Complexity Dimensions and Learnability

S.H. Nienhuys-Cheng and M. Polman
Dept. of Computer Science
Erasmus University of Rotterdam
The Netherlands
e-mail: cheng@cs.few.eur.nl

Abstract

In machine learning theory, problem classes are distinguished because of differences in complexity. In [6], a stochastic model of learning from examples was introduced. This *PAC-learning* model (PAC = probably approximately correct) reflects differences in complexity of *concept classes*, i.e. very complex classes are not efficiently PAC-learnable. Blumer et al. [1] found, that efficient PAC-learnability depends on the size of the *Vapnik Chervonenkis dimension* ([7]) of a class. In Section 2 we will discuss this dimension and give an algorithm to compute it, in order to provide the reader with the intuitive idea behind it. In [3] a new, equivalent dimension is defined for *well-ordered* classes. These well-ordered classes happen to satisfy a general condition, that is sufficient for the possible construction of a number of equivalent dimensions. We will give this condition, as well as a generalized notion of an equivalent dimension. Also, a relatively efficient algorithm for the calculation of one such dimension for well-ordered classes is given.

1 Introduction

To avoid confusion about terms and results used in this paper, the most important ones are stated formally.

Let X be the domain of our interest: a set of finite strings over some finite alphabet Σ. X can be infinite. X_n is the set of all strings in X of length at most n. A *concept* f is a subset of X.

A number of concepts with distinct features can be grouped in a *class* of concepts F. The concept in F an algorithm is required to learn is called the *target* concept.

A PAC-algorithm for a class of concepts F learns a target concept $f \in F$ from positive and negative *examples* for it. An *example* for a concept f is a pair (x, y), where $x \in X$ and $y = 1$ if $x \in f$ and $y = 0$ if $x \notin f$. The value of y for any x is a function of x, determined by the concept in discussion and will therefore be denoted by $f(x)$.

Examples are chosen according to some unknown probability distribution P on X_n, where n is one of the input parameters of the algorithm. Eventually, the algorithm outputs a concept $g \in F$, such that with high probability, g is a good approximation of f. Since we are solely interested in the number of examples needed, neither the way in which the algorithm finds such a concept, nor the way in which it is presented in the end, are specified.

A formal definition of a PAC-algorithm is as follows:

- **Definition:** A learning algorithm A is a *PAC-algorithm* for a class of concepts F over X if

    1. A takes as input ε, $\delta > 0$ and $n \in \mathrm{N}$, where ε is the *error* parameter, δ is the *confidence* parameter and n is the *length* parameter.

2. A may call the procedure **example**, which returns examples for some concept $f \in F$, according to an arbitrary and unknown probability distribution P on X_n.
3. For all concepts $f \in F$ and for all probability distributions P on X_n, A outputs a concept $g \in F$, such that with probability at least $1 - \delta$, $P(f \triangle g) \leq \varepsilon$, where $f \triangle g$ is the symmetric difference between f and g.

We are interested in the number of examples an algorithm needs, to learn concepts from a class probably approximately correctly. The following complexity measure for learning-algorithms plays an important role.

- **Definition:** Let A be a learning algorithm for concept class F. The *sample complexity* of A is a function s with parameters ε, δ and n. It returns the maximum number of calls of **example** by A, for all runs of $A(\varepsilon, \delta, n)$, for all $f \in F$ and all P on X_n. s is infinite if no finite maximum exists.

- **Definition:** Class F is said to be *polynomial sample learnable* if there exists a learning algorithm for F, with a sample complexity that is bounded by some polynomial p in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$ and n.

In the following sections, the Vapnik Chervonenkis dimension [7] of a class is introduced, followed by a discussion of a number of aspects of it, including its importance to PAC-learning, as well as an algorithm for its calculation. After that, Natarajans dimension [3] is discussed, and we introduce a condition for a class, that is sufficient for the definition of a number of dimensions over this class, which are all *equivalent* to the Vapnik Chervonenkis dimension. An algorithm involving the efficient calculation of one such dimension, is given.

2 Shattering and the Vapnik Chervonenkis Dimension

An important notion in PAC-learning is *shattering.*

- **Definition:** A class of concepts F on X *shatters* a set $S \subseteq X$ if the set given by $\{f \cap S | f \in F\}$ is the power set of S (denoted by 2^S).

Shattering is used in the definition of a complexity measure for concept classes, the *Vapnik Chervonenkis dimension.*

- **Definition:** The *Vapnik Chervonenkis dimension* of a concept class F on X is the greatest integer d such that there exists a set $S \subseteq X$ of cardinality d that is shattered by F. It is denoted by $\mathbf{D_{vc}}(F)$. If no greatest d exists, $\mathbf{D_{vc}}(F)$ is infinite.

By the following discussion an attempt is made to provide the reader with an intuitive idea behind $\mathbf{D_{vc}}$.

If concept class F shatters S, then F is partitioned by S in the following way: two concepts $f, g \in F$ are equivalent iff $f \cap S = g \cap S$. The total number of equivalence classes is $2^{|S|}$. This gives us a surjective mapping from F to the power set of S, where a concept is mapped to its intersection with S. The more elements there are in S, the closer this mapping is to an injection.

For example if $S = \emptyset$ we have only one equivalence class. If $S = \{x\}$, then F is divided into two equivalence classes: the set of all concepts containing x, and the set of those not containing x. If $|S| = \mathbf{D_{vc}}(F)$, then we have the most refined mapping.

Notice, that if S is shattered by F, then F has to contain at least $2^{|S|}$ concepts, so $2^{|S|} \leq |F|$. If $|S| = \mathbf{D_{vc}}(F)$ then we can use $\mathbf{D_{vc}}(F)$ to find an upper bound of $|F|$ (see Lemma 2). Notice also, that if S is shattered by F, then every $S_1 \subseteq S$, is also shattered by F. Three questions now come to mind:

1. If F shatters S and $x \in \cup F - S$, is there an easy way to check whether or not $S \cup \{x\}$ is also shattered by F?
2. If there does not exist an $x \in \cup F - S$ such that $S \cup \{x\}$ is shattered by F (we say S is *maximal*) does that imply that $|S| = \mathbf{D_{vc}}(F)$?
3. Is there an algorithm, which finds $\mathbf{D_{vc}}(F)$?

These questions are answered by the following lemmas/examples:

- **Definition:** Let F shatter S. An element $x \in \cup F - S$ is said to be an *extending* element of S if for every $A \subseteq S$, there are $f, g \in F$ such that $x \notin f$, $x \in g$ and $f \cap S = g \cap S = A$.

- **Lemma 1:** Let F shatter S. There exists a set T, with $T \supset S$, $T \neq S$ and T shattered by F if and only if $\cup F - S$ contains an extending element for S.

Proof:
$\rightarrow$: Let there exist a set $T \supset S$, $T \neq S$, such that T is shattered by F. Let $T - S$ contain x and let A be a subset of S. Since T is shattered by F, there is an $f \in F$ such that $f \cap T = A$ and also $f \cap S = A$. Clearly, $x \notin f$. There is also a $g \in F$ with $g \cap T = A \cup \{x\}$. Notice that $g \cap S = A$ and that $x \in g$. It follows that x is an extending element of S.
$\leftarrow$: Let $x \in \cup F - S$ be an extending element for S. Let $T = S \cup \{x\}$. Let $A \subseteq T$. If $x \notin A$, then $A \subseteq S$, so there is an $f \in F$ such that $f \cap T = f \cap S = A$. If $x \in A$ let $B = A - \{x\}$. Since $B \subseteq S$, there is a $g \in F$ with $x \in g$ and $g \cap S = B$. But then, $g \cap T = A$.

From Lemma 1 it follows that if $\cup F$ is finite, then for any S shattered by F, there is a maximal S' such that $S \subseteq S'$ and F shatters S'.

From the following example it can be seen that two maximal sets are not always of the same cardinality. Let S and T be two nonempty sets such that $S \cap T = \emptyset$ and $|S| > |T|$. Also, let $F = \{h | h \subseteq S \vee h \subseteq T\}$. F shatters S because every subset of S is a concept in F. Similarly F shatters T. Let $x \in \cup F - S$. Then $x \in T$. Let $A \subseteq S$, A nonempty. Since F contains only subsets of T or S and $T \cap S = \emptyset$, no $f \in F$ exists, such that $(A \cup \{x\}) \subseteq f$. So x can never be an extending element of S. It follows that S is maximal and similarly is T. Hence the claim follows.

It is our interest to find an algorithm generating $\mathbf{D_{vc}}(F)$ for any concept class F with finite $\cup F$. Notice the following:

Let F shatter S and let x be an extending element for S. Suppose, that $f \in F$ and $g \in F$ are equivalent with respect to $S \cup \{x\}$. Then $f \cap S = g \cap S$ and hence f and g are also equivalent with respect to S. Using this an extending element x of S can be found in the following way: for every $A \subseteq S$ let $F_A \subseteq F$ be the equivalence class defined by A. Then x is an extending element of S iff for every A there are f, g in F_A such that $x \in f$ and $x \notin g$. So every F_A can be divided into two classes with respect to $S \cup \{x\}$: the set $\{f | f \in F_A \wedge x \in f\}$ and the set $\{f | f \in F_A \wedge x \notin f\}$.

The above can be used in the following algorithm:

- **Algorithm:**
  1. Let $d = 0$. Start with the empty set $\emptyset$; for all $f, g \in F$, f is equivalent with g with respect to $\emptyset$. With regard to $\emptyset$ the only equivalence class is F.
  2. Suppose $d = n$. Suppose also that we have constructed $S_1, ..., S_k$ where every S_i is shattered by F and contains n elements. Now, the above discussion can be applied to every S_i. For every $A \subseteq S_i$, the equivalence class constructed in

the previous iteration was $F_A = \{f \in F | f \cap S_i = A\}$. Use the above discussion and these F_A to find extending elements $x_1, ..., x_m$. The x_j can be found by checking whether every F_A can be divided into two nonempty classes with regard to $S_i \cup \{x_j\}$. These new classes can be used in the next iteration. If an extending element exists for some S_i, then let $d = n + 1$. Every $S_i \cup \{x_j\}$ is a shattered set of $n + 1$ elements.

3. Repeat step 2 until no extending elements can be found for any S_i. Then we have $\mathbf{D_{vc}}(F) = d$.

Of course, this algorithm can be improved, if we eliminate the possibility of the algorithm generating duplicate sets. To illustrate the above, an example follows, in which the natural numbers are taken as a domain. Let class F contain the following concepts:
$f_1 = \{0, 2, 3\}$, $f_2 = \{0, 3, 4\}$, $f_3 = \{1, 2, 3\}$, $f_4 = \{0, 1, 3, 4\}$, $f_5 = \{0, 1, 2, 3\}$, $f_6 = \{2, 3, 5\}$, $f_7 = \{1, 3, 4\}$, $f_8 = \{3, 4\}$. Observe, that $\cup F = \{0, 1, 2, 3, 4, 5\}$. $\mathbf{D_{vc}}(F)$ can be found in the following way:

1. Start with $\emptyset$ and class F.
2. $\emptyset$ can be extended to $\{0\}$ because F can be divided into two subclasses: $F_1 = \{f_1, f_2, f_4, f_5\}$ and $F_2 = \{f_3, f_6, f_7, f_8\}$.
   Hence $\{0\}$ is also shattered by F. Similar divisions can be made for $\{1\}, \{2\}, \{4\}, \{5\}$.
3. $\{0\}$ can be extended to $\{0, 1\}$ because:
   F_1 can be divided into the classes $F_3 = \{f_4, f_5\}$ and $F_4 = \{f_1, f_2\}$.
   F_2 can be divided into the classes $F_5 = \{f_3, f_7\}$ and $F_6 = \{f_6, f_8\}$.
   Similar divisions can be performed for $\{0, 2\}, \{0, 4\}, \{1, 2\}, \{1, 4\}$.
4. $\{0, 1\}$ can be extended to $\{0, 1, 2\}$ because F_3 can be divided into $\{f_5\}$ and $\{f_4\}$. F_4 can be divided into $\{f_1\}$ and $\{f_2\}$ etc. Hence, $\{0, 1, 2\}$ is shattered by F.
   No 4-element shattered set can be found, so $\mathbf{D_{vc}}(F) = 3$.

An intuitive notion of the importance of $\mathbf{D_{vc}}(F)$ in machine learning can be given by the following.

Let S be a set shattered by a concept class F and let $|S| = \mathbf{D_{vc}}(F)$. Also, let $x \in \cup F - S$. Then there is an $A \subseteq S$, such that for all $f \in F$, with $f \cap S = A$, $f(x)$ is the same. Otherwise x is an extending element for S.

So, for all $x \in \cup F - S$, the value of $f(x)$ for some $f \in F$ can be predicted to some extent, which would speed up the learning process.

Two new definitions and an important lemma, due to Vapnik and Chervonenkis [7] now follow.

- **Definition:** The *projection* f_n of a concept f on X_n is the set of all strings of X in f with length at most n.
  The *projection* F_n of a concept class F on X_n is the set given by $\{f_n | f \in F\}$.
- **Definition:** A concept class F is said to be of *polynomial* Vapnik Chervonenkis dimension if $\mathbf{D_{vc}}(F_n)$ is $O(p(n))$ for some polynomial p.
- **Lemma 2:** Let F be a class of concepts on some finite domain X. Then $2^{d_{vc}} \leq |F| \leq (|X| + 1)^{d_{vc}}$, where $d_{vc} = \mathbf{D_{vc}}(F)$.
- **Remark:** Notice that the projection F_n of a class F on X_n is also a concept class over the (finite) domain X_n.

Lemma 2 is used to prove an important result, which formalizes the relation between learnability and the Vapnik Chervonenkis dimension. The proof can be found in Natarajan [4] (see also [2]).

- **Theorem 1:** A class of concepts F is polynomial sample learnable *if and only if* F is of polynomial Vapnik Chervonenkis dimension.

As an example, take class F to be the class of monotone monomials. These are boolean functions consisting of the conjunction of positive boolean variables a_i, e.g. $a_1 \wedge a_3$. For clarity, we assume these functions to be preceded by a tautology in all variables, e.g. in the case of 3 variables this would be $a_1 \vee \neg a_1 \vee a_2 \vee \neg a_2 \vee a_3 \vee \neg a_3$. A concept in this class is a set of $(0,1)$-strings that all satisfy the same monotone monomial. The number of monotone monomials in exactly n variables for some n is bounded from above by 2^n. Therefore, $|F_n| \leq \sum_{i=0}^{n} 2^i$. So, by Lemma 2, $2^{d_{vc}} \leq \sum_{i=0}^{n} 2^i$, where $d_{vc} = \mathbf{D_{vc}}(F_n)$. It follows that $d_{vc} \leq n+1$ and thus F is polynomial sample learnable. Actually, F_n always shatters a set of n elements. For examples, in the case of F_4 this set could be $\{0111, 1011, 1101, 1110\}$. We conclude that $n \leq d_{vc} \leq n+1$.

For another complexity result, concerning this dimension, we refer to Subsection 7.3.

3 Alternative Dimension

In a variant of the PAC-learning model, concerned with learning boolean functions, it is required, that a learning algorithm always outputs a *subset* of the concept to learn, when fed with a number of *positive* examples. If for a concept class F such an algorithm exists, then F is called PAC-learnable with *omission-only error* from positive examples. In this setting, according to a result in [3], there are two requirements for an algorithm to be polynomial sample learnable, namely polynomial Vapnik Chervonenkis dimension (as before) and *well-ordered*ness. We will first introduce the notion of $graph(f)$ and *consistency*.

- **Definition:** Let F be a concept class. For any concept $f \in F$, $graph(f)$ is the set of all examples for f (positive *and* negative).
- **Definition:** A concept f is *consistent* with a set of examples S if $S \subseteq graph(f)$.
- **Definition:** A class F is *well-ordered* if, for any set of positive examples S for some concept f in F, there exists a concept $g \in F$ such that g is consistent with S and g is a subset of any concept in F consistent with S (we call g the *least concept consistent with* S).
- **Remark:** In the following discussion, positive examples will play an important role. Notice, that a set of positive examples for some $f \in F$ corresponds with a subset, say S, of f. Now, to avoid unnecessary complications in our discussion, we will also speak of the least concept consistent with S.

In this variant however, the condition of having a polynomial *Vapnik Chervonenkis dimension* can be replaced by having a polynomial *dimension*. This dimension was introduced by Natarajan [3], who argues that it is intuitively more appealing than the Vapnik Chervonenkis dimension. It is defined as follows:

- **Definition:** The *dimension* of a well-ordered class of concepts F, denoted by $dim(F)$, is the least integer d such that for every concept $f \in F$, there exists a set S_f of d or fewer elements such that f is the least concept in F consistent with S_f.

- **Remark:** To find $dim(F)$ for some F, we can use the following approach: for every $f \in F$, we consider the sets S of elements in f such that f is the least concept consistent with S. Any such S of minimal cardinality may be chosen as S_f. Now, let f range over the whole F. Then we have:

$$dim(F) = max\{|S_f| \, | f \in F\}$$

- **Remark:** From now on, it is implicitly assumed that any concept class we discuss contains as one of its concepts the empty concept $\emptyset$. This is the least concept consistent with an empty set of (positive) examples.

We will proceed with a number of propositions concerning properties of well-ordered classes. We use the symbol $\subset$ to correspond with a *proper* subset.

- **Proposition 1:** ([3]) A finite class of concepts F is well-ordered iff for any two concepts $f, g \in F$, there exists a concept $h \in F$ such that $h = f \cap g$.

For any set A of elements within some concept, let $M(A)$ denote the least concept consistent with A. Notice that if $|F|$ is finite, then $M(A) = \cap\{f | f \supseteq A\}$. For any two sets A and B:

- **Proposition 2:** ([3])

$$M(A \cup B) = M(M(A) \cup M(B))$$

- **Proposition 3:** Let F be a well-ordered class of concepts and let $f \in F$. f is the least concept consistent with a set of elements $S \subset f$ iff there is no $g \in F$ such that $S \subseteq g \subset f$.

- **Proposition 4:** Let f be the least concept consistent with a set of elements S. Let $S' \supset S$ be a set of elements within f. Then f is the least concept consistent with S'.

With Proposition 2, 3 and 4 the following theorem can be proved:

- **Theorem 2:** Let F be a well-ordered class over domain X. If $f \in F$ and S is any set such that

  - f is the least concept consistent with S
  - There is no $S' \subset S$ such that f is the least concept consistent with S'

  then S is shattered by F and has no extending elements within f. (We call S a *minimal set* of f).

Proof: f is not the least concept consistent with any proper subset of S. Suppose F does not shatter S. Then there exists a set $S' \subset S$ such that there is no concept $g \in F$ with $g \cap S = S'$. Let h be the least concept consistent with S'. Let $T = h \cap S$. Then $T \neq S'$: $T \cap (S - S')$ is nonempty and $|T| > |S'|$. Notice, that $S = (S - T) \cup S' \cup T$. By Proposition 2:

$$M(S) = M(M(S - T) \cup M(S') \cup M(T))$$

By Proposition 4: $M(S') = M(T)$. Therefore,

$$M(S) = M(M(S - T) \cup M(S'))$$

and, again by Proposition 2:

$$M(S) = M((S - T) \cup S')$$

So, $f = M(S) = M((S - T) \cup S')$, but $|S| > |(S - T) \cup S'|$, so f is the least concept consistent with a proper subset of S, which gives us a contradiction: $h \cap S = S'$. It follows that for every subset S' of S there is a concept $h_{S'}$ such that $h_{S'} \cap S = S'$: S is shattered by F.

S has no extending elements in f. Suppose it has an extending element in f, say x. Then there must be a concept g such that $g \supseteq S$, but $x \notin g$. But then g is consistent with S and yet, f, the least concept consistent with S, is not a subset of g. This is a contradiction, which completes the proof of Theorem 2.

From Theorem 2 it follows, that:

- **Lemma 3:** If F is a well-ordered class of concepts *over a finite domain* X, then $d \leq d_{vc} \leq d\,^2\log(|X|+1)$, where $d = dim(F)$ and $d_{vc} = \mathbf{D_{vc}}(F)$.

Proof: Since S_f, as defined in the definition of a well-ordered class, is a minimal set of f, it follows from Theorem 2 that S_f is shattered for each $f \in F$. Therefore, $|S_f| \leq d_{vc}$ for all f and thus, by the definition of d: $d \leq d_{vc}$. Now, every concept $f \in F$ is the least concept consistent with a set of d or fewer elements. Of course, no two different concepts are both the least consistent with the same set. Therefore, the number of concepts in F is always smaller then the number of sets of at most d elements. This number is bounded from above by $(|X|+1)^d$. Now we have $2^{d_{vc}} \leq |F| \leq (|X|+1)^d$ and thus $d_{vc} \leq d\,^2\log(|X|+1)$. So, $d \leq d_{vc} \leq d\,^2\log(|X|+1)$.

The following theorem follows almost immediately from Lemma 3. It is a result found by Natarajan [3], who gives an alternative proof. We have chosen to state and prove it again using Lemma 3 (and thus Theorem 2 implicitly), because this is useful to and illustrative to the results in the following sections.

- **Theorem 3:** A well-ordered class F is polynomial sample learnable iff $dim(F_n)$ is $O(p(n))$ for some polynomial p.

Proof: Of course if F over X is well-ordered, then so is F_n over X_n for each n. So, since X_n is finite, it is perfectly legitimate to read F_n for F and X_n for X in Lemma 3. Thus, we get $dim(F_n) \leq \mathbf{D_{vc}}(F_n) \leq dim(F_n)\,^2\log(|X_n|+1)$. Since $^2\log(|X_n|+1)$ grows only polynomially in n, it follows that $dim(F_n)$ is $O(p(n))$ for some polynomial p if and only if $\mathbf{D_{vc}}(F_n)$ is $O(q(n))$ for some polynomial q. From this, Theorem 3 follows immediately.

For an example, we return to the monotone monomials of Section 2. Let f be any such function in, say n, variables. Consider the string x having 1's for every variable that appears in f and 0's elsewhere. Then it is easy to see, that f is the least concept consistent with $\{x\}$. For example, if f is a function in 4 variables, given by $a_1 \wedge a_3$, then it is the least concept consistent with $\{1010\}$. It follows, that $dim(F_n) = 1$ for each n. Furthermore, it is easy to see, that F is well-ordered. It follows that F is polynomial sample learnable with omission-only error.

- **Remark:** Unfortunately, Theorem 2 cannot be reversed. This can be seen by the following example: let F consist of four concepts: $f_1 = \{a, b, c\}$, $f_2 = \{a, b\}$, $f_3 = \{b, c\}$, $f_4 = \{b\}$, $f_5 = \{d\}$ and $f_6 = \emptyset$. Notice, that F is well-ordered. We can see that $S_{f_1} = \{a, c\}$. However, the set $\{b\}$ is a maximal shattered set within f_1 (i.e it has no extending elements within f).

4 Equivalent Dimensions

We have seen that for any class F over domain X to be efficiently learnable, it has to be of polynomial Vapnik Chervonenkis dimension. If F is well-ordered, this requirement can be

replaced by polynomial dimension; a dimension notion, which is equivalent to $\mathbf{D_{vc}}$. In this section, we will, to some extent, generalize this equivalence. That is, we will give a more general property of concept classes than that of well-orderedness; for classes that have this property, a number of alternative dimensions can be constructed, which are all equivalent to $\mathbf{D_{vc}}$. It appears that *dim* is an example of such a dimension. We hope that this more general property leads us to the definition of a dimension that is computable in a more efficient way than $\mathbf{D_{vc}}$ by an intuitively appealing algorithm.

Consider this: let F be a concept class over finite domain X, such that there exists a function μ, defined as follows:

- $\mu : F \rightarrow 2^X$
- μ is injective
- $|\mu(f)| \leq \mathbf{D_{vc}}(F)$ for every $f \in F$

Then we can associate with μ a number $\mathbf{D}_\mu(F) = max\{|\mu(f)| \mid f \in F\}$. It can be proved, that

- **Theorem 4:** $d_\mu \leq d_{vc} \leq d_\mu \, {}^2\log(|X|+1)$, where $d_\mu = \mathbf{D}_\mu(F)$ and $d_{vc} = \mathbf{D_{vc}}(F)$.

Proof: It is easy to see that $d_\mu \leq d_{vc}$. Furthermore, since μ is injective, $|F|$ equals the number of $\mu(f)$'s. This number is bounded from above by $(|X|+1)^{d_\mu}$. So, $2^{d_{vc}} \leq |F| \leq (|X|+1)^{d_\mu}$ and thus $d_{vc} \leq d_\mu \, {}^2\log(|X|+1)$.

The essence of the above is this: suppose F is such, that for each concept $f \in F$, there is a set of elements of cardinality less than or equal to $\mathbf{D_{vc}}(F)$, that is somehow uniquely related to f. Then we can define some function μ, which gives one such set for every f. If dimension $\mathbf{D}_\mu$ is then defined as the cardinality of the largest set, then it has the properties of Theorem 4.

Now suppose that F is such, that there exists an injective function μ, which generates (for every $n \in$ N) for every $f_n \in F_n$ a set of elements from X_n smaller than $\mathbf{D_{vc}}(F_n)$. Then our result would change to:

$$\mathbf{D}_\mu(F_n) \leq \mathbf{D_{vc}}(F_n) \leq \mathbf{D}_\mu(F_n) \, {}^2\log(|X_n|+1)$$

Since ${}^2\log(|X_n|+1)$ grows only polynomially in n, it follows that $\mathbf{D_{vc}}(F_n)$ is $O(p(n))$ for some polynomial p if and only if $\mathbf{D}_\mu(F_n)$ is $O(q(n))$ for some polynomial q: $\mathbf{D}_\mu$ is *equivalent* to $\mathbf{D_{vc}}$.

So, the general property we were looking for turns out to be the existence of a function μ as specified above. With any such μ we can associate a new dimension, equivalent to the Vapnik Chervonenkis dimension.

Clearly, well-ordered classes are an example of classes having this general property. The existence of a 'μ' for such classes, has namely been proved in the previous section: μ could be such that $\mu(f_n) = S_{f_n}$. In this case $\mathbf{D}_\mu(F_n)$ would be $dim(F_n)$, which is, indeed, equivalent to $\mathbf{D_{vc}}(F_n)$. In the next section we will present another 'μ' for well-ordered classes by an efficient algorithm.

- **Remark:** As we know, if F shatters a set S, we have $|F| \geq 2^{|S|}$. Therefore, $|F| \geq 2^{d_{vc}}$, where $d_{vc} = \mathbf{D_{vc}}(F)$. In this section we have proved, that $|F| \leq (|X|+1)^{d_\mu}$, where d_μ is as in Theorem 4. Now, $d_\mu \leq d_{vc}$. So, $|F| \leq (|X|+1)^{d_{vc}}$. This gives us another proof of Lemma 2 [7] for classes for which equivalent dimensions can be constructed. We can combine all these results in the following way:

$$2^{d_\mu} \leq 2^{d_{vc}} \leq |F| \leq (|X|+1)^{d_\mu} \leq (|X|+1)^{d_{vc}}$$

5 An Algorithm For An Alternative Equivalent Dimension

In this section, we will construct an algorithm to find, for any concept f in a well-ordered class F over domain X, a set of elements R_f in X, such that f is the least concept consistent with R_f. Any subset R_f^- of R_f, that is a minimal set (as defined in Theorem 2) of f, is shattered by F (and has no extending elements within f). Furthermore, if we define a new dimension as the cardinality of the greatest R_f^- for all f, then this dimension is equivalent to the Vapnik Chervonenkis dimension (just as $dim(F)$). We need the following definitions, in which the symbol $\subset$ is again used to denote a proper subset.

- **Definition:** $\emptyset$ is said to have 0 layers. Let $f \in F$. f is said to have k *layers* if every $g \subset f$ has less than k layers and there is at least one $g \subset f$ that has $k-1$ layers.
- **Definition:** For every $f \in F$, a *representation set* R_f is defined as follows:
  1. If f is $\emptyset$, then $R_f = \emptyset$.
  2. Suppose R_g is defined for every concept g with less than k layers. Consider the set H, being $\{h|\ \nexists g, f \supset g \supset h\}$. Let $H = \{h_1, ..., h_n\}$. If $f \neq \cup h_i$, then pick any $a \in f - \cup h_i$ and let $R_f = \{a\}$. If $f = \cup h_i$, then define $R_f = \cup R_{h_i}$.

For an example, see Figure 1: In this figure, the concepts are represented by ellipses; the elements of the concepts by the numbers inside these ellipses. The most outer concept has 4 layers, and a representation set for it is $\{9\}$. A representation set for the concept $\{1,2,3,4\}$, is $\{2,3,4\}$. For the concept $\{4\}$ we have $\{4\}$ as a representation set, and for $\{4,7,8\}$ we have $\{7\}$.

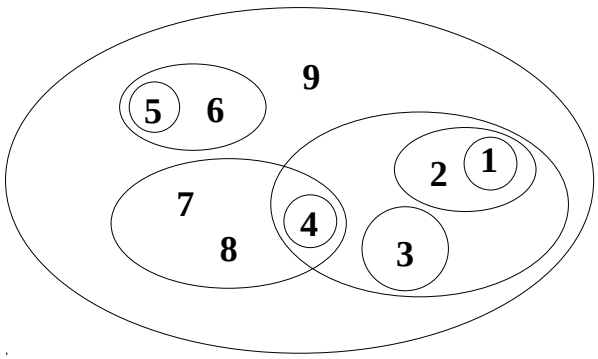

Figure 1

- **Proposition 5:** For any $f \in F$, $\nexists g \in F$, with $R_f \subseteq g \subset f$. So f is the least concept consistent with R_f.

Proof (by induction over the number of layers): The proposition is trivial for concepts with 0 or 1 layer. Suppose that the proposition is valid for every g with k layers. Now, let f have $k+1$ layers. Then we can distinguish two situations:

1. If $a \in f - \cup h_i$ and $R_f = \{a\}$, then there is no $g \subset f$ containing a, which is trivial.
2. If $f = \cup h_i$, then
   $M(R_f) = M(\cup R_{h_i})$
   $= M(\cup M(R_{h_i}))$ (by Proposition 2 of section 3)
   $= M(\cup h_i)$ (by the induction proposition)
   $= M(f) = f$

From Theorem 2 of Section 3, it follows that any minimal set $R_f^- \subseteq R_f$ is shattered by F. Proposition 5 guarantees that such an R_f^- exists. Now, if we choose for every $f \in F$ an R_f and if $\mu(f) \subseteq R_f$ is any minimal set of f, then we have the following property for the corresponding dimension:

- **Theorem 5:** Let F be well-ordered, and let $\mathbf{D}(F_n) = max\{|R_{f_n}^-| \mid f_n \in F_n\}$ for each n. Then $\mathbf{D}$ is a dimension equivalent to $\mathbf{D_{vc}}$.

Proof: The results of this section are still valid if we limit the discussion to F_n over domain X_n. Then it is easy to see that this new $\mathbf{D}$ is constructed using a function μ, that gives an $R_{f_n}^-$ for each $f_n \in F_n$. So, by arguments similar to those used in section 4, it is equivalent to $\mathbf{D_{vc}}$.

All of the above in this section can be used immediately in an algorithm to find representation sets of all $f \in F$: we start with the concepts in F with 1 layer and construct their representative sets. Then we proceed with the concepts of 2 layers, then 3 layers, etc. until every $f \in F$ has a representative set. The efficiency lies in the fact that every R_f is built up from at most $|H|$ (as defined in the definition of R_f) sets of elements, which are already known by the time R_f is being calculated. Furthermore, the total amount of elements involved in the calculation never exceeds the number of concepts in F. The next thing to be done is to find a set $R_f^- \subseteq R_f$, that is a minimal set of f. The largest such R_f^- over the whole F gives us our dimension.

- **Remark:** In Theorem 5, it does not matter, in what way the $R_{f_n}^-$ are chosen. Suppose, that we take $R_{f_n}^-$ to be a minimal set of f with as few as possible elements. Then it can be proved that the corresponding dimension is equal to dim, i.e. R_f^- is a valid S_f.

Proof: In this proof we will use the following definition:

- **Definition:** Let g be the least concept consistent with $\{x\}$ as well as with $\{y\}$, where $x \neq y$. Then x and y are called *peers*.

For example, in Figure 1, the elements 7 and 8 are peers. Now, let S_f be a set of minimal cardinality, such that f is the least concept consistent with S_f. Let R_f be chosen by the above algorithm. Furthermore, let x be an element of S_f that is not in R_f. We will show that we can allways replace elements in S_f that are not in R_f by elements of R_f. The resulting set will contain the same number of elements as the original S_f, and f will still be the least concept consistent with it. There are two reasons why an element x of S_f might not be contained in R_f.

1. Let $g \subset f$ be the least concept consistent with $\{x\}$. Suppose x has a peer, say y_1 (see Figure 2). It follows that any concept containing x has to contain y_1 as well and vice versa. Suppose $\{y_1\}$ is (in a previous iteration of the algorithm) chosen as R_g. Then x will not appear in R_f. Now, let $S'_f = (S_f - \{x\}) \cup \{y_1\}$. Let h be the least concept consistent with S'_f. Now, since h contains y_1 it has to contain x as well. It follows that h contains S_f. Therefore, $h \supseteq f$. We also have, since f contains S'_f, that $f \supseteq h$. It follows that $f = h$: f is the least concept consistent with S'_f.

2. Let $g \subset f$ be the least concept consistent with $\{x\}$. Suppose $\{x\}$ is, in contrast to the above, indeed chosen as R_g. Let concept h be such that $g \subset h \subseteq f$. (Notice that $x \in h$). Suppose $\cup R_{h_i}$, as defined in the definition of a representative set, is not equal to h. In this case, x will not be an element of R_h anymore (and will not be 're-chosen' in following iterations). Instead, some $y_2 \in h - \cup h_i$ will be chosen (see also Figure 2). Notice, that h is the least concept consistent with $\{y_2\}$. Now, we can see, that any concept containing y_2, has to contain x as well: suppose there is a concept h' such that $y_2 \in h'$ but $x \notin h'$. Then the concept $h' \cap h$ would be a proper subset of h containing y_2, which is impossible,

since h is the least concept consistent with y_2. Let $S'_f = (S_f - \{x\}) \cup \{y_2\}$. Then the least concept consistent with S'_f (which of course, contains y_2), has to contain x as well: it contains the whole of S_f. Therefore, by the same arguments as those used in the first reason, it must be f itself.

So, if an element x that is in an S_f but not in R_f and if x was never chosen in any R_g for some $g \subseteq f$, then x can always be replaced by a chosen peer y_1. The resulting set is a valid S_f as well. Now, elements in an S_f that were once chosen in an R_g for some $g \subset f$ but do not appear in R_f can always be replaced by an element in a proper superset of g. The resulting set is again a valid S_f. Since there is only a finite number of layers between a $g \subset f$ and f itself, sooner or later, we will arrive at an S_f consisting of elements that are all in R_f. If we choose this set as our R_f^-, then we have the desired result. This completes the proof of the above remark.

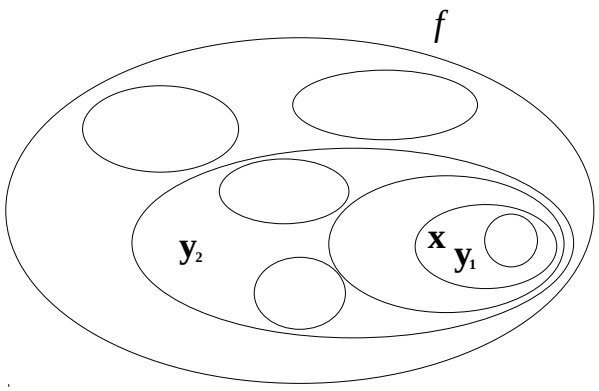

Figure 2

As an example of the above, consider Figure 3, where the representative set of the most outer concept is $\{1, 2, 3, 4\}$. A minimal subset of this set is $\{1, 2, 3\}$. We could also choose $\{1, 4\}$ as a minimal subset. This is one of the minimal sets of minimal cardinality.

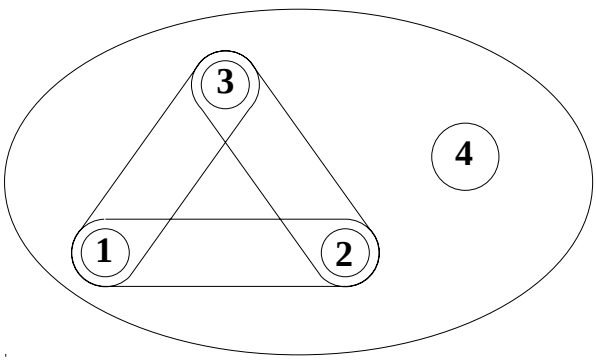

Figure 3

6 Additional Remarks

6.1 Positive and Negative Examples

In the variant of the PAC-learning model from Section 3, it was stated that a learning algorithm should output a subset of the concept to learn from *positive* examples. Now, consider the following definition:

- **Definition:** A class of concepts F is called *minimally consistent* if for every $f \in F$ and for every subset S of $graph(f)$, there is a concept $g \in F$, such that g is consistent with S and g is a subset of every concept consistent with S.

Natarajan proves (see [4]), that a class is learnable with omission only error from positive as well as negative examples iff it is of polynomial Vapnik Chervonenkis dimension *and* minimally consistent.

Now, it is easy to see that any minimally consistent class over domain X is also well-ordered. The reverse is also true: let F be well-ordered, and let S be a set of examples of some $f \in F$, being the union of S^+ (the positive examples of S) and S^- (the negative ones). If S^- is empty, the result is trivial. If S^+ is empty, then $\emptyset$ is the least concept consistent with S. Suppose, that both sets are nonempty. Let g be the least concept consistent with S^+. Suppose that g is not consistent with S. Then S^- contains an example $(x, 0)$ such that $x \in g$. However $x \notin f$, so $g \not\subseteq f$, but f is consistent with S^+. But then g is not the least concept consistent with S^+, which gives us a contradiction. So g is consistent with S. Furthermore, any $h \in F$ consistent with S is consistent with S^+. Therefore, g is also the least concept consistent with S. It follows that a class is well-ordered iff it is minimally consistent.

The consequence of all this, is that if an algorithm learns a class with omission-only error from a polynomial number of examples, the negative examples are of no importance to this algorithm. The possibility of feeding a learning algorithm with negative examples does not contribute to the learnability (with omission-only error) of the class.

6.2 Negative Well-Ordered Classes

We can construct a theory, which is in some way the dual of the above material. Herein we define a class F to be *negative well-ordered* if for every set of negative examples S for some $f \in F$ there is a $g \in F$ such that g is a *superset* of every concept in F consistent with S. An PAC-algorithm A is said to learn F with *inclusion only error* if it always outputs a superset of the target concept. It can be proved that a class F is PAC-learnable with *inclusion only error* if and only if it is negative well-ordered.

6.3 Uniform Learnability

In a more general definition of a PAC-algorithm, the parameter n is not included; the probability distributions according to which examples are chosen range over the entire example set. In this setting a class is called *uniformly learnable* [1] if, globally, the number of examples needed to PAC-learn concepts in this class, is bounded from above by an integer-valued function of ε and δ only. It can be proved [1] that a class F is uniformly learnable if and only if its Vapnik Chervonenkis dimension is finite. So, if a class has an infinite Vapnik Chervonenkis dimension (which goes for a lot of classes), it is not uniformly learnable. Therefore, uniform learnability is replaced by the less strong requirement of polynomial sample learnability; the sample size is allowed to grow polynomially in the maximum length of the input strings (as in the definition used in this article). In the Sections 3, 4, and 5 results were presented concerning the equivalence of some alternative dimensions and the Vapnik Chervonenkis dimension with

respect to polynomial sample learnability. From the results in these sections, we can see, that if we can define for some class F over a finite domain X a dimension $\mathbf{D}$ as in Section 4, then $\mathbf{D_{vc}}(F)$ is infinite if $\mathbf{D}(F)$ is infinite.

7 Conclusion

In this article we have made an attempt to give the reader an intuitive idea of the Vapnik Chervonenkis dimension by discussing a number of its properties and an algorithm for its computation. By abstracting Natarajan's dimension, we can define new dimensions with respect to functions over concept classes, that are equivalent to the Vapnik Chervonenkis dimension. After proving that a minimal set for some concept in a well-ordered class is shattered, we can indeed compute such a dimension by a relatively efficient algorithm, using a representation set R_f for every concept f and a minimal subset $R_f^- \subseteq R_f$ of f.

Acknowledgements We thank Patrick van der Laag for reading and commenting on this article.

References

[1] Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M. (1989), *Learnability and the Vapnik-Chervonenkis Dimension*, Journal of the Association for Computing Machinery, Vol.36, No.4, 929-965.

[2] Ehrenfeucht, A., Haussler, D., Kearns, M., Valiant, L. (1989), *A General Lower Bound on the Number of Examples Needed for Learning*, Information and Computation, Vol.82, 247-261.

[3] Natarajan, B. (1987), *On Learning Boolean Functions*, Proceedings of the 19th Annual ACM Symposium on Theory of Computation, 269-304.

[4] Natarajan, B. (1991), *Machine Learning, a Theoretical Approach*, Morgan Kaufman Publishers, Inc.

[5] Polman, M., Nienhuys-Cheng, S.H. (1992), *Some Topics Related to PAC-Learning*, (To appear in) Proceedings of CSN'92, Utrecht, Nehterlands, Nov. 1992.

[6] Valiant, L.G. (1984), *A theory of the learnable*, Communications of the ACM, Vol.27 , No.11, 1134-1142.

[7] Vapnik, V., and Chervonenkis, A. (1971), *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability and its Applications, Vol.16, No.2, 264-280.