in Computer Science 6220, pp. 71–84. Springer.

[27] Cowan, N. (2001) The magical number four in short-term memory. A reconsideration of mental storage capacity. *Behav. Brain Sci.*, **24**, 87–185.

[28] Miller, G.A. (1956) The magical number seven, plus or minus two. *Psychol. Rev.*, **63**, 81–97.

[29] French, S. (1986) *Decision Theory*. Ellis Horwood, New York.

[30] Chang, C.-C. and Lin, C.-J. (2011) LibSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 1–39 (see also http://www.csie.ntu.edu.tw/∼cjlin/libsvm).

[31] Wang, K., Wang, X. and Zhong, Y. (2010) A Weighted Feature Support Vector Machines Method for Semantic Image Classification. *Proc. Int. Conf. Measuring Technology and Mechatronics Automation*, Changsha, China, pp. 377–380.

[32] Hsu, C.-H. and Lin, C.-J. (2002) A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Netw.*, **13**, 415–425.

[33] Platt, J.C., Cristianini, N. and Shawe-Taylor, J. (2000) Large margin dags for multiclass classification. *Adv. Neural Inf. Process. Syst.*, **12**, 547–553.