

Some observations about covering arrays and VC-dimension

Robert Bailey
University of Regina

Joint work with Karen Meagher, Pavel Semukhin and Sandra Zilles

CMS Winter Meeting
Toronto
12 December 2011

Covering arrays

- ▶ Let N, t, k, v, λ be positive integers.

Covering arrays

- ▶ Let N, t, k, v, λ be positive integers.
- ▶ A *covering array* $CA_\lambda(N; k, s, t)$ is an $N \times k$ array A , where:
 - ▶ the entries are from an alphabet of size s ;
 - ▶ in every set of t columns, each t -tuple occurs in at least λ rows.

Covering arrays

- ▶ Let N, t, k, v, λ be positive integers.
- ▶ A *covering array* $CA_\lambda(N; k, s, t)$ is an $N \times k$ array A , where:
 - ▶ the entries are from an alphabet of size s ;
 - ▶ in every set of t columns, each t -tuple occurs in at least λ rows.
- ▶ The parameter t is called the *strength* of the array.

Covering arrays

- ▶ Let N, t, k, v, λ be positive integers.
- ▶ A *covering array* $CA_\lambda(N; k, s, t)$ is an $N \times k$ array A , where:
 - ▶ the entries are from an alphabet of size s ;
 - ▶ in every set of t columns, each t -tuple occurs in at least λ rows.
- ▶ The parameter t is called the *strength* of the array.
- ▶ In this talk, we are only concerned with the case $\lambda = 1$ (so we omit the subscript λ), and where the alphabet has size 2.

Covering arrays: an example

- ▶ A covering array $CA(5; 4, 2, 2)$:

0	0	0	0
1	1	1	0
1	1	0	1
1	0	1	1
0	1	1	1

Covering arrays: an example

- ▶ A covering array $CA(5; 4, 2, 2)$:

0	0	0	0
1	1	1	0
1	1	0	1
1	0	1	1
0	1	1	1

- ▶ In each pair of columns, each of the 2^2 possible combinations 00, 01, 10, 11 appears in at least one row.

A hypothetical situation

- ▶ Suppose someone asked you for a (binary) covering array of strength t , which you willingly provided.

A hypothetical situation

- ▶ Suppose someone asked you for a (binary) covering array of strength t , which you willingly provided.
- ▶ Unfortunately, after they've finished their software-testing exercise (or whatever), they tell you they didn't use your array after all!

A hypothetical situation

- ▶ Suppose someone asked you for a (binary) covering array of strength t , which you willingly provided.
- ▶ Unfortunately, after they've finished their software-testing exercise (or whatever), they tell you they didn't use your array after all!
- ▶ They then ask you “how strong was what we actually used”?

Covering dimension

- ▶ Let A be an $N \times k$ (binary) array, with (wlog) no repeated rows.

Covering dimension

- ▶ Let A be an $N \times k$ (binary) array, with (wlog) no repeated rows.
- ▶ **Definition:** The *covering dimension* of A , denoted $\text{CD}(A)$, is the largest integer t such that A is a covering array of strength t (with $\lambda = 1$).

Covering dimension

- ▶ Let A be an $N \times k$ (binary) array, with (wlog) no repeated rows.
- ▶ **Definition:** The *covering dimension* of A , denoted $\text{CD}(A)$, is the largest integer t such that A is a covering array of strength t (with $\lambda = 1$).
- ▶ In other words, $\text{CD}(A)$ is the largest t such that, for all t -subsets of columns, each of the 2^t binary t -tuples appears at least once in a row.

Vapnik–Chervonenkis dimension

- ▶ Again, let A be an $N \times k$ binary array, with no repeated rows.

Vapnik–Chervonenkis dimension

- ▶ Again, let A be an $N \times k$ binary array, with no repeated rows.
- ▶ **Definition:** The *Vapnik–Chervonenkis dimension* (or *VC-dimension*) of A , denoted $\text{VCD}(A)$, is the largest t such that *there exists* a t -subset of columns where each of the 2^t binary t -tuples appears at least once in a row.

Vapnik–Chervonenkis dimension

- ▶ Again, let A be an $N \times k$ binary array, with no repeated rows.
- ▶ **Definition:** The *Vapnik–Chervonenkis dimension* (or *VC-dimension*) of A , denoted $\text{VCD}(A)$, is the largest t such that *there exists* a t -subset of columns where each of the 2^t binary t -tuples appears at least once in a row.
- ▶ This definition is very similar to covering dimension, except that \forall has been replaced by \exists .

Vapnik–Chervonenkis dimension

- ▶ Again, let A be an $N \times k$ binary array, with no repeated rows.
- ▶ **Definition:** The *Vapnik–Chervonenkis dimension* (or *VC-dimension*) of A , denoted $VCD(A)$, is the largest t such that *there exists* a t -subset of columns where each of the 2^t binary t -tuples appears at least once in a row.
- ▶ This definition is very similar to covering dimension, except that \forall has been replaced by \exists .
- ▶ Frequently, this is defined in terms a of set system rather than a binary array (i.e. its incidence matrix). In that context, the t -subset is said to be *shattered*.

Vapnik–Chervonenkis dimension

- ▶ Again, let A be an $N \times k$ binary array, with no repeated rows.
- ▶ **Definition:** The *Vapnik–Chervonenkis dimension* (or *VC-dimension*) of A , denoted $VCD(A)$, is the largest t such that *there exists* a t -subset of columns where each of the 2^t binary t -tuples appears at least once in a row.
- ▶ This definition is very similar to covering dimension, except that \forall has been replaced by \exists .
- ▶ Frequently, this is defined in terms a of set system rather than a binary array (i.e. its incidence matrix). In that context, the t -subset is said to be *shattered*.
- ▶ VC-dimension is widely used, particularly in computational learning theory.

Some simple observations

- ▶ It is clear from the definitions that $CD(A) \leq VCD(A)$, for any binary array A .

Some simple observations

- ▶ It is clear from the definitions that $CD(A) \leq VCD(A)$, for any binary array A .
- ▶ For both parameters, we have the same lower bound on N :

Some simple observations

- ▶ It is clear from the definitions that $CD(A) \leq VCD(A)$, for any binary array A .
- ▶ For both parameters, we have the same lower bound on N :
 - ▶ if $CD(A) = t$, then $N \geq 2^t$;

Some simple observations

- ▶ It is clear from the definitions that $\text{CD}(A) \leq \text{VCD}(A)$, for any binary array A .
- ▶ For both parameters, we have the same lower bound on N :
 - ▶ if $\text{CD}(A) = t$, then $N \geq 2^t$;
 - ▶ if $\text{VCD}(A) = t$, then $N \geq 2^t$.

Some simple observations

- ▶ It is clear from the definitions that $\text{CD}(A) \leq \text{VCD}(A)$, for any binary array A .
- ▶ For both parameters, we have the same lower bound on N :
 - ▶ if $\text{CD}(A) = t$, then $N \geq 2^t$;
 - ▶ if $\text{VCD}(A) = t$, then $N \geq 2^t$.
- ▶ For example, if A is an orthogonal array of strength t and index 1, then $N = 2^t$, and so $\text{VCD}(A) = t$ also.

Some simple observations, II

- ▶ In general, the two parameters need not coincide.

Some simple observations, II

- ▶ In general, the two parameters need not coincide.
- ▶ For example, the following is an 8×4 array, with $CD = 2$ and $VCD = 3$:

0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	0

A conflict

- ▶ There is a fundamental conflict between the two parameters!

A conflict

- ▶ There is a fundamental conflict between the two parameters!
- ▶ In applications of covering arrays, for a given strength t , generally one wants to *minimize* the size of the array.

A conflict

- ▶ There is a fundamental conflict between the two parameters!
- ▶ In applications of covering arrays, for a given strength t , generally one wants to *minimize* the size of the array.
- ▶ In learning theory, for a given VC-dimension t , one wants to *maximize* the size of the array.

Sauer's Lemma

- ▶ A classic result of Sauer is as follows:

Lemma: If an $N \times k$ binary array (with no repeated rows) has VC-dimension t , then

$$N \leq \sum_{i=0}^t \binom{k}{i}.$$

Sauer's Lemma

- ▶ A classic result of Sauer is as follows:

Lemma: If an $N \times k$ binary array (with no repeated rows) has VC-dimension t , then

$$N \leq \sum_{i=0}^t \binom{k}{i}.$$

- ▶ An array which meets this bound is said to be *maximum*.

Sauer's Lemma

- ▶ A classic result of Sauer is as follows:

Lemma: If an $N \times k$ binary array (with no repeated rows) has VC-dimension t , then

$$N \leq \sum_{i=0}^t \binom{k}{i}.$$

- ▶ An array which meets this bound is said to be *maximum*.
- ▶ **Fact:** If A is maximum with $\text{VCD}(A) = t$, then $\text{CD}(A) = t$ also.

Sauer's Lemma

- ▶ A classic result of Sauer is as follows:

Lemma: If an $N \times k$ binary array (with no repeated rows) has VC-dimension t , then

$$N \leq \sum_{i=0}^t \binom{k}{i}.$$

- ▶ An array which meets this bound is said to be *maximum*.
- ▶ **Fact:** If A is maximum with $VCD(A) = t$, then $CD(A) = t$ also.
- ▶ (But as a covering array it's enormous....)

Pushing apart

- ▶ One can concoct examples where CD and VCD are arbitrarily far apart.

Pushing apart

- ▶ One can concoct examples where CD and VCD are arbitrarily far apart.
- ▶ For all t , $\exists A$ with $CD = 0$ and $VCD = t$. (Append an all-zero column to any array with $VCD = t$.)

Pushing apart

- ▶ One can concoct examples where CD and VCD are arbitrarily far apart.
- ▶ For all t , $\exists A$ with $CD = 0$ and $VCD = t$. (Append an all-zero column to any array with $VCD = t$.)
- ▶ For all t , $\exists A$ with $CD = 1$ and $VCD = t$. (As above, but include a single 1 in the extra column.)

Pushing apart

- ▶ One can concoct examples where CD and VCD are arbitrarily far apart.
- ▶ For all t , $\exists A$ with $CD = 0$ and $VCD = t$. (Append an all-zero column to any array with $VCD = t$.)
- ▶ For all t , $\exists A$ with $CD = 1$ and $VCD = t$. (As above, but include a single 1 in the extra column.)
- ▶ For all t , $\exists A$ with $CD = 2$ and $VCD = t$. (This actually requires some effort.)

Pushing apart, II

- ▶ Let $N = 2^t$ and $k = \binom{2^t - 1}{2^{t-1}}$. Make an $N \times k$ binary array A as follows:

Pushing apart, II

- ▶ Let $N = 2^t$ and $k = \binom{2^t - 1}{2^{t-1}}$. Make an $N \times k$ binary array A as follows:
- ▶ Top row: all 1s.

Pushing apart, II

- ▶ Let $N = 2^t$ and $k = \binom{2^t - 1}{2^{t-1}}$. Make an $N \times k$ binary array A as follows:
 - ▶ Top row: all 1s.
 - ▶ Fill the columns with all ways to have 2^{t-1} 1s, rest 0s.

Pushing apart, II

- ▶ Let $N = 2^t$ and $k = \binom{2^t - 1}{2^{t-1}}$. Make an $N \times k$ binary array A as follows:
 - ▶ Top row: all 1s.
 - ▶ Fill the columns with all ways to have 2^{t-1} 1s, rest 0s.
 - ▶ Then $\text{VCD}(A) = t$, but $\text{CD}(A) = 2$.

Pushing apart: Example

An example where $t = 3$, $N = 8$, $k = 35$:

$$\begin{array}{cccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots & 1 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & & 1 \end{array}$$

Strong VC dimension

- ▶ Also originating in learning theory is another parameter, the *strong VC dimension*.

Strong VC dimension

- ▶ Also originating in learning theory is another parameter, the *strong VC dimension*.
- ▶ For an $N \times k$ array A , this is the largest t such that, in any set of t columns, each of the 2^t binary t -tuples appears at least once in a row, and in those 2^t rows, the remaining $k - t$ entries are the same each time.

Strong VC dimension

- ▶ Also originating in learning theory is another parameter, the *strong VC dimension*.
- ▶ For an $N \times k$ array A , this is the largest t such that, in any set of t columns, each of the 2^t binary t -tuples appears at least once in a row, and in those 2^t rows, the remaining $k - t$ entries are the same each time.
- ▶ Clearly, $\text{SVCD}(A) \leq \text{CD}(A) \leq \text{VCD}(A)$.

Example

The following array has $SVCD = 2$:

0	0	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	1	0	0
1	0	1	0
1	0	0	1
0	1	1	0
0	1	0	1
0	0	1	1

Example

The following array has $SVCD = 2$:

0	0	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	1	0	0
1	0	1	0
1	0	0	1
0	1	1	0
0	1	0	1
0	0	1	1

Example

The following array has $SVCD = 2$:

0	0	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	1	0	0
1	0	1	0
1	0	0	1
0	1	1	0
0	1	0	1
0	0	1	1

Strong VC dimension, II

- ▶ Having $SVCD = t$ is a very strong condition, so it is not surprising that such arrays must be large.

Strong VC dimension, II

- ▶ Having $\text{SVCD} = t$ is a very strong condition, so it is not surprising that such arrays must be large.
- ▶ **Fact:** (Semukhin/Zilles) If an $N \times k$ binary array A has $\text{SVCD}(A) = t$, then

$$N \geq \sum_{i=0}^t \binom{k}{i},$$

i.e. the Sauer upper bound.

Strong VC dimension, II

- ▶ Having $\text{SVCD} = t$ is a very strong condition, so it is not surprising that such arrays must be large.
- ▶ **Fact:** (Semukhin/Zilles) If an $N \times k$ binary array A has $\text{SVCD}(A) = t$, then

$$N \geq \sum_{i=0}^t \binom{k}{i},$$

i.e. the Sauer upper bound.

- ▶ It follows that if $\text{SVCD}(A) = \text{CD}(A) = \text{VCD}(A)$, A is necessarily maximum.

Pushing apart, again

- ▶ One can easily construct arrays with $SVCD(A) = t$ and $CD(A) = VCD(A) = t + 1$, by appending an all-1s row to a maximum array of $VCD = t$.

Pushing apart, again

- ▶ One can easily construct arrays with $SVCD(A) = t$ and $CD(A) = VCD(A) = t + 1$, by appending an all-1s row to a maximum array of $VCD = t$.
- ▶ Example:

0	0	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	1	0	0
1	0	1	0
1	0	0	1
0	1	1	0
0	1	0	1
0	0	1	1
1	1	1	1

Some questions

- ▶ For which integers $a \leq b \leq c$ can we construct arrays with $\text{SVCD}(A) = a$, $\text{CD}(A) = b$ and $\text{VCD}(A) = c$?

Some questions

- ▶ For which integers $a \leq b \leq c$ can we construct arrays with $\text{SVCD}(A) = a$, $\text{CD}(A) = b$ and $\text{VCD}(A) = c$?
- ▶ What can we say about arrays where $\text{CD}(A) = \text{VCD}(A)$?

Some questions

- ▶ For which integers $a \leq b \leq c$ can we construct arrays with $\text{SVCD}(A) = a$, $\text{CD}(A) = b$ and $\text{VCD}(A) = c$?
- ▶ What can we say about arrays where $\text{CD}(A) = \text{VCD}(A)$?
- ▶ Can we find any useful applications?

THE END

*See you at the CMS Summer Meeting in Regina, 2–4 June 2012:
special session on Combinatorics (organizers: K. Meagher, M.
Mishna)*