The Science of Pattern Recognition.
Achievements and Perspectives.

Robert P.W. Duin[1] and Elżbieta Pękalska[1,2]

[1] ICT group, Faculty of Electr. Eng., Mathematics and Computer Science
     Delft University of Technology, The Netherlands
     `r.duin@ieee.org, e.m.pekalska@tudelft.nl`
[2] School of Computer Science, University of Manchester, United Kingdom

Summary. Automatic pattern recognition is usually considered as an engineering area study-ing the development and evaluation of systems that imitate or assist the human ability of recognizing patterns. It may, however, also be considered as a science that studies the natural phenomenon that human beings (and possibly other biological systems) are able to discover, distinguish and characterize patterns in their environment, and identify new observations ac-cordingly. The engineering approach to pattern recognition is in this view an attempt to build systems that simulate this phenomenon. By that, scientific understanding is achieved of what is needed in order to recognize patterns.

Like in any science understanding can be gained from different, sometimes opposite view-points. We will introduce the main approaches to the science of pattern recognition as two di-chotomies of complementary scenarios, giving rise to four different schools. These schools are roughly defined under the terms of expert systems, neural networks, structural and statistical pattern recognition. We will briefly describe what has been achieved by these schools, what is common and what is specific, which limitations are encountered and which perspectives arise for the future. Finally, we will focus on the challenges facing pattern recognition in the decen-nia to come. They deal mainly with weaker assumptions to make procedures for learning and recognition wider applicable, others need to develop new formalisms.

1 Introduction

We are very familiar with the human ability of pattern recognition. Since our early years we have been able to recognize voices, faces, animals, fruits, or inanimate ob-jects. Before the speaking ability is developed, an object like a ball is recognized, even when it just resembles the balls seen before. So, except for the memory, the skills of abstraction and generalization are essential to find our way in the world. In later years we are able to deal with much more complex patterns that may not be directly based on sensorial observations. Examples are the pattern of an underlying theme in a discussion or the subtle pattern in human relations. The latter may be-come apparent e.g. only by listening to somebody's complaints about his personal problems at work that are also present in a new job. Without a direct participation in

the events, we are able to see analogy and similarity in examples as complex as the social interaction between people. Here we learn to distinguish the pattern from just two examples.

The pattern recognition ability may also be found in other biological systems: the cat knows the way home, the dog recognizes his boss from the footsteps or the bee finds the delicious flower. In these examples a direct connection can be made to sensory experiences. Memory alone is insufficient; an important role is that of generalization from observations which are similar but not identical to the previous ones. A scientific challenge is to find out how this (may) work.

Scientific questions may be approached by building models and, more explicitly, by creating simulators, i.e. artificial systems that roughly exhibit the same phenomenon as the object under study. Understanding will be gained while pursuing in a construction of such a system and its evaluation with respect to the real object. Such systems may be used to replace the original ones and may even improve some of their properties, but also perform worse in other aspects. For instance, planes fly faster than birds but are far from being autonomous. We should realize, however, that what is studied may not be the bird itself, but more importantly, the ability to fly. Much can be learned about flying in an attempt to imitate the bird, but also when differentiating from its exact appearance. By constructing fixed wings instead of freely movable ones, the insight in how to fly grows. Finally, there are engineering aspects that may gradually deviate from the original scientific question. These are concerned with how to fly a long time, or with heavy loads, or with less noise, and slowly shift the point of attention to other domains of knowledge.

The above shows that a distinction can be made between the scientific study of pattern recognition as the ability to abstract and generalize from observations and the applied technical area of the design of artificial pattern recognition devices, without neglecting the fact that they may highly profit from each other. Note that patterns can be distinguished on many levels, starting from simple characteristics of structural elements like strokes, through features of an individual towards a set of qualities in a group, to a composite of traits of concepts and their possible generalizations. Pattern may also denote a single individual as a representative for its population, model or concept. Pattern recognition deals, therefore, with patterns, regularities, characteristics or qualities that can be discussed on a low level of sensory measurements (such as pixels in an image) as well as on a high level of the derived, meaningful concepts (such as faces in images). In this paper we will focus on the scientific aspects, i.e. what we know about the way pattern recognition works and especially, what can be learned from our attempts to build artificial recognition devices.

A number of authors have already discussed the science of pattern recognition starting from perspectives achieved through their attempts of simulation and modeling. One of the first, in the beginning of the sixties, was Sayre [64], who presented a philosophical study on perception, pattern recognition and classification. He made clear that classification is a task that can be fulfilled with some success, but recognition either happens or not. We can stimulate the recognition by focussing on some aspects of the question. Although we cannot set out to fully recognize something, we can at least start to classify objects on command. The way Sayre distinguishes recog-
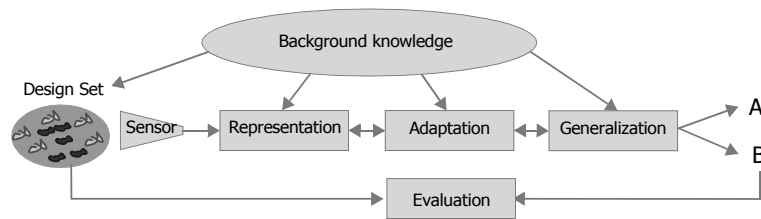
Fig. 1. Components of a pattern recognition system.

nition and classification is related to the two subfields discussed in general texts on pattern recognition, namely unsupervised and supervised learning. They fulfill two complementary tasks acting as automatic tools in the hand of a scientist who sets out to find the regularities in nature.

Unsupervised learning (also related to exploratory analysis or cluster analysis) gives the scientist an automatic system to suggest the presence of yet unspecified patterns (regularities) in the observations. They have to be confirmed (verified) by him. Here, in the terms of Sayre, a pattern is recognized. **Supervised learning** is an automatic system that verifies (confirms) the patterns described by the scientist based on a representation defined by him. This is done by an automatic classification followed by an evaluation.

In spite of Sayre's discussion, the concepts of pattern recognition and classification are still frequently mixed. In our discussion classification is a significant component of the pattern recognition system, but unsupervised learning may also play a role there. Typically, such a system is first presented with a set of known objects, the training set, in some convenient representation. Learning relies on finding the data descriptions such that the system can correctly characterize, identify or classify novel examples. After appropriate preprocessing and adaptations, various mechanisms are employed to train the entire system well. Various models and techniques are used and their performances are evaluated and compared by suitable criteria. If the final goal is prediction, the findings are validated by applying the best model to unseen data. If the final goal is characterization, the findings may be validated by complexity of organization (relations between objects) as well as by interpretability of the results.

Fig. 1 shows the three main stages of pattern recognition systems: Representation, Generalization and Evaluation, and an intermediate stage of Adaptation [20]. The system is trained and evaluated by a set of examples, the Design Set:

- **Design Set.** It is used both for training and validating the system. Given the background knowledge this set has to be chosen such that it is representative for the set of objects to be recognized by the trained system. There are various approaches how to split it into suitable subsets for training, validation and testing; see e.g. [22, 62, 68, 77].

- **Representation.** Real world objects have to be represented in a formal way in order to be analyzed and compared by mechanical means, such as a computer. Moreover, the observations derived from the sensors or other formal represen-

tations have to be integrated with the existing, explicitly formulated knowledge either on the objects themselves or on the class they may belong to. The issue of representation is an essential aspect of pattern recognition and is different from classification. It largely influences the success of the stages to come.

- **Adaptation.** In this intermediate stage between Representation and Generalization representations, learning methodology or problem statement are adapted or extended in order to enhance the final recognition. This step may be neglected as being transparent, but its role is essential. It may reduce or simplify the representation, or it may enrich it by emphasizing particular aspects, e.g. by a nonlinear transformation that simplifies the next stage. Background knowledge may be appropriately (re)formulated and incorporated into a representation. If needed, additional representations may be considered to reflect other aspects of the problem. Exploratory data analysis (unsupervised learning) may be used to guide the choice of suitable learning strategies.

- **Generalization or Inference.** In this stage we learn from a training set, the set of known and appropriately represented examples, in such a way that predictions can be made on some unknown properties of new examples. We either generalize towards a concept or infer a set of general rules that describe the qualities of the training data. The most common property is the class or pattern it belongs to, which is the above mentioned classification task.

- **Evaluation.** In this stage we estimate how our system performs on known training and validation data while training the entire system. If the results are unsatisfactory, then the previous steps have to be reconsidered.

Different disciplines emphasize, or just exclusively study, different parts of this system. For instance, perception and computer vision deal mainly with the representation aspects [21], while books on artificial neural networks [62], machine learning [4, 53] and pattern classification [15] are usually restricted to generalization. It should be noted that these and other studies with the words "pattern" and "recognition" in the title very often almost entirely neglect the issue of representation. We think, however, that *the main goal of the field of pattern recognition is to study generalization in relation to representation* [20].

In the context of representations, especially images, generalization has been thoroughly studied by Grenander [36]. What is very specific and worthwhile is that he deals with infinite representations (say, unsampled images), thereby avoiding the frequently returning discussions on dimensionality and directly focussing on a high, abstract level of pattern learning. We like to mention two other scientists that present very general discussions on the pattern recognition system: Watanabe [75] and Goldfarb [31, 32]. They both emphasize the structural approach to pattern recognition that we will discuss later on. Here objects are represented in a form that focusses on their structure. A generalization over such structural representations is very difficult if one aims to learn the *concept*, i.e. the underlying, often implicit definition of a pattern class that is able to generate possible realizations. Goldfarb argues that traditionally used numeric representations are inadequate and that an entirely new, structural representation is necessary. We judge his research program as very ambitious, as he

wants to learn the (generalized) structure of the concept from the structures of the examples. He thereby aims to make explicit what usually stays implicit. We admit that a way like his has to be followed if one ever wishes to reach more in concept learning than the ability to name the right class with a high probability, without having built a proper understanding.

In the next section we will discuss and relate well-known general scientific approaches to the specific field of pattern recognition. In particular, we like to point out how these approaches differ due to fundamental differences in the scientific points of view from which they arise. As a consequence, they are often studied in distinct traditions based on different paradigms. We will try to clarify the underlying cause for the pattern recognition field. In the following sections we sketch some perspectives for pattern recognition and define a number of specific challenges.
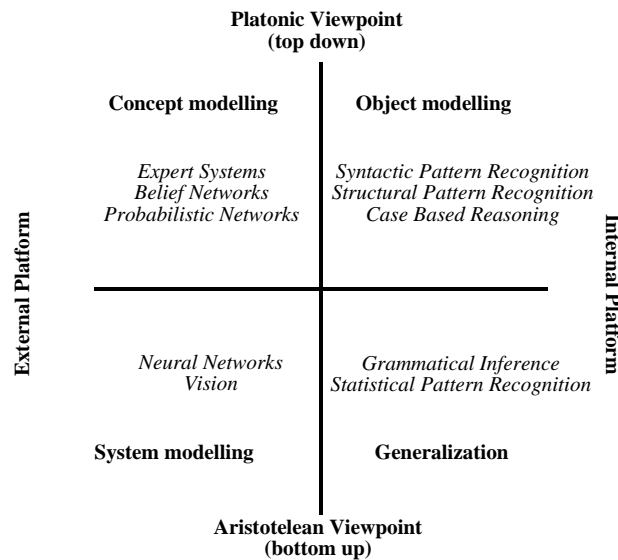
2 Four approaches to pattern recognition

In science new knowledge is phrased in terms of existing knowledge. The starting point of this process is set by generally accepted evident views, or observations and facts that cannot be explained further. These foundations, however, are not the same for all researchers. Different types of approaches may be distinguished originating from different starting positions. It is almost a type of taste from where a particular researcher begins. As a consequence, different 'schools' may arise. *The point of view, however, determines what we see*. In other words, staying within a particular framework of thought we cannot achieve more than what is derived as a consequence of the corresponding assumptions and constraints. To create more complete and objective methods, we may try to integrate scientific results originating from different approaches into a single pattern recognition model. It is possible that confusion arises on how these results may be combined and where they essentially differ. But the combination of results of different approaches may also appear to be fruitful, not only for some applications, but also for the scientific understanding by the researcher that broadens the horizon of allowable starting points. This step towards a unified or integrated view is very important in science, as only then a more complete understanding is gained or a whole theory is built.

Below we will describe four approaches to pattern recognition, which arise from two different dichotomies of the starting points. Next, we will present some examples illustrating the difficulties of their possible interactions. This discussion is based on our earlier publications [16, 17].

2.1 Platonic and Aristotelian viewpoints

Two principally different approaches to almost any scientific field rely on the so-called Platonic and Aristotelean viewpoints. In a first attempt they may be understood as top-down and bottom-up ways of building knowledge. They are also related to deductive (or holistic) and inductive (or reductionistic) principles. These aspects will be discussed in Section 4.

Fig. 2. Four approaches to Pattern Recognition.

The Platonic approach starts from generally accepted concepts and global ideas of the world. They constitute a coherent picture in which many details are undefined. The primary task of the Platonic researcher is to recognize in his[3] observations the underlying concepts and ideas that are already accepted by him. Many theories of the creation of the universe or the world rely on this scenario. An example is the drifts of the continents or the extinction of the mammoths. These theories do not result from a reasoning based on observations, but merely from a more or less convincing global theory (depending on the listener!) that seems to extrapolate far beyond the hard facts. For the Platonic researcher, however, it is not an extrapolation, but an adaptation of previous formulations of the theory to new facts. That is the way this approach works: existing ideas that have been used for a long time have been gradually adapted to new incoming observations. The change does not rely on an essential paradigm shift in the concept, but on finding better, more appropriate relations with the observed world in definitions and explanations. The essence of the theory is constant for a long time. So, in practise the Platonic researcher starts from a theory which can be stratified into to a number of hypotheses that can be tested. Observations are collected to test theses hypotheses and, finally, if the results are positive, the theory is confirmed.

The observations are of primary interest in the Aristotelian approach. Scientific reasoning stays as closely as possible to them. It is avoided to speculate on large, global theories that go beyond the facts. The observations are always the foundation on which the researcher builds his knowledge. Based on them, patterns and regularities are detected or discovered, which are used to formulate some tentative hypothe-

---

[3]For simplicity, we refer to researchers in a male form; we mean both women and men.

ses. These are further explored in order to arrive at general conclusions or theories. As such, the theories are not global, nor do they constitute high level descriptions. A famous guideline here is the so-called Occams razor principle that urges one to avoid theories that are more complex than strictly needed for explaining the observations. Arguments may arise, however, since the definition of complexity depends, e.g. on the mathematical formalism that is used.

The choice for a particular approach may be a matter of preference or determined by non-scientific grounds, such as upbringing. Nobody can judge what the basic truth is for somebody else. Against the Aristotelians may be held that they do not see the overall picture. The Platonic researchers, on the other hand, may be blamed for building castles in the air. Discussions between followers of these two approaches can be painful as well as fruitful. They may not see that their ground truth is different, resulting in a pointless debate. What is more important is the fact that they may become inspired by each other's views. One may finally see real world examples of his concepts, while the other may embrace a concept that summarizes, or constitutes an abstraction of his observations.

2.2 Internal and the external observations

In the contemporary view, science is *'the observation, identification, description, experimental investigation, and theoretical explanation of phenomena'*[4] or *'any system of knowledge that is concerned with the physical world and its phenomena and that entails unbiased observations and systematic experimentation.*[5] So, the aspect of observation that leads to a possible formation of a concept or theory is very important. Consequently, the research topic of the science of pattern recognition, which aims at the generalization from observations for knowledge building, is indeed scientific. Science is in the end a brief explanation summarizing the observations achieved through abstraction and their generalization.

Such an explanation may primarily be observed by the researcher in his own thinking. Pattern recognition research can thereby be performed by introspection. The researcher inspects himself how he generalizes from observations. The basis of this generalization is constituted by the primary observations. This may be an entire object (I just see that it is an apple) or its attributes (it is an apple because of its color and shape). We can also observe pattern recognition in action by observing other human beings (or animals) while they perform a pattern recognition task, e.g. when they recognize an apple. Now the researcher tries to find out by experiments and measurements how the subject decides for an apple on the basis of the stimuli presented to the senses. He thereby builds a model of the subject, starting from senses to decision making.

Both approaches result into a model. In the external approach, however, the senses may be included in the model. In the internal approach this is not possible, or just very partially. We are usually not aware of what happens in our senses. Introspection thereby start by what they offer to our thinking (and reasoning). As a

---

[4]http://dictionary.reference.com/
[5]http://www.britannica.com/

consequence, models based on the internal approach have to be externally equipped with (artificial) senses, i.e. with sensors.

2.3 The Four Approaches

By combining the two dichotomies as presented above, the following four approaches can be distinguished:

(1) Introspection by a Platonic viewpoint: object modeling.
(2) Introspection by an Aristotelian viewpoint: generalization.
(3) Extrospection by an Aristotelian viewpoint: system modeling.
(4) Extrospection by a Platonic viewpoint: concept modeling.

These four approaches will now be discussed separately. We will identify some known procedures and techniques that may be related to these. See also Fig. 2.

Object modeling. This is based on introspection from a Platonic viewpoint. The researcher thereby starts from global ideas on how pattern recognition systems may work and tries to verify them in his own thinking and reasoning. He thereby may find, for instance, that particular color and shape descriptions of an object are sufficient for him to classify it as an apple. More generally, he may discover that he uses particular reasoning rules operating on a fixed set of possible observations. The so-called syntactic and structural approaches to pattern recognition [26] thereby belong to this area, as well as case based reasoning [3]. There are two important problems in this domain: how to constitute the general concept of a class from individual object descriptions and how to connect particular human qualitative observations such as "sharp edge" or "egg shaped" with physical sensor measurements.

Generalization. Let us leave the Platonic viewpoint and consider a researcher who starts from observations, but still relies on introspection. He wonders what he should do with just a set of observations without any framework. An important point is the nature of observations. Qualitative observations as "round", "egg shaped" or "gold colored" can be judged as recognitions in themselves based on low-level outcomes of senses. It is difficult to neglect them and to access the outcomes of senses directly. One possibility for him is to use artificial senses, i.e. of sensors, which will produce quantitative descriptions. The next problem, however, is how to generalize from such numerical outcomes. The physiological process is internally unaccessible. A researcher who wonders how he himself generalizes from low level observations given by numbers may turn for help to statistics. This approach thereby includes the area of statistical pattern recognition.

If we consider low-level inputs that are not numerical, but expressed in attributed observations as red, egg-shaped, then the generalization may be based on logical or grammatical inference. As soon, however, as the structure of objects or attributes is not generated from the observations, but derived (postulated) from a formal global description of the application knowledge, e.g. by using graph matching, the approach is effectively top-down and thereby starts from object or concept modeling.

System modeling. We now leave the internal platform and concentrate on research that is based on the external study of the pattern recognition abilities of humans and animals or their brains and senses. If this is done in a bottom-up way, the Aristotelian approach, then we are in the area of low-level modeling of senses, nerves and possibly brains. These models are based on the physical and physiological knowledge of cells and the proteins and minerals that constitute them. Senses themselves usually do not directly generalize from observations. They may be constructed, however, in such a way that on a higher level this process is strongly favored. For instance, the way the eye, and in particular the retina is constructed, is advantageous for the detection of edges and movements and for finding interesting details in a global, overall picture. The area of vision thereby profits from this approach. On a level closer to the brain it is studied how nerves process the signals they receive from the senses. Somehow this is combined to a generalization of what is observed by the senses. Models of systems of multiple nerves are called neural networks. They appeared to have a good generalization ability and are thereby also used in technical pattern recognition applications in which the physiological origin is not relevant [4, 62].

Concept modeling. In the external platform, the observations in the starting point are replaced by ideas and concepts. Here one still tries to model externally given pattern recognition systems, but now in a top-down manner. An example is the field of expert systems: by interviewing experts in a particular pattern recognition task, it is attempted to investigate what rules they use and in what way they are using observations. Also belief networks and probabilistic networks belong to this area as far as they are defined by experts and not learned from observations. This approach can be distinguished from the above discussed system modeling by the fact that it is in no way attempted to model a physical or physiological system in a realistic way. The building blocks are the ideas, concepts and rules, as they live in the mind of the researcher. They are adapted to the application by external inspection of an expert, e.g. by interviewing him. If this is done by the researcher internally by introspection, we have closed the circle and are back to what we have called object modeling, as the individual observations are our internal starting point. We admit that the difference between the two Platonic approaches is minor here (in contrast to the physiological level) as we can also try to interview ourselves to create an objective (!) model of our own concept definitions.

2.4 Examples of interaction

The four presented approaches are four ways to study the science of pattern recognition. Resulting knowledge is valid for those who share the same starting point. If the results are used for building artificial pattern recognition devices, then there is, of course, no reason to restrict oneself to a particular approach. Any model that works well may be considered. There are, however, certain difficulties in combining different approaches. These may be caused by differences in culture, in assumptions or in targets. We will present two examples, one for each of the two dichotomies.

Artificial neural networks constitute an alternative technique to be used for generalization within the area of statistical pattern recognition. It took, however, almost

ten years since their introduction around 1985 before neural networks were fully acknowledged in this field. In that period, the neural network community suffered from lack of knowledge on the competing classification procedures. One of the basic misunderstandings in the pattern recognition field was caused by its dominating paradigm, stating that learning systems should never be larger than strictly necessary, following the Occams razor principle. It could have not been understood how largely oversized systems as neural networks would have ever been able to generalize without adapting to peculiarities in the data (the so-called overtraining). At the same time, it was evident in the neural network community that the larger neural network the larger its flexibility, as a brain with more neurons would perform better in learning than a brain with less neurons. When this contradiction was finally solved (an example of Kuhn's paradigm shifts [48]), the area of statistical pattern recognition was enriched with a new set of tools. Moreover, some principles were formulated towards understanding of pattern recognition that otherwise would only have been found with great difficulties.

In general, it may be expected that the internal approach profits from the results in the external world. It is possible that thinking, the way we generalize from observations, changes after it is established how this works in nature. For instance, once we have learned how a specific expert solves his problems, this may be used more generally and thereby becomes a rule in structural pattern recognition. The external platform may thereby be used to enrich the internal one.

A direct formal fertilization between the Platonic and Aristotelian approaches is more difficult to achieve. Individual researchers may build some understanding from studying each other's insights, and thereby become mutually inspired. The Platonist may become aware of realizations of his ideas and concepts. The Aristotelian may see some possible generalizations of the observations he collected. It is, however, still one of the major challenges in science to formalize this process.

How should existing knowledge be formulated such that it can be enriched by new observations? Everybody who tries to do this directly encounters the problem that observations may be used to reduce uncertainty (e.g. by the parameter estimation in a model), but that it is very difficult to formalize uncertainty in existing knowledge. Here we encounter a fundamental 'paradox' for a researcher summarizing his findings after years of observations and studies: he has found some answers, but almost always he has also generated more new questions. Growing knowledge comes with more questions. In any formal system, however, in which we manage to incorporate uncertainty (which is already very difficult), this uncertainty will be reduced after it has absorbed some observations. In order to generate new questions, we need automatic hypothesis generation. How should the most likely ones be determined? For a breakthrough or toward building a more complete theory, we need to look from different perspectives in order to stimulate the creative process and bring sufficient inspiration and novelty to hypothesis generation. This is the cause of the computational complexity mentioned in the literature [60] when the Platonic structural approach to pattern recognition has to be integrated with the Aristotelian statistical approach.

The same problem may also be phrased differently: how can we express the uncertainty in higher level knowledge in such a way that it may be changed (upgraded?)

by low level observations? Knowledge is very often structural and has thereby a qualitative nature, while on the lowest level observations are often treated as quantities, certainly in automatic systems equipped with physical sensors. And here the Platonic – Aristotelian polarity meets the internal – external polarity: by crossing the border between concepts and observations we also encounter the border between qualitative symbolic descriptions and quantitative measurements.

3 Achievements

In this section we will sketch in broad terms the state of the art in building systems for generalization and recognition. In practical applications it is not the primary goal to study the way of bridging the gap between observations and concepts in a scientific perspective. Still, we can learn a lot from the heuristic solutions that are created to assist the human analyst performing a recognition task. There are many systems that directly try to imitate the decision making process of a human expert, such as an operator guarding a chemical process, an inspector supervising the quality of industrial production or a medical doctor deriving a diagnosis from a series of medical tests. On the basis of systematic interviews the decision making can be made explicit and imitated by a computer program: an **expert system** [54]. The possibility to improve such a system by learning from examples is usually very limited and restricted to logical inference to make the rules as general as possible, and the estimation of the thresholds on observations. The latter is needed as the human expert is not always able to define exactly what he means, e.g. by 'an unusually high temperature'.

In order to relate knowledge to observations, which are measurements in automatic systems, it is often needed to relate **knowledge uncertainty** to imprecise, noisy, or not generally valid measurements. Several frameworks have been developed to this end, e.g. fuzzy systems [74], Bayesian belief networks [42] and reasoning under uncertainty [82]. Characteristic for these approaches is that the given knowledge is already structured and needs explicitly defined parameters of uncertainty. New observations may adapt these parameters by relating them to observational frequencies. The knowledge structure is not learned; it has to be given and is hard to modify. An essential problem is that the variability of the external observations may be probabilistic, but the uncertainty in knowledge is based on 'belief' or 'fuzzy' definitions. Combining them in a single mathematical framework is disputable [39].

In the above approaches the general knowledge, or the concept underlying a class of observations, is directly modeled. In **structural pattern recognition** [26, 65] the starting point is the description of the structure of a single object. This can be done in several ways, e.g. by strings, contour descriptions, time sequences, or other order-dependent data. Grammars can be inferred for a collection of strings, resulting in a syntactical approach to pattern recognition [26]. The incorporation of probabilities, e.g. needed for modeling the measurement noise, is not straightforward. Another possibility is the use of graphs. This is in fact already a reduction since objects are decomposed into highlights or landmarks, possibly given by attributes and their relations, which may also be attributed. Inferring a language for graphs is already much

more difficult than for strings. Consequently, the generalization from a set of objects to a class is usually done by finding typical examples, prototypes, followed by graph matching [5, 78] for classifying new objects.

Generalization in structural pattern recognition is not straightforward. It is often based on the comparison of object descriptions using the entire available training set (the nearest neighbor rule) or a selected subset (the nearest prototype rule). Application knowledge is needed for defining the representation (strings, graphs) as well as for the dissimilarity measure to perform graph matching [7, 51]. The generalization may rely on an analysis of the matrix of dissimilarities, used to determine prototypes. More advanced techniques using the dissimilarity matrix will be described later.

The **1-Nearest-Neighbor Rule** (1-NN) is the simplest and most natural classification rule. It should always be used as a reference. It has a good asymptotic performance for metric measures [10, 14], not worse than twice the Bayes error, i.e. the lowest error possible. It works well in practice for finite training sets. Fig. 3 shows how it performs on the Iris data set in comparison to the linear and quadratic classifiers based on the assumption of normal distributions [27]. The k-NN rule, based on a class majority vote over the k nearest neighbors in the training set is, like the Parzen classifier, even Bayes consistent. For increasing training sets they approximate the Bayes error [14, 27].
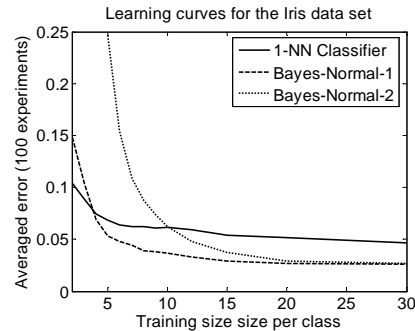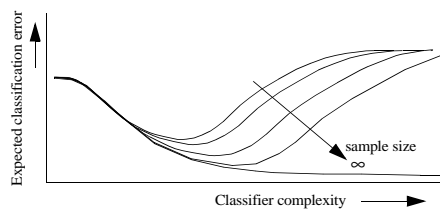

Fig. 3. Learning curves for Iris data.

These results are heavily rely on the assumption that the training examples are identically and independently drawn (iid) from the same distribution as the (future) objects to be tested. This assumption of a fixed and stationary distribution is very strong, but it yields the best possible classifier. There are, however, other reasons, why it cannot be claimed that pattern recognition is solved by these statistical tools. The 1-NN and k-NN rules have to store the entire training set. The solution is thereby based on a comparison with all possible examples, including ones that are very similar, and asymptotically identical to the new objects to be recognized. By this, a class or a concept is not learned, as the decision relies on memorizing all possible instances. Hence, there is no generalization.

Other classification procedures, such as the two **learning curves** shown in Fig. 3, are based on some model assumptions. The classifiers may well perform when the assumptions hold, or may entirely fail, otherwise. An important observation is that such models for statistical learning procedures have almost necessarily a statistical formulation. Human knowledge, however, certainly in daily life, has almost nothing to do with statistics. Perhaps it is hidden in the human learning process, but it is not explicitly available in the context of human recognition. As a result, there is a need to look for effective model assumptions that are not phrased in statistical terms.

We can see from Fig. 3 that a more complex quadratic classifier performs initially worse than the other two, but it behaves like a simpler linear classifier for large training sets. In general, complex problems may be better solved by complex procedures. This is illustrated in Fig. 4, in which the resulting error curves are shown as functions of com-

Fig. 4. Curse of dimensionality.

plexity and training size. Like in Fig. 3, small training sets require simple classifiers. Larger training sets may be used to train more complex classifiers, but the error will increase, if pushed too far. This is a well-known and frequently studied phenomenon in relation with the dimensionality (complexity) of the problem. Objects described by many features often rely on complex classifiers, which may thereby lead to worse results if the number of training examples is insufficient. This is the **curse of dimensionality**, also known as the Rao's paradox or the peaking phenomenon [44, 45]. It is caused by the fact that the classifiers badly generalize, due to a poor estimation of their parameter or their focus/adaptation to the noisy information or or irrelevant details in the data. The same phenomenon can be observed while training complex neural networks without proper precautions. As a result, they will adapt to accidental data configurations, hence they will *overtrain*. This phenomenon is also well known outside the pattern recognition field. For instance, it is one of the reasons one has to be careful with extensive mass screening in health care: the more diseases and their relations are considered (the more complex the task), the more people will we be unnecessarily sent to hospitals.

An important conclusion from this research is that the cardinality of the set of examples from which we want to infer a pattern concept bounds the complexity of the procedure used for generalization. This method should be simple if we have just a few examples. If we still want to learn a somewhat complicated concept this can only be done if sufficient prior knowledge is available, incorporated in such a way that the simple procedure is able to contribute.

An extreme consequence of the lack of prior knowledge is given by Watanabe as the **Ugly Duckling Theorem** [75]. Assume that objects are described by a set of atomic properties and we consider predicates consisting of all possible logic combinations of these properties in order to train a pattern recognition system. Then, all pairs of objects are equally similar in terms of the number of predicates they share. This is caused by the fact that all atomic properties, their presence as well as their absence, have initially equal weights. As a result, the training set is of no use. Summarized briefly, if we do not know anything about the problem we cannot learn (generalize and/or infer) from observations.

An entirely different reasoning pointing to the same phenomenon is the **No-Free-Lunch Theorem** formulated by Wolpert [81]. It states that all classifiers perform equally well if averaged over all possible classification problems. This also includes a random assignment of objects to classes. In order to understand this theorem it should be realized that the considered set of all possible classification problems includes all

possible ways a given data set can be distributed over a set of classes. This again emphasizes that learning cannot be successful without any preference or knowledge.

In essence, it has been established that without prior or background knowledge, no learning, no generalization from examples is possible. In specific applications using strong models for the classes, it has been shown that additional observations may fill well the specified gaps or solve uncertainties in these models. In addition, if these uncertainties can be formulated in statistical terms, it is well possible to diminish their influence by a statistical analysis of the training set. It is, however, unclear which the minimum prior knowledge is that is necessary to make the learning from examples possible. This is of interest if we want to uncover the roots of concept formation, such as learning a class from examples. There exists one principle, formulated at the very beginning of the study of automatic pattern recognition, which may point to a promising direction. This is the principle of **compactness** [1], also phrased as a hypothesis. It states that we can only learn from examples or phenomena if their representation is such that small variations in these examples cause small deviations in the representation. This demands that the representation is based on a continuous transformation of the real world objects or phenomena. Consequently, it is assumed that a sufficiently small variation in the original object will not cause the change of its class membership. It will still be a realization of the same concept. Consequently, we may learn the class of objects that belong to the same concept by studying the domain of their corresponding representations.

The Ugly Duckling Theorem deals with discrete logical representations. These cannot be solved by the compactness hypothesis unless some metric is assumed that replaces the similarity measured by counting differences in predicates. The No-Free-Lunch Theorem clearly violates the compactness assumption as it makes object representations with contradictory labelings equally probable. In practice, however, we encounter only specific types of problems.

Building proper **representations** has become an important issue in pattern recognition [20]. For a long time this idea has been restricted to the reduction of overly large feature sets to the sizes for which generalization procedures can produce significant results, given the cardinality of the training set. Several procedures have been studied based on selection and linear as well as nonlinear extraction [45]. A pessimistic result was found that about any hierarchical ordering of (sub)space separability that fulfills the necessary monotonicity constraints can be constructed by an example based on normal distributions only [11]. Very advanced procedures are needed to find such 'hidden' subspaces in which classes are well separable [61]. It has to be doubted, however, whether such problems arise in practice, and whether such feature selection procedures are really necessary in problems with finite sample sizes. This doubt is further supported by an insight that feature reduction procedures should rely on global and not very detailed criteria if their purpose is to reduce the high dimensionality to a size which is in agreement with the given training set.

Feed-forward neural networks are a very general tool that, among others, offer the possibility to train a single system between sensor and classification [4, 41, 62]. They thereby cover the representation step in the input layer(s) and the generalization step in the output layer(s). These layers are optimized simultaneously. The number of

neurons in the network should be sufficiently enough to make the interesting optima tractable. This, however, brings the danger of overtraining. There exist several ways to prevent that by incorporating some regularization steps into the optimization. This replaces the adaptation step in Fig. 1. A difficult point here, however, is that it is not sufficiently clear how to choose regularization of an appropriate strength. The other important application of neural networks is that the use of various regularization techniques enables one to control the nonlinearity of the resulting classifier, leading also This leads moderately nonlinear functions. Neural networks are thereby one of the most general tools for building pattern recognition systems.

In statistical learning, Vapnik has rigorously studied the problem of adapting the complexity of the generalization procedure to a finite training set [72, 73]. The resulting **Vapnik-Chervonenkis** (VC) dimension, a complexity measure for a family of classification functions, gives a good insight into the mechanisms that determine the final performance (which depends on on the training error and the VC dimension). The resulting error bounds, however, are too general for a direct use. One of the reasons is that, like in the No-Free-Lunch Theorem, the set of classification problems (positions and labeling of the data) is not restricted to the ones that obey the compactness assumption.

One of the insights gained by studying the complexity measures of polynomial functions is that they have to be as simple as possible in terms of their number of free parameters to be optimized. This was already realized by Cover in 1965 [9]. Vapnik extended this finding around 1994 to arbitrary non-linear classifiers [73]. In that case, however, the number of free parameters is not necessarily indicative for the complexity of a given family of functions, but the VC dimension is. In Vapnik's terms, the VC dimension reflects the flexibility of a family of functions (such as polynomials or radial basis functions) to separate arbitrarily labeled and positioned n-element data in a vector space of a fixed dimension. This VC dimension should be finite and small to guarantee the good performance of the generalization function.

This idea was elegantly incorporated to the **Support Vector Machine** (SVM) [73], in which the number of parameters is as small as a suitably determined subset of the training objects (the support vectors) and independent of the dimensionality of the vector space. One way to phrase this principle is that the structure of the classifier itself is simplified as far as possible (following the Occam's razor) So, after a detour along huge neural networks with sometimes many more parameters than training examples, pattern recognition was back to the *small-is-beautiful* principle, but now better understood and elegantly formulated.

The use of **kernels** largely enriched the applicability of the SVM to nonlinear decision functions [66, 67, 73]. The kernel approach virtually generates nonlinear combinations of the existing features. By using the representer theorem, a linear classifier in the non-linear feature space can be constructed, since the kernel encodes generalized inner products of the original vectors only. Consequently, well-performing nonlinear classifiers between training sets of almost any size in almost any feature space can be computed, using the SVM in combination with the 'kernel trick' [66].

This method has still a few limitations. It was originally designed for separable classes, hence it suffers from high class overlaps, as the use of slack variables leads

to a large number of support vectors, and consequently to a large VC dimension. In such cases, other learning procedures have to be preferred. Another difficulty is that the class of admissible kernels is very narrow. Kernels K have to be (conditionally) positive semidefinite (cpd) functions of two variables as only then they are interpreted as generalised inner products in Hilbert spaces induced by K. Although kernels are now designed for more general representations, they are still mainly considered as functions in Euclidean vector spaces. For a number of applications it is difficult to find good features, as in text processing and shape recognition. **Special purpose kernels** are designed in these cases. They use background knowledge from the application in which similarities between objects are defined in such a way that a proper kernel is constructed. The difficulty is, again, the strong requirement of kernels as being cpd.

The next step is the so-called **dissimilarity representation** [56] in which general proximity measures between objects can be used for their representation. The measure itself may be arbitrary, provided that it is meaningful for the considered problem. Proximity plays a key role in the quest for an integrated structural and statistical learning model, since it is a natural bridge between these two approaches [6, 56]. The reason is that proximity is the basic quality to capture what makes a set objects to be considered as a group. It can be defined in various ways and contexts, based on sensory measurements, numerical descriptions, sequences, graphs, relations and other non-vectorial representations, as well as their combinations. A representation based on proximities is, therefore, universal.

Although some foundations are laid down [56], the ways for effective learning from general proximity representations are still to be developed. Since measures may not belong to the class of permissable kernels, the SVM, as such, can not be used. There exist alternative interpretations of such indefinite kernels and their relation to pseudo-Euclidean and Krein spaces [38, 50, 55, 56, 58], in which learning is possible for **non-Euclidean representations**. In general, such representations are embedded into suitable vector spaces equipped with a generalized inner product or norm, in which numerical techniques can be either developed or adapted from the existing ones. It has been experimentally shown that many classification techniques may perform well for general dissimilarity representations.

4 Perspectives

Pattern recognition deals with discovering, distinguishing, detecting or characterizing patterns present in the surrounding world. It relies on extraction and representation of information from the observed data, such that after integration with background knowledge, it ultimately leads to a formulation of new knowledge and concepts. The result of learning is that the knowledge already captured in some formal terms is used to describe the present interdependencies such that the relations between patterns are better understood (interpreted) or used for generalization. The latter means that a concept, e.g. of a class of objects, is formalized such that it can be applied to unseen examples of the same domain, inducing new information, e.g. the

class label of a new object. In this process, new examples should obey the same deduction process and follow the same reasoning as applied to the original examples.

In the following subsections we will first recapitulate the elements of logical reasoning that contribute to learning. Next, this will be related to the Platonic and Aristotelian scientific approaches discussed in Section 2. Finally, two novel pattern recognition paradigms are placed in this view.

4.1 Learning by logical reasoning

Learning from examples is as an active process of concept formation that relies on abstraction (focus on important characteristics or reduction of detail) and analogy (comparison between different entities or relations focussing on some aspect of their similarity). Learning often requires dynamical, multi-level (seeing the details which, when united, lead to concepts, which further build higher level concepts) and possibly multi-strategy actions (e.g. in order to support good predictive power as well as interpretability). A learning task is basically defined by input data (design set), background knowledge or problem context and a learning goal [52]. Many inferential strategies need to be synergetically integrated to be successful in reaching this goal. The most important are inductive, deductive and abductive principles, which are briefly presented next. More formal definitions can be sought in the literature on formal logic, philosophy or e.g. in [23, 40, 52, 83].

Inductive reasoning is the *synthetic* inference process of arriving at a conclusion or a general rule from a limited set of observations. This relies on a formation of a concept or a model, given the data. Although such a derived inductive conclusion cannot be proved, its reliability is supported by empirical observations. As along as the related deductions are not in contradiction with experiments, the inductive conclusion remains valid. If, however, future observations lead to contradiction, either an adaption or a new inference is necessary to find a better rule. To make it more formal, induction learns a general rule R (concerning A and B) from numerous examples of A and B. In practice, induction is often quantitative. Its strength often relies on probability theory and the law of large numbers, in which given a large number of cases, one can describe their properties in the limit and the corresponding rate of convergence.

Deductive reasoning is the *analytic* inference process in which existing knowledge of known facts or agreed-upon rules is used to derive a conclusion. Such a conclusion does not yield 'new' knowledge , since it is a logical consequence of what has already been known, but implicitly (it is not of a greater generality than the premises). Deduction, therefore, uses a logical argument to make explicit what has been hidden. It is also a valid form of proof provided that one starts from true premises. It has a predictive power, which makes it complementary to induction. In a pattern recognition system, both evaluation and prediction rely on deductive reasoning. To make it more formal, let us assume that A is a set of observations, B is a conclusion and R is a general rule. Let B be a logical consequence of A and R, i.e. $(A \wedge R) \models B$, where $\models$ denotes entailment. In a deductive reasoning, given A and using the rule R, the consequence B is derived.

Abductive reasoning is the *constructive* process of deriving the most likely or best explanations of known facts. This is a creative process, in which possible and feasible hypotheses are generated for a further evaluation. Since both abduction and induction deal with incomplete information, induction may be be viewed in some aspects as abduction and vice versa, which leads to some confusion between these two [23, 52]. Here, we assume they are different. Concerning the entailment $(A \wedge R) \models B$, having observed the consequence B in the context od the rule R, A is derived to *explain B*.

In all learning paradigms there is an interplay between inductive, abductive and deductive principles. Both deduction and abduction make possible to conceptually understand a phenomenon, while induction verifies it. More precisely, abduction generates or reformulates new (feasible) ideas or hypotheses, induction justifies the validity of these hypothesis with observed data and deduction evaluates and tests them. Concerning pattern recognition systems, *abduction* explores data, transforms the representation and suggests feasible classifiers for the given problem. It also generates new classifiers or reformulates the old ones. Abduction is present in an initial exploratory step or in the Adaptation stage; see Fig. 1. Induction trains the classifier in the Generalization stage, while deduction predicts the final outcome (such as label) for the test data by applying the trained classifier in the Evaluation stage.

Since abduction is hardly emphasized in learning, we will give some more insights. In abduction, a peculiarity or an artifact is observed and a hypothesis is then created to explain it. Such a hypothesis is suggested based on existing knowledge or may extend it, e.g. by using analogy. So, the abductive process is creative and works towards new discovery. In data analysis, visualization facilitates the abductive process. In response to visual observations of irregularities or bizarre patterns, a researcher is inspired to look for clues that can be used to explain such an unexpected behavior. Mistakes and errors can therefore greatly serve the purpose of discovery when odd results are inquired with a critical mind. Note, however, that this process is very hard to implement into automatic recognition systems as it would require to encode not only the detailed domain knowledge, but also techniques that are able to detect 'surprises' as well as strategies for their possible use. In fact, this requires a conscious interaction. Ultimately, only a human analyst can interactively respond in such cases, so abduction can be incorporated into semi-automatic systems well. In traditional pattern recognition systems, abduction is usually defined in the terms of data and works over pre-specified set of transformations, models or classifiers.

### 4.2 Logical reasoning related to scientific approaches

If pattern recognition (learning from examples) is merely understood as a process of concept formation from a set of observations, the inductive principle is the most appealing for this task. Indeed, it is the most widely emphasized in the literature such that the 'learning' is implicitly understood as 'inductive learning'. Such a reasoning leads to inferring new knowledge (rule or model) which is hopefully valid not only for the known examples, but also for novel, unseen objects. Various validation measures or adaptation steps are taken to support the applicability of the

determined model. Additionally, care has to be taken that the unseen objects obey the same assumptions as the original objects used in training. If this does not hold, such an empirical generalization becomes invalid. One should therefore exercise in critical thinking while designing a complete learning system. It means that one has to be conscious which assumptions are made and be able to quantify their sensibility, usability and validity with the learning goal.

On the other hand, deductive reasoning plays a significant role in the Platonic approach. This top-down scenario starts from a set of rules derived from expert knowledge on problem domain or from a degree of belief in a hypothesis. The existing prior knowledge is first formulated in appropriate terms. These are further used to generate inductive inferences regarding the validity of the hypotheses in the presence of observed examples. So, deductive formalism (description of the object's structure) or deductive predictions (based on the Bayes rule) precede inductive principles. A simple example in the Bayesian inference is the well-known Expectation-Maximization (EM) algorithm used in problems with incomplete data [13]. The EM algorithm iterates between the E-step and M-step until convergence. In the E-step, given a current (or initial) estimate of the unknown variable, a conditional expectation is found, which is maximized in the M-step and derives a new estimate. The E-step is based on deduction, while the M-step relies on induction. In the case of Bayesian nets, which model a set of concepts (provided by an expert) through a network of conditional dependencies, predictions (deductions) are made from the (initial) hypotheses (beliefs over conditional dependencies) using the Bayes theorem. Then, inductive inferences regarding the hypotheses are drawn from the data. Note also that if the existing prior knowledge is captured in some rules, learning may become a simplification of these rules such that their logical combinations describe the problem.

In the Aristotelian approach to pattern recognition, observation of particulars and their explanation are essential for deriving a concept. As we already know, abduction plays a role here, especially for data exploration and characterization to explain or suggest a modification of the representation or an adaptation of the given classifier. Aristotelian learning often relies on the Occam's razor principle which advocates to choose the simplest model or hypothesis among otherwise equivalent ones and can be implemented in a number of ways [8].

In summary, the Platonic scenario is dominantly inductive-deductive, while the Aristotelian scenario is dominantly inductive-abductive. Both frameworks have different merits and shortcomings. The strength of the Platonic approach lies in the proper formulation and use of subjective beliefs, expert knowledge and possibility to encode internal structural organization of objects. It is model-driven. In this way, however, the inductive generalization becomes limited, as there may be little freedom in the description to explore and discovery of new knowledge. The strength of the Aristotelian approach lies in a numerical induction and a well-developed mathematical theory of vector spaces in which the actual learning takes place. It is data-driven. The weakness, however, lies in the difficulty to incorporate the expert or background knowledge about the problem. Moreover, in many practical applications, it is known that the implicit assumptions of representative training sets, identical and identically distributed (iid) samples as well as stationary distributions do not hold.
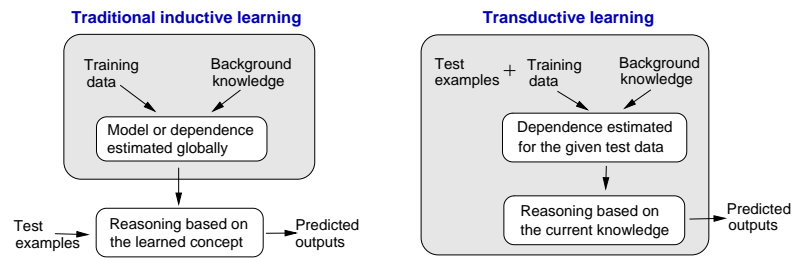
Fig. 5. Inductive (left) and transductive (right) learning paradigms. Background knowledge is here understood in terms of properties of the representations and the specified assumptions on a set of learning algorithms and related parameters; see also [8].

4.3 Two new pattern recognition paradigms

Two far-reaching novel paradigms have been proposed that deal with the drawbacks of the Platonic and Aristotelian approaches. In the Aristotelian scenario, Vapnik has introduced *transductive* learning [73], while in the Platonic scenario Goldfarb has advocated a new structural learning paradigm [31, 32]. We think these are two major perspectives of the science of pattern recognition.

Vapnik [73] formulated the main learning principle as: *'If you posses a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step.'* In the traditional Aristotelian scenario, the learning task is often transformed to the problem of function estimation, in which a decision function is determined *globally* for the entire domain (e.g. for all possible examples in a feature vector space). This is, however, a solution to a more general problem than necessary to arrive at a conclusion (output) for *specific* input data. Consequently, the application of this common-sense principle requires a reformulation of the learning problem such that novel (unlabeled) examples are considered in the context of the given training set. This leads to the transductive principle which aims at estimating the output for a *given* input only when required and may differ from an instance to instance. The training sample, considered either globally, or in the local neighborhoods of test examples, is actively used to determine the output. As a result, this leads to confidence measures of single predictions instead of globally estimated classifiers. It provides ways to overcome the difficulty of iid samples and stationary distributions. More formally, in a transductive reasoning, given an entailment $A \models (B \cup C)$, if the consequence B is observed as the result of A, then the consequence C becomes more likely.

The truly transductive principle requires an active synergy of inductive, deductive and abductive principles in a conscious decision process. We believe it is practised by people who are analyzing complex situations, learn, validate and deduce in order to make decisions in novel ways. Examples are medical doctors, financial advisers, strategy planners or leaders of large organizations. In the context of automatic learning, transduction has applications to learning from partially labeled sets and otherwise missing information, information retrieval, active learning and all types of

diagnostics. Some proposals can be found e.g. in [34, 46, 47, 73]. Although many researchers recognize the importance of this principle, many remain also reluctant. This may be caused by unfamiliarity with this idea, little existing procedures, or by the accompanying computational costs as a complete decision process has to be constantly inferred anew.

In the Platonic scenario, Goldfarb and his colleagues have developed structural inductive learning, realized by the so-called evolving transformation systems (ETS) [31, 32]. Goldfarb first noticed the intrinsic and impossible to overcome inadequacy of vector spaces to truly learn from examples [30]. The reason is that such quantitative representations loose all information on object structure; there is no way an object can be generated given its numeric encoding. The second crucial observation was that all objects in the universe have a formative history. This led Goldfarb to the conclusion that an object representation should capture the object's formative evolution, i.e. the way the object is created through a sequence of suitable transformations in time. The creation process is only possible through structural operations. So, *'the resulting representation embodies temporal structural information in the form of a formative, or generative, history'* [31]. Consequently, objects are treated as evolving structural processes and a class is defined by structural processes, which are 'similar'. This is an inductive structural/symbolic class representation, the central concept in ETS. This representation is learnable from a (small) set of examples and has the capability to generate objects from the class.

The generative history of a class starts from a single progenitor and is encoded as a multi-level hierarchical system. On a given level, the basic structural elements are defined together with their structural transformations, such that both are used to constitute a new structural element on a higher level. This new element becomes meaningful on that level. Similarity plays an important role, as it is used as a basic quality for a class representation as a set of similar structural processes. Similarity measure is learned in training to induce the optimal finite set of weighted structural transformations that are necessary on the given level, such that the similarity of an object to the class representation is large. 'This mathematical structure allows one to capture dynamically, during the learning process, the compositional structure of objects/events within a given inductive, or evolutionary, environment' [31].

Goldfarb's ideas bear some similarity to the ones of Wolfram, presented in his book on 'a new kind of science' [80]. Wolfram considers computation as the primary concept in nature; all processes are the results of cellular-automata[6] type of computational processes, and thereby inherently numerical. He observes that repetitive use of simple computational transformations can cause very complex phenomena and even more if computational mechanisms are used at different levels. Goldfarb also discusses dynamical systems, in which complexity is built from simpler structures, through hierarchical folding up (or enrichment). The major difference is that he considers structure of primary interest, which leads to evolving temporal structural processes instead of computational ones.

---

[6]Cellular automata are discrete dynamical systems that operate on a regular lattice in space and time, and are characterized by 'local' interactions.

In summary, Goldfarb proposes a revolutionary paradigm: an ontological model of a class representation in an epistemological context, as it is learnable from examples. This is a truly unique unification. We think it is the most complete and challenging approach to pattern recognition to this date, a breakthrough. By including the formative history of objects into their representation, Goldfarb attributes them some aspects of human consciousness. The far reaching consequence of his ideas is a generalized measurement process that will be one day present in sensors. Such sensors will be able to measure 'in structural units' instead of numerical units (say, meters) as it is currently done. The inductive process over a set of structural units lies at the foundation of new inductive informatics. The difficulty, however, is that the current formalism in mathematics and related fields is not yet prepared for adopting these far-reaching ideas. We, however, believe, they will pave the road and be found anew or rediscovered in the next decennia.

5 Challenges

A lot of research effort is needed before the two novel and far-reaching paradigms are ready for practical applications. So, this section focuses on several challenges that naturally come in the current context, an will be summarized for the design of automatic pattern recognition procedures. A number of fundamental problems, related to the various approaches, have already been identified in the previous sections and some will return here on a more technical level. Many of the points raised in this section have been more extensively discussed in [17]. We will emphasize these which have only been touched or are not treated at all in the standard books [15, 71, 76] or in the review by Jain et al. [45]. The issues to be described are just a selection of the many which are not yet entirely understood. Some of them may be solved in the future by the development of novel procedures or by gaining an additional understanding. Others may remain an issue of concern to be dealt with in each application separately. We will be systematically describe them, following the line of advancement of a pattern recognition system; see also Fig. 1:

- **Representation and background knowledge.** This is the way in which individual real world objects and phenomena are numerically described or encoded such that they can be related to each other in some meaningful mathematical framework. This framework has to allow the generalization to take place.
- **Design set.** This is the set of objects available or selected to develop the recognition system.
- **Adaptation.** This is usually a transformation of the representation such that it becomes more suitable for the generalization step.
- **Generalization.** This is the step in which objects of the design set are related such that classes of objects can be distinguished and new objects can be accurately classified.
- **Evaluation.** This is an estimate of the performance of a developed recognition system.

5.1 Representation and background knowledge

The problem of representation is a core issue for pattern recognition [18, 20]. Representation encodes the real world objects by some numerical description, handled by computers in such a way that the individual object representations can be interrelated. Based on that, later a generalization is achieved, establishing descriptions or discriminations between classes of objects. Originally, the issue of representation was almost neglected, as it was reduced to the demand of having discriminative features provided by some expert. Statistical learning is often believed to start in a given feature vector space. Indeed, many books on pattern recognition disregard the topic of representation, simply by assuming that objects are somehow already represented [4, 62]. A systematic study on representation [20, 56] is not easy, as it is application or domain-dependent (where the word *domain* refers to the nature or character of problems and the resulting type of data). For instance, the representations of a time signal, an image of an isolated 2D object, an image of a set of objects on some background, a 3D object reconstruction or the collected set of outcomes of a medical examination are entirely different observations that need individual approaches to find good representations. Anyway, if the starting point of a pattern recognition problem is not well defined, this cannot be improved later in the process of learning. It is, therefore, of crucial importance to study the representation issues seriously. Some of them are phrased in the subsequent sections.

The use of vector spaces. Traditionally, objects are represented by vectors in a feature vector space. This representation makes it feasible to perform some generalization (with respect to this linear space), e.g. by estimating density functions for classes of objects. The object structure is, however, lost in such a description. If objects contain an inherent, identifiable structure or organization, then relations between their elements, like relations between neighboring pixels in an image, are entirely neglected. This also holds for spatial properties encoded by Fourier coefficients or wavelets weights. These original structures may be partially rediscovered by deriving statistics over a set of vectors representing objects, but these are not included in the representation itself. One may wonder whether the representation of objects as vectors in a space is not oversimplified to be able to reflect the nature of objects in a proper way. Perhaps objects might be better represented by convex bodies, curves or by other structures in a metric vector space. The generalization over sets of vectors, however, is heavily studied and mathematically well developed. How to generalize over a set of other structures is still an open question.

The essential problem of the use of vector spaces for object representation is originally pointed out by Goldfarb [30, 33]. He prefers a structural representation in which the original object organization (connectedness of building structural elements) is preserved. However, as a generalization procedure for structural representations does not exist yet, Goldfarb starts from the evolving transformation systems [29] to develop a novel system [31]. As already indicated in Sec. 4.3 we see this as a possible direction for a future breakthrough.

Compactness. An important, but seldom explicitly identified property of representations is compactness [1]. In order to consider classes, which are bounded in their

domains, the representation should be constraint: objects that are similar in reality should be close in their representations (where the closeness is captured by an appropriate relation, possibly a proximity measure). If this demand is not satisfied, objects may be described capriciously and, as a result, no generalization is possible. This compactness assumption puts some restriction on the possible probability density functions used to describe classes in a representation vector space. This, thereby, also narrows the set of possible classification problems. A formal description of the probability distribution of this set may be of interest to estimate the expected performance of classification procedures for an arbitrary problem.

In Sec. 3, we pointed out that the lack of a formal restriction of pattern recognition problems to those with a compact representation was the basis of pessimistic results like the No-Free-Lunch Theorem [81] and the classification error bounds resulting from the VC complexity measure [72, 73]. One of the main challenges for pattern recognition to find a formal description of compactness that can be used in error estimators the average over the set of possible pattern recognition problems.

Representation types. There exists numerous ways in which representations can be derived. The basic 'numerical' types are now distinguished as:

- *Features.* Objects are described by characteristic attributes. If these attributes are continuous, the representation is usually compact in the corresponding feature vector space. Nominal, categorical or ordinal attributes may cause problems. Since a description by features is a reduction of objects to vectors, different objects may have identical representations, which may lead to class overlap.

- *Pixels or other samples.* A complete representation of an object may be approximated by its sampling. For images, these are pixels, for time signals, these are time samples and for spectra, these are wavelengths. A pixel representation is a specific, boundary case of a feature representation, as it describes the object properties in each point of observation.

- *Probability models.* Object characteristics may be reflected by some probabilistic model. Such models may be based on expert knowledge or trained from examples. Mixtures of knowledge and probability estimates are difficult, especially for large models.

- *Dissimilarities, similarities or proximities.* Instead of an absolute description by features, objects are relatively described by their dissimilarities to a collection of specified objects. These may be carefully selected prototypes or representatives for the problem, but also random subsets of the training set may work well [56]. The dissimilarities may be derived from raw data, such as images, spectra or time samples, from original feature representations or from structural representations such as strings or relational graphs. If the dissimilarity measure is nonnegative and zero only for two identical objects, always belonging to the same class, the class overlap may be avoided by dissimilarity representations.

- *Conceptual representations.* Objects may be related to classes in various ways, e.g. by a set of classifiers, each based on a different representation, training set or model. The combined set of these initial classifications or clusterings constitute

a new representation [56]. This is used in the area of combining clusterings [24, 25] or combining classifiers [49].

In the structural approaches, objects are represented in qualitative ways. The most important are strings or sequences, graphs and their collections and hierarchical representations in the form of ontological trees or semantic nets.

Vectorial object descriptions and proximity representations provide a good way for generalization in some appropriately determined spaces. It is, however, difficult to integrate them with the detailed prior or background knowledge that one has on the problem. On the other hand, probabilistic models and, especially, structural models are well suited for such an integration. The later, however, constitute a weak basis for training general classification schemes. Usually, they are limited to assigning objects to the class model that fits best, e.g. by the nearest neighbor rule. Other statistical learning techniques are applied to these if given an appropriate proximity measure or a vectorial representation space found by graph embeddings [79].

It is a challenge to find representations that constitute a good basis for modeling object structure and which can also be used for generalizing from examples. The next step is to find representations not only based on background knowledge or given by the expert, but to learn or optimize them from examples.

5.2 Design Set

A pattern recognition problem is not only specified by a representation, but also by the set of examples given for training and evaluating a classifier in various stages. The selection of this set and its usage strongly influence the overall performance of the final system. We will discuss some related issues.

Multiple use of the training set. The entire design set or its parts are used in several stages during the development of a recognition system. Usually, one starts from some exploration, which may lead to the removal of wrongly represented or erroneously labeled objects. After gaining some insights into the problem, the analyst may select a classification procedure based on the observations. Next, the set of objects may go through some preprocessing and normalization. Additionally, the representation has to be optimized, e.g. by a feature/object selection or extraction. Then, a series of classifiers has to be trained and the best ones need to be selected or combined. An overall evaluation may result in a re-iteration of some steps and different choices.

In this complete process the same objects may be used a number of times for the estimation, training, validation, selection and evaluation. Usually, an expected error estimation is obtained by a cross-validation or hold-out method [68, 77]. It is well known that the multiple use of objects should be avoided as it biases the results and decisions. Re-using objects, however, is almost unavoidable in practice. A general theory does not exist yet, that predicts how much a training set is 'worn-out' by its repetitive use and which suggests corrections that can diminish such effects.

Representativeness of the training set. Training sets should be representative for the objects to be classified by the final system. It is common to take a randomly

selected subset of the latter for training. Intuitively, it seems to be useless to collect many objects represented in the regions where classes do not overlap. On the contrary, in the proximity of the decision boundary, a higher sampling rate seems to be advantageous. This depends on the complexity of the decision function and the expected class overlap, and is, of course, inherently related to the chosen procedure.

Another problem are the unstable, unknown or undetermined class distributions. Examples are the impossibility to characterize the class of non-faces in the face detection problem, or in machine diagnostics, the probability distribution of all casual events if the machine is used for undetermined production purposes. A training set that is representative for the class distributions cannot be found in such cases. An alternative may be to sample the *domain* of the classes such that all possible object occurrences are approximately covered. This means that for any object that could be encountered in practice there exists a sufficiently similar object in the training set, defined in relation to the specified class differences. Moreover, as class density estimates cannot be derived for such a training set, class posterior probabilities cannot be estimated. For this reason such a type of domain based sampling is only appropriate for non-overlapping classes. In particular, this problem is of interest for non-overlapping (dis)similarity based representations [18].

Consequently, we wonder whether it is possible to use a more general type of sampling than the classical iid sampling, namely the domain sampling. If so, the open questions refer to the verification of dense samplings and types of new classifiers that are explicitly built on such domains.

5.3 Adaptation

Once a recognition problem has been formulated by a set of example objects in a convenient representation, the generalization over this set may be considered, finally leading to a recognition system. The selection of a proper generalization procedure may not be evident, or several disagreements may exist between the realized and preferred procedures. This occurs e.g. when the chosen representation needs a non-linear classifier and only linear decision functions are computationally feasible, or when the space dimensionality is high with respect to the size of the training set, or the representation cannot be perfectly embedded in a Euclidean space, while most classifiers demand that. For reasons like these, various adaptations of the representation may be considered. When class differences are explicitly preserved or emphasized, such an adaptation may be considered as a part of the generalization procedure. Some adaptation issues that are less connected to classification are discussed below.

Problem complexity. In order to determine which classification procedures might be beneficial for a given problem, Ho and Basu [43] proposed to investigate its complexity. This is an ill-defined concept. Some of its aspects include data organization, sampling, irreducibility (or redundancy) and the interplay between the local and global character of the representation and/or of the classifier. Perhaps several other attributes are needed to define complexity such that it can be used to indicate a suitable pattern recognition solution to a given problem; see also [2].

Selection or combining. Representations may be complex, e.g. if objects are represented by a large amount of features or if they are related to a large set of prototypes. A collection of classifiers can be designed to make use of this fact and later combined. Additionally, also a number of representations may be considered simultaneously. In all these situations, the question arises on which should be preferred: a selection from the various sources of information or some type of combination. A selection may be random or based on a systematic search for which many strategies and criteria are possible [49]. Combinations may sometimes be fixed, e.g. by taking an average, or a type of a parameterized combination like a weighted linear combination as a principal component analysis; see also [12, 56, 59].

The choice favoring either a selection or combining procedure may also be dictated by economical arguments, or by minimizing the amount of necessary measurements, or computation. If this is unimportant, the decision has to be made according to the accuracy arguments. Selection neglects some information, while combination tries to use everything. The latter, however, may suffer from overtraining as weights or other parameters have to be estimated and may be adapted to the noise or irrelevant details in the data. The sparse solutions offered by support vector machines [67] and sparse linear programming approaches [28, 35] constitute a way of compromise. How to optimize them efficiently is still a question.

Nonlinear transformations and kernels. If a representation demands or allows for a complicated, nonlinear solution, a way to proceed is to transform the representation appropriately such that linear aspects are emphasized. A simple (e.g. linear) classifier may then perform well. The use of kernels, see Sec. 3, is a general possibility. In some applications, indefinite kernels are proposed as being consistent with the background knowledge. They may result in non-Euclidean dissimilarity representations, which are challenging to handle; see [57] for a discussion.

5.4 Generalization

The generalization over sets of vectors leading to class descriptions or discriminants was extensively studied in pattern recognition in the 60's and 70's of the previous century. Many classifiers were designed, based on the assumption of normal distributions, kernels or potential functions, nearest neighbor rules, multi-layer perceptrons, and so on [15, 45, 62, 76]. These types of studies were later extended by the fields of multivariate statistics, artificial neural networks and machine learning. However, in the pattern recognition community, there is still a high interest in the classification problem, especially in relation to practical questions concerning issues of combining classifiers, novelty detection or the handling of ill-sampled classes.

Handling multiple solutions. Classifier selection or classifier combination. Almost any more complicated pattern recognition problem can be solved in multiple ways. Various choices can be made for the representation, the adaptation and the classification. Such solutions usually do not only differ in the total classification performance, they may also make different errors. Some type of combining classifiers will thereby be advantageous [49]. It is to be expected that in the future most pattern

recognition systems for real world problems are constituted of a set of classifiers. In spite of the fact that this area is heavily studied, a general approach on how to select, train and combine solutions is still not available. As training sets have to be used for optimizing several subsystems, the problem how to design complex systems is strongly related to the above issue of multiple use of the training set.

Classifier typology. Any classification procedure has its own explicit or built-in assumptions with respect to data inherent characteristics and the class distributions. This implies that a procedure will lead to relatively good performance if a problem fulfils its exact assumptions. Consequently, any classification approach has its problem for which it is the best. In some cases such a problem might be far from practical application. The construction of such problems may reveal which typical characteristics of a particular procedure are. Moreover, when new proposals are to be evaluated, it may be demanded that some examples of its corresponding typical classification problem are published, making clear what the area of application may be; see [19].

Generalization principles. The two basic generalization principles, see Section 4, are probabilistic inference, using the Bayes-rule [63] and the minimum description length principle that determines the most simple model in agreement with the observations (based on Occam's razor) [37]. These two principles are essentially different[7]. The first one is sensitive to multiple copies of an existing object in the training set, while the second one is not. Consequently, the latter is not based on densities, but just on object differences or distances. An important issue is to find in which situations each of these principle should be recommended and whether the choice should be made in the beginning, in the selection of the design set and the way of building a representation, or it should be postpone until a later stage.

The use of unlabeled objects and active learning. The above mentioned principles are examples of statistical inductive learning, where a classifier is induced based on the design set and it is later applied to unknown objects. The disadvantage of such approach is that a decision function is in fact designed for all possible representations, whether valid or not. Transductive learning, see Section 4.3, is an appealing alternative as it determines the class membership only for the objects in question, while relying on the collected design set or its suitable subset [73]. The use of unlabeled objects, not just the one to be classified, is a general principle that may be applied in many situations. It may improve a classifier based on just a labeled training set. If this is understood properly, the classification of an entire test set may yield better results than the classification of individuals.

Classification or class detection. Two-class problems constitute the traditional basic line in pattern recognition, which reduces to finding a discriminant or a binary decision function. Multi-class problems can be formulated as a series of two-class problems. This can be done in various ways, none of them is entirely satisfactory. An entirely different approach is the description of individual classes by so-called

---

[7]Note that Bayesian inference is also believed to implement the Occam's razor [8] in which preference for simpler models is encoded by encouraging particular prior distributions. This is, however, not the primary point as in the minimum description length principle.

one-class classifiers [69, 70]. In this way the focuss is given to class description instead of to class separation. This brings us to the issue of the structure of a class.

Traditionally classes are defined by a distribution in the representation space. However, the better such a representation, the higher its dimensionality, the more difficult it is to estimate a probability density function. Moreover, as we have seen above, it is for some applications questionable whether such a distribution exist. A class is then a part of a possible non-linear manifold in a high-dimensional space. It has a structure instead of a density distribution. It is a challenge to use this approach for building entire pattern recognition systems.

5.5 Evaluation

Two questions are always apparent in the development of recognition systems. The first refers to the overall performance of a particular system once it is trained, and has sometimes a definite answer. The second question is more open and asks which good recognition procedures are in general.

Recognition system performance. Suitable criteria should be used to evaluate the overall performance of the entire system. Different measures with different characteristics can be applied, however, usually, only a single criterion is used. The basic ones are the average accuracy computed over all validation objects or the accuracy determined by the worst case scenario. In the first case, we again assume that the set of objects to be recognized is well defined (in terms of distributions). Then, it can be sampled and the accuracy of the entire system is estimated based on the evaluation set. In this case, however, we neglect the issue that after having used this evaluation set together with the training set, a better system could have been found. A more interesting point is how to judge the performance of a system if the distribution of objects is ill-defined or if a domain based classification system is used as discussed above. Now, the largest mistake made becomes a crucial factor for this type of judgements. One needs to be careful, however, as this may refer to an unimportant outlier (resulting e.g. from invalid measurements).

Practice shows that a single criterion, like the final accuracy, is insufficient to judge the overall performance of the whole system. As a result, multiple performance measures should be taken into account, possibly at each stage. These measures should not only reflect the correctness of the system, but also its flexibility to cope with unusual situations in which e.g. specific examples should be rejected or misclassification costs incorporated.

Prior probability of problems. As argued above, any procedure has a problem for which it performs well. So, we may wonder how large the class of such problems is. We cannot state that any classifier is better than any other classifier, unless the distribution of problems to which these classifiers will be applied is defined. Such distributions are hardly studied. What is done at most is that classifiers are compared over a collection of benchmark problems. Such sets are usually defined ad hoc and just serve as an illustration. The collection of problems to which a classification procedure will be applied is not defined. As argued in Section 3, it may be as large as all problems with a compact representation, but preferably not larger.

6 Discussion and conclusions

Recognition of patterns and inference skills lie at the core of human learning. It is a human activity that we try to imitate by mechanical means. There are no physical laws that assign observations to classes. It is the human consciousness that groups observations together. Although their connections and interrelations are often hidden, by the attempt of imitating this process, some understanding might be gained. The human process of learning patterns from examples may follow along the lines of trial and error. By freeing our minds of fixed beliefs and petty details we may not only understand single observations but also induce principles and formulate concepts that lie behind the observed facts. New ideas can be born then. These processes of abstraction and concept formation are necessary for development and survival. In practice, (semi-)automatic learning systems are built by imitating such abilities in order to gain understanding of the problem, explain the underlying phenomena and develop good predictive models.

It has, however, to be strongly doubted whether statistics play an important role in the human learning process. Estimation of probabilities, especially in multivariate situations is not very intuitive for majority of people. Moreover, the amount of examples needed to build a reliable classifier by statistical means is much larger than it is available for humans. In human recognition, proximity based on relations between objects seems to come before features are searched and may be, thereby, more fundamental. For this reason and the above observation, we think that the study of proximities, distances and domain based classifiers are of great interest. This is further encouraged by the fact that such representations offer a bridge between the possibilities of learning in vector spaces and the structural description of objects that preserve relations between objects inherent structure. We think that the use of dissimilarities for representation, generalization and evaluation constitute the most intriguing issues in pattern recognition.

The existing gap between structural and statistical pattern recognition partially coincides with the gap between knowledge and observations. Prior knowledge and observations are both needed in a subtle interplay to gain new knowledge. The existing knowledge is needed to guide the deduction process and to generate the models and possible hypotheses needed by induction, transduction and abduction. But, above all, it is needed to select relevant examples and a proper representation. If and only if the prior knowledge is made sufficiently explicit to set this environment, new observations can be processed to gain new knowledge. If this is not properly done, some results may be obtained in purely statistical terms, but these cannot be integrated with what was already known and have thereby to stay in the domain of observations. The study of automatic pattern recognition systems makes perfectly clear that learning is possible, only if the Platonic and Aristotelian scientific approaches cooperate closely. This is what we aim for.

References

[1] A.G. Arkadev and E.M. Braverman. *Computers and Pattern Recognition.* Thompson, Washington, DC, 1966.

[2] M. Basu and T.K. Ho, editors. *Data Complexity in Pattern Recognition.* Springer, 2006.

[3] R Bergmann. *Developing Industrial Case-Based Reasoning Applications.* Springer, 2004.

[4] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[5] H. Bunke. Recent developments in graph matching. In *Int. Conf. on Pattern Recognition*, volume 2, pages 117–124, 2000.

[6] H. Bunke, S. Günter, and X. Jiang. Towards bridging the gap between statistical and structural pattern recognition: Two new concepts in graph matching. In *Int. Conf. on Advances in Pattern Recognition*, pages 1–11, 2001.

[7] H. Bunke and K. Shearer. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19(3-4):255–259, 1998.

[8] V.S. Cherkassky and F. Mulier. *Learning from data: Concepts, Theory and Methods.* John Wiley & Sons, Inc., New York, NY, USA, 1998.

[9] T.M. Cover. Geomerical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. on Electronic Computers*, EC-14:326–334, 1965.

[10] T.M. Cover and P.E. Hart. Nearest Neighbor Pattern Classification. *IEEE Trans. on Information Theory*, 13(1):21–27, 1967.

[11] T.M. Cover and J.M. van Campenhout. On the possible orderings in the measurement selection problem. *IEEE Trans. on Systems, Man, and Cybernetics*, SMC-7(9):657–661, 1977.

[12] I.M. de Diego, J.M. Moguerza, and A. Muñoz. Combining kernel information for support vector classification. In *Multiple Classifier Systems*, pages 102–111. Springer-Verlag, 2004.

[13] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society*, Series B, 39(1):1–38, 1977.

[14] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, 1996.

[15] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification.* John Wiley & Sons, Inc., 2nd edition, 2001.

[16] R.P.W. Duin. Four scientific approaches to pattern recognition. In *Fourth Quinquennial Review 1996-2001. Dutch Society for Pattern Recognition and Image Processing*, pages 331–337. NVPHBV, Delft, 2001.

[17] R.P.W. Duin and E. Pękalska. Open issues in pattern recognition. In *Computer Recognition Systems*, pages 27–42. Springer, Berlin, 2005.

[18] R.P.W. Duin, E. Pękalska, P. Paclík, and D.M.J. Tax. The dissimilarity representation, a basis for domain based pattern recognition? In L. Goldfarb, editor, *Pattern representation and the future of pattern recognition, ICPR 2004 Workshop Proceedings*, pages 43–56, Cambridge, United Kingdom, 2004.

[19] R.P.W. Duin, E. Pękalska, and D.M.J. Tax. The characterization of classification problems by classifier disagreements. In *Int. Conf. on Pattern Recognition*, volume 2, pages 140–143, Cambridge, United Kingdom, 2004.

[20] R.P.W. Duin, F. Roli, and D. de Ridder. A note on core research issues for statistical pattern recognition. *Pattern Recognition Letters*, 23(4):493–499, 2002.

[21] S. Edelman. *Representation and Recognition in Vision*. MIT Press, Cambridge, 1999.

[22] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, London, 1993.

[23] P. Flach and A. Kakas, editors. *Abduction and Induction: essays on their relation and integration*. Kluwer Academic Publishers, 2000.

[24] A. Fred and A.K. Jain. Data clustering using evidence accumulation. In *Int. Conf. on Pattern Recognition*, pages 276–280, Quebec City, Canada, 2002.

[25] A. Fred and A.K. Jain. Robust data clustering. In *Conf. on Computer Vision and Pattern Recognition*, pages 442 –451, Madison - Wisconsin, USA, 2002.

[26] K.S. Fu. *Syntactic Pattern Recognition and Applications*. Pretice-Hall, 1982.

[27] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[28] G.M. Fung and O.L. Mangasarian. A feature selection newton method for support vector machine classification. *Computational Optimization and Aplications*, 28(2):185–202, 2004.

[29] L. Goldfarb. On the foundations of intelligent processes – I. An evolving model for pattern recognition. *Pattern Recognition*, 23(6):595–616, 1990.

[30] L. Goldfarb, J. Abela, V.C. Bhavsar, and V.N. Kamat. Can a vector space based learning model discover inductive class generalization in a symbolic environment? *Pattern Recognition Letters*, 16(7):719–726, 1995.

[31] L. Goldfarb and D. Gay. What is a structural representation? fifth variation. Technical Report TR05-175, University of New Brunswick, Fredericton, Canada, 2005.

[32] L. Goldfarb and O. Golubitsky. What is a structural measurement process? Technical Report TR01-147, University of New Brunswick, Fredericton, Canada, 2001.

[33] L. Goldfarb and J. Hook. Why classical models for pattern recognition are not pattern recognition models. In *Int. Conf. on Advances in Pattern Recognition*, pages 405–414, 1998.

[34] T. Graepel, R. Herbrich, and K. Obermayer. Bayesian transduction. In *Advances in Neural Information System Processing*, pages 456–462, 2000.

[35] T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K.-R. Müller, K. Obermayer, and R. Williamson. Classification on proximity data with LP-machines. In *Int. Conf. on Artificial Neural Networks*, pages 304–309, 1999.

[36] U. Grenander. *Abstract Inference*. John Wiley & Sons, Inc., 1981.

[37] P. Grünwald, I.J. Myung, and Pitt M., editors. *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2005.

[38] B. Haasdonk. Feature space interpretation of svms with indefinite kernels. *IEEE Trans. on PAMI*, 25(5):482–492, 2005.

[39] I. Hacking. *The emergence of probability*. Cambridge University Press, 1974.

[40] G. Harman and S. Kulkarni. *Reliable Reasoning: Induction and Statistical Learning Theory*. MIT Press, to appear.

[41] S. Haykin. *Neural Networks, a Comprehensive Foundation, second edition*. Prentice-Hall, 1999.

[42] D. Heckerman. A tutorial on learning with bayesian networks. In M. Jordan, editor, *Learning in Graphical Models*, pages 301–354. MIT Press, Cambridge, MA, 1999.

[43] T.K. Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE Trans. on PAMI*, 24(3):289–300, 2002.

[44] A. K. Jain and B. Chandrasekaran. Dimensionality and sample size considerations in pattern recognition practice. In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of Statistics*, volume 2, pages 835–855. North-Holland, Amsterdam, 1987.

[45] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. on PAMI*, 22(1):4–37, 2000.

[46] T. Joachims. Transductive inference for text classification using support vector machines. In I. Bratko and S. Dzeroski, editors, *Int. Conf. on Machine Learning*, pages 200–209, 1999.

[47] T. Joachims. Transductive learning via spectral graph partitioning. In *Int. Conf. on Machine Learning*, 2003.

[48] T.S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1970.

[49] L.I. Kuncheva. *Combining Pattern Classifiers. Methods and Algorithms*. Wiley, 2004.

[50] J. Laub and K.-R. Müller. Feature discovery in non-metric pairwise data. *J. of Machine Learning Research*, pages 801–818, 2004.

[51] A. Marzal and E. Vidal. Computation of normalized edit distance and applications. *IEEE Trans. on PAMI*, 15(9):926–932, 1993.

[52] R.S. Michalski. Inferential theory of learning as a conceptual basis for multistrategy learning. *Machine Learning*, 11:111–151, 1993.

[53] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[54] Richard E. Neapolitan. *Probabilistic reasoning in expert systems: theory and algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 1990.

[55] C.S. Ong, S. Mary, X.and Canu, and Smola A.J. Learning with non-positive kernels. In *Int. Conf. on Machine Learning*, pages 639–646, 2004.

[56] E. Pękalska and R.P.W. Duin. *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications*. World Scientific, Singapore, 2005.

[57] E. Pękalska, R.P.W. Duin, S. Günter, and H. Bunke. On not making dissimilarities Euclidean. In *Joint IAPR Int. Workshops on SSPR and SPR*, pages 1145–1154. Springer-Verlag, 2004.

[58] E. Pękalska, P. Paclík, and R.P.W. Duin. A Generalized Kernel Approach to Dissimilarity Based Classification. *J. of Machine Learning Research*, 2:175–211, 2002.

[59] E. Pękalska, M. Skurichina, and R.P.W. Duin. Combining Dissimilarity Representations in One-class Classifier Problems. In *Multiple Classifier Systems*, pages 122–133. Springer-Verlag, 2004.

[60] L.I. Perlovsky. Conundrum of combinatorial complexity. *IEEE Trans. on PAMI*, 20(6):666–670, 1998.

[61] P. Pudil, J. Novovićova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

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

[63] C.P. Robert. *The Bayesian Choice*. Springer-Verlag, New York, 2001.

[64] K.M. Sayre. *Recognition, a study in the philosophy of artificial intelligence*. University of Notre Dame Press, 1965.

[65] M.I. Schlesinger and Hlavác. *Ten Lectures on Statistical and Structural Pattern Recognition*. Kluwer Academic Publishers, 2002.

[66] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[67] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, UK, 2004.

[68] M. Stone. Cross-validation: A review. *Mathematics, Operations and Statistics*, (9):127–140, 1978.

[69] D.M.J. Tax. *One-class classification. Concept-learning in the absence of counter-examples*. PhD thesis, Delft University of Technology, The Netherlands, 2001.

[70] D.M.J. Tax and R.P.W. Duin. Support vector data description. *Machine Learning*, 54(1):45–56, 2004.

[71] F. van der Heiden, R.P.W. Duin, D. de Ridder, and D.M.J. Tax. *Classification, Parameter Estimation, State Estimation: An Engineering Approach Using MatLab*. Wiley, New York, 2004.

[72] V. Vapnik. *Estimation of Dependences based on Empirical Data*. Springer Verlag, 1982.

[73] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

[74] L.-X. Wang and J.M. Mendel. Generating fuzzy rules by learning from examples. *IEEE Trans. on Systems, Man, and Cybernetics*, 22(6):1414–1427, 1992.

[75] S. Watanabe. *Pattern Recognition, Human and Mechanical*. John Wiley & Sons, 1985.

[76] A. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, Ltd., 2002.

[77] S.M. Weiss and C.A. Kulikowski. *Computer Systems That Learn*. Morgan Kaufmann, 1991.

[78] R.C. Wilson and E.R. Hancock. Structural matching by discrete relaxation. *IEEE Trans. on PAMI*, 19(6):634–648, 1997.

[79] R.C. Wilson, B. Luo, and E.R. Hancock. Pattern vectors from algebraic graph theory. *IEEE Trans. on PAMI*, 27(7):1112–1124, 2005.

[80] S. Wolfram. *A new kind of science*. Wolfram Media, 2002.

[81] D.H. Wolpert. *The Mathematics of Generalization*. Addison-Wesley, 1995.

[82] R.R. Yager, M. Fedrizzi, and J. (Eds) Kacprzyk. *Advances in the Dempster-Shafer Theory of Evidence*. Wesley, 1994.

[83] C.H. Yu. Quantitative methodology in the perspectives of abduction, deduction, and induction. In *Annual Meeting of American Educational Research Association*, San Francisco, CA, 2006.