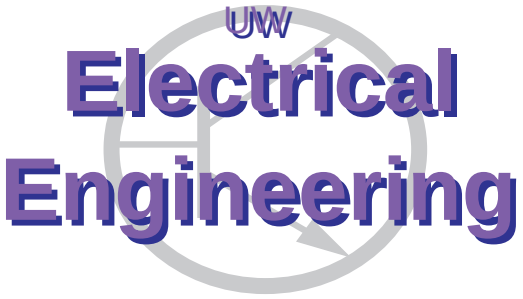
Intransitive Classification and Choice: A Review

Jeff Bilmes
Department of Electrical Engineering
University of Washington
Seattle, WA 98195-2500
{bilmes}@ee.washington.edu

Marina Meilă
Department of Statistics
University of Washington
Seattle, WA 98195-4322
mmp@stat.washington.edu



UWEE Technical Report
Number UWEETR-2006-0021
October 2006

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

Intransitive Classification and Choice: A Review

Jeff Bilmes
Department of Electrical Engineering
University of Washington
Seattle, WA 98195-2500
{*bilmes*}*@ee.washington.edu*

Marina Meilă
Department of Statistics
University of Washington
Seattle, WA 98195-4322
mmp@stat.washington.edu

University of Washington, Dept. of EE, UWEETR-2006-0021

October 2006

Abstract

A classifier is a machine that takes a description of an object, and then labels that object as being of a particular category. Classifiers have a rich history in the fields of statistics, artificial intelligence, and machine learning. This paper gives a brief background in statistical pattern recognition, multi-class classification, transitive games, and presents a brief history of intransitive preferences in a variety of fields.

Project Summary: Intransitive Classification and Choice

1 Introduction

In today's world, one of the crucial problems facing humankind is that of building machines capable of automatically making intelligent decisions in potentially adverse or uncooperative environments. The applications for such an ability span across many different fields, ranging from commercial manufacturing and automated robotics control, to defense related machinery, space exploration and satellite navigation, elections, intelligent human-computer interfaces such as speech and vision systems, security and biometrics, and to assistive devices for the handicapped and the aged — and this is just a few. In all cases, the goal is to produce a machine that appears to act both "robustly" and "intelligently." While a difficult term to define precisely, to act "intelligently" might mean to act in a way such that it appears a human being, after having undergone careful deliberation, has made a conscious decision to achieve some pre-specified and precisely stated goal or plan.

There are many different forms of intelligent decision making. In sequential decisions making, for example, the set of decisions are made one after another in a multi-stage process. A decision made at a given time can have an affect and possibly limit decisions made at some time in the future. A decision is typically made based on (sometimes partial) knowledge of the world, and based on optimizing an expected reward function over some time period spanning into the future. Markov decision processes [139, 147, 100, 185, 161] are one way to model such a sequential decision maker, where knowledge of the world at any given time is either complete or only partially known [112, 119, 144, 158].

Another style of decision making is well captured by the notion of a classifier. In this proposal, a *classifier* means a machine that takes as input an object encoded in some way, and based on that encoding makes a decision, such as deciding the class, category, or type of that object, and might also include rank information as well. In the simplest of cases, a decision is made once and the classifier must accept the consequence of having made that decision. No future decision can be considered as a weight or bias to influence current decisions, and no past set of decisions or function of a world state is available to help the classifier in making its decisions. Therefore, a classifier can be seen as a stateless machine with an input representation of an object or environment and output which corresponds to a decision to be made. Of course, one might augment the input to include both the state of the environment, the machine, and past decisions, and thereby produce sequential decision procedures, as in a recurrent neural network [86].

Such classifiers are a crucial part of many intelligent decision making machines. This includes statistical pattern recognition [50, 51], discriminant analysis [126], and regression [72, 84]. These, in turn, encompass a wide variety of applications such as speech [203], face [197], and hand-writing recognition[113], texture identification, land cover classification [152], security-based biometrics such as retinal and iris scanning, finger-print recognition, voice identification, speaker verification, speaker [151] and language identification [105], and so on. The list is enormous.

A critical research problem in all such systems is to discover methods that will reduce errors with only a negligible increase in computational and parameter complexity. It is often the case that an error penalty can be extremely large, so any methodology that is capable of producing an appreciable error reduction can achieve huge cost savings. Indeed, much theoretical and practical research has recently occurred to produce better performing classifiers [126, 193, 51, 84].

2 Review

This section gives a brief background in statistical pattern recognition, multi-class classification, transitive games, and presents a brief history of intransitive preferences [101, 122, 171, 182, 134, 3, 189, 67, 172, 125, 60, 62, 65, 7, 13, 120, 163, 184, 2, 88, 132, 188, 34, 155, 154, 196]. A small part of the material presented here was initially presented at [19, 91].

2.1 Statistical Pattern Classification

The foundation behind all pattern classification involves the identification of the *class* or category of some unknown object x. There are a set of K classes $C_1, \ldots, C_K$ and the object is presumably one of those classes. The *input* x typically (but not necessarily) is a vector with real or integer components. A pair (x, y) is called an *example* or *data point*, and y is called the *label* of x. The true association between inputs and labels can be deterministic or stochastic.

A *classifier* is a system that takes an input x and outputs $\hat{y}(x)$, a guess of its true class label y. The theory of pattern classification [50, 51, 126, 193] shows that even in the case of a stochastic dependence, it is optimal for a classifier to output a single y as response to an x. Therefore, we focus on classifiers that take x as input and output a single label y from the set $\{1, 2, \ldots K\}$.

A classifier is *trained* on a particular problem (e.g., digit recognition, identifying tumors in MRI images, identifying the category of a news article) by using a set of *examples* $\mathcal{D} = \{(x^n, y^n)\}_n$. There are many different types of classifiers, including logistic regression [126], multi-logistic regression [126], multi-layered perceptrons [28, 153], support vector machines (SVMs) [193, 170], classification trees [31] and there are many ways to train once a model family has been chosen [6, 97, 98, 199, 54, 53, 55].

Generative vs discriminative classifiers: It is possible to view classification purely statistically, where each class C_i is a stochastic process generating vectors x according to the distribution $p(x|i)$. The prior probability of the next example coming from class C_i is $P(i)$. Therefore, by Bayes rule, one can compute the posterior probability given x of label $y = i$ as $P(i|x) = P(i)p(x|i)/\sum_j P(j)p(x|j)$ The classifier then outputs $\hat{y}(x) = \text{argmax}_i P(i|x)$.

The models $p(x|i)$ are called *generative* and classifiers employing them are called *generative classifiers* (Naive Bayes being one example). Thus, a generative classifier models explicitly the process that generates data from each class. By contrast, a *discriminative* classifier (e.g., nearest neighbors, SVM, perceptron) focuses on modeling decision boundaries between classes. The latter has the advantage of utilizing information optimally; however, generative classifiers are important especially in tasks involving complex inputs because they allow a user a more flexible way of entering prior knowledge about the problem through the generative model. Moreover, generative classifiers can be trained [54, 53, 55] and induced [22, 26] discriminatively.

Binary vs multiway classification: A classification problem with $K = 2$ classes is a *binary* problem, while when $K > 2$ we say that we have a K*-way* or a *multiway* classification task. Binary classifiers (BCs) form a group apart for two major reasons: First, being simpler BCs can be more easily analyzed. In fact, one can cast any binary classification problem in the form $\hat{y} = \text{sign} g(x) \in \{\pm 1\}$, where g is a real-valued function called *discriminant*. For example, binary classification with generative models is expressed as $\hat{y}(x) = \text{sign} \log \{P(y = +1)p(x|y = +1)/P(y = -1)p(x|y = -1)\}$. We also define the log *likelihood ratio* between two classes i and j as $L_{ij}(x) = \log p(x|C_i)/p(x|C_j)$, and a general real-valued "preference" $S_{ij}(x)$ for class i over class j, either of which can be used as a binary classifier between class i and j.

While some classifiers can directly incorporate any number of classes (nearest neighbors, multi-layer perceptrons, generative classifiers, decision trees), others are naturally defined only for the $K = 2$ case (SVMs, decision stumps, linear perceptrons). Significant work has gone into constructing K-way classifiers using several binary classifiers as building blocks (see below). This direction is a main objective of our proposal, so in the next section we survey relevant work.

2.2 K-way classification as coding

K-way classifiers can be constructed using binary classifiers as building blocks. This is an active and important subfield since: 1) some of the best overall performing classifiers (voted perceptron, SVM) are naturally defined as binary classifiers; and 2) theoretical guarantees for K-way classifiers can sometimes be derived from the existing results for the binary ones.

One of the most popular ways of *binarizing* a K-class problem is the *all-pairs* (or *pairwise coupling* or *round-robin*) approach [71, 85, 146, 107, 41, 42, 133, 123, 124, 204] described presently. Let $\mathcal{D}_k$ represent all the examples representing class k. Binarization proceeds as follows: in the **learning stage**, for each pair of classes (k, l) train binary classifier g_{kl} on only the $\mathcal{D}_k \cup \mathcal{D}_l$ subset of the training set. In the **classification stage**, we label an unknown input x as $\hat{y}(x) = \text{sign}[h(\bar{g}(x))]$ where $\bar{g}(x) = (\, g_{kl}(x)\,)_{k<l}$ is the vector containing the outputs of all binary classifiers and the *decoding* function h is some function of the $M = K(K-1)/2$ variables in $\bar{g}$. In this paradigm, the BCs are treated as black boxes and the challenge is to design and analyze the function h.

The *one-vs-all* method, often used with neural networks [28, 153], constructs one classifier g_k to discriminate between examples in class k and all other classes. A possible problem with this method is that more than one $g_k(x)$ can be positive for an input x (note that *base classifiers* g_k or g_{kl} can have either discrete ± 1 or continuous real-valued outputs). Therefore, the label is typically calculated as $\hat{y}(x) = \text{argmax}_k\, g_k(x)$.

A powerful approach, the *error correcting code* (ECC) method [47] and the classifiers inspired by them are best described in the unified framework proposed by [1]. Define a $K \times M$ *code matrix* H, where K is the number of classes and M is the number of binary classifiers used, as follows: Column m of H is a vector $H_{:m}$ containing the

values $\{0, +1, -1\}$. Denote by C_{m+} (resp. C_{m-}) the set of class indices represented by a "$+1$" (resp. "-1") in $H_{:m}$. Classifier g_m is trained to discriminate between the classes in C_{m+} vs. the classes in C_{m-} on the corresponding subsets of training data $\bigcup_{k \in (C_{m+} \cup C_{m-})} \mathcal{D}_k$.

For example, in the one-vs-all approach, $M = K$, $\forall m \ H_{mm} = 1$ and $H_{km} = -1$ for $k \neq m$. In pairwise coupling, $M = K(K-1)/2$, the variable m indexes the pairs $(k, l),\ 1 \leq k < l \leq K$ and for an m corresponding to (k, l), we have $H_{km} = +1,\ H_{lm} = -1,\ H_{k'm} = 0$ for all $k' \neq k, l$. In the ECC approach [47], H is the matrix of an error correcting code (thus typically $\log_2 K \leq M \leq K$). This ensures that the rows of H (each row representing one class) can be well separated. The label of an unknown input x represents the row $H_{k:}$ of H that is closest in Hamming distance to the vector $\bar{g}(x) = [g_1(x)\ g_2(x)\ \ldots\ g_m(x)]$, i.e., $\hat{y}(x) = \operatorname{argmin}_k d_H(\bar{g}(x),\ H_{k:})$. In general, classification is seen as "encoding" an input x by using M functions g_m, followed by a phase where the vector $\bar{g}(x)$ is "decoded" according to the code matrix H. In the original ECC paper, the classifiers g_m output ± 1. Further work of [39] showed that constructing a coding matrix with ± 1 entries given the M base classifiers is NP-hard and generalized ECC to code matrices with real-entries and real-valued g_m classifiers. The papers of [44, 159, 1] explored the optimal design and learning of H and g.

Boosting is a method of constructing a classifier $\hat{y}$ by sequentially training the blackbox classifiers $g_1,\ g_2, \ldots\ g_m, \ldots$ on differently weighted versions of the same training set $\mathcal{D}$. If one assimilates the columns of H above with weighting functions over the data, then boosting can be seen as part of the same encoding/decoding family of algorithms. Boosting was originally designed for binary classification [177] but it was soon extended in various ways to multiway classification [169, 70, 177, 82, 166].

2.3 Preferences, Tournaments, Transitivity, and Likelihood Ratios

The theory of transitive preferences and tournament games is well studied [116, 184, 132]. In general, there are K objects, $A_k \quad k = 1, \ldots, K$ and a (resp. strict) binary preference relation $\succeq$ (resp. $\succ$) which encodes the preference patterns of an organism for each pair of objects. For example, if $A_i \succeq A_j$ (resp. $A_i \succ A_j$) then A_i is either preferable or indifferent to A_j (resp. A_i is strictly preferable to A_j).

It is common to encode preferences using utility functions $u(A_i)$ [116] that map each object to a real number such that standard comparison operators reflect the exact same preferences as does the $\succ$ relation. For example, if $u(\cdot)$ is such a utility function, then $u(A_i) \geq u(A_j)$ if and only if $A_i \succeq A_j$. It is often noted that utilities are not "true numbers", in that there are many possible values for each utility function which encode exactly the same preferences. Therefore, meaning should not be given to the absolute values of and relative differences between utility values (i.e., use of utilities is insensitive to monotone transformations).

Preference relations and utilities can be used to encode tournament style games. In such a game, there are K players. Pairs of players compete with each other, and there is either a winner or looser (or a tie). If A_i and A_j are players, and if $A_i \succ A_j$, then A_i has won over A_j. Tournaments can be formally defined as a binary relation over a set of objects [132] and can be used to encode tournament-style games such as tennis, football, world-cup soccer, and so on.

It is often assumed that preference relations are transitive. This means that if $A_i \succeq A_j$ and $A_j \succeq A_k$, then $A_i \succeq A_k$. Transitive relations go hand-in-hand with utilities; a set of single real-valued utility functions can be defined that encodes preferences iff these preferences are transitive. In other words, strict transitivity means that the objects being compared can be placed into rank order, where an object's relative rank is used to determine preference. Once objects have a rank, they have a strictly sequential ordering and transitivity results. Transitive tournament style games can be encoded using directed acyclic graphs, where a directed edge points from a looser to a winner. In such a graph, the final winner has no outward pointing edges.

Note that many standard statistical pattern classification methods can be cast in terms of transitive preferences and transitive tournament-style games (this idea was originally presented by the PIs in [19]). The function $L_{ij}(x)$ then codes, in the context of an object represented by x, a preference relationship between either class i and class j as being the category for object x. That preference relation is $C_i \succ C_j \iff L_{ij}(x) > 0$. Given a particular ordering of the classes, $\{C_1, C_2, ..., C_K\}$, a game strategy proceeds by evaluating $L_{C_1 C_2}$ which if positive means that C_1 has "won" so it is followed by evaluating $L_{C_1 C_3}$. If C_1 lost, the procedure is followed instead by evaluating L_{C_3, C_2}. This procedure continues until a final winner is found. If posterior probabilities $p(C_j|x)$ are used rather than likelihoods, the final winner is the same as what is decided by Bayes decision rule. Neither a class order permutation nor a monotone transformation on the probabilities affects the ultimate winner. This is because underlying the relationships are utility functions $u(C_i) = \log p(x|C_i)$ corresponding to the likelihoods (or log posteriors), thus implying transitivity.

2.4 Intransitivity in Decision Making

A given preference relation need not be transitive. An *intransitive relation* arises when there is at least one instance of three objects A_i, A_j, and A_k such that $A_i \succeq A_j$, $A_j \succeq A_k$ but $A_k \succeq A_i$. Perhaps the simplest and most widely known intransitive game is that of rock-paper-scissors where rock wins over (crushes) scissors who wins over (cuts) paper who wins over (covers) rock, thereby creating a loop. Intransitive games can be depicted by directed graphs that posses directed cycles.

There has been much debate in the past regarding whether it is logical for intransitive preferences to exist in natural organisms [116, 122, 3, 2]. Transitivity, for example, is sometimes seen as a consequence of "rational thought", as any rational being would not permit itself to have intransitive preferences. One such argument against intransitivity makes this case quite clearly. Suppose we encounter a person who prefers apples over bananas, bananas over oranges, and oranges over apples. Suppose also that the preferences are strict, such that there is no pairing in which there is an indifference between any two fruit. When such a person is in possession of an apple, he would be willing to exchange it plus a small sum of money for a banana. With a banana, he would exchange it plus money for an orange, and given an orange he would exchange it plus money for an apple. The cycle would then continue indefinitely, leading to what has been called a "money pump." [2, 163, 40]. Some have claimed that such preferences would indicate irrational judgment by the person since economic survival requires that individuals behave as if they had transitive preferences. Others have claimed that intransitive relations (indicating friendship between individuals) are inherently unstable (short lived) [57]. In fact, there are many such discussions against intransitive preferences, but most of these arguments are quite specialized and philosophical in nature. In recent years most of them have fallen out of favor [69, 120, 2, 173], and intransitive preferences are now even seen as being sophisticated [163].

Regardless of the rationality of transitivity, intransitive relations represent many real-world phenomena such as sports tournaments (team A wins over team B, who wins over team C, who then wins over team A). It has also been demonstrated that intransitivity can represent preference patterns in both humans [189, 114, 115, 205, 30] and non-human animals [118, 196, 176, 150, 89, 198, 136, 8].

Intransitive preferences may arise from one of a number of causes. First, preferences can potentially be formed by the use of non-optimal "heuristics and biases" [190], where a decision is formed that can have a high probability of errors — intransitivity results from these errors. On the other hand, intransitive preferences can arise by ever changing preferences, where the act of choosing itself determines a preference [34]. Alternatively, intransitive preferences can result from ecologically advantageous "fast and frugal" or simple heuristics which are, essentially, both computationally cheap but perform well in nature [80, 68]. These heuristics fall under the category of satisficing procedures [81, 80], an instance of "bounded rationality" [181, 180], where the minimum satisfactory condition or outcome is chosen, and where past experience is used to gauge which choices are those that satisfice. Yet a third viewpoint espouses the view that intransitivities can result from context dependences [88]. In this case, the preference depends on a (sometimes hidden or unknown) context, and it is only by not controlling for these contexts do intransitivities arise.

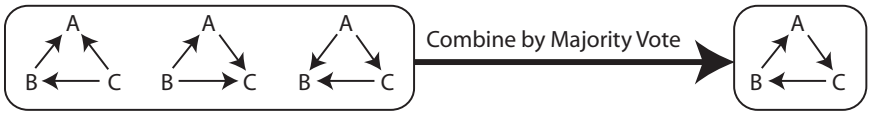


Figure 1: The combination by majority vote of three individuals with transitive preferences (left three graphs) into intransitive preferences (right most graph). Each directed graph depicts pairwise preferences, where an arrow points to the more preferred from the less preferred.

One particularly compelling explanation in favor of intransitivity arises from preference combination in majority voting schemes [3, 125, 163, 164]. In particular, it is possible to combine pairwise preferences of three agents using a majority vote for each pair, and arrive at intransitive preferences. This is shown by the graphs in Figure 1. This is an important example as it suggests that multiple individuals, each of which exhibit purely "transitive" preferences, will exhibit intransitivity when their preferences are jointly and democratically represented (i.e., what has been called "social choice"). Moreover, Arrow's [3] famous theorem shows that, under several reasonable assumptions, no system of elections with more than two candidates is "rational" in this sense except for a dictatorship [3, 163]. This implies that intransitivity can be explained by the presence of multiple "hidden" agents whose preferences are combined in some way An individual's intransitive preferences might also be explained by multi-agent descriptions of mind [131].

Another compelling example of mathematical intransitivity is the notion of *non-transitive dice*, where given three fair cubic 6-faced dice A, B, and C, it is possible to label the faces such that $P(A > B) > 1/2$, $P(B > C) > 1/2$,

and $P(C > A) > 1/2$ [77, 78, 187, 96]. While potentially appearing paradoxical, clearly this example demonstrates that intransitivity can easily arise via a probabilistic model.

Intransitivity also occurs in coevolution [58, 137, 33, 95], a branch of evolutionary science where one population evolves directly in response to the evolution of other (competitive) populations. Coevolutionary algorithms are often used for the evaluation of populations in pairwise comparisons with these other populations, but the existence of intransitivity in such evaluations can often lead to cycles, where previously visited evaluative states are re-visited over and over.

In fact, there are a number of mathematical and statistical models of preference that can result in intransitivity. These include multi-attribute or multi-dimensional preference models [62] including a lexicographic semi-order [189], non-transitive additive models [62], and additive difference models [189]. In the field of consumer demand theory, it is also possible to prove the existence of competitive equilibrium without assuming transitivity [183] and that such intransitive orderings can be represented by a continuous numerical function [175]. Many other models are also possible.

Without a doubt, at this point intransitive preferences have been demonstrated to exist in both natural organisms and mathematical models. Therefore, any understanding of natural decision making should strive to understand the reasons behind and the modeling of such intransitive preferences. Proposed below is methodology that represents multi-class classification as an intransitive game and that can exploit these preferences both for classification and understanding group choice. As will be seen, research is then proposed to represent and exploit this intransitivity in many novel ways.

3 Classification As Intransitive Games

Classification can be seen as an intransitive game, where intransitivity can occur either by "corrected" likelihood-ratio classifiers [19, 91] (Section 3.1), or by a set of binary classifiers used to construct a multiway classifier. We use the notation $S_{ij}(x)$ to indicate a preference for class i over class j given unknown object x. With K classes, there are at most $K(K-1)/2$ such preferences.

3.1 Intransitivity via Augmented Likelihood Ratio

In previous work [19, 91], it was shown how significant error rate reductions can be obtained by making corrections to the *approximate* likelihood ratio defined as $\hat{L}_{ij}(x) \stackrel{\Delta}{=} \log \frac{\hat{p}(x|c_i)}{\hat{p}(x|c_j)}$, so that it is a better estimation of the true likelihood ratio $L_{ij}(x)$. Using this, a preference relation $S_{ij}(x)$ was obtained by likelihood ratio augmentation. In general, there will be error between $\hat{L}_{ij}(x)$ and $L_{ij}(x)$ for each x. The goal is to find the best x-independent correction term α_{ij} that minimizes the error between the corrected likelihood ratio $f(\hat{L}_{ij}(x) + \alpha_{ij})$ and the true likelihood ratio $f(L_{ij}(x))$, where both ratios are first processed by a function $f(\cdot)$ (see below). Note that for simplicity, the techniques as presented here are introduced using likelihoods $p(x|C_k)$ and their estimations $\hat{p}(x|C_k)$, but the methodology may use any of the discriminant functions, BCs, and preference relations as described below, in Sections 2.1, 2.3, and in [91].

The goal is to find the α_{ij} that minimizes the error, defined as the expected squared difference between the corrected approximation and the true term modulo an importance function $f()$, or:

$$J(\alpha_{ij}) = \frac{1}{2} \int \left(f(\hat{L}_{ij}(x) + \alpha_{ij}) - f(L_{ij}(x)) \right)^2 p(x) dx, \tag{1}$$

the minimization of which leads to the optimal alphas $\alpha^*_{ij} = \text{argmin}_{\alpha_{ij}} J(\alpha_{ij})$.

Given these best α_{ij} coefficients, a new preference relation can be defined as $S_{ij}(x) = \hat{L}_{ij}(x) + \alpha_{ij}$, so that $C_i \succ C_j \iff S_{ij}(x) > 0$. A key issue is that $S_{ij}(x)$ can become an intransitive preference relation. First, if the correction terms α_{ij} are perfect, so that $S_{ij}(x) = L_{ij}(x)$, then the true likelihood ratios are obtained, and therefore $S_{ij}(x)$ would be transitive since likelihood-ratio based preferences are always transitive. It is unlikely, however, that prefect correction occurs or is possible. Therefore, since the α_{ij} coefficients are a function of both i and j and there is no longer an underlying utility to rank-order the categories, intransitivity can result. Note that due to the potential intransitivity, the resulting preference relations $S_{ij}(x)$ have escaped from the family of possible models representable purely by a likelihood ratio.

The resulting preference relation, itself, was seen to significantly improve classification accuracy [19], and when seen as a intransitive game, further improvements were found. Specifically, in our previous work [19], the PIs optimized this function using $f(a) = a$ (the identity function), and obtained, under certain assumptions, that the best α_{ij} is a *difference* between two Kullback-Leibler divergence terms.

This corrected term when used as a preference relationship produces significant improvements in classification accuracy over the baseline likelihood ratio on an isolated-word speech recognition task, where the likelihood functions were hidden Markov models (HMMs) [148, 16]. We demonstrated empirically that S_{ij} is in many cases indeed intransitive. More importantly, there were several beneficial strategies for exploiting such intransitivity. These included playing a multi-player tournament a number of times, each with a different random permutation of players. The final class chosen was the most frequent winner among the tournaments, additionally improving classification accuracy. Further, we found that the degree of intransitivity, defined via detected cycles in the winner-looser directed graph, is a good indicator of error: if a large number of different cycles is found, the chance of error occurs is large. Therefore, the degree of intransitivity appears to be a useful *confidence* or *inverse robustness* measure.

In subsequent work [91], it was demonstrated that the methodology presented above is general beyond just isolated-word speech recognition, and showed improvements on the UCI machine learning data sets [29, 87]. This was done using both $f(a) = a$ as above, and $f(a) = \text{sign}(a)$ and various of its continuous differentiable approximations. The sign function case was called "necessary corrections" since it asks for the least change to $\hat{L}_{ij}(x)$ to achieve sign equality with $\hat{L}_{ij}(x)$, the minimum that is necessary to obtain perfect classification. Moreover, improvements were found for two baseline classifiers: 1) multi-layer perceptrons trained using a KL-divergence error function [28, 153] where the term corrected was a ratio of posterior probabilities; and 2) Naive Bayes classifiers [73, 49, 108]. In both cases, there were significant improvements over the baseline.

References

[1] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. In *Proc. 17th International Conf. on Machine Learning*, pages 9–16. Morgan Kaufmann, San Francisco, CA, 2000.

[2] P. Anand. The philosophy of intransitive preference. *Economic Journal*, 103(417):337–346, March 1993.

[3] K. J. Arrow. *Social Choice and Individual Values (2nd ed.)*. Yale University Press, 1963.

[4] A. Artale. A model of stability and persistence in a democracy. *Games and Economic Behavior*, 33:20–40, 2000.

[5] P. Avery. An algorithmic proof that semiorders are representable. *Journal of Algorithms*, 13:144–147, 1992.

[6] L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer. Maximum mutual information estimation of HMM parameters for speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 49–52, Tokyo, Japan, December 1986.

[7] M. Bar-Hillel and A. Margalit. How vicious are cycles of intransitive choice? *Theory and Decision*, 24(2):119—145, March 1988.

[8] M. Bateson, S. Healy, and T. A. Hurly. Irrational choices in hummingbird foraging behavior. *Animal Behaviour*, 63:587–596, 2002.

[9] R. Battiti and A.M. Colla. Democracy in neural nets: voting schemes for classification. *Neural Networks*, 7:691–707, 1994.

[10] Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: bagging, boosting and variants. *MAchine learning*, 36:105–142, 1999.

[11] E. Bax. Validation of voting committees. *Neural Computation*, 11(4):975–986, 1998.

[12] Yakov Ben-Haim and Keith W.Hipel. The graph model for conflict resolution with information-gap uncertainty in preferences. *Applied Mathematics and Communication*, pages 319C–340, 2002.

[13] J. J. Bernardo and D. E. Upton. Stochastic and intransitive behavior in a state-preference model of asset choice. *Decision-Sciences*, 23(5):1114–1126, 1992.

[14] J. Bilmes. Pattern recognition II: Introduction to graphical models. http://www.ee.washington.edu/~bilmes/pr2/.

[15] J. Bilmes. The GMTK documentation, 2002. `http://ssli.ee.washington.edu/~bilmes/gmtk`.

[16] J. Bilmes. What HMMs can do. Technical Report UWEETR-2002-003, University of Washington, Dept. of EE, 2002.

[17] J. Bilmes. Graphical model research in audio, speech, and language processing. In *Uncertainty in Artificial Intelligence: Nineteenth Conference (UAI-2003), Tutorial Program*. Morgan Kaufmann Publishers, 2003. `http://research.microsoft.com/uai2003/TutorialAbstract.htm`.

[18] J. Bilmes and C. Bartels. On triangulating dynamic graphical models. In *Uncertainty in Artificial Intelligence: Proceedings of the Nineteenth Conference (UAI-2003)*, pages 47–56. Morgan Kaufmann Publishers, 2003.

[19] J. Bilmes, G. Ji, and M. Meilă. Intransitive likelihood-ratio classifiers. In *Neural Information Processing Systems (NIPS)*, 14, Vancouver, Canada, December 2001.

[20] J. Bilmes and K. Kirchhoff. Generalized rules for combination and joint training of classifiers. *Pattern Analysis and Applications*, (in press). Springer.

[21] J. Bilmes and G. Zweig. The Graphical Models Toolkit: An open source software system for speech and time-series processing. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 2002.

[22] J. Bilmes, G. Zweig, T. Richardson, K. Filali, K. Livescu, P. Xu, K. Jackson, Y. Brandman, E. Sandness, E. Holtz, J. Torres, and B. Byrne. Discriminatively structured graphical models for speech recognition: JHU-WS-2001 final workshop report. Technical report, CLSP, Johns Hopkins University, Baltimore MD, 2001. `http://www.clsp.jhu.edu/ws2001/groups/gmsr/GMRO-final-rpt.pdf`.

[23] J. A. Bilmes. Graphical models and automatic speech recognition. In R. Rosenfeld, M. Ostendorf, S. Khudanpur, and M. Johnson, editors, *Mathematical Foundations of Speech and Language Processing*. Springer-Verlag, New York, 2003.

[24] J.A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report TR-97-021, ICSI, 1997.

[25] J.A. Bilmes. Buried Markov models for speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, March 1999.

[26] J.A. Bilmes. Dynamic Bayesian Multinets. In *Proceedings of the 16th conf. on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 2000.

[27] J.A. Bilmes and K. Kirchhoff. Directed graphical models of classifier combination: Application to phone recognition. In *Proc. Int. Conf. on Spoken Language Processing*, Beijing, China, 2000.

[28] C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[29] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[30] D. Bouyssou and P. Vincke. Introduction to topics on preference modelling. *Annals Operations Research*, 80:U11—U24, 1998.

[31] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[32] Leo Breiman. Bagging predictors. *Machine learning*, 26(2):123–140, 1996.

[33] A. Bucci and J.B. Pollack. Focusing versus intransitivity. geometrical aspects of coevolution. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-03)*, Berlin, 2003. Springer.

[34] D. J. Butler. A choice-rule formulation of intransitive utility theory. *Economics Letters*, 59:323–329, 1998.

[35] J. Chipman, L. Hurwicz, M. Richter, and H. Sonnenschein, editors. *Preferences, Utility, and Demand*. Harbrace Series in Business and Economics, 1971.

[36] Citeseer: Scientific literature digital library. NEC. `http://citeseer.nj.nec.com`.

[37] W. W. Cohen. Fast effective rule induction. In A. Prieditis and S. Russell, editors, *Proceedings of the 12th International Conference on Machine Learni ng (ML-95)*, pages 115–123. Morgan Kaufmann, 1995.

[38] A.H. Copeland. A reasonable social welfare function. Technical report, University of Michigan, 1951. Seminar on Applications of Mathematics to Social Sciences.

[39] Koby Crammer and Yoram Singer. On the learnability and design of output codes for multiclass problems. In *Proceedings of the Thirteen Annual Conference on Computational Learning Theory (COLT)*, pages 35–46, 2000.

[40] R. P. Cubitt and R. Sugden. On money pumps. *Games and Economic Behavior*, 37:121–160, 2001.

[41] Florin Cutzu. Polychotomous classification with pairwise classifiers: a new voting principle. *Lecture Notes in Computer Science*, 2709/2003:115–124, January 2003.

[42] Florin Cutzu. Polychotomous classification with pairwise classifiers: a new voting principle. Technical report, Computer Science, Indiana University, January 2003.

[43] C. Mitchell Dayton. Information criteria for the paired-comparisons problem. *The American Statistician*, 52(2):144–151, May 1998.

[44] O. Dekel and Y. Singer. Multiclass learning by probabilistic embeddings. In *Advances in Neural Information Processing Systems 15*, 2002.

[45] T. G. Dietterich. An experimental comaprison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization. *Machine learning*, pages 1–22, 1999.

[46] T. G. Dietterich. Ensemble methods in machine learning. In F. Roli, editor, *First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science*, New York, 2000. Springer Verlag.

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

[48] Discrete applied mathematics: Submodularity. Elsevier Science Publishers B. V., September 2003. Special Issue, ISSN: 0166-218X, 131(2), ACM.

[49] Pedro Domingos and Michael J. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

[50] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, Inc., 1973.

[51] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley and Sons, Inc., 2000.

[52] Ivo D untsch. *Relational Methods in Computer Science: 6th International Conference*, chapter Tangent Circle Algebras, pages 300–313. Springer-Verlag, 2002.

[53] Y. Ephraim, A. Dembo, and L. Rabiner. A minimum discrimination information approach for HMM. *IEEE Trans. Info. Theory*, 35(5):1001–1013, September 1989.

[54] Y. Ephraim and L. Rabiner. On the relations between modeling approaches for information sources. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 24–27, 1988.

[55] Y. Ephraim and L. Rabiner. On the relations between modeling approaches for speech recognition. *IEEE Trans. Info. Theory*, 36(2):372–380, September 1990.

[56] `http://ekI.go.jp/~etIcdb`. Small database of Hiraganas.

[57] L. Festinger. *A Theory of Cognitive Dissonance*. Stanford Univ Press, 1957.

[58] S.G. Ficici and J. B. Pollack. Pareto optimality in coevolutionary learning. In J. Kelemen and P. Sosik, editors, *Advances in Artificial Life: 6th European Conference (ECAL 2001)*. Springer, 2001.

[59] J.G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, California, 1997.

[60] P. C. Fishburn. Intransitive indifference in preference theory: A survey. *Operations Research*, 18(2):202–228, 1970.

[61] P. C. Fishburn. Intransitive individual indifference and transitive majorities. *Econometrica*, 38(3):482–489, 1970.

[62] P. C. Fishburn. Nontransitive preferences in decision theory. *Journal of Risk and Uncertainty*, 4, 1991.

[63] P. C. Fishburn and I. H. LaValle. A nonlinear, nontransitive and additive-probability model for decisions under uncertainty. *The Annals of Statistics*, 15(2):830–844, 1987.

[64] P. C. Fishburn and I. H. LaValle. Context-dependent choice with nonlinear and nontransitive preferences. *Econometrica*, 56(5):1221–1239, 1988.

[65] P.C. Fishburn and I.H. LaValle. Subjective expected lexicographic utility. *Annals of Operations Research*, 80:183—206, 1998.

[66] L. Fleischer. Recent progress in submodular function minimization. *Optima: Mathematical Programming Society Newsletter*, September 2000.

[67] M. M. Flood. Implicit intransitivity under majority rule with mixed motions. *Management Science*, 26(3):312–321, March 1980.

[68] M. R. Forster. How do simple rules 'fit to reality' in a complex world. *Minds and Machines*, 9(1999):543–564, 1999.

[69] J. Fountain. Consumer surplus when preferences are intransitive: Analysis and interpretation. *Econometrica*, 49(2), 1981.

[70] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

[71] J. Friedman. Another approach to polychotomous classification. Technical report, Department of Statistics, Stanford University, 1996.

[72] J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141, 1991.

[73] J.H. Friedman. On bias, variance, 0/1–loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

[74] K. Fukunaga. *Introduction to Statistical Pattern Recognition, 2nd Ed.* Academic Press, 1990.

[75] Johannes Fürnkranz. Round robin rule learning. In C. E. Brodley and A. P. Danyluk, editors, *Proceedings of the 18th International Conference on Machine Learning (ICML-01)*, pages 146–153, Williamstown, MA, 2001. Morgan Kaufmann Publishers.

[76] Johannes Fürnkranz. Round robin ensembles. *Intelligent Data Analysis*, 7(5), 2003.

[77] M. Gardner. The paradox of nontransitive dice and the elusive principle of indifference. *Scientific American*, 223:110–114, 1970.

[78] M. Gardner. On the paradoxical situations that arise from nontransitive relations. *Scientific American*, 231:120–125, 1974.

[79] S. I. Gass. Tournaments, transitivity and pairwise comparison matrices. *Journal of the Operational Research Society*, 49:616–624, 1998.

[80] G. Gigerenzer. *Adaptive Thinking: Rationality in the Real World*. Oxford University Press, 2000.

[81] M.A. Goodrich, W.C. Stirling, and E.R. Boer. Satisficing revisited. *Minds and Machines*, 10:79–100, 2000.

[82] V. Guruswami and A. Sahai. Multiclass learning, boosting, and errorcorrecting codes. In *In Proc. of the Twelfth Annual Conference on Comptational Learning Theory*, pages 145–155. ACM Press, 1999.

[83] F. Harary and E. M. Palmer. *Graphical Enumeration*. Academic Press, 1973.

[84] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2001.

[85] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[86] J. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the Theory of Neural Computation*. Allan M. Wylde, 1991.

[87] S. Hettich and S.D. Bay. The UCI KDD archive. University of California, Department of Information and Computer Science, 1999. Irvine, CA.

[88] A. I. Houston. Natural selection and context-dependent values. *Proceedings of the Royal Society B*, 264:1539–1541, 1977.

[89] T.A. Hurly and M.D. Oseen. Context-dependent risk-sensitive foraging preferences in wild rufous hummingbirds. *Animal Behavior*, 58:59–66, 1999.

[90] Tommi Jaakkola, Marina Meilă, and Tony Jebara. Maximum entropy discrimination. In Sara A. Solla, Todd K. Leen, and Klaus-R. Müller, editors, *Neural Information Processing Systems*, volume 12, pages 470–476. MIT Press, 2000.

[91] G. Ji and J. Bilmes. Necessary intransitive likelihood-ratio classifiers. In *Neural Information Processing Systems (NIPS)*, 16, Vancouver, Canada, December 2003.

[92] D. B. Johnson. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing*, 4(1):77–84, March 1975.

[93] Bradford Jones, Benjamin Radcliff, Charles Taber, and Richard Timpone. Condorcet winners and the paradox of voting: probability calculations for weak preference orders. *American Political Science Review*, 89(1):137–144, March 1995.

[94] D. S. Jones. *Generalised functions.* McCraw-Hill Publishing Company Limited, 1966.

[95] E.D. De Jong. Intransitivity in coevolution. In *Proceedings of the 8th International Conference on Parallel Problem Solving from Nature (PPSN-04)*, pages 843—851, 2004.

[96] R.P. Savage Jr. The paradox of nontransitive dice. *American Mathematical Monthly*, 101:429–436, May 1994.

[97] B.-H. Juang, W. Chou, and C.-H. Lee. Minimum classification error rate methods for speech recognition. *IEEE Trans. on Speech and Audio Signal Processing*, 5(3):257–265, May 1997.

[98] B-H Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Trans. on Signal Processing*, 40(12):3043–3054, December 1992.

[99] J. Junkins, Z. Rahman, and H. Bang. Near minimum-time maneuvers of distributed parameter systems: analytical and experimental results. *Journal of Guidance, Control, and Dynamics*, 14:406–415, 1991.

[100] L. P. Kaelbling, M. L. Littman, and A. P. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[101] M.G. Kendall and B. Babington Smith. On the method of paired comparisons. *Biometrika*, 31(3/4):131–139, 1940.

[102] D. M. Kilgour, L. Fang, and K.W. Hipel. General preference structures in the graph model for conflicts. *Information and Decision Technologies*, 16:291–300, 1990.

[103] K. Kirchhoff and J. Bilmes. Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values. *Proceedings ICASSP-99*, pages 693–696, 1999.

[104] K. Kirchhoff and J. Bilmes. Generalized acoustic classifier combination for speech recognition. In *Proc. of the Automatic Speech Recognition (ASR) Workshop*, Paris, FR, October 2000.

[105] K. Kirchhoff, S. Parandekar, and J. Bilmes. Mixed-memory Markov models for automatic language identification. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Orlando, Florida, 2002.

[106] J. Kittler, M. Hataf, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[107] U.H.-G. Kreßel. Pairwise classification and support vector machines. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods: Support Vector Learning, 15*, pages 255–268. MIT Press, Cambridge, MA, 1998.

[108] David D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 4–15, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[109] Qunhua Li and Marina Meilă. Clustering by intersection-merging. Technical Report 451, University of Washington, Department of Statistics, 2004.

[110] S. K. Lioukas. Thresholds and transitivity in stochastic consumer choice: a multinomial logit analysis. *Management Science*, 30(1):110–122, Jan 1984.

[111] T. Lissack and K.S. Fu. Error estimation in pattern recognition via l^{α}-distance between posterior density functions. *IEEE Transactions on Information Theory*, IT-22(1):34–45, January 1976.

[112] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling. Learning policies for partially observable environments: Scaling up. In A. Prieditis and S. Russell, editors, *Proceedings of the Twelfth International Conference on Machine Learning*, pages 362–370, San Francisco, CA, USA, 1995. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

[113] Z.-Q. Liu, J.H. Cai, and R. Buse. *Handwriting recognition : soft computing and probabilistic approaches*. Springer-Verlag, 2003.

[114] G. Loomes, C. Starmer, and R. Sugden. Preference reversal: Information-processing effect or rational non-transitive choice? *The Economic Journal*, 99(395):149–151, 1989. Supplement: Conference Papers.

[115] G. Loomes and C. Taylor. Non-transitive preferences over gains and losses. *The Economic Journal*, 102(411):357–365, 1992.

[116] R. D. Luce and H. Raiffa. *Games and Decisions*. Dover, 1957.

[117] R.D. Luce. Semiorders and a theory of utility discrimination. *Econometrica*, 24(2):178–191, 1956.

[118] R.D. Luce. *Individual Choice Behavior*. Wiley, New York, 1959.

[119] O. Madani, S. Hanks, and A. Gordon. On the undecidability of probabilistic planning and infinite-horizon partially observable markov decision processes. In *In Proceedings of the Sixteenth National Conference in Artificial Intelligence*, 1999.

[120] P. Maher. *Betting on Theories*. Cambridge University Press, New York, 1993.

[121] A. Mas-Colell. An equilibrium existence theorem without complete or transitive preferences. *Journal of Mathematical Economics*, 1:237–246, 1974.

[122] K. O. May. Intransitivity, utility, and the aggregation of preference patterns. *Econometrica*, 22(1):1–13, January 1954.

[123] Eddy Mayoraz and Ethem Alpaydin. Support vector machines for multi-class classification. In *IWANN (2)*, pages 833–842, 1999.

[124] Eddy Mayoraz and Miguel Moreira. On the decomposition of polychotomies into dichotomies. In *Proc. 14th International Conference on Machine Learning*, pages 219–226. Morgan Kaufmann, 1997.

[125] R. D. McKelvey. General conditions for global intransitivites in formal voting models. *Econometrica*, 47(5):1085–1112, September 1979.

[126] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Statistics, 1992.

[127] Marina Meilă and William Pentney. Minimizing normalized cuts in asymmetric graphs. Department of Statistics TR (in preparation), University of Washington, 2005.

[128] Marina Meilă and Jianbo Shi. Learning segmentation by random walks. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 873–879, Cambridge, MA, 2001. MIT Press.

[129] Marina Meilă, Susan Shortreed, and Liang Xu. Regularized spectral learning. In Robert Cowell and Zoubin Ghahramani, editors, *Proceedings of the Artificial Intelligence and Statistics Workshop(AISTATS 05)*, 2005.

[130] V. Merlin and D. Saari. Copeland method ii: Manipulation, monotonicity, and paradoxes. *Journal of Economic Theory*, 72:148–172, 1997.

[131] M. Minsky. *Society of Mind*. Simon and Schuster, New York, 1985.

[132] H. Monsuur and T. Storcken. Measuring intransitivity. *Mathematical Social Sciences*, 34(2):125–152, 1997.

[133] Miguel Moreira and Eddy Mayoraz. Improved pairwise coupling classification with correcting classifiers. In *European Conference on Machine Learning*, pages 160–171, 1998.

[134] H. W. Morrison. *Intransitivity of paired comparison choices*. PhD thesis, University of Michigan, 1962.

[135] M. Narasimhan and J. Bilmes. PAC-learning bounded tree-width graphical models. In *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI-2004)*. Morgan Kaufmann Publishers, 2004.

[136] D.J. Navarick and E. Fantino. Transitivity as a property of choice. *Journal of the Experimental Analysis of Behavior*, 18(3):389–401, 1972.

[137] J. Noble and R. A. Watson. Pareto coevolution: Using performance against coevolved opponents in a game as dimensions for pareto selection. In L. Spector et. al., editor, *Proc. 2001 Genetic and Evo. Comp. Conf.* Morgan Kaufmann, 2001.

[138] G. Owen. *Game Theory: 3rd Ed.* Academic Press, 2001.

[139] C. Papadimitriou and J. Tsitsiklis. The complexity of markov decision processes. *Mathematics of Operation Research*, 12(3), 1987.

[140] B. Parhami. Voting algorithms. *IEEE Trans. Reliability*, 43(4):617–629, 1994.

[141] E. A. Patrick and F. P. Fischer II. Nonparametric feature selection. *IEEE Transactions on Information Theory*, IT-15(5):577–584, September 1969.

[142] Anne Patrikainen and Marina Meilă. Comparing subspace clusterings. *IEEE Transactions on Knowledge and Data Engineering*, page (submitted), 2005.

[143] William Pentney and Marina Meilă. Spectral clustering of biological sequence data. In Manuela Veloso and Subbarao Kambhampati, editors, *Proceedings of Twentieth National Conference on Artificial Intelligence (AAAI-05)*, 2005.

[144] J. Pineau, G. Gordon, and S. Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In *In International Joint Conference on Artificial Intelligence*, 2003.

[145] J. Pitrelli, C. Fong, S.H. Wong, J.R. Spitz, and H.C. Lueng. PhoneBook: A phonetically-rich isolated-word telephone-speech database. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1995.

[146] J. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In *Advances in Neural Information Processing Systems*, volume 12, pages 546–553, 1999.

[147] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, 1994.

[148] L.R. Rabiner and B.H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 1986.

[149] R.C. Read and R. E. Tarjan. Bounds on backtrack algorithms for listing cycles, paths and spanning trees. *Networks*, 5:237–252, 1975.

[150] L. Real. Paradox, performance, and the architecture of decision- making in animals. *American Zoologist*, 36(4):518–529, 1996.

[151] D. Reynolds, B. Peskin, J. Navratil, J. Campbell, W. Andrews, D. Klusacek, A. Adami, Q. Jin, J. Abramson, R. Mihaescu, J. Goddfrey, D. Jones, and B. Xiang. Supersid: Exploiting high-level information for high-performance speaker recognition. Technical report, CLSP, Johns Hopkins University, Baltimore MD, 2002.

[152] J.A. Richards and X. Jia. *Remote Sensing Digital Image Analysis: An Introduction, 3rd Ed.* Springer-Verlag, New York, 1993.

[153] B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[154] P. H. M. P. Roelofsma and D. Read. Intransitive intertemporal choice. *Journal of Behavioral Decision Making*, 13(2):161–177, April-June 2000.

[155] A. W. Roscoe and M. H. Goldsmith. What is intransitive noninterference? In *Proceedings of the 12th IEEE Computer Security Foundations Workshop*, pages 228–238, 1999.

[156] V. Roth and K. Tsuda. Pairwise coupling for machine recognition of handprinted japanese characters. In *In CVPR*, pages 1120–1125, 2001.

[157] Volker Roth. Probabilistic discriminative kernel classifiers for multi-class problems. *Lecture Notes in Computer Science*, 2191:246–266, 2001.

[158] N. Roy and G. Gordon. Exponential family pca for belief compression in pomdps. In *Advances in Neural Information Processing Systems*, 2003.

[159] Gunnar Rtsch, Alexander J. Smola, and Sebastian Mika. Adapting codes and embeddings for polychotomies. In S. Becker, Sebastian Thrun, and David Cohn, editors, *Neural Information Processing Systems*, volume 15. MIT Press, 2003.

[160] Salvador Ruiz-Correa, Linda Shapiro, Marina Meilă, and Gabriel Berson. Discriminating deformable shape classes. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Neural Information Processing Systems*, volume 15, Cambridge, MA, 2004. MIT Press.

[161] S.J. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach, 2nd Ed*. Prentice Hall, 2003.

[162] D. Saari and D. Merlin. The copeland method i; relationships and the dictionary. *Economic Theory*, 1996.

[163] Donald G. Saari. *Geometry of Voting*. Springer-Verlag, 1994.

[164] Donald G. Saari. *Decisions and Elections, Explaining the Unexpected*. Cambridge, 2001.

[165] Petr Savicky and Johannes Frnkranz. Combining pairwise classifiers with stacking. *Lecture notes in Computer Science (LNCS)*, 2810:219–229, September 2003.

[166] R. E. Schapire. The boosting approach to machine learning: An overview. In *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

[167] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 27(5):1651—1686, 1998.

[168] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26(5):1651–1686, 1998.

[169] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. In *Computational Learing Theory*, pages 80–91, 1998.

[170] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[171] D. Scott and P. Suppes. Foundational aspects of theories of measurement. *Journal of Symbolic Logic*, 23:113–128, 1958.

[172] U. Segal. Stochastic transitivity and quadratic representation functions. *Journal of Mathematical Psychology*, 38:102–114, 1994.

[173] A. K. Sen. Rationality and uncertainty. *Theory and Decision*, 18:109–27, 1985.

[174] W. Shafer and H. Sonnenschein. Equilibrium in abstract economies without ordered preferences. *Journal of Mathematical Economics*, 2:345–348, 1975. Corrections in 1979, Vol. 6, p.297- 8.

[175] Wayne J. Shafer. The nontransitive consumer. *Econometrica*, 42(5):913–919, September 1974.

[176] S. Shafir. Intransitivity of preferences in honey bees: support for "comparative" evaluation of foraging options. *Animal Behaviour*, 48:55—67, 1994.

[177] R.E. Shapire. The strength of weak learnability. *Machine Learning*, 5:1990, 197–227.

[178] Susan Shortreed and Marina Meilă. Unsupervised spectral learning. In *Proceedings of the 21st Conference on Uncertainty in AI*, page (submitted), 2005.

[179] R. Shostak. Deciding linear inequalities by computing loop residues. *Journal of the Association for Computing Machinery*, 28(4):769–779, October 1981.

[180] H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63:129–138, 1956.

[181] H. A. Simon. *Models of Bounded Rationality*. MIT Press, 1982.

[182] P. Slater. Inconsistencies in a schedule of paired comparisons. *Biometrika*, 48(3/4):303–312, 1961.

[183] H. Sonnenschein. Demand theory without transitive preferences, with applications tto the theory of competitive equilibrium. In J.S. Chipman, L. Hurwicz, M.K. Richter, and H.F. Sonnenschein, editors, *Preferences, Utility and Demand*. Harcourt Brace Jovanovich, Inc., New York, 1971.

[184] P.D. Straffin. *Game Theory and Strategy*. The Mathematical Association of America, 1993.

[185] R. Suton and A. Barto. *Reinforcement Learning*. MIT Press, 1998.

[186] A. Tarski. On the calculus of relations. *J. Symbolic Logic*, 6(73–89), 1941.

[187] R. Tenney and C.C. Foster. Non-transitive dominance. *Mathematics Magazine*, 49(3), 1976.

[188] Jarle Tufto, Erling Johan Solberg, and Thor-Harald Ringsby. Statistical models of transitive and intransitive dominance structures. *Animal Behaviour*, 55(6):1489–1498, June 1998.

[189] A. Tversky. Intransitivity of preferences. *Psychological Review*, 76:31–48, 1969.

[190] A. Tversky. Judgement under uncertainty: heuristics and biases. *Science*, 185:1124–1131, 1974.

[191] A. Tversky and I. Simonson. Context-dependent preferences. *Management Science*, 39(10), 1993.

[192] A.B. Urken. Social choice theory and distributed decision making. In R. Allen, editor, *Proceedings of the IEEE/ACM Conference on Office Information Systems.*, 1988.

[193] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[194] Xavier Vila. On the intransitivity of preferences consistent with similarity relations. *Journal of Economic Theory*, 79(2):281–287, April 1998.

[195] M. Voorneveld. Numerical representation of incomplete and nontransitive preferences and indierences on a countable set. Technical Report JEL Classification: D11, Department of Econometrics and CentER, Tilburg University, the Netherlands, 1999.

[196] T. A. Waite. Intransitive preferences in hording gray jays (Perisoreus canadensis). *Behavioral Ecology and Sociobiology*, 50(2):116–121, 2001.

[197] H. Wechsler, J. P. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors. *Face Recognition: From Theory To Applications*, June/July 1997.

[198] D.D. Wiegmann, D.A. Wiegmann, J. MacNeal, and J.Gafford. Transposition of flower height by bumble bee foragers (bombus impatiens). *Animal Cognition*, 3:85–89, 2000.

[199] P.C. Woodland and D. Povey. Large scale discriminative training for speech recognition. In *ICSA ITRW ASR2000*, 2000.

[200] J. Wu. On sorting an intransitive total ordered set using semi-heap. In IEEE, editor, *14th International Parallel and Distributed Processing Symposium (IPDPS'00)*, Cancun, Mexico, May 2000.

[201] Jie Wu. On finding a hamiltonian path in a tournament using semi-heap. *Parallel Processing Letters*, 10(4):279–294, Dec 2000.

[202] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. In Sebastian Thrun, Lawrence Saul, and Bernhad Scholkofp, editors, *Neural Information Processing Systems*, volume 16. MIT Press, 2004.

[203] S. Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine*, 13(5):45–56, September 1996.

[204] B. Zadrozny. Reducing multiclass to binary by coupling probability estimates. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Neural Information Processing Systems*, volume 14, pages 1041–1048, Cambridge, MA, 2001. MIT Press.

[205] T. Van Zandt. Hidden information acquisition and static choice. *Theory and Decision*, 40:235–247, 1996.