The Goodman-Kruskal Coefficient and Its Applications in Genetic Diagnosis of Cancer

Szymon Jaroszewicz*, Dan A. Simovici†, Winston P. Kuo‡, Lucila Ohno-Machado§

July 30, 2003

Abstract

Increasing interest in new pattern recognition methods has been motivated by bioinformatics research. The analysis of gene expression data originated from microarrays constitutes an important application area for classification algorithms, and illustrates the need for identifying important predictors. We show that the Goodman-Kruskal coefficient can be used for constructing minimal classifiers for tabular data and we give an algorithm that construct such classifiers.

1 Introduction

Global gene expression profiling using DNA microarrays has emerged as a rapid means to explore, classify, and predict the biological processes underlying human diseases. In the field of cancer, this revolutionary technology permits the simultaneous measurement of the transcription of tens of thousands of genes, and of their relative expression between normal, dysplastic and malignant cells. Since the first report of microarrays [SSDB95], the number of cancer and microarray related publications has increased exponentially, from approximately 30 publications in the first four years (1995-1999), to almost 600 publications in the past 24 months. Microarrays have evolved from depositing DNA onto a solid support to measure gene expression profiles to DNA copy number analysis

*University of Massachusetts Boston, Dept. of Computer Science, Boston, MA 02125, sj@cs.umb.edu

†University of Massachusetts Boston, Dept. of Computer Science, Boston, MA 02125, dsim@cs.umb.edu

‡Decision System Group, Thorn 310, Brigham and Womens' Hospital, 75 Francis Street, Boston, MA 02115 and Division of Health Sciences and Technology, Harvard and MIT, wkuo@dsg.bwh.harvard.edu

§Decision System Group, Thorn 310, Brigham and Womens' Hospital, 75 Francis Street, Boston, MA 02115 and Division of Health Sciences and Technology, Harvard and MIT, machado@dsg.bwh.harvard.edu

(comparative genomic hybridization arrays, CGH) [PSS+98] and more recently protein and antibody spotted microarrays [HDB01].

The construction of gene expression databases requires technologies that can accurately and reproducibly measure changes in global mRNA expression levels. Ideally, these technologies should be able to screen all gene transcripts, be applicable across a wide range of cell and tissue types, require minimal amounts of biological material, and be capable of processing large number of samples. There are currently two commonly used types of DNA microarrays: spotted complementary DNA (cDNA) microarrays [AKG+91] and short oligonucleotide arrays (Affymetrix GeneChips)[LFGL99]. cDNA arrays represent a popular platform in which double-stranded polymerase chain reaction (PCR) products amplified from expressed sequence tag (EST) clones are robotically spotted onto glass slides. The average product size ranges from 100-1,000 nucleotides in length. cDNA arrays offer versatility where project specific arrays can be custom designed. For example, it is possible to build cancer-specific and chromosome-specific arrays, and they can be employed in the simultaneous analysis of two different biological samples (normal versus malignant tissues, tumors of different stages, cancer cell lines, etc.). More importantly is the opportunity to discover novel genes since, in addition to well-characterized genes, expressed sequence tags (ESTs) of unknown function can be spotted on the DNA arrays. This approach has been used by several groups for cancer research [AED+00, BMC+00, PJvdR+99]. GeneChips on the other hand are manufactured by synthesizing one nucleotide at a time onto a glass slide and consist of short oligonucleotides. These oligonucleotide arrays have also been used widely in cancer research[ABN+99, GST+99, LYC+02]. In general, there are advantages and disadvantages to both microarray technology platforms; however, the crucial difference between the way cDNA and oligonucleotide microarrays are commonly used is that cDNA experiments return the amount of each transcript relative to another sample, whereas oligonucleotide experiments return an absolute amount of each transcript. This implies a major difference in the ability to group and universally compare across different microarray platforms [KJB+02]. Regardless of the technology platform chosen, microarray experiments yield far more information than we were used to process in biological experiments.

Classical empirical approaches to anticancer therapy strategies are still being used, although global approaches to identify new targets for anticancer drugs represent a faster strategy. However, several bottlenecks still exist and translating microarrays findings into clinical applications remains a work in progress. Recent microarray studies involving cancer research have focused in tumor classification, outcome prediction. and progression of disease. Combining expression analysis with emerging technologies, such as CGH microarrays, laser capture microdissection (LCM) [TLK02], and single-nucleotide polymorphisms (SNPs) and proteomics analyses may provide new possibilities to understand the genetic changes associated with cancer etiology and development. Ultimately, this may allow the discovery of new biomarkers for disease diagnosis and prognosis prediction, and development of therapeutic strategies for cancer treatment.

Tumor classification using microarray data can be implemented using algorithms that identify minimal classification criteria for tabular data. We present such an algorithm starting with the Goodman-Kruskal association index (see [GK80, Lie83]), which is one of two measures of association introduced in [GK80] that can be naturally interpreted as a probability of misclassification.

Let X, Y be two discrete random variables. We assume that we deal with a finite probability space where the elementary events are pairs of values (a_i, b_j), where a_i is a value of X and b_j is a value of Y. The *Goodman-Kruskal coefficient* of X and Y is defined by

$$\begin{aligned} \mathsf{GK}(X,Y) &= \sum_{i=1}^{l} P(X=a_i)\left(1 - \max_{1\leq j\leq k} P(Y=b_j|X=a_i)\right) \\ &= 1 - \sum_{i=1}^{l} P(X=a_i) \max_{1\leq j\leq k} P(Y=b_j|X=a_i). \end{aligned}$$

Goodman and Kruskal adopt the classification rule that prescribes that an elementary event is to be classified in the class that has the maximal probability. In the absence of any knowledge about X, an elementary event will be classified in the Y-class b_j if b_j corresponds to the highest value among the probabilities $P(Y=b_j)$ for $1 \leq j \leq k$. If $P(Y=b_j|X=a_i)$ is the probability of predicting the value b_j for Y when $X=a_i$, then an event that has the component $X=a_i$ will be classified in the Y-class b_j if j is the number for which $P(Y=b_j|X=a_i)$ has the largest value. The probability of misclassification committed by applying this rule is $1-\max_{1\leq j\leq k} P(Y=b_j|X=a_i)$. Thus, $\mathsf{GK}(X,Y)$ is the expected probability that in a randomly chosen case the value of Y will be incorrectly predicted from X.

The *Goodman-Kruskal association index* $\lambda_{Y|X}$, commonly used in literature, is the relative reduction in the probability of prediction error:

$$\lambda_{Y|X} = 1 - \frac{\mathsf{GK}(X,Y)}{1-\max_{1\leq j\leq k} P(Y=b_j)}$$

In other words, $\lambda_{Y|X}$ is the proportion of the relative error in predicting the value of Y that can be eliminated by knowledge of the X-value.

In the rest of the paper we use exclusively the Goodman-Kruskal coefficient.

In the next section we formulate a definition of the Goodman-Kruskal coefficient GK within a purely algebraic setting, using partitions of finite sets. The main advantage of this formulation is that we can use the properties of the partially ordered set of partitions of a set in our considerations.

Starting from the properties of GK discussed in Section 2, we formulate an algorithm that identifies minimal classification criteria for tabular data. We applied this algorithm to two well-known sets of data discussed in [KWR+01] and in [GST+99]. The data set of [KWR+01] is used to differentiate between four types of childhood tumors known collectively as round blue cell tumors. The second set, presented in [GST+99] is used to distinguish between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Both data sets were intensively explored in recent publications [MTS+98, OMVW02, YRT+01] using fuzzy sets, neural networks, and support vector machines.

2 The Goodman-Kruskal Coefficient for Partitions

Let S be a finite set. A partition of S is a family of its subsets $\{B_1, B_2, \ldots, B_k\}$ such that $\bigcup_{i=1}^{k} B_i = S$ and $B_i \cap B_j = \emptyset$ for all $1 \leq i < j \leq k$. Let $\mathsf{PART}(S)$ be the set of partitions of S. For $\pi, \sigma \in \mathsf{PART}(S)$ we write $\pi \leq \sigma$ if for every block B of π there exists a block C of σ such that $B \subseteq C$. It is easy to verify that the relation "$\leq$" is a partial order on $\mathsf{PART}(S)$. For example if $S = \{1, 2, \ldots, 5\}$, then $\{\{1,2\},\{3\},\{4,5\}\} \leq \{\{1,2,3\},\{4,5\}\}$.

Consider two partitions $\pi = \{B_1, \ldots, B_l\}$ and $\sigma = \{C_1, \ldots, C_k\}$ in $\mathsf{PART}(S)$. Define the *Goodman-Kruskal coefficient* of these partitions $\mathsf{GK}(\pi, \sigma)$ as the number:

$$\mathsf{GK}(\pi, \sigma) = 1 - \sum_{i=1}^{l} \max_{1 \leq j \leq k} \frac{|C_j \cap B_i|}{|S|}.$$

The partitions π, σ define two random variables

$$X : \begin{pmatrix} 1 & \cdots & l \\ \frac{|B_1|}{|S|} & \cdots & \frac{|B_l|}{|S|} \end{pmatrix} \text{ and } Y : \begin{pmatrix} 1 & \cdots & k \\ \frac{|C_1|}{|S|} & \cdots & \frac{|C_k|}{|S|}. \end{pmatrix}$$

such that conditional probability $P(Y = j | X = i)$ is given by:

$$P(Y = j | X = i) = \frac{P(Y = j \wedge X = i)}{P(X = i)} = \frac{|C_j \cap B_i|}{|B_i|}.$$

Thus, for a fixed i, the largest error in predicting Y is

$$1 - \max_{1 \leq j \leq k} P(Y = j | X = i) = 1 - \max_{1 \leq j \leq k} \frac{|C_j \cap B_i|}{|B_i|}.$$

The expected value of this error is

$$\sum_{i=1}^{l} \frac{|B_i|}{|S|} \cdot \left(1 - \max_{1 \leq j \leq k} \frac{|C_j \cap B_i|}{|B_i|}\right) = 1 - \sum_{i=1}^{l} \max_{1 \leq j \leq k} \frac{|C_j \cap B_i|}{|S|},$$

which is exactly the Goodman-Kruskal coefficient $\mathsf{GK}(X, Y)$.

Several properties of GK important from the point of view of classification are given next. The proofs can be found in Appendix A.

Theorem 2.1 *Let S be a finite set, and let $\pi, \sigma \in \mathsf{PART}(S)$, where $\sigma = \{C_1, \ldots, C_k\}$. We have:*

$$\mathsf{GK}(\pi, \sigma) \leq 1 - \max_{1 \leq j \leq k} \frac{|C_j|}{|S|}.$$

Proof. See Appendix A.1. ∎

Theorem 2.2 *Let S be a finite set and let $\pi = \{B_1, \ldots, B_l\}$ and $\sigma = \{C_1, \ldots, C_k\}$ be two partitions of the set S. We have $\mathsf{GK}(\pi, \sigma) = 0$ if and only if $\pi \leq \sigma$.*

Proof. See Appendix A.2. ∎

Theorem 2.3 *The function GK is monotonic in its first argument and dually monotonic in its second. In other words, if π, π', σ are three partitions of the set S such that $\pi \leq \pi'$, then $\mathsf{GK}(\pi, \sigma) \leq \mathsf{GK}(\pi', \sigma)$, and if π, σ', σ are three partitions of the set S such that $\sigma \leq \sigma'$, then $\mathsf{GK}(\pi, \sigma) \geq \mathsf{GK}(\pi, \sigma')$.*

Proof. See Appendix A.3. ∎

The Goodman-Kruskal coefficient allows us to define a metric on $\mathsf{PART}(S)$. Consider the function $d_{GK} : \mathsf{PART}(S) \times \mathsf{PART}(S) \longrightarrow \mathbb{R}$ given by

$$d_{GK}(\pi, \sigma) = \mathsf{GK}(\pi, \sigma) + \mathsf{GK}(\sigma, \pi).$$

for $\pi, \sigma \in \mathsf{PART}(S)$.

Theorem 2.4 *The function d_{GK} is a metric on the set $\mathsf{PART}(S)$.*

Proof. See Appendix A.4. ∎

3 An Algorithm for Predictor Identification

We use standard relational database terminology, as it appears, for example in [ST95]. A database table τ is a triple $\tau = (T, H, \rho)$, where T is table's name, H its header, and ρ its contents. The header is a set of symbols $H = A_1 \cdots A_n$ called attributes, which serve as labels of the columns of the table (see Figure 1). For each attribute A_i we have a set called the domain of A_i that is denoted by $\mathrm{Dom}(A_i)$. Only values of $\mathrm{Dom}(A_i)$ may appear in the column labeled A_i. The content of the table consists of a set of *rows*, $\{t_1, \ldots, t_m\}$ that belong to the set product $\mathrm{Dom}(A_1) \times \cdots \times \mathrm{Dom}(A_n)$. Every attribute set $K \subseteq H$ induces a

T

	A_1	A_2	$\cdots$	A_n
t_1	a_{11}	a_{12}	$\cdots$	a_{1n}
t_2	a_{21}	a_{22}	$\cdots$	a_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_m	a_{m1}	a_{m2}	$\cdots$	a_{mn}

Figure 1: The structure of a table

partition π_K on the contents of the table. Namely, two tuples t and t' belong to the same block of the partition π_K if they have equal values on the attributes of

K, that is, $t[K] = t'[K]$ (see Figure 2). Thus, the tuples t_1, t_2, t_3 belong to the same block of the partition π_K because they have equal values for the attributes of K, etc. Therefore terms *attribute set* and *partition* are used interchangeably.

T

	$\cdots$	$\longleftarrow K \longrightarrow$	$\cdots$
t_1	$\cdots$	k_1	$\cdots$
t_2	$\cdots$	k_1	$\cdots$
t_3	$\cdots$	k_1	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_l	$\cdots$	k_p	$\cdots$
t_{l+1}	$\cdots$	k_p	$\cdots$
t_{l+2}	$\cdots$	k_p	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{n-1}	$\cdots$	k_r	$\cdots$
t_n	$\cdots$	k_r	$\cdots$

Figure 2: The partition of the content induced by a set of rows

Let $\tau = (T, H, \rho)$ be a table and let K, L be two sets of attributes, $K, L \subseteq H$. The Goodman-Kruskal coefficient $\mathsf{GK}(K, L)$ of the sets of attributes K, L is defined as $\mathsf{GK}(\pi_K, \pi_L)$ and can be interpreted as the expected error that occurs when we try to predict the value of $t[L]$ from the value of $t[K]$. If $K_1 \subseteq K_2$, then, by Theorem 2.3, $\pi_{K_2} \leq \pi_{K_1}$, hence $\mathsf{GK}(K_2, L) \leq \mathsf{GK}(K_1, L)$, which has an intuitive interpretation: the expected error of a larger set of attributes is smaller than the expected error of a small set of attributes.

Using a mechanism similar the metric d_{GK} can be transferred to the collection of sets of attributes of a table by defining $d_{GK}(K, L) = d_{GK}(\pi_K, \pi_L)$ for any two sets of attributes K, L. The new metric can be used for constructing classifiers, for performing discretization of continuous attributes, and for attribute clustering.

Definition 3.1 An *ϵ-predictor* for a set of attributes Y is a set of attributes K such that $\mathsf{GK}(K, Y) \leq \epsilon$. □

If K is an ϵ-predictor for Y, then any superset K' of K is also a ϵ-predictor for Y. An ϵ-predictor such that none of its proper subsets is an ϵ-predictor is called *minimal*.

To find ϵ-predictors we use the following algorithm:

Input: A set of attributes H, a target attribute Y, $Y \notin H$ and an error level ϵ.
Output: Set P of all minimal ϵ-predictors from H.
(1) $\mathsf{Cand} = \{\{A\} : A \in H\}$;
(2) $P = \emptyset$;
(3) $\mathsf{P} = \mathsf{P} \cup \{K \in \mathsf{Cand} : \mathsf{GK}(K, Y) \leq \epsilon\}$;
(4) $\mathsf{Cand} = \mathsf{Cand} \setminus \mathsf{P}$;
(5) $\mathsf{Cand} = \{L \subseteq H : \text{ for all } K \subset L,\ |K| = |L| - 1 \text{ we have } K \in \mathsf{Cand}\}$;
(6) goto (3);

The algorithm works in a manner very similar to the well-known data-mining Apriori algorithm [HK01], using the fact that if a set is a nonminimal predictor, so are all of its supersets, which can thus be skipped. It begins by constructing a candidate set of predictors Cand in which all one-set attributes are initially included. The set of minimal predictors P is constructed starting from Cand. Initially, we include in P all one-attribute predictors whose error is below the threshold ϵ. These attributes are removed from P and the search for minimal two-attribute predictors makes use of the remaining candidate attributes, etc. The stopping condition could be exceeding the maximum predictor size or finding a predictor with desired prediction error.

4 Experimental Results

As experimental data we used the blood cell tumor microarray data of Khan et al. [KWR+01] and the leukemia data of Golub et al. [GST+99]. In both datasets gene expression levels have been normalized.

Full descriptions of experimental procedures and normalization methods used can be found in the original paper [KWR+01]; details on normalization methods applied to the second set of data are available at the web site `www-genome.wi.mit.edu/mpr/publications/projects/Leukemia`, which is a companion of [GST+99].

The first data set (in [KWR+01]) involves microarray data that is used for differential diagnosis of four round blue cell tumors of childhood (SRBCTs) : neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt lymphoma (BL), and the Ewing family of sarcomas (EWS). The expression levels of 2308 genes were measured using cDNA microarrays and the results of a predictive model based on a single layer neural network architecture [KWR+01] and a logistic regression model [WVOM02] were previously reported.

There are 63 training (12 NB, 20 RMS, 8 BL, and 23 EWS) and 25 test cases (6 EWS, 5 RMS, 6 NB, 3 BL, and 5 non-SRBCTs) in the dataset, which includes samples obtained from cell lines and biopsies . The test cases include 5 cases which do not belong to any of the predicted SRBCT types. Such cases are not present in the training set.

In the initial preprocessing step, we replace each class attribute with 4 binary attributes, one for each cancer type. Each of the 4 binary attributes has value 1 if and only if the corresponding caner type is present. The table below illustrates this

original attribute	computed attributes			
Cancer type	NB	RMS	BL	EWS
NB	1	0	0	0
RMS	0	1	0	0
BL	0	0	1	0
EWS	0	0	0	1
other	0	0	0	0

Then a separate predictor is built for each of the new binary attributes, i.e. we try to predict every tumor type separately. The main reason for this is to allow for handling of cases of type 'other' present in the test set, but absent in the training set. We expect that for 'other' cancer type all of the predictors will give the value of 0 thus indicated that none of the 4 cancer types is present.

A disadvantage is that predictors may contradict each other, however the experiments have shown that such cases are infrequent due to low error rate of individual classifiers. If however, presence of more than one cancer type is predicted, we assume that the prediction cannot be made in such a case, and consider it misclassified.

Since the number of training cases is very small, there is a serious risk of the so called overfitting, that is building a complex classifier that fits particular training data very well, but does not generalize to unseen cases. Building a separate predictor for each cancer type makes the predictors very simple and thus less likely to overfit the data.

Next, every gene expression level X attribute is discretized into two intervals: $X \leq T$ and $X > T$ for some threshold T, thus changing the continuous gene expression level X into a binary attribute X'. The threshold T is chosen such that the Shannon entropy $H(Y|X')$ of the target Y conditional on the discretized attribute X' is minimal. A separate discretization has been performed with respect to each cancer type. To identify the thresholds used in discretization we applied the Fayyad-Irani algorithm presented in [Fay91, FI93].

Afterwards, the algorithm given above is used to find all predictors with 1 or 2 attributes, allowing up to one misclassified instance on the training set. Thus, the stopping rule for the learning algorithm is reaching the maximum prescribed size of the predictor, or obtaining an error rate less than to $\frac{1}{t}$, where t is the size of the training set. Building classifiers with more attributes would incur a large risk of overfitting due to small number of training cases (see [Mit97]). Since the number of attributes is huge, there is an extra attribute selection step involved; namely, before the algorithm is applied, all but 30 most predictive attributes are discarded.

For each class (cancer type) the first predictor with minimal training error is manually picked at random (without looking at its test set performance to avoid bias in the choice). The results are summarized in table below

Cancer type	selected predictor	image ids	mtr	mte	1GP	2GP
BL	WAS $\leq 0.69 \Rightarrow$ BL	236282	0	1	15	5
EWS	FCGRT $\leq 1.59 \Rightarrow$ EWS	770394	1	3	2	10
NB	MAP1B > 2.17 or RCV1 $> 1.98 \Rightarrow$ NB	629896 - 383188	0	0	2	28
RMS	TNNT2 > 0.55 or SGCA $> 0.44 \Rightarrow$ RMS	298062 - 796258	0	2	0	25

Legend:

mte	misclassified cases in test set
mtr	misclassified cases in training set
1GP	number of one-gene predictors
2GP	number of two-gene predictors

The results show that a fairly large number (12–30) of very simple predictors have been found for each cancer type. Each of those predictors has very good classification rate on the training set: up to one misclassified case is allowed.

For each cancer type number of errors on the test set is also given for selected predictors. The predictions on the test are slightly worse than on the training set, but still fairly accurate. This suggests that overfitting was not a serious issue here.

The results show that there are many genes based on which a diagnosis can be made for each cancer type.

Of note, all genes except for the one that predicts BL were reported among the 96 selected in [KWR$^+$01]. The probability that a single gene expression predicts BL perfectly on training set when there is no correlation between the gene and the tumor type:

$$2 \cdot \frac{55! \cdot 8!}{63!} = 5.16 \cdot 10^{-10}$$

much less than $0.05/2308 = 2.16 \cdot 10^{-5}$, the 5% significance level after Bonferroni correction. This shows that selected gene is with very high probability related to BL. Similar statements hold for other cancer types. Bonferroni correction has been used since the number of attributes is large and there could be a possibility that one of them passes the statistical test due to chance rather than true correlation. In the absence of the attribute independence the Bonferroni correction is overly conservative. However, the inferences using this correction are correct.

We combined the classifiers for each tumor type in the following way: if a classifier for only one type of tumor gave a positive prediction, then the instance was classified as this type of tumor. If none of them gave positive prediction we declared the case as ‘other tumor type’. If more than one classifier was active the case was considered a prediction error.

The combined classifier used a total of 6 genes and classified correctly 18 out of 25 cases. Out of the 6 misclassified cases, 2 gave classifications when the real outcome was ‘other’, 3 SRBCT cases were undetected, and there was 1 conflict.

The leukemia training data set contains 38 cases (27 acute lymphoblastic leukemia and 11 acute myelocytic leukemia) and the test set contains 34 cases (20 ALL and 14 AML); the data refer to 6817 genes. We discretized the gene expression levels using the same procedure as for the [KWR$^+$01] data. We retained 20 genes for which the Goodman-Kruskal coefficient was below 0.04. Five single-genes predictors and 66 two-gene predictors were identified.

It is worth noting that our technique identified two two-genes predictors (MGST1, APLP2 and CD33, CystatinA) for which the errors on the test set are 0 and 0.0294118, respectively. CD33 was among the 50 genes selected in [GST+99]. The distribution of the errors on the test set for the remaining set of minimal two-genes predictors is shown below:

Error Interval	Number of 2-attribute predictors
$[0.0, 0.05]$	2
$(0.05, 0.10]$	9
$(0.10, 0.15]$	10
$(0.15, 0.20]$	7
$(0.20, 0.25]$	13
$(0.25, 0.30]$	14
$(0.30, 0.35]$	3
$(0.35, 0.40]$	4
$(0.40, 0.45]$	3

To obtain a robust classifier, we used a voting mechanism on the second data set. Initially, we extracted 19 one-attribute predictors whose prediction error on the training set did not exceed 5.3% (that is, two errors out of the 38 training cases). A vote was taken, and the instance was classified according to the majority vote. We obtained 3 errors on the test set of 34 cases. Namely, the errors occurred on the 57th, 60th and 66th cases of the original Golub test set. It is worth mentioning that all errors occurred on cases that had low prediction strengths and were considered "unclassifiable" in [GST+99].

5 Conclusions and Future Research

The Goodman-Kruskal dissimilarity GK is a simple, but powerful measure of predictive power that can be used to produce robust classifiers that are comparable with those obtained using more complex techniques. In our opinion, the small number of training cases makes reliable construction of more complex models like Bayesian networks or C4.5 trees very hard or even impossible. On the other hand, methods related to Naive Bayesian classifiers suffer from independence assumptions which may not be satisfied in the microarray setting where most gene measurements are correlated with each other. As we mentioned earlier, the use of the Bonferroni correction, although conservative, yields valid results.

It would be interesting to see how some more sophisticated attribute selection techniques influence the quality of classification for microarray data, see e.g. [ILES00, KJ97, LM98].

The fact that GK generates a metric suggests one could use this metric to cluster attributes such that those that belong to a cluster have similar predictive power and, thus, are interchangeable in classifiers. The clustering structure could be used in forming the voting committees, and in simplifying and increasing the robustness of predictive algorithms. This approach dovetails with other metrics on partitions that we explored in [SJ03].

6 Acknowledgement

The authors wish to acknowledge the judicious observations made by Dr. Stephan Dreiseitl from FSH Hagenberg, Austria, and the helpful comments of the anonymous reviewers. Dr. Ohno-Machado was partially supported by the Taplin award from the Division of Health Sciences and Technology, Harvard-MIT.

A Lemmas and Proofs of Theorems

Lemma A.1 *Let $\{x_{ij} | 1 \leq i \leq l, 1 \leq j \leq k\}$ be a collection of lk real numbers. We have:*

$$\max_{1 \leq j \leq k} \sum_{i=1}^{l} x_{ij} \leq \sum_{i=1}^{l} \max_{1 \leq j \leq k} x_{ij}.$$

Proof. Observe that $x_{ij} \leq \max_{1 \leq j \leq k} x_{ij}$ for every i, j, which implies $\sum_{i=1}^{l} x_{ij} \leq \sum_{i=1}^{l} \max_{1 \leq j \leq k} x_{ij}$ for every j. Thus, we obtain

$$\max_{1 \leq j \leq k} \sum_{i=1}^{l} x_{ij} \leq \sum_{i=1}^{l} \max_{1 \leq j \leq k} x_{ij}.$$

∎

A.1 Proof of Theorem 2.1

The definition of $\mathsf{GK}(\pi, \sigma)$ gives

$$\mathsf{GK}(\pi, \sigma) = 1 - \sum_{i=1}^{l} \max_{1 \leq j \leq k} \frac{|C_j \cap B_i|}{|S|},$$

where $\pi = \{B_1, \ldots, B_l\}$. By Lemma A.1 we can write:

$$\mathsf{GK}(\pi, \sigma) \leq 1 - \max_{1 \leq j \leq k} \sum_{i=1}^{l} \frac{|C_j \cap B_i|}{|S|} = 1 - \max_{1 \leq j \leq k} \frac{|C_j|}{|S|},$$

which is the desired inequality. ∎

A.2 Proof of Theorem 2.2

Suppose that $\mathsf{GK}(\pi, \sigma) = 0$, which means that

$$\sum_{i=1}^{l} \max_{1 \le j \le k} \frac{|C_j \cap B_i|}{|S|} = 1.$$

This is possible only if for every B_i there exists C_j such that $B_i \subseteq C_j$, that is, if $\pi \le \sigma$. The reverse implication is clear. ∎

A.3 Proof of Theorem 2.3

To prove the monotonicity of GK in its first argument it suffices to show that if π' is obtained from π by fusing two blocks B_i and B_h, then $\mathsf{GK}(\pi, \sigma) \le \mathsf{GK}(\pi', \sigma)$.

We claim that for every i, h we have:

$$\max_{1 \le j \le k} |C_j \cap (B_i \cup B_h)| \le \max_{1 \le j \le k} |C_j \cap B_i| + \max_{1 \le j \le k} |C_j \cap B_h|.$$

Indeed, observe that $|C_j \cap (B_i \cup B_h)| = |C_j \cap B_i| + |C_j \cap B_h| \le \max_{1 \le j \le k} |C_j \cap B_i| + \max_{1 \le j \le k} |C_j \cap B_h|$. This gives the desired inequality. Thus,

$$\begin{aligned}
&\mathsf{GK}(\pi, \sigma) \\
&= 1 - \left(\sum_{m=1, m \ne i,h}^{l} \max_{1 \le j \le k} \frac{|C_j \cap B_m|}{|S|} + \max_{1 \le j \le k} \frac{|C_j \cap B_i|}{|S|} + \max_{1 \le j \le k} \frac{|C_j \cap B_h|}{|S|} \right) \\
&\le 1 - \left(\sum_{m=1, m \ne i,h}^{l} \max_{1 \le j \le k} \frac{|C_j \cap B_m|}{|S|} + \max_{1 \le j \le k} \frac{|C_j \cap (B_i \cup B_h)|}{|S|} \right) = \mathsf{GK}(\pi', \sigma).
\end{aligned}$$

For the second part of the theorem it suffices to show that if σ' is obtained from σ by fusing two blocks C_i and C_h, then the above inequality holds.

For a block B_i of π we have:

$$\begin{aligned}
&\max\{|B_i \cap C_1|, \ldots, |B_i \cap C_i|, \ldots, |B_i \cap C_h|, \ldots, |B_i \cap C_k|\} \le \\
&\max\{|B_i \cap C_1|, \ldots, |B_i \cap (C_i \cup C_h)|, \ldots, |B_i \cap C_k|\}.
\end{aligned}$$

Therefore,

$$\sum_{i=1}^{k} \max_{1 \le j \le k} |B_i \cap C_j| \le \sum_{i=1}^{k} \max\{|B_i \cap C_j| | C_j \in \sigma'\},$$

which gives the inequality of the theorem. ∎

A.4 Proof of Theorem 2.4

It is obvious that the function d_{GK} is non-negative and symmetric.

Suppose that $d_{GK}(\pi,\sigma) = 0$, which implies $\mathsf{GK}(\pi,\sigma) = \mathsf{GK}(\sigma,\pi) = 0$. Thus, by Theorem 2.2 we have both $\pi \leq \sigma$ and $\sigma \leq \pi$, which implies $\pi = \sigma$. Since $d_{GK}(\pi,\pi) = 0$ for any partition $\pi \in \mathsf{PART}(S)$ it follows that $d_{GK}(\pi,\sigma) = 0$ if and only if $\pi = \sigma$.

Let (b_{jh}) be a $k \times g$ matrix where $b_{jh} \geq 0$ for $1 \leq j \leq k$ and $1 \leq h \leq g$. Let j_0 be a integer between 1 and k and let h_0 be the least integer in $\{1,\ldots,g\}$ such that $b_{j_0h_0} = \max_{1\leq h\leq g} b_{j_0h}$. We claim that

$$\sum_{h=1}^{g} b_{j_0h} + \sum_{j=1}^{k} \max_{1\leq h\leq g} b_{jh} \leq \sum_{j=1}^{k}\sum_{h=1}^{g} b_{jh} + \sum_{j=1}^{k} b_{jh_0}. \tag{1}$$

Inequality (1) can be written as:

$$\sum_{j=1}^{k} \max_{1\leq h\leq g} b_{jh} \leq \sum_{j=1,j\neq j_0}^{k} \sum_{h=1}^{g} b_{jh} + \sum_{j=1}^{k} b_{jh_0},$$

or

$$\sum_{j=1,j\neq j_0}^{k} \max_{1\leq h\leq g} b_{jh} + \max_{1\leq h\leq g} b_{j_0h} \leq \sum_{j=1,j\neq j_0}^{k} \sum_{h=1}^{g} b_{jh} + \sum_{j=1}^{k} b_{jh_0},$$

which follows from the fact that $\max_{1\leq h\leq g} b_{j_0h}$ is one of the members of the sum $\sum_{j=1}^{k} b_{jh_0}$ in view of the definition of h_0. This proves Inequality (1). Consequently, we have

$$\sum_{h=1}^{g} b_{j_0h} + \sum_{j=1}^{k} \max_{1\leq h\leq g} b_{jh} \leq \sum_{j=1}^{k}\sum_{h=1}^{g} b_{jh} + \max_{1\leq h\leq g} \sum_{j=1}^{k} b_{jh}$$

for every j_0 which, in turn, implies:

$$\max_{1\leq j\leq k} \sum_{h=1}^{g} b_{jh} + \sum_{j=1}^{k} \max_{1\leq h\leq g} b_{jh} \leq \sum_{j=1}^{k}\sum_{h=1}^{g} b_{jh} + \max_{1\leq h\leq g} \sum_{j=1}^{l} b_{jh}. \tag{2}$$

Consider the partitions $\pi = \{B_1,\ldots,B_i,\ldots,B_l\}$, $\sigma = \{C_1,\ldots,C_j,\ldots,C_k\}$, and $\tau = \{D_1,\ldots,D_h,\ldots,D_g\}$ on the set S and define $a_{ijh} = |B_i \cap C_j \cap D_h|$ for $1 \leq i \leq l$, $1 \leq j \leq k$, and $1 \leq h \leq g$. Inequality (2) implies

$$\max_{1\leq j\leq k} \sum_{h=1}^{g} a_{ijh} + \sum_{j=1}^{k} \max_{1\leq h\leq g} a_{ijh} \leq \sum_{j=1}^{k}\sum_{h=1}^{g} a_{ijh} + \max_{1\leq h\leq g} \sum_{j=1}^{k} a_{ijh}.$$

for $1 \leq i \leq l$. Summing on i we have:

$$\sum_{i=1}^{l} \max_{1\leq j\leq k} \sum_{h=1}^{g} a_{ijh} + \sum_{i=1}^{l}\sum_{j=1}^{k} \max_{1\leq h\leq g} a_{ijh} \leq \sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{h=1}^{g} a_{ijh} + \sum_{i=1}^{l} \max_{1\leq h\leq g} \sum_{j=1}^{l} a_{ijh}.$$

Equivalently,

$$\sum_{i=1}^{l} \max_{1\le j\le k} |B_i \cap C_j| + \sum_{i=1}^{l}\sum_{j=1}^{k} \max_{1\le h\le g} a_{ijh} \le |S| + \sum_{i=1}^{l} \max_{1\le h\le g} |B_i \cap D_h|. \tag{3}$$

Observe that:

$$\begin{aligned}
&\sum_{i=1}^{l} \max_{1\le j\le k} |B_i \cap C_j| + \sum_{j=1}^{k} \max_{1\le h\le g} |C_j \cap D_h| \\
&= \sum_{i=1}^{l} \max_{1\le j\le k} |B_i \cap C_j| + \sum_{j=1}^{k} \max_{1\le h\le g} \sum_{i=1}^{l} |B_i \cap C_j \cap D_h| \\
&\le \sum_{i=1}^{l} \max_{1\le j\le k} |B_i \cap C_j| + \sum_{j=1}^{k}\sum_{i=1}^{l} \max_{1\le h\le g} |B_i \cap C_j \cap D_h| \\
&\quad \text{(by Lemma A.1)} \\
&= \sum_{i=1}^{l} \max_{1\le j\le k} |B_i \cap C_j| + \sum_{j=1}^{k}\sum_{i=1}^{l} \max_{1\le h\le g} a_{ijk} \\
&\le |S| + \sum_{i=1}^{l} \max_{1\le h\le g} |B_i \cap D_h| \\
&\quad \text{(due to Inequality (3)).}
\end{aligned}$$

After an elementary transformation the last inequality can be written as:

$$1-\frac{1}{|S|}\sum_{i=1}^{l} \max_{1\le j\le k} |B_i\cap C_j| + 1-\frac{1}{|S|}\sum_{j=1}^{k} \max_{1\le h\le g} |C_j\cap D_h| \ge 1-\frac{1}{|S|}\sum_{i=1}^{l} \max_{1\le h\le g} |B_i\cap D_h|,$$

which is equivalent to:

$$\mathsf{GK}(\pi,\sigma) + \mathsf{GK}(\sigma,\tau) \ge \mathsf{GK}(\pi,\tau). \tag{4}$$

Similarly, we can write

$$\mathsf{GK}(\tau,\sigma) + \mathsf{GK}(\sigma,\pi) \ge \mathsf{GK}(\tau,\pi). \tag{5}$$

Adding the inequalities (4) and (5) yields the triangular inequality: $d_{GK}(\pi,\sigma) + d_{GK}(\sigma,\tau) \ge d_{GK}(\pi,\tau)$. This allows us to conclude that d_{GK} is a metric.

References

[ABN+99] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96(12):6745–6750, 1999.

[AED+00] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[AKG+91] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, and R. F. Moreno. Complementary dna sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656, 1991.

[BMC+00] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–540, 2000.

[Fay91] U. M. Fayyad. *On the Induction of Decision Trees for Multiple Concept Learning*. PhD thesis, University of Michigan, 1991.

[FI93] U. M. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1027. Morgan-Kaufmann, 1993.

[GK80] L. A. Goodman and W. H. Kruskal. *Measures of Association for Cross-Classification*, volume 1. New York, Springer-Verlag, 1980.

[GST+99] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Dowining, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery by gene expression monitoring. *Science*, 286:531–537, 1999.

[HDB01] B. B. Haab, M. J. Dunham, and P. O. Brown. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol.*, 2(2), 2001. RESEARCH0004.

[HK01] J. Han and M. Kamber. *Data Mining – Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2001.

[ILES00] I. Inza, P. Larranaga, R. Etxeberria, and B. Sierra. Feature subset selection by bayesian network-based optimization. *Artificial Intelligence*, 123(1-2):157–184, 2000.

[KJ97] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[KJB+02] W. P. Kuo, T. K. Jenssen, A. J. Butte, L. Ohno-Machado, and I. S. Kohane. Analysis of matched mrna measurements from two different microarray technologies. *Bioinformatics*, 18(3):405–412, 2002.

[KWR+01] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westerman, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7:673–679, 2001.

[LFGL99] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nat. Genet.*, 21(1 Suppl):20–24, 1999.

[Lie83] A. M. Liebtrau. *Measures of Association.* SAGE, Beverly Hills, CA, 1983.

[LM98] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining.* Kluwer Academic Publishers, 1998.

[LYC+02] J. H. Luo, Y. P. Yu, K. Cieply, F. Lin, P. Deflavia, R. Dhir, S. Finkelstein, G. Michalopoulos, and M. Becich. Gene expression analysis of prostate cancers. *Mol. Carcinog*, 33(1):25–35, 2002.

[Mit97] T. M. Mitchell. *Machine Learning.* McGraw and Hill, Boston, 1997.

[MTS+98] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. Mesirov, and T. Poggio. Support vector machine classification of microarray data. AI MEMO CBCL Paper 1677, MIT Artificial Intelligence Laboratory, 1998.

[OMVW02] L. Ohno-Machado, S. Vinterbo, and G. Weber. Classification of gene expression data using fuzzy logic. *Journal of Intelligent and Fuzzy Systems*, 12:19–24, 2002.

[PJvdR+99] C. M. Perou, S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. Lee, D. Lashkari, D. Shalon, P. O. Brown, and D. Botstein. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA*, 96(16):9212–9217, 1999.

[PSS+98] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M.

Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, 20(2):207–211, 1998.

[SJ03] D. A. Simovici and S. Jaroszewicz. Generalized conditional entropy and decision trees. In *Proceedings of EGC 2003*, pages 369–380, Lyon, France, 2003.

[SSDB95] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[ST95] D. A. Simovici and R. L. Tenney. *Relational Database Systems*. Academic Press, New York, 1995.

[TLK02] R. Todd, M. W. Lingen, and W. P. Kuo. Gene expression profiling using laser capture microdissection. *Expert Rev. Mol. Diagn*, 2(5):497–507, 2002.

[WVOM02] G. Weber, S. Vinterbo, and L. Ohno-Machado. Building an asynchronous web-based tool for machine learning classification. In *Proceedings of AMIA Symposium*, pages 869–873, Chicago, 2002.

[YRT+01] C. Yeang, S. Ramaswami, P. Tamayo, S. Mukherjee, R. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub. Molecular classification of multiple tumor types. *Bioinformatics, 1 ISMB 2001*, 17:316–323, 2001.