Does Cognitive Science Need Kernels?

Frank Jäkel
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, 02139, MA, USA

Bernhard Schölkopf
MPI für biologische Kybernetik
Spemannstr. 38
72076 Tübingen, FRG

Felix A. Wichmann
Technische Universität Berlin
FR 6-4, Franklinstr. 28/29
10587 Berlin, FRG

Kernel methods are among the most successful tools in machine learning and are used in challenging data-analysis problems in many disciplines. Here we provide examples where kernel methods have proven to be powerful tools for analyzing behavioral data, especially for identifying features in categorization experiments. We also demonstrate that kernel methods relate to perceptrons and exemplar models of categorization. Hence, we argue that kernel methods have neural and psychological plausibility and theoretical results about their behavior are therefore potentially relevant for human category learning. In particular, we think that kernel methods show the prospect of providing explanations ranging from the implementational via the algorithmic to the computational level.

Learning in Humans and Machines

Researchers in the field of machine learning study algorithms that are able to learn from data. Since learning is an important aspect of intelligent behavior, machine learning has become a central aspect of research in artificial intelligence. If machines are to behave intelligently in real-world scenarios they will have to adapt autonomously to uncertain environments. Modern machine learning is therefore deeply rooted in probability theory and statistics; fields that deal with modeling of and reasoning with uncertainty. Machine learning has some of its roots in cognitive science, too. Since the clearest examples of learning happen in humans and animals, early research in machine learning was heavily influenced by ideas from psychology and neuroscience. Whole subfields of machine learning deal with reinforcement learning (Sutton & Barto, 1998) or learning in artificial neural networks (Bishop, 1995; Hinton, Osindero, & Teh, 2006). Biologically inspired algorithms have, however, become less and less popular in machine learning as the theoretical understanding of the statistical aspects of learning progressed. At the same time, researchers in machine learning have broadened their interests in applications considerably. Machine learning algorithms are now successfully applied in such diverse areas as bioinformatics (Schölkopf, Tsuda, & Vert, 2004) or collaborative filtering (Marlin, 2004). Many of these applications are far removed from the core interests of cognitive scientists. Since the computational constraints that these applications impose on learning in terms of space, time and data are very different from the constraints that humans or animals have, most of the algorithms that are used do not seem to be psychologically or biologically plausible as models for human or animal learning. Hence, it is no surprise that today interactions between researchers interested in machine learning and researchers interested in human or animal learning are fewer and less intense as they were during the heydays of neural networks, be they perceptrons (Rosenblatt, 1958) or parallel distributed processing models (Rumelhart & McClelland, 1986).

There are, nevertheless, good reasons why cognitive scientists should care about the path that machine learning has taken in the time since neural networks went out of fashion. First, machine-learning methods are used for challenging data-analysis problems in many fields and they can be used for data-analysis in cognitive science, too. Second, many of the problems that machine-learning techniques try to address are still similar to the problems that a human learner faces—even if, on first glance, these techniques do not seem to be plausible psychological models for learning. If there is progress in the theoretical understanding of the core problems of learning in machines this will very likely have an im-

This work was funded, in part, by the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research and the DFG. We would like to thank Y. Katz and P. Battaglia for their helpful comments.

Notice: This is the authors version of a work accepted for publication by Elsevier. Changes resulting from the publishing process, including peer review, editing, corrections, structural formatting and other quality control mechanisms, may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. The definitive version has been published in *Trends in Cognitive Sciences*, Vol. 13, No. 9, 2009, DOI: 10.1016/j.tics.2009.06.002.

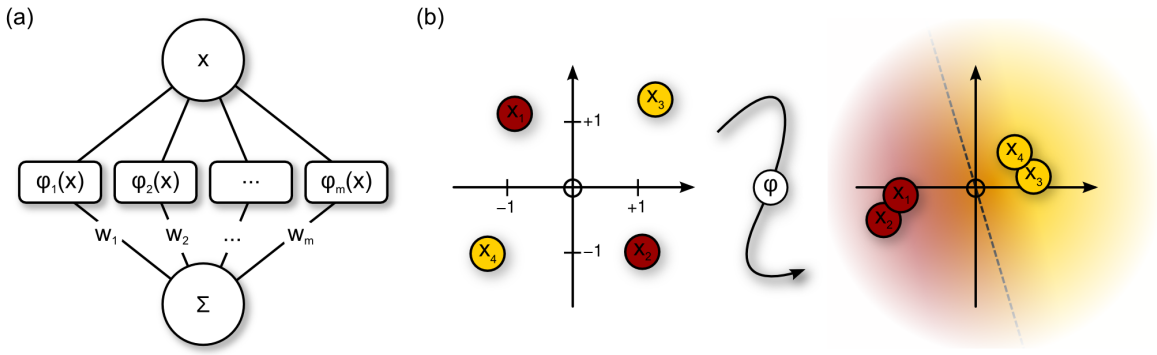

Figure 1. A perceptron for discriminating two categories A and B is defined by a set of feature detectors φ_1, ... φ_m. For a stimulus x the responses for all the feature detectors are calculated and a weighted sum, using weights w_1 to w_m, is formed. This process is illustrated in panel (a). Each of the feature detectors is represented by a node in the network. The output node collects the responses of all the feature detectors. If the output is greater than a threshold the perceptron responds with category A otherwise it responds with category B. Graphically, the perceptron maps all stimuli, for example the four stimuli on the left side of panel (b), to a feature space where the perceptron defines a linear decision boundary (dashed line on the right). Learning a new category distinction can be accomplished by adapting the weights and thereby the linear decision boundary. Note that the four stimuli in panel (b) form a XOR pattern that cannot be separated by a linear decision boundary in the original representation but can be learned after mapping the stimuli to the feature space.

pact on our theoretical understanding of learning in humans, and vice versa. After all, on the computational level (Marr, 1982) humans and machines both face the same problem: they ought to learn from data.

During recent years both of these roles that machine learning can play within cognitive science, data-analysis and computational-level modeling, have been articulated clearly in a Bayesian framework (Lee, 2008; Chater, Tenenbaum, & Yuille, 2006). Here, we will focus on a different set of techniques from machine learning, called kernel methods (Jäkel, Schölkopf, & Wichmann, 2007; Hofmann, Schölkopf, & Smola, 2008; Schölkopf & Smola, 2002). Contrary to multilayer neural networks, kernel methods are linear methods, in a way we will describe in more detail below. They combine the simplicity of linear methods with the flexibility of nonlinear models. We will give some examples where kernel methods have proven to be powerful tools for analyzing behavioral data, especially for identifying features in categorization experiments. We will also show that kernel methods can naturally be linked to perceptrons and exemplar models of categorization (Medin & Schaffer, 1978; Nosofsky, 1986). Hence, we argue that kernel methods have neural and psychological plausibility and theoretical results about their behavior are therefore potentially relevant for human category learning.

Overview of Kernel Methods

There are two complementary views of kernel methods. In the context of category learning cognitive scientists might call them the perceptron view and the exemplar view. Imagine a subject is presented with a stimulus and faces the task of putting this stimulus in one of two categories. The subject will note certain aspects of the stimulus and map the stimulus to an internal representation using (usually non-linear) feature detectors. A perceptron is a linear combination of the extracted features together with a threshold element (Rosenblatt, 1958). The feature detectors play an important role in determining the categorization behavior and the learning capabilities of the perceptron (see Fig. 1).

Usually, one does not know in advance which feature detectors will be useful for learning a new category. Hence, one would like to allow a great number of different feature detectors. However, in the extreme, if the number of independent feature detectors m is greater than the number of stimuli n that a subject encounters one will have to estimate m parameters (the sign and relative importance of each feature detector) from $n < m$ data points (the stimuli). In this case it seems hard to learn anything from the training examples. It is nevertheless possible by using a technique called regularization (see Box 1).

The so-called representer theorem (Schölkopf & Smola, 2002) states that the optimal weights in a regularized learning problem are a linear combination of the training examples in feature space. Classic learning rules like the perceptron rule or the delta rule share this property. The response of the perceptron to a stimulus can then be expressed as a weighted combination of the similarity of the stimulus to the training exemplars. This is what exemplar models in cognitive psychology do (Kruschke, 1992; Nosofsky, 1992; Jäkel, Schölkopf, & Wichmann, 2008a) (see Box 2 for details and Fig. 2 for an illustration of an exemplar model). Hence, the response of the perceptron in a potentially high-dimensional feature space can equivalently be expressed as an exemplar model with a suitably defined similarity measure. The similarity measure is called a *kernel* in machine learning, hence the term kernel methods. Interestingly, many similarity measures that are used in cognitive psychology are kernels in the machine learning sense (Jäkel et al., 2007, 2008a; Jäkel, Schölkopf, & Wichmann, 2008b). In particular, radial basis

functions (RBFs)—for example Gaussians or Laplacians—are used in models of categorization (Nosofsky, 1986; Kruschke, 1992; Love, Medin, & Gureckis, 2004) and object recognition (Poggio & Edelman, 1990; Bülthoff & Edelman, 1992). In theoretical neuroscience RBFs are used to model tuning curves of neurons and there the connection with kernels and regularization theory has long been known (Poggio & Girosi, 1989; Poggio, 1990; Poggio & Bizzi, 2004).

Kernel Methods for Feature Identification

A central goal for behavioral scientists is to find out how the different aspects or features of a stimulus influence behavior. For simple tasks and stimuli it is frequently possible to predict a subjects' response from the physical properties of the stimulus. In many laboratory tasks experimenters *impose* on participants which features he or she can use to solve the task. This is particularly true for experiments and models of categorization where the stimuli "wear their features on their sleeves" (Schyns, Goldstone, & Thibaut, 1998). For natural categorization tasks, however, there are typically a multitude of potential features that a subject may be using. Without knowing the features, or internal representations that subjects use, it is hard to develop models of how subjects reach a decisions that lead to the observed response.

Over the last years, we and several colleagues used kernel methods to identify those features that best predict a subject's response in psychophysical tasks with natural stimuli (Wichmann, Graf, Simoncelli, Bülthoff, & Schölkopf, 2005; Kienzle, Wichmann, Schölkopf, & Franz, 2007; Kienzle, Franz, Schölkopf, & Wichmann, 2009; Yovel, Franz, Stilz, & Schnitzler, 2008). Like other black-box methods, these methods substitute a very hard to analyze complex system—the complete human observer—with a less complex system that is sufficiently sophisticated to re-create human decisions during a psychophysical task but is still amenable to mathematical analysis. The central idea is to use networks of the

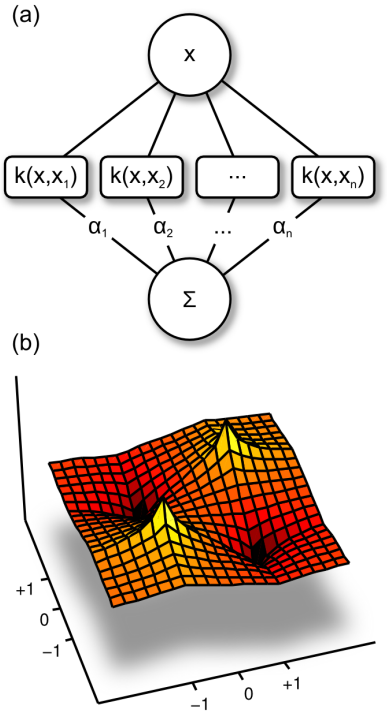

Figure 2. An exemplar model has a similar structure like a perceptron: For a stimulus x it forms a weighted sum, using the weights α_1 to α_n, of the responses of a set of neurons. The nodes in the network calculate the similarity $k(x, x_i)$ of x with each of the exemplars x_1 to x_n that were shown during training. This is illustrated in panel (a). Kernel methods in machine learning do the same. They use special similarity measures, called kernels. In cognitive psychology and theoretical neuroscience one often uses so-called radial-basis-functions (RBFs) as similarity measures, for example Gaussians or Laplacians. Panel (b) shows the response of such an exemplar model to the stimuli from Fig. 1. Close to the stimuli from the yellow category the response is high and close to the stimuli from the red category the response is low. Learning in kernel methods means adapting the contribution of each exemplar to the overall response.

Box 1 Perceptrons and Regularization

A perceptron is defined as a weighted sum of the responses of a set of m feature detectors φ_1, ..., φ_m (see Fig. 1):

$$f(x) = \sum_{i=1}^{m} w_i \varphi_i(x). \tag{1}$$

If the response $f(x)$ is larger than a threshold, x will be classified as category A otherwise as B. Learning in a perceptron is understood as adapting the weights. One can think about learning as the statistical problem of trying to estimate the weights that best predict the correct category labels for new stimuli.

If the feature detectors are chosen restrictively then there will be category distinctions that cannot be learned by the perceptron because the categories cannot be separated by a linear decision function in the feature space. One restriction is the functional form of the feature detectors (for example, linear or quadratic in the inputs, diameter-limited, etc.). Another restriction is the number of feature detectors. Both of these determine a perceptron's capabilities. For example, it is well-known that four points that form an XOR pattern in a two-dimensional space cannot be separated by a linear decision function in this space (Nilsson, 1965; Minsky & Papert, 1967).

In order to give the perceptron a lot of flexibility one would like to allow a large number m, potentially an infinite number, of feature detectors. In this case, it seems hard to generalize to new, previously unseen stimuli. Learning—or in statistical terms: estimating—the weights w_1, ..., w_m can still be successful, even if the number of stimuli n is a lot smaller than m, if regularization techniques are used. The optimal weights are found by trading-off the fit to the data, that is how well the perceptron replicates all the category labels on the training examples, with the magnitude of the weights. This extra constraint reduces the effective dimensionality of the learning problem. By using regularization techniques it is possible to use high-dimensional feature spaces and, in many cases, still be able to generalize. Similar techniques for trading-off model fit with a penalty term are used in model selection. For more details see (Jäkel et al., 2007; Hofmann et al., 2008; Schölkopf & Smola, 2002).

kind shown in Fig. 1 and Fig. 2 as a statistical model to predict the category responses provided by human subjects. In this way a network is used to re-create the internal decision space of individual human subjects.

For example, this approach was used to predict observers' categorization responses when indicating the gender of human faces shown repeatedly during a gender categorization task (Wichmann et al., 2005). The human category responses, male or female, are systematically different from the actual gender of the faces shown. In the simplest case a perceptron (as shown in Fig. 1) is used and restricted to linear features φ_1, ..., φ_m. The goal is to find the weights of the network that best predict the category responses of the subjects. Since the network calculates a linear combination of the features and a linear combination of linear functions is a linear function, this is equivalent to trying to find the best linear feature. In the gender categorization example the best linear feature that can be calculated from the pixel values as inputs is itself an image that puts most weight on the pixels that best predict subjects' responses. If the stimuli consist of 256x256 pixel images a learning algorithm will have to estimate as many weights. Since the number of stimuli that one is able to show to participants in an experiment is usually smaller than this, it is important to use regularization techniques (see Box 1).

There are other methods that try to find the stimuli that best predict responses, for example the so-called *bubbles* technique (Gosselin & Schyns, 2001; Dupuis-Roy, Fortin, Fiset, & Gosselin, 2009) or *classification image* methods (Ahumada & Lovell, 1971; Abbey & Eckstein, 2006) that are closely related to reverse-correlation as used in single-cell physiology (Marmarelis & Marmarelis, 1978; Neri & Levi, 2006). In contrast to the approach described above, these methods disrupt or at least change the stimulus dramatically: either stimuli are embedded in visual noise in case of "traditional" classification images or they are severely windowed in case of the bubbles technique. Also, generalizations of these methods to non-linear features or even combinations of non-linear features are not straightforward.

Recently, Kienzle et al. (2007, 2009) showed one natural extension of our approach to identifying non-linear features of a certain form. This work also illustrates that kernel methods are, of course, not limited to predicting subject's behavior in standard categorization tasks. In the study, subjects' eye-movements were tracked while viewing natural images. Small image patches around each fixation were extracted and compared to equally big image patches from the same images that were not fixated. The goal was to identify the features in an image patch that lead an observer to fixate it. Using an exemplar model (as shown in Fig. 2 and explained in Box 2) with a Gaussian kernel—that is, an RBF-network—fixations on new images were predicted in terms of weighted similarities to previous fixations. The resulting network could effectively be approximated by an RBF-network with just four nodes. This smaller network can be interpreted as extracting the similarity to four distinct stimuli as non-linear features and led to four neurophysiologically very plausible features (Fig. 3 illustrates how the method works).

Box 2 Kernel Methods and Exemplar Models

The so-called representer theorem assures that the optimal weights for the regularized learning problem (see Box 1) can be expressed as a linear combination of the training examples x_1, ..., x_n:

$$w_i = \sum_{j=1}^{n} \alpha_j \varphi_i(x_j). \tag{2}$$

Also many classic learning rules (for example, the perceptron rule or the delta rule) guarantee that the weights at each step during learning are always expressible in this form. Plugging this form of the weights into the definition of the perceptron (see Box 1) and reordering terms one obtains

$$f(x) = \sum_{j=1}^{n} \alpha_j \sum_{i=1}^{m} \varphi_i(x)\varphi_i(x_j). \tag{3}$$

Let us define a shorthand $k(x, y)$ for the sum over the features:

$$k(x, y) = \sum_{i=1}^{m} \varphi_i(x)\varphi_i(y). \tag{4}$$

This symmetric function is defined on pairs of stimuli x and y and it is called a *kernel*. It calculates the inner product of x and y in feature space. Intuitively speaking, the inner product measures the similarity between x and y. The response of the perceptron is then equivalently written as

$$f(x) = \sum_{j=1}^{n} \alpha_j k(x, x_j), \tag{5}$$

a weighted sum of the similarities to the exemplars x_1, ..., x_n (see Fig. 2). This function is linear in the weights α_j. As in the perceptron, learning in this representation can be understood as estimating weights. In cognitive science such a model is called an exemplar model (Kruschke, 1992; Nosofsky, 1992; Jäkel et al., 2008a). Hence, each set of feature detectors defines a natural similarity measure k such that any perceptron can equivalently be expressed as an exemplar model. It is often more natural to define the similarity k between stimuli directly, rather than defining a large set of feature detectors first. In these cases the exemplar formulation of the perceptron can be more useful (and computationally more efficient, too). Under certain conditions on k, that is k has to be a positive definite kernel, it is guaranteed that k can be expressed as an inner product between feature vectors (similar to (4), but with a potentially infinite number of features) and therefore an exemplar model using k has an equivalent formulation as a perceptron. A common way of regularizing the solution in kernel methods is by the norm of the weight vector in the feature space. It can be shown that even in the infinite dimensional case, this norm can be evaluated using the kernel. For more details see (Jäkel et al., 2007; Hofmann et al., 2008; Schölkopf & Smola, 2002).

Kernel Methods as Computational Models

Ultimately, we need models that are neurally and psychologically plausible. We also want to understand how a model solves the problem that a subject faces in an experiment. Models in cognitive science are hence often catego-

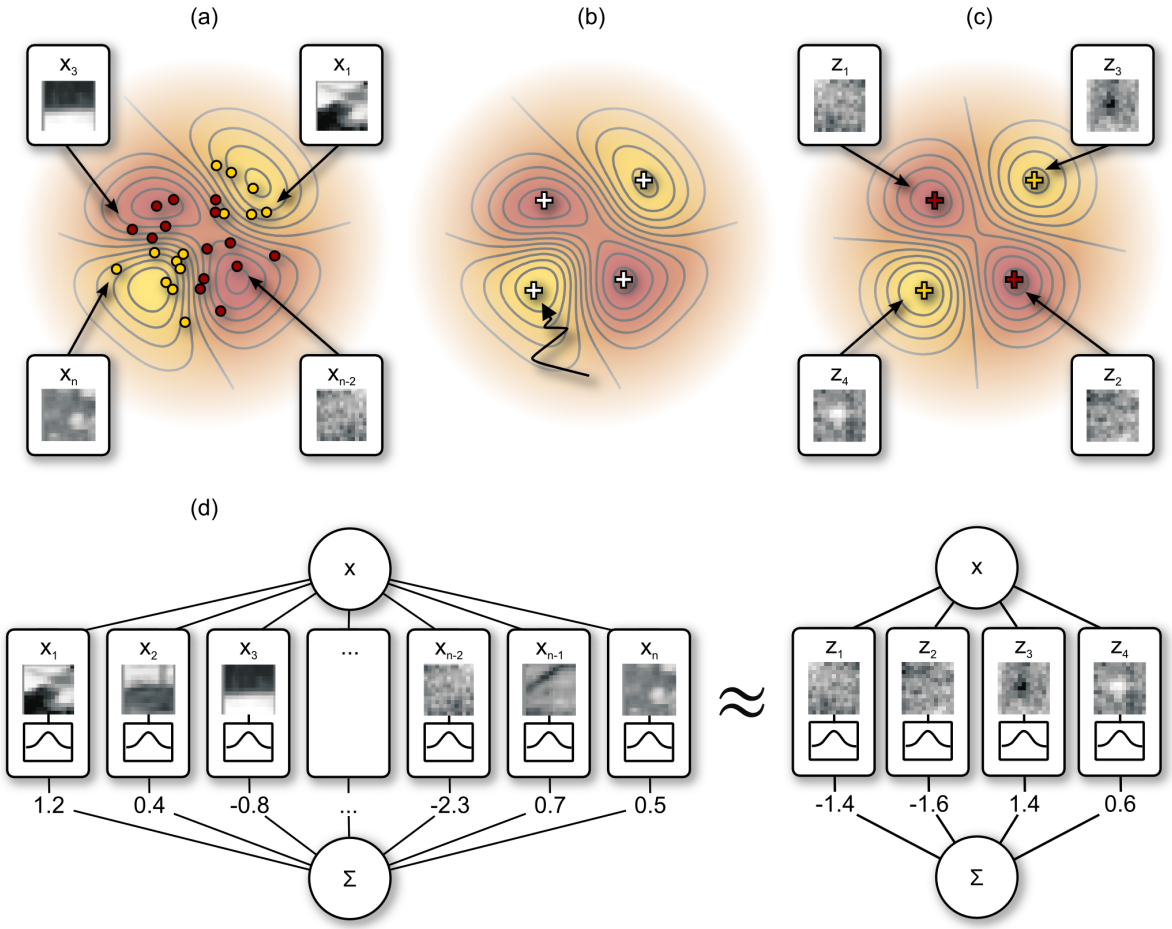

Figure 3. One way to estimate non-linear features was suggested by Kienzle et al. (2009): As an example consider their problem of predicting for a 13x13 image patch whether it is likely to be a saccade target or not. Formally, this is a categorization problem, too. Panels (a) to (d) illustrate how the method works in principle. Panel (a) depicts a set of image patches that have (yellow) or have not (red) been saccade targets for a set of training patches in a high-dimensional image space. The yellow-to-red gradient in the background and the contour lines represent the response of a kernel method with a Gaussian kernel after having been trained on the training patches. The resulting network is shown on the left of panel (d). Panel (b): In a second step gradient descent is performed repeatedly with different starting points on the function that the trained network implements. In this case, four extrema (crosses) were identified. Panel (c): Using the positions of the extrema a network with four kernels centered on the extrema is constructed in order to approximate the original network. Note the change in the contour lines compared to panel (b). Panel (d) shows the original network (left) and its approximation (right). The features that were recovered in this example are neurophysiologically plausible: The more similar an image patch is to a center-surround structure (z_3 and z_4) and the less similar it is to a ramp (z_1 and z_2) the more likely it will be a saccade target. This figure is adapted from Kienzle et al. (2009).

rized according to Marr's levels of explanation: the computational, the algorithmic, and the implementational level (Marr, 1982). Kernel methods are related to RBF networks on the implementational level and exemplar models on the algorithmic level. Hence, we think that theory developed for kernel methods in machine learning is of great interest to cognitive scientist because it potentially offers explanations on the computational level. In this triangle of related models and methods from machine learning (kernel methods), cognitive psychology (exemplar models), and theoretical neuroscience (RBF networks) we see the prospect of bridging Marr's levels. The role that machine learning could play in this enterprise is clear: help us to better understand what learning is, what the core problems are, and how models of human, animal, and neural learning solve these.

In a categorization task, like gender categorization, a subject has to solve a categorization problem and the same problem can be given to a machine classifier, for example a kernel machine as shown in Fig. 2. Instead of using the network as a statistical model for a subject's responses, by training it on the subject's responses as in the research discussed in the previous section, one can train the network on the same inputs and the same feedback that a subject is given during category learning—as is usually done in cognitive modeling. Instead of providing a statistical model for the subject's responses one treats the subject as trying to statistically model the data

Box 3 Questions for Future Research

• There is a large literature on feature selection in machine learning, especially in conjunction with kernel methods (Guyon & Elisseeff, 2003). Also, there are attempts to learn the kernel, that is the similarity measure, without making strong assumptions (Bousquet & Herrmann, 2003; Lanckriet, Cristianini, Bartlett, Ghaoui, & Jordan, 2004). Hence, there are many more methods that could potentially be useful for identifying the features, or corresponding similarity measures, that best predict subjects' responses. Furthermore, in many situations a human category learner has to learn the right features—or alternatively the right similarity measure—at the same time as he or she learns the categories (Schyns et al., 1998). Machine learning methods could provide us with hypotheses on how a human learner might achieve this.
• A major concern for exemplar models in cognitive science is whether all stimuli have to be stored in memory. Often the same categorization performance can be achieved with a smaller network that does not remember all exemplars (that is, most of the exemplars have a zero weight). Such solutions are called *sparse* in machine learning and statistics. For feature identification, we also want a solution that is sparse, so we can interpret the results. In the methods described here we obtained sparseness by approximating the original network by a smaller network with different nodes (see Fig. 3). There are other ways to achieve sparseness, for example by reduced-set metods (Schölkopf & Smola, 2002). Different regularization mechanisms also lead to different degrees of sparseness (Weston, Elisseeff, B., & Tipping, 2003), as exploited in the Lasso (Hastie, Tibshirani, & Friedman, 2009) and recently in the field of compressive sensing (Candes & Wakin, 2008).
• One major advantage of kernel methods is that the same techniques can be used irrespective of what the kernel is. Here, we have mainly considered radial basis functions because of their psychological and neural plausibility. However, polynomial kernels have some plausibility, too (Jäkel et al., 2007) and they can potentially be used for identifying critical features via Wiener and Volterra Theory (Franz & Schölkopf, 2006). There are also kernels that can deal with non-vectorial stimuli, like strings, trees or graphs (Hofmann et al., 2008). Such kernels might be useful for modelling categorization of interestingly structured stimuli, like sentences or visual objects. Of particular interest in this context are recursively defined kernels (Haussler, 1999; Smale, Rosasco, Bouvire, Caponnetto, & Poggio, 2008).
• A lot of theoretical work in cognitive science and machine learning has focused on either supervised or unsupervised learning, that is scenarios where either the category labels for all of the stimuli or for none of the stimuli are provided. However, in the real world only some of the stimuli might be labeled. This scenario is called semi-supervised learning in machine learning (Chapelle, Schölkopf, & Zien, 2006) and such scenarios can be studied in human category learning, too (Zhu, Rogers, Qian, & Kalish, 2007; Vandist, De Schryver, & Rosseel, 2009). Similarly, a category label might refer to one of several possible stimuli. Imagine a parent uttering the word *dog* when there are plenty objects in a scene and a child does not know which of these the parent refers to. This scenario is called multiple instance learning in machine learning (Andrews, Tsochantaridis, & Hofmann, 2003).

he or she observes. Using the tools-to-theories heuristic (Gigerenzer, 1991) the computational problem that the subject tries to solve can be formalized in the same way that it is formalized in machine learning.

At the computational level, most experimental tasks are set up as a two-category problem. As in machine learning, the problem of learning a new category distinction in an experiment can be conceptualized as a matter of generalization. In this view, the computational problem that a learning algorithm tries to solve is to generalize well. If a subject is confronted with a previously unseen stimulus will this stimulus be categorized correctly? Kernel methods with regularization techniques are one particularly well-understood way of assuring a good generalization performance (Schölkopf & Smola, 2002). Since kernel methods are related to exemplar models and RBF networks one may hope that insights from machine learning can be transferred to cognitive psychology and theoretical neuroscience. However, how much potential for cognitive science one sees in such an approach—that has yet to be spelled out in detail—crucially depends on how one assesses the relevance of exemplar models and RBF networks for cognitive science in the first place.

Exemplar models are not only discussed as models for categorization but also as models for perceptual expertise, object recognition, or automaticity (Palmeri, Wong, & Gauthier, 2004; Palmeri & Gauthier, 2004). As classic psychological models, exemplar models are specified on the algorithmic level. Exemplars are stored in memory and a new stimulus is compared to old stimuli before a response is made (Medin & Schaffer, 1978). While this basic idea links exemplar models and kernel methods (Ashby & Alfonso-Reese, 1995), exemplar models also account for additional aspects of behavioral data, like attention shifts, learning curves, and response times (Nosofsky, 1986; Kruschke, 1992; Palmeri, 2001). Despite the success of exemplar models in fitting data from a large number of laboratory experiments, there are many effects in the categorization literature—especially those involving background-knowledge, rule-based categories, and abstraction—that go beyond the capabilities of basic exemplar models (Murphy, 2002). Some of these effects seem to require the specification of additional mechanisms on top of a simple exemplar model (Heit, 1994; Lamberts, 1994; Nosofsky & Johansen, 2000; Rodrigues & Murre, 2007) or the postulation of multiple categorization systems (Ashby, Alfonso-Reese, Turken, & Waldron, 1998) and hybrid models of which exemplar models are a subpart (Erickson & Kruschke, 1998; Denton, Kruschke, & Erickson, 2008).

Since exemplar models lack abstraction mechanisms they are often contrasted with prototype models. Even for those laboratory experiments that have traditionally been considered to provide good evidence for exemplar effects there has been some debate about whether the data can be modeled without prototype abstraction (Smith & Minda, 1998, 2000; Nosofsky & Zaki, 2002; Navarro, 2007). This debate sometimes obscures, however, that prototype models and exemplar models are almost identical in all but one respect: the choice of representatives. Recent modeling efforts have blurred the distinction between the two types of models. There are models that allow for multiple prototypes and a continuum of abstractions (Love et al., 2004; Edelman, 1998; Rosseel, 2002; Vanpaemel & Storms, 2008) and there are

models that reduce the number of exemplars (De Schryver, Vandist, & Rosseel, 2009). These models are similar in spirit to the approach that is illustrated in Fig. 3: find a set of stimuli smaller than the set of all training exemplars but equally representative of the category (or category distinction) in question. Furthermore, due to the correspondence between perceptrons and exemplar models that is explained in Box 2 the simplest exemplar model can be seen as a prototype model in a suitable feature space—at least formally (Jäkel et al., 2007). In any case, exemplar models have been and still are extremely important as a well-developed null-hypothesis and a theoretical starting point for all research on categorization.

Part of the attractiveness of exemplar models, in our view, is that linking them to the implementational level seems straightforward, since they can be expressed as neural networks (Kruschke, 1992). An exemplar is represented by a pool of neurons. The similarity of this exemplar to other stimuli is implemented by the tuning curves of the neurons. Stimuli that are similar to an exemplar will make the corresponding neurons fire but at a lower rate. Tuning curves thus implement the kernel. The response of the animal is obtained by integrating excitatory and inhibitory responses for many pools of neurons. Related ideas have long been discussed in the object recognition literature as neurally plausible mechanisms where it may be possible to link the neural level with the computational level (Poggio & Edelman, 1990; Bülthoff & Edelman, 1992; Poggio, 1990; Poggio & Bizzi, 2004). These studies motivated electrophysiological work looking for respective tuning curves (Logothetis, Pauls, & Poggio, 1995). More recently, multidimensional psychological spaces and attention shifts that have been crucial aspects of models in cognitive psychology have been investigated electrophysiologically, too (Palmeri & Gauthier, 2004; Sigala & Logothetis, 2002; Beeck, Wagemans, & Vogels, 2001).

Box 4 Glossary

Generalization: The response of a learner to previously unseen stimuli. Generalization is a crucial aspect of learning. For example, in categorization the correct category labels for a set of training stimuli can be learnt by heart. Successful learning of a category should therefore also manifest itself in correct categorization of new stimuli.

Kernel: Intuitively speaking, a kernel measures the similarity of two stimuli. Many of the similarity measures used in psychological models are kernels in the machine learning sense. Formally, a so-called positive definite kernel is a function of two arguments that represents an inner product (dot product) in some feature space.

Learning: In machine learning, the term *learning* is often used to mean *estimating* parameters in a model. The estimation problem often takes the form of an optimization problem, in which case learning amounts to finding the optimal parameters.

Radial-basis-function (RBF): A function that depends on the distance of a stimulus to a reference stimulus. Usually the distance is given by a norm on a vector space. This is analogous to tuning-curves where the response of the neuron depends on the "distance" of the stimulus to the characteristic stimulus of the neuron. For example, a Gaussian RBF is an exponential function of the squared Euclidean distance between the stimulus and a reference stimulus. Likewise, an exponential of the city-block distance is sometimes called Laplacian (but so is an exponential of the Euclidean distance).

Regularization: A technique to deal with high-dimensional estimation problems where the amount of data is not sufficient to effectively estimate all the parameters. Regularization imposes additional constraints on the parameters, for example a preference for small parameter values.

Training Data: In order to get a realistic estimate of a learning algorithm's generalization performance the available data is split up into a training and a test set. During training the learning algorithm is given the training data, however the generalization performance is measured on new data, the test set, that the algorithm has never seen before.

Weights: In a linear model the parameters are often referred to as weights because the output is a weighted combination of the response of the feature detectors.

The Future of Kernel Methods in Cognitive Science

In a recent series of papers, we have spelled out explicitly the relationships between certain kernel methods in machine learning and common exemplar models in cognitive science (Jäkel et al., 2007, 2008a, 2008b). These studies provide groundwork for transferring insights from kernel methods to exemplar models. At the moment the role for machine learning in the enterprise of bridging Marr's levels of explanation for categorization models remains admittedly promissory. We think, however, that there are some promising directions to take.

If one accepts that at the computational level humans as well as machine learning algorithms try to generalize well, then cognitive scientists will be able to make use of the wealth of theory on generalization that has been developed in machine learning (Vapnik, 2000; Devroye, Györfi, & Lugosi, 1996). In fact, there are already attempts to use insights on generalization for understanding human category learning (Jäkel et al., 2008a; Love et al., 2004; Briscoe & Feldman, 2006). We believe that future work will have to bridge the gap between concrete learning algorithms and the computational problem of generalization (Bousquet & Elisseeff, 2002; Poggio, Rifkin, Mukherjee, & Niyogi, 2004). It could be useful if concrete and successful suggestions for learning algorithms in the literature on human category learning, such as ALCOVE (Kruschke, 1992), could be backed up by a formal analysis of their generalization performance—so that it becomes clear that they actually solve the problem that the learner faces. More suggestions for future research can be found in Box 3.

We have shown that kernel methods can be useful for identifying features in categorization tasks. We have further argued that there is great potential for transferring insights from kernel methods to human category learning. More generally, one could hope that increased interest in kernel methods from cognitive scientists could increase cross-talk between theoretical neuroscience, cognitive psychology, and

machine learning. While it is exciting to see that all three fields converge on similar ideas, it is important to note that these developments were not independent of each other. In addition to sharing intuitions on categorization and similarity, researchers across fields share influences from the early days of artificial intelligence and cognitive science. The popularity of these ideas may also partly be explained by the fact that kernel methods are linear methods and are therefore simple enough to be handled with widely- and well-understood mathematical tools. This does not have to be a disadvantage. Kernel methods are promising tools for cognitive scientists because they are simple enough to be analyzed thoroughly but at the same time they are powerful enough to tackle realistic learning problems.

References

Abbey, C., & Eckstein, M. (2006). Classification images for detection, contrast discrimination, and identification tasks with a common ideal observer. *Journal of Vision*, *6*(4), 335-355.

Ahumada, A., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America*, *49*(6B), 1751-1756.

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2003). Support vector machines for multiple-instance learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, p. 561-568). Cambridge, MA: MIT Press.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216-233.

Ashby, F. G., Alfonso-Reese, L. A., Turken, U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442-481.

Beeck, H. Op de, Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, *4*(12), 1244-52.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*, 499-526.

Bousquet, O., & Herrmann, D. (2003). On the complexity of learning the kernel matrix. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, p. 415-422). Cambridge, MA: MIT Press.

Briscoe, E., & Feldman, J. (2006). Conceptual complexity and the bias-variance tradeoff. In R. Sun, N. Miyake, & C. Schunn (Eds.), *Proceedings of the 28th annual conference of the cognitive science society.*

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences USA*, *89*(1), 60-4.

Candes, E., & Wakin, M. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine*, *25*(2), 21-30.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning*. Cambridge, MA: MIT Press.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*, 287-291.

Denton, S. E., Kruschke, J. K., & Erickson, M. A. (2008). Rule-based extrapolation: A continuing challenge for exemplar models. *Psychonomic Bulletin & Review*, *15*(4), 780-786.

De Schryver, M., Vandist, K., & Rosseel, Y. (2009, Apr). How many exemplars are used? Explorations with the Rex Leopold I model. *Psychonomic Bulletin & Review*, *16*(2), 337–343.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer.

Dupuis-Roy, N., Fortin, I., Fiset, D., & Gosselin, F. (2009). Uncovering gender discrimination cues in a realistic setting. *Journal of Vision*, *9*(2), 1-8.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*, 449-498.

Erickson, M. A., & Kruschke, J. K. (1998, Jun). Rules and exemplars in category learning. *Journal Experimental Psychology: General*, *127*(2), 107–140.

Franz, M., & Schölkopf, B. (2006). A unifying view of Wiener and Volterra theory and polynomial kernel regression. *Neural Computation*, *18*(12), 3097-3118.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*(2), 254-267.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261-71.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157-1182.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Second ed.). New York: Springer.

Haussler, D. (1999, July). *Convolution kernels on discrete structures* (Tech. Rep. No. UCSC-CRL-99-10). Santa Cruz, CA: Department of Computer Science, University of California at Santa Cruz.

Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning Memory & Cognition*, *20*(6), 1264-1282.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527-1554.

Hofmann, T., Schölkopf, B., & Smola, A. (2008). Kernel methods in machine learning. *Annals of Statistics*, *36*(3), 1171-1220.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology*, *51*, 343-358.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008a). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, *15*(2), 256-271.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008b). Similarity, kernels and the triangle inequality. *Journal of Mathematical Psychology*, *52*(5), 297-303.

Kienzle, W., Franz, M. O., Schölkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, *9*, 1-15.

Kienzle, W., Wichmann, F. A., Schölkopf, B., & Franz, M. O. (2007). A nonparametric approach to bottom-up visual saliency. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems* (Vol. 19, p. 689-696.). Cambridge, MA: MIT Press.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Lamberts, K. (1994). Flexible tuning of similarity in exemplar-based categorization. *Journal of Experimental Psychology: General*, *20*(5), 1003-1021.

Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L. E., & Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, *5*, 27-72.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1-15.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*(5), 552-63.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, *111*(2), 309-32.

Marlin, B. (2004). *Collaborative filtering: A machine learning perspective*. Unpublished master's thesis, University of Toronto.

Marmarelis, P., & Marmarelis, V. (1978). *Analysis of physiological systems: The white-noise approach*. New York: Plenum Press.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207-238.

Minsky, M., & Papert, S. (1967). *Linearly unrecognizable patterns* (AIM No. 140). MIT.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology*, *51*, 85-98.

Neri, P., & Levi, D. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision Research*, *46*(16), 2465-2474.

Nilsson, N. J. (1965). *Learning machines*. New York: McGraw-Hill.

Nosofsky, R. M. (1986). Attention, similarity, and the indentification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Nosofsky, R. M. (1992). Exemplar-based approach to relating categorization, identification, and recognition. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition.* Hilldale, NJ: Erlbaum.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychon Bull Rev*, *7*(3), 375-402.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*(5), 924-40.

Palmeri, T. J. (2001). The time course of perceptual categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (p. 193-224). Oxford University Press.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, *5*(4), 291-303.

Palmeri, T. J., Wong, A.-N., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *Trends in Cognitive Sciences*, *8*(8), 378-286.

Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, *LV*, 899-910.

Poggio, T., & Bizzi, E. (2004). Generalization in vision and motor control. *Nature*, *431*(7010), 768-774.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*(6255), 263-6.

Poggio, T., & Girosi, F. (1989). *A theory of networks for approximation and learning* (Tech. Rep. No. A. I. Memo No. 1140). Cambridge, MA: MIT AI LAB and Center for Biological Information Processing Whitaker College.

Poggio, T., Rifkin, R., Mukherjee, S., & Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature*, *428*, 419-422.

Rodrigues, P. M., & Murre, J. M. J. (2007). Rules-plus-exception tasks: A problem for exemplar models? *Psychonomic Bulletin & Review*, *14*(4), 640-646.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386-408.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*, 178-210.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Schölkopf, B., Tsuda, K., & Vert, J.-P. (Eds.). (2004). *Kernel methods in computational biology*. Cambridge, MA: MIT Press.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, *21*(1), 1-17.

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*, 318-320.

Smale, S., Rosasco, L., Bouvire, J., Caponnetto, A., & Poggio, T. (2008). *Mathematics of the neural response* (Tech. Rep.). MIT CBCL.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411-1436.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 3-27.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Vandist, K., De Schryver, M., & Rosseel, Y. (2009, Feb). Semisupervised category learning: the impact of feedback in learning the information-integration task. *Attenention, Perception, & Psychophysics*, *71*(2), 328–341.

Vanpaemel, W., & Storms, G. (2008, Aug). In search of abstraction: the varying abstraction model of categorization. *Psychonomic Bulletin Review*, *15*(4), 732–749.

Vapnik, V. N. (2000). *The nature of statistical learning theory* (2nd ed.). New York: Springer.

Weston, J., Elisseeff, A., B., S., & Tipping, M. (2003). Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, *3*, 1439-1461.

Wichmann, F. A., Graf, A. B. A., Simoncelli, E. P., Bülthoff, H. H., & Schölkopf, B. (2005). Machine learning applied to perception: decision-images for gender classification. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 17, p. 1489-1496). Cambridge, MA: MIT Press.

Yovel, Y., Franz, M. O., Stilz, P., & Schnitzler, H.-U. (2008). Plant classification from bat-like echolocation signals. *PLoS Computational Biology*, *4*(3), e1000032.

Zhu, X., Rogers, T., Qian, R., & Kalish, C. (2007). Humans perform semi-supervised classification too. In R. Holte & A. Howe (Eds.), *Twenty-second aaai conference on artificial intelligence* (p. 864-870). Menlo Park, CA: AAAI Press.