Prediction Games in Infinitely Rich Worlds

[Extended Abstract]

Omid Madani
madani@yahoo-inc.com
Yahoo! Research
3333 Empire Ave
Burbank, CA 91504

ABSTRACT

In order to familiarize oneself with a rich new world, one needs to learn a great deal. This massive learning includes learning to recognize myriad categories and learning to make use of them for learning other categories. The process of playing *prediction games* may make this massive learning possible. In this paper, we describe prediction games and present a discussion of properties of algorithms and systems that would play them well. The great potential of prediction games is in pointing to ways of achieving powerful large-scale learning without the need for human supervision.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning - Induction

General Terms

Algorithms

1. INTRODUCTION

We introduce the learning process of *playing prediction games* in an *infinitely rich world*. Prediction games may be a step toward making powerful massive unsupervised learning possible. The learned outcome could find tremendous use.

The games are played by a *system* that takes its sequence of inputs from the world and makes learning episodes out of it. Our running example will be *prediction in text*, *i.e.*, playing prediction games in the world of natural language, as available in the online text. One such game is *fill-in-the-blank*, played as follows: the system repeatedly inputs a sentence and hides a portion of it such as a word or phrase. Several components of the system then become active and make predictions about what is missing using the available context. The activated components may then be updated according to the answer, which is available. The system moves on to the next learning opportunity.

UBDM'06 August 20, 2006, Philadelphia, Pennsylvania, USA

The important point is that the amount of training instances is viewed as unbounded. All the learning subprocesses work to improve prediction. The utility is in terms of the operationality of the prediction system: coverage, depth, accuracy, and speed. In the case of prediction in text, the learned system may lead to significant practical improvements in statistical language modeling or complement current techniques [16]. Byproducts of such processes, as we explain, include learning associations and discovering new concepts. We next describe infinitely rich worlds and prediction games in more detail. We motivate a *systems* approach for playing these games and discuss desiderata on solution techniques and some of the challenges we see, such as scalability and prevalence of noise. We briefly describe the work we have begun in this direction. Besides prediction in language, we mention possible games in other worlds. We give the wider context and motivation for prediction games, and situate prediction games with respect to existing learning frameworks.

2. PREDICTION GAMES

Prediction games are proposed to make massive learning possible. By *massive learning*, we mean learning in the order of millions of categories and beyond, and estimating values for variables in the order of billions, and beyond. Prediction games consist of an infinitely rich world and a system that plays the games in that world.

We will use prediction in text as our main example, and in particular the fill-in-the-blank game, as explained in the introduction: every sentence or passage in the online text can serve as a source of one or more learning episodes. In each game the system hides a portion of an input sentence, say a word or a phrase, and sees how well it can predict it, using the context derived from the rest of the sentence and possibly the broader context such as the passage, the page, and so on. For example, in the sentence, "I rode my bike to school", the word "bike" could be covered, and then the question posed by the system to itself is what can replace X in "I rode my X to school". The answer may be in the form of a single phrase, or several candidate phrases ranked (*e.g.*, "bike, vehicle, motor bike, car, horse, table,..") or assigned probabilities. If we constrain the context to the words and sentences occurring before X, we obtain the game of *predict-the-next-word*, which is also the typical problem addressed by statistical language modeling [16, 7].

2.1 The Fundamental Role of Concepts

In describing the world and the system we make use of

the work on categories or concepts[1] in cognitive psychology [14]. An infinitely rich world has a few properties. First, an infinitely rich world is a source of unbounded experience. Ample learning experience is a prereqisite for achieving massive learning. Second, the world is also infinitely rich in that it enjoys infinitely many regularities. These regularities are in the form of the many categories that entities in the world can be grouped into. For us, a category is a very general notion and in our every day world includes groupings of physical things, events, actions, feelings, and so on. Example categories include: "red", "happy", [2] "office chair", "my father", "southern lakes", and so on. In prediction in text, categories are sets of words. From the view point of the system, we may think of concepts (hardwired or learned categories) as an implicit set of vector values (feature values or stimuli) for which membership in the set can be determined by some components of the system at adequate accuracy and efficiency levels.

There is an infinity of ways in which to carve up a rich world into various categories. Learned new categories for the most part should serve a purpose for the system. Here, the main purpose is improving overall prediction ability. Categories serve as great abstractors. Without them, one could not learn from the past, every experience would be brand new, and one would not be able to predict. Categories are fundamental in dealing with a rich world. If they are effectively discovered and recognized, the problems of sparsity in natural language and handling invariance in vision are adequately addressed. Categories are necessary for intelligence.

The regularities in the world are also in the form of the many types of relations that the categories tend to enjoy with one another, such as temporal or spatial co-occurrence, part-of, friends, or generalization and specialization relations (*e.g.*, the category "animal" is a generalization of the category "dog"). Relations help in predicting categories and learning new ones.

Each learning episode contains a few categories that comprise the context (the features) to be used for prediction, and a category to be predicted. Many categories, the raw categories, are hard-wired: their detection in a learning episode is achieved by some component of the system that was part of the system from the outset. Many categories will be constructed (discovered) by the system. A new category can be a grouping of several existing categories, or may be a refinement of an existing category. For example, a category corresponding to the to-be family of verbs may be discovered at some point: it is recognized whenever any of the to-be verbs ("is", "are", "was",..) in their various forms are present. The category "bank" can be refined into the financial-institution sense, the physical building sense (that houses the financial institution), and the side-of-river sense.

Improving prediction performance is the main driver for learning new categories. Other factors include the difficulty of discovering the category and recognizing it. A category has to have some use for the system to spend resources for detecting it. New categories, once learned, are treated the same way as raw categories: they serve as features to define context and they are treated as targets to be predicted, though their detection may take longer time than raw categories.

In text prediction, each possible phrase is treated as a raw category. Detection of these categories may be regarded as hard-wired (NOT learned). Simple features include those corresponding to presence of phrases, and those corresponding to relative position of the phrase with respect to the phrase to be predicted (*e.g.*, a feature can correspond to the word "drink" appearing in two positions before the word to be predicted). More complex features that may be only evaluated part of the time include part of speed tagging and parsing output. Thus, the number of features (learning dimensionality) and classes can be in the millions and beyond. After some learning, the features extracted from the sentence "He flies frequently to X" includes raw features such as the word "flies" appears 3 positions before, as well as more abstract features such as one corresponding to the composite category "traveling-action" appearing before target with confidence 0.6. The system's predictions for X may include the names of a few locations, such as cities or travel destinations, as well as categories or groupings that correspond to travel destinations.

2.2 An Abstraction

We may view prediction games as a table or matrix of instances and target values, $< x_i, y_i >$, just as is common in supervised learning. However the table has infinitely many rows, and the dimensionality and the set of categories to be predicted are unbounded as well. Each row corresponds to an instance or an episode and the last column, the target column to be predicted, corresponds to a category.

In text prediction, each phrase serves as a category, so new phrases will be seen as the online text is processed. More importantly, new categories will be constructed out of old ones and these categories will be equal citizens. Newly constructed categories address the sparsity problem. Efficient discovery of new useful categories is a very important problem. Features are some functions of the categories, such as whether the category is present in the context, or was recently seen, or the location of the category, or conjunctive and relational features involving multiple categories. Therefore, the dimensionality of the problem in practice starts out very high, and furthermore it is growing, as more categories are constructed out of the old ones. The table is highly sparse: only a finite number of features will have a nonzero value in each row. In prediction in text, we do not see a need for negative feature values. Some functionalities may best be achieved if the feature values reflect a measure of confidence or probability.

2.3 Other Games

Other worlds include the world-wide-web, our own every day visual/physical world (*e.g.*, the problem of vision), and the social world and predicting people's actions.

The world-wide-web may be viewed as a superset of the text online. Web pages include text, but are part of a larger structure of a website, and contain additional content or structure that could be utilized, such as tables and pictures, html layout, and links to other pages.

Real-time prediction in a changing visual scene provides

[1] In cognitive psychology, a distinction is made between a category in the external world and the concept of it, as represented internally in the mind [14]. This is an important distinction, but in this paper to simplify the presentation, we don't make a distinction.

[2] The system playing the game is viewed here as part of the world. Therefore, parts of the system may learn to predict the current or subsequent state of other parts.

ample learning episodes. The changes can be due to moving objects, camera movement, and changes in attention. One game can be to predict what will be seen next, based on what has been seen recently. What is learned include temporal or spatial co-occurrence of categories (physical objects) as well as recognizing a whole based on seeing some parts and inferring the remaining parts.

The task of predicting people's actions and reactions, *i.e.*, obtaining a sense of what a person or groups of people may do in various circumstances, for example in social interactions, or in cooperative team work or in competitions, may turn out to be the most challenging of prediction games.

2.4 A Systems Approach

The diversity of categories and relations that could hold between them, and the evolving nature of the world, makes statistical modeling challenging and likely inadequate. Similarly, a single algorithm cannot do the job either. Search in the space of *systems* may be the best strategy for finding effective solutions. A prediction system does not try to model the world, in the sense of building replicas of it internally, nor estimate parameter values for a model within some constrained family of statistical models. It only strives to predict some of the world's aspects well. A system, for example the operating system of a computer, consists of a number of interacting components and is driven by a number of different algorithms responsible for tasks such as paging, process scheduling, and I\O. Similar to an operating system, issues of operationality or maintaining some level of service also holds for a prediction system. The difference from an operating system is that the prediction system learns and adapts and self-organizes extensively, and grows in its functionality over time. Learning is central to its operationality.

2.5 Applications

A powerful predicting system and the byproducts of what is learned should find a number of applications. For the sake of the the system itself, as the system improves at predicting, it does not have to spend as much time and resources into verifying the presence of certain categories such as certain important events. This can have life saving consequences. Based on sounds alone, one may assess the danger level, and quickly finish crossing the street when a fast approaching car is heard. In other situations, partial context is the only information the system has practical access to (*e.g.*, due to occlusion or poor lighting conditions). However, in much of the day to day activity, since the system can predict and classify robustly based on partial context, the system is in general faster and more accurate and is freed to spend its resources to focus on other learning activity.

A prediction system for text may lead to significant practical improvements over current language modeling technology or complement it in ways. Another application is answer to questions of the type "what does the phrase X mean?": find contexts that X appears in by querying the web, hide the phrase, and see what categories (single phrases or sets of phrases) are predicted. Those categories or phrases likely have the same type as X.

3. HOW?

Prediction games involve predicting a large number of categories, and while there is noise, the system has access to an unbounded sequence of learning episodes. The large-scale aspects of prediction games motivate the following desiderata on candidate solutions (systems and algorithms):

- Online algorithm that are time and memory efficient
- Handle large number of features: sample size efficiency
- Handle large numbers of categories (prediction as well as discovery of new categories)
- Robust to imperfections, uncertainty, and variety

Several online linear learning algorithms satisfy some of the above desiderata. The issue of very large numbers of classes appears not to have been addressed adequately before. A basic immediate question is how to efficiently learn and recognize myriad categories? It appears that this problem has found good solutions in nature [17, 10].

A related problem is what may be referred to as the *recall problem*: how to quickly narrow down the possible candidate true positive categories when the system is presented with an instance? An effective solution to the recall problem may play an important part in addressing the challenge of large number of classes. We recently proposed the framework of a *recall system* for this purpose [13]. The ideal recall system, given an instance (a feature vector), quickly outputs a small set of candidate categories (say tens), and does not miss the true categories of the instance. We realized this functionality by an *index* that maps (connects) features to categories. An efficient algorithm for learning the index was given, although superior algorithms may yet to be discovered. The recall system, while high on recall and drastically reducing the number of candidate categories, may still be poor on precision[3]. Online learning algorithms could be used to efficiently learn classifiers for each category. The classifiers corresponding to the retrieved concepts can be applied to the instance to precisely determine the right categories. Alternatively, or in conjunction, learning weighted mappings can make possible adequate ranking of the recalled categories (*i.e.*, without the need for classifier training and application). The recall system may also facilitate other learning activities, such as concept discovery and learning concepts in terms of other learned concepts.

Next, we discuss constraints and desiderata that we see important for any would be candidate architecture and algorithm in effectively playing prediction games.

3.1 Scalability: Information vs. Computation

Scalability is paramount. The system is to operate in an information rich environment, and what is to be learned, an operational predicting system, requires by definition an information hungry process. In some scenarios, for example learning to predict by processing the online text, there is much data available, and the data is relatively static. It is the algorithms' speed that determines how fast it can learn from this abundance. In real-time scenarios such as in vision, the system is bombarded with information, and may

[3]Here, precision and recall are from the point of view of an instance, and not the category. In machine learning literature, recall and precision are often used from the point of view of the category (*e.g.*, recall is the portion of true positive instances that were classified positive for a single category. Our use is similar to information retrieval, were documents are categories here.

have to ignore (drop) much incoming and possibly valuable information due to its processing limits.

There is a trade-off here: on one hand the system can spend much time on the current instance, or it can revisit past instances, perhaps optimize some measure of accuracy over them, or, on the other hand, to spend the computational resources for further exploring the world and acquiring possibly new useful information. A natural question is what determines a good balance and how to achieve it. The ultimate objective, whether the emphasis is competence more on breadth of prediction versus depth, may play a role in the answer. Obtaining insights into the nature of this tradeoff would be very valuable. Recent work in large-scale text mining has raised some of these issues [15].

3.1.1 Learn in an Online Fashion

Batch learning or optimization can provide significant improved accuracy in typical learning problems, when compared to online methods. However, batch techniques are designed inherently with finite data in mind, and prediction games are about unbounded data. Bottou and Le Cun point to both theoretical and empirical advantages of properly designed online algorithms over batch, in terms of accuracy achieved, when training data is abundant [4]. Furthermore, as features are induced (new concepts are discovered), the whole learning activity takes on an evolving nature. While one can imagine subsampling and instance selection (*e.g.*, via some measure of instance utility) to keep things sufficiently small, and using incremental or staged batch learning and optimization, this solution appears complex, and may not be the best utilization of learning time (see also Section 3.3 on code complexity).

A quick calculation demonstrates the considerable potential advantages of light linear time online learning. Consider algorithms A, B and C taking respectively n, $n \log_{10} n$, and n^2 steps to learn from n instances. Then in the time that algorithm B takes to process a million instances, algorithm A can learn from six millions instances, while algorithm C has only processed a 1000 instances. And of course, the advantages of linear time processing grows with increasing n. One has to ask whether the accuracy benefits of a more time consuming algorithm is worth the opportunity loss in learning from more data.

Even within linear time processing settings, one should keep the tradeoffs in mind: imagine a system B taking 10 times longer per instance than a simpler system A as system B uses more sophisticated feature extractors or many more newly discovered concepts or deploys sophisticated inference processes. The lighter system A may still exhibit superiority since it can learn from ten times as many training instances in the same period. More sophistication, for example in improved feature extraction or inference, is definitely needed at some point to extend the reach of the prediction system, but the question is how to determine a good balance. For best results, a system may have to learn to deploy increasingly sophisticated processing only as needed. The field of perceptual leaning in cognitive science has investigated learning of new features for improved recognition as well as learning to speed up task accomplishment [12].

In general, the constraint of a finite memory will also be a great constraining factor, informing the design of systems and algorithms. The significant advantage of online algorithms here is that one need not store the instances: instances (learning episodes) are processed and translated into a feature vector, used for learning, and then discarded (onto the next instance). Most of the memory will be in terms of the categories stored and the connections among them. Thus considerations need to be made regarding for example the number of (nonzero) weights used in the system, and the memory consumption during learning those weights. Research on online computations and mining data streams share some of the objectives and challenges, and insights and algorithmic and analysis techniques developed there may be applicable [8, 3].

Therefore, our view is that the primary solution methods will be online. In order to achieve effective learning from a rich world, a number of online algorithms for various subtasks will have to be devised. These subtasks include:

- online feature selection and reduction
- feature or concept discovery and clustering
- speeding up of feature extraction and classification.

Speed ups in overall concept detection are important so that the system can move on to effectively learning more complex concepts.

3.2 Robustness to Imperfections, Uncertainty, Variety

There are numerous sources of imperfections and uncertainty or noise. In prediction in text, features include phrases or categories discovered, and their relations. But phrases are ambiguous and there can be misspellings, or missing values, or imperfect segmenters, or inaccurate passages. The newly discovered categories are also imperfectly recognized (poor precision or recall). In vision (and other perceptual domains), sensors are noisy and objects may occur in different lighting conditions, or may be partially occluded or be viewed from different angles and distances. There may be considerable variety within the same object class. Therefore, uncertainty and variability may be even more problematic. Prediction games are played for developing robustness in a variety of conditions, to detect/predict categories in many of their guises. Therefore effective solution algorithms will yield a system that is robust in numerous ways. The prediction system should learn from the sum of all its experience, rather than putting too much weight on any single instance. A learning strategy may include ignoring those instances for which the target category to be predicted, the observed outcome, is very uncertain. A similar strategy goes in using feature values. Igoring difficult episodes, at least temporarily, may be an effective partial strategy. Of course, this assumes basically that the system knows (at least for the most part) when it doesn't know: it has access to accurate confidence values. This is also related to a problem we may refer to as the *grounding problem*. The first time a system starts out, how could it be sure of any thing? We assume that the system is endowed with many sensors or feature extractors that work adequately, and they are sufficient to start the games.

3.3 Program Simplicity

In designing such learning systems, low code complexity and uniformity of architecture is an important guideline not to deviate from. Perhaps most of the program complexity and diversity may be concentrated into the preprocessing

and raw feature extraction components. The system will be learning a variety of things to improve overall prediction ability (in breadth, depth, and speed), but we hope that the number of distinct central algorithms utilized will be relatively few. Much of the functionality and diversity in the system should be the result of learning and experience, as opposed to explicit programming. "Keep it simple, as much as possible" should inform the design and search for algorithms.

3.4 Other Aspects

There are a number of other issues that can lead to fruitful research. For example, for the most part, we have assumed that the system will be passive in obtaining its learning episodes. However, increasing coverage (number of categories learned) motivates an active exploration of the world, at least after some period of purely passive observation. In the scenario of learning from the online text on the web, passive learning corresponds to crawling the web randomly, while active crawling can correspond to activities such as looking for specific pages and passages that contain certain phrases, perhaps using a search engine.

Another general area is defining the problems and discovering the algorithms for learning the control mechanisms necessarily for managing the complexity of the system. How are decisions regarding whether to deploy more costly feature extractors or inference algorithms made? How are these decisions streamlined over time? In general, what are the organizational principles and the organizing processes? Much is unclear regarding even the nature of the problems, but we think processes such as prediction games are promising candidates for raising these issues and may inspire useful problem formulations. Prediction games also provide the ample learning experience that appears necessary to achieve effective operational systems.

4. DISCUSSION

Prediction games liberate us from the labeling bottleneck, *i.e.*, explicit human supervision. The world serves as the teacher. Considerations of early learning, in infants and babies, is specially valuable. This stage provides the crucial foundation for learning and development in later years. Understanding how this foundation is developed from a computational point of view, *i.e.*, the nature of the major algorithms and organizing principles at play, is very important. Considering how infants and babies develop, in animals as well as in humans, one may conclude that:

1. There is massive learning taking place during first months and years, and
2. It does not involve (explicit) supervision.

We assume the above two statements are true. This massive learning includes recognizing myriad categories in various conditions (*e.g.*, physical objects, such as faces), learning the dynamics of the physical world, and becoming adept at physical movement. The infant, in the first few months of its birth, may have indeed mined its world very effectively!

There is much that remains unclear regarding the nature of this learning stage, *i.e.*, what problems are being solved and what kind of algorithms are at play. Research on processes similar to prediction games is a promising avenue for pointing to ways on how massive learning of different kinds may take place in the young brain. The outcome of this process may ultimately be a system that has developed a feel or a sense of a world that was once very unfamiliar.

The brain likely implements a variety of algorithms, but it has also been referred to as a prediction machine. Hawkins claims that making predictions is central to all of intelligence [11]. For example, he states that the nature of understanding or knowing may be explained by the ability to predict. He describes at a high level how predictions may be taking place by the brain's circuitry and how the ability may be aquired and developed [11].

Valiant proposes and explores a network or graph model of the neocortex with locally programmable elements [18]. He stresses the importance of paying attention to resource constraints, such as limited connectivity and processing speed, and shows how a number of learning algorithms could be implemented on his model. The field of bounded rationality [9], in studying human behavior, has also emphasized the roles of uncertainty, the issues of resource constraints, and the variety of implicit frequently competing objectives shaping observed human behavior. In playing prediction games, a system has to contend with significant resource constraints and uncertainty. We also expect that achieving some kind of optimization of the overall prediction objective is not the best place to put one's research efforts on. A more useful goal is in understanding what determines satisfactory operationality, *i.e.*, prediction capability. As we have mentioned, prediction ability involves satisfying the desiderata of speed/efficiency as well as accuracy in breadth and depth (how precise the predictions are). Research has found that considerations of category and feature utility can explain certain observed phenomena regarding speed of categorization and category naming in humans [6].

4.1 Relation to Other Learning Frameworks

Prediction games involve unsupervised learning, and in particular the outputs of the components may be interpreted as confidence values or probabilities for random variables (categories). In this sense, they are akin to density estimation or distribution learning. The use of graphical models has made density estimation tasks more efficient, due to the focusing on constraints on the type of relations that the variables may actually have (*e.g.*, the actual dependencies in case of Bayes networks). Graphical models may turn out to be too constraining for learning the myriad categories and the variety of relations between them, which apriori may not be anticipated or programmed. On the other hand, graphical models allow powerful types of inference. The benefits of sophisticated inference versus its computational costs is a subject under the theme of computation versus information: the tradeoff between extensive computation on the current situation, or learning instances accumulated so far, versus foraging for further information via new learning episodes.

Prediction games involve both supervised tasks and techniques, in the sense of learning connections between predictor features and existing categories, as well as unsupervised tasks and techniques, in the sense of forming new categories out of existing ones. The difference with most traditional clustering would be that the grouping or new category building should serve the objective of improving predictions. Another big difference, as raised in Section 3.1.1, is the requirement of scalability. To be most effective, online clustering algorithms need to be devised. Most existing clustering al-

gorithms are too costly and were designed with finite data in mind. Also, the successful clustering techniques may not be based on similarity between instances, but instead may rely on techniques such as observing co-occurrence between existing categories (inside the system, or as observed in the external world). The discovered categories may not necessarily be constrained to form a structure such as a tree or a dag. The major condition is that their benefits (improving prediction) outweigh the costs (memory requirements, time to recognize).

Supervised learning has enjoyed much success, but the issue of obtaining training data has always been a significant bottleneck. This has motivated much research on topics such as query learning, active learning, and semisupervised learning, in order to reduce the need for human/teacher involvement [1, 5]. Researches have even developed social games to motivate humans to label [2]. Playing prediction games involve learning myriad categories and their many relationships. The aim is to learn in the order of billions of variable values, and beyond. We think that the amount of training information required to make sophisticated systems makes classic explicit supervision infeasible.[4]

Prediction games complement reinforcement learning. Learning from prediction games can occur at a larger scale than reinforcement learning, as reinforcement learning requires taking action and attaining (possibly delayed) rewards. In the physical world, actions take time and energy. Prediction games involve mostly observations and information processing. However, from time to time, they also involve information seeking actions such as moving the camera. Acquiring prediction ability serves the goal of familiarization to ones world, and directly or indirectly is geared towards helping the intelligent agent predict and obtain rewards, and in general avoid bad situations and obtain pleasant ones (for example, for a baby, sensing that mom is about to feed her).

For predicting the next word, statistical language modeling uses techniques based on n-gram models [16]. For each possible length-limited history seen so far (sequence of words), these techniques keep counts for each of the different words that occur afterwards. They later use the counts to compute probabilities. The approach proposed here shifts the focus from history to the word to be predicted. It treats each word or phrase as a class to be learned. Statistics on candidate predicting features are kept for each target word rather than the history. In this respect, it is akin to the work of Even-Zohar and Roth [7], who showed that the online classification (discriminative) approach has flexibility in utilizing diverse and sophisticated features, but focused on discriminating between a relatively small number of classes.

Learning tasks such as time series prediction and learning to solve puzzles and to play games such as chess also share similarities, but differ mainly on the aspect of richness of the world and objectives.

5. CONCLUSION: LETS PLAY THE GAME!

In this paper we have been concerned with an abstraction of tasks and processes that would lead to massive learning, and proposed prediction games for that purpose. In the process, we raised a number of issues and potential research directions. At a broad level, a major message of this paper is stressing the importance of thinking about practical massive ongoing learning that can be achieved without supervision.

Prediction games are about having much to learn and plenty to learn from. Domains such as the online text, the web, and vision provide the richness that would enable playing prediction games in infinitely rich worlds.

[4]Note that we are making a distinction between supervised techniques, which will be utilized in playing prediction games, and explict supervision or training signal via a human teacher.

Acknowledgments

Many thanks to Michael Connor, Yann Le Cun, Dennis DeCoste, Mark Gluck, Wiley Greiner, Rosie Jones, Gregory Murphy, and Russ Poldrack for valuable discussions or pointers.

6. REFERENCES

[1] D. Angluin. Comp. learning theory: survey and selected bibl. In *Proc. 24th Annu. ACM Sympos. Theory Comput.*, pages 351–369, 1992.

[2] L. V. Anh and L. Dabbish. Labeling images with a computer game. In *ACM CHI 2004*, 2004.

[3] A. Borodin and R. El Yaniv. *Online Computation and Competitive Analysis*. Cambridge University Press, 1998.

[4] L. Bottou and Y. L. Cun. Large scale online learning. In *NIPS*, 2003.

[5] D. A. Cohn, L. Atlas, and R. E. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[6] J. E. Corter and M. A. Gluck. Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, 111(2):291–303, 1992.

[7] Y. Even-Zohar and D. Roth. A classification approach to word prediction. In *Annual meeting of the North Amercian Association of Computational Linguistics (NAACL)*, 2000.

[8] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy. Mining data streams: A review. *SIMOD Record*, 34(2), June 2005.

[9] G. Gigerenzer and R. Selten, editors. *Bounded Rationality*. The MIT Press, 2002.

[10] K. Grill-Spector and N. Kanwisher. Visual recognition, as soon as you know it is there, you know what it is. *Psychological Science*, 16(2):152–160, 2005.

[11] J. Hawkins. *On Intelligence*. Owl Books, 2004.

[12] P. J. Kellman. *Handbook of Experimental Psychology*, chapter 7: Perceptual Learning. NY, Wiley, 2002.

[13] O. Madani and W. Greiner. Learning when concepts abound. Technical report, Yahoo! Research, 2006.

[14] G. L. Murphy. *The Big Book of Concepts*. MIT Press, 2002.

[15] D. Ravichandran, P. Pantel, and E. Hovy. The terascale challenge. In *KDD Workshop on Mining for and from the Semantic Web*, 2004.

[16] R. Rosenfeld. Two decades of statistical language modeling: Where do we go from here? *IEEE*, 88(8), 2000.

[17] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

[18] L. G. Valiant. *Circuits of the Mind*. New York: Oxford University Press, 1994.