

Evaluating the Performance of Association Mining Methods in 3-D Medical Image Databases*

Vasileios Megalooikonomou[†]

Abstract. Evaluation of the process of mining associations is an important problem in database systems and especially those that store critical data and are used for making critical decisions. In the context of spatial databases and in particular in 3-D medical image databases we present an evaluation framework in which we use probability distributions to model the spatial regions of interest (ROIs), and Bayesian networks to model the joint probability distribution among ROIs and observed deficits or medical conditions. By controlling these parameters, we evaluate, as example, the Fisher exact test of independence, one of the methods currently available for detection of associations. We obtain measures of recovery of known associations as a function of the number of samples used, the strength and number of associations in the statistical model, the number of spatial ROIs associated with a particular deficit, the prior probabilities of spatial regions being of interest, the conditional probabilities of the deficits and the spatial normalization error.

1 Introduction

Mining in 3-D image databases [13, 46, 48, 45, 37] can be seen as a special case of spatial data mining. Spatial data mining refers to the discovery of spatial relation-

*This work was supported in part by National Science Foundation grant IIS 0083423, by National Institutes of Health grant R01 AG13743-03, and by an *Engines for Innovation* grant from Illustra.

[†]The author is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA (email: vasilis@cis.temple.edu). The author wishes to thank C. Davatzikos and E. Herskovits for assisting in developing an earlier version of the simulator and for providing image registration and medical expertise, J. Gerring for providing the FLIC study data set, and D. Kontos, D. Pokrajac and D. Margaritis for providing constructive comments.

ships and relationships between spatial and non-spatial data, or other interesting patterns not explicitly stored in spatial databases [31]. Recent advances in spatial databases [21, 43, 22, 9] led to the study of spatial data mining. Statistical methods [16] were the first applied to spatial data. Initially these methods had difficulty working with incomplete data, they were computationally expensive in large databases and they typically made statistical independence assumptions that do not hold among “neighboring” spatial objects. General data mining methods [42, 1] were also extended towards spatial data. The algorithms for spatial data mining include statistical cluster analysis and generalization-based methods for mining spatial characteristic and discriminant rules [23, 39], two-step spatial computation technique for mining spatial association rules [31], aggregate proximity technique for finding characteristics of spatial clusters [30], spatial clustering [11, 47], etc. Challenging issues in mining of spatial databases [24] include the exploration of statistical dependence that exists between neighboring objects, the large dimensionality of the data involved, and the efficient management of topological and/or distance information through spatial reasoning, spatial knowledge representation techniques and multidimensional spatial access methods (SAM).

Although a lot of effort has been directed towards the development of spatial data mining methods, the evaluation of the mining process in spatial data as a function of several parameters including the sample size, has not been addressed sufficiently so far. This problem becomes even more important when dealing with critical data coming from medical studies, environmental studies, navigation, remote sensing, etc. Needs for performing this evaluation are the following: (a) the ground truth, i.e., the set of true associations is not known and therefore, it is difficult to evaluate the validity of associations that are discovered, (b) the variability in data sets (i.e., different formats, resolutions, etc), data collection procedures, and analysis approaches makes it difficult to compare results from different studies, (c) the relative merits and limitations of each method employed cannot be objectively evaluated, and (d) the analytic methods used require a relatively large number of samples to produce results with high confidence.

Some related work has been done on non-spatial data. Megiddo and Srikant [38], used simulations to determine the threshold for the p-value for association rules. They empirically showed that the support and confidence threshold prune out most rules that are not statistically significant. In addition, several researchers have studied systematically the problem of sample size corresponding to power of statistical tests [5, 19, 18] (such as the chi-square and Fisher exact tests of independence), compared with other tests [50] and performed simulations to study the power of statistical tests in sample spaces of much higher dimensionality [40, 51, 50] (similar to those expected from epidemiological studies).

Evaluating mining methods that are applied on spatial data and in particular on 3-D image data is more complex than evaluating mining methods that are applied on non-spatial data such as market basket data. First, the high dimensionality of the data involved and the need to evaluate multiple spatial relationships among a large number of objects, complicates the analysis (e.g., yielding very sparse contingency tables in chi-square tests). Second, the degree of exploitation of the statistical dependence between neighboring objects must also be measured. Finally, the effect

of the data preprocessing (e.g., image registration and segmentation) on the performance of the mining methods and the mining methods themselves must also be characterized as a function of the sample size, and the strength and complexity of the associations. Due to the complexity of this domain, an analytical solution for the evaluation problem is not possible.

In this paper, we do not propose new data mining methods; we propose a framework for evaluating different mining methods (statistical and non-statistical) used for detection of associations between 3-D image data and non-image data. This work is motivated by 3-D medical image database systems [17, 12, 3, 34] that were built to discover structure-function relationships in the human brain although the methods presented can be applied in general to 3-D medical image databases. Our approach is to design a simulator in order to generate a large number of artificial patients (including image data, deficits, and associations among them), simulate the error of a given spatial normalization (i.e., registration) method and apply this nonlinear error to the image data, apply a mining method, and compare the true associations with the ones that the method discovered. The simulator provides the ability to create unlimited amounts of data and allows us to explore in a controllable way the effects of various parameters and preprocessing steps in the performance of the data mining algorithms.

The rest of the paper is organized as follows: Section 2 presents background information. In Section 3 we describe the method used for the evaluation of the spatial mining procedure including the various components of a simulator. In Section 4, we demonstrate the use of our evaluation framework by presenting, as a case study, experimental results from the evaluation of the Fisher exact test (one of the statistical methods available for detection of associations). Finally, the paper concludes with a discussion in Section 5.

2 Background

Typically 3-D medical image databases [17, 12, 3, 27, 34] consist of a large collection of studies that include 3-D images from different medical imaging modalities that capture structural (e.g., CT ¹, MRI ²), functional/physiological (e.g., PET ³, fMRI ⁴), and other information about the tissue and internal structures as well as a set of anatomical templates. The image data resulting from scanning of the patient are multiple layers of images which are combined into a voxel-based 3-D representation. Each such study, typically involves hundreds or even thousands of patients. Depending on the kind of study (e.g., fMRI requires time-sequences for each patient and each task performed) and the spatial and contrast resolutions used, the amount of data per patient can easily reach hundreds of megabytes. Adding these numbers together we find that in such environments terabytes of data are not unusual.

In brain mapping, behavioral and image data are collected from patients and analyzed in order to detect associations among spatial regions of the brain and

¹Computed Tomography: shows hard-tissue structural information.

²Magnetic Resonance Imaging: shows soft-tissue structural information.

³Positron Emission Tomography: shows physiological activity.

⁴functional-Magnetic Resonance Imaging: shows physiological activity.

Table 1. *Symbol Table.*

<i>symbol</i>	<i>definition</i>
\mathcal{D}	real data set
\mathcal{P}	set of patients $\{\rho_1, \rho_2, \dots, \rho_L\}$
\mathcal{S}	set of spatial regions $\{s_1, s_2, \dots, s_K\}$
\mathcal{C}	set of deficits or conditions $\{c_1, c_2, \dots, c_M\}$
d	number of spatial regions affecting a deficit
f_{s_i, ρ_j}	fraction of abnormal volume for spatial region (structure) s_i and patient ρ_j
t_a	% of existing associations discovered
f_a	% of false positive associations discovered

their functions. The analysis is performed through the study of either abnormal regions and associated deficits or activated regions and tasks performed. In the following presentation we will follow the first paradigm. The first step in this process is to make data comparable across patients. In particular, for image data, the ROIs must be identified (segmented), and image registration must be performed to map homologous anatomical regions to the same location in a common spatial standard (such as the Talairach anatomical atlas [49]). An atlas models the exact shapes and positions of anatomical structures. An MR image does not identify the structure to which each voxel (volume element) belongs, but an atlas can provide this information, with the accuracy of the registration methods, when overlaid on the image. Several linear and nonlinear spatial normalization (registration) methods have been developed; in this work we used a nonlinear method based on a three-dimensional elastically deformable model [8]. In the discussion that follows, spatial regions correspond to anatomical structures of the atlas and ROIs correspond to fractions of spatial regions that are abnormal.

The current paper is motivated by previously reported work on the development of a brain image database (BRAID) [34], a large-scale archive of normalized 3-D image and clinical (behavioral) data with an analytical query mechanism. Megalooikonomou et al. [37] have presented a data mining process to discover brain structure–function relationships through the study of spatial ROIs and associated conditions. In this framework, several methods to detect spatial region–deficit associations have been proposed, including the use of the chi-square test, Fisher exact test, voxel-based logistic regression, and clustering analysis.

Although the evaluation framework we propose in this paper can be applied to the study of any mining method, either statistical or non-statistical, in the results Section we present as a case study the analysis of a method based on the Fisher exact test. Here, we provide the necessary background regarding this statistical test.

The chi-squared test was recently proposed in the data mining community to measure significance of associations [6]. This test is used in the analysis as follows: for each spatial region–deficit pair⁵ a contingency table is constructed and

⁵Since computing a statistic for many pairwise tests creates the multiple-comparison prob-

Table 2. A contingency table for a spatial region s_i and a deficit c_k .

	s_i is abnormal	
	No	Yes
c_k is present	No	Yes
	a	b
	Yes	Yes
	c	d

the variables are examined to see whether they are independent of each other. A 2×2 contingency table (consisting of four cells) for a spatial region s_i and a deficit c_k is shown in Table 2. The chi-square statistic is calculated as: $\chi^2 = \sum \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$ where n_{ij} and m_{ij} are the observed and expected cell counts for cell (i, j) . The observed cell count for the cell (1, 1) of Table 2 is a . The expected cell count for cell (i, j) is: $m_{ij} = (n_{i.}n_{.j})/n$. where $n_{i.}$ is the i -th row total, $n_{.j}$ is the j -th column total, and n is the grand total. Greater differences between n_{ij} and m_{ij} produce larger χ^2 values and stronger evidence against the hypothesis of independence. The p -value is calculated from the chi-squared distribution and it is the right-hand or two tail probability above the observed χ^2 value. The p -value tells you how rarely you could observe a difference as large or larger than the one observed in the contingency table if the two variables were independent.

Due to its limitations[6] the chi-squared test is usually replaced by the Fisher exact test[14] where one calculates the actual (hypergeometric) probability $\frac{(n_1)! (n_2)! (n-1)! (n-2)!}{n! a b c d}$ of the observed 2×2 contingency table with respect to all other possible 2×2 contingency tables with the same column and row totals. The probabilities of all such tables that are each more likely than the observed table are added and if the sum is less than or equal to the specified significance level, the hypothesis of independence is rejected. In previous work a more restricted evaluation framework based on a lesion-deficit simulator[36] was used to evaluate the Fisher exact test. Here, we present a generalization of this framework for any type of spatial ROIs and deficits in 3-D medical image databases.

3 Proposed method

We design a spatial simulator to generate a large number of samples (patients) with 3-D images that include spatial ROIs and with deficits and associations among them all conforming to predefined distributions. Then, we analyze the generated spatial ROIs and deficits by applying data mining methods to determine the associations. Comparing the results of the analysis to the known associations in our simulation model allows us to quantify the accuracy of the mining methods. The number of spatial regions can range from the number of spatially distinct structures of an anatomical atlas ($\sim 10^2$) to the number of voxels in a 3-D image ($\sim 10^7$). An atlas-based approach is more sensitive than the voxel-based approach since the atlas provides significant prior knowledge. It is also less computationally demanding.

lem [28], i.e., a certain portion of tests will be positive by chance, the Bonferroni correction [2] is usually applied.

On the other hand, results from an atlas-based approach are as good as the atlas being used. Although our simulator can model spatial regions of any resolution range, in the Results section we show results from using the simulator to evaluate an atlas-based approach. In the next section we describe in more detail the spatial simulator.

3.1 The spatial simulator

Our goal in designing a simulator is to characterize the power of spatial data mining methods used to detect multivariate associations among spatial regions and deficits, taking into account the effects of registration, noise, and region segmentation. To accomplish this task, we must be able to produce artificial patients, each of which consisting of a 3-D image with abnormal regions (e.g., lesions) and a set of deficits (or abnormal conditions), so that both deficits and regions conform to predefined distributions, which can be determined from previously analyzed data sets, or empirical distributions based on clinical experience or information in the clinical literature. The data set \mathcal{D} that our simulator conforms to is a parameter of the simulator; any study or combination of studies can be used.

First, the simulator must generate the abnormal regions for each simulated patient. In order to account for the error that is introduced when the images are registered to the atlas, we displace the abnormal regions according to the distribution of registration error for the spatial-normalization method that we use. We then find the abnormal ROIs (structures) for each patient by superimposing the registered atlas on each patient’s image data; a spatial region is considered to be abnormal (ROI) if a certain fraction of its volume is abnormal. We generate a BN model of associations between spatial regions and deficits. After determining for each spatial region whether it is abnormal or not, the corresponding deficits are produced stochastically, using the BN model’s conditional probability distributions for spatial region-deficit associations. Thus, the major components of the simulation are: a) generation of simulated ROIs, b) modeling of spatial normalization error, and c) generation of synthetic associations.

Generation of simulated ROIs

In order to construct the spatial distribution of ROIs, the simulator enables us to collect statistics on a) the number of ROIs per patient, b) their sizes, and c) their spatial distribution from the sample data set \mathcal{D} . Then, for each patient, we draw samples from these distributions, generating the number of ROIs, and, for each ROI, its centroid and size.

The calculation of the spatial distribution of ROI centroids is quite interesting and it is presented here in more details. For each voxel of the 3-D image, we compute the probability that it will be the centroid of a ROI as follows: After computing the centroids for all ROIs in \mathcal{D} , we calculate the number of ROI centroids, o_l , that coincide with every point, l , of the 3-D image, i. e., we form a 3-D histogram of centroids for all these points. Since most voxels do not coincide with any ROI centroids the non-zero values are very sparse, incorrectly excluding most voxels

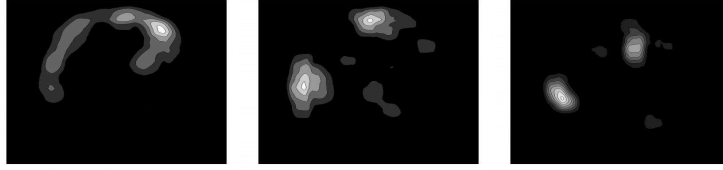


Figure 1. *Sample distributions of spatial ROIs.*

from the set of possible ROI centroids. To overcome this difficulty, we perform *smoothing*, by calculating the probability of observing a ROI at location l using a spherical window of radius r centered on l and counting the number of ROI centroids, o_l , that fall within that window. We chose the radius, r , by visualization of the distributions for spheres of different radius and by examining the fraction of the points in the 3-D image with $o_l = 0$. We next perform *padding* by replacing the zero values by very small positive values and *normalization* of the scalar probability field so that it forms a proper probability density function. Finally, we create an array where each voxel is repeated as many times as the probability density function denotes. An element of this array corresponding to a voxel in the 3-D image is then selected randomly using a uniform random number generator to form a centroid of a ROI. Sample 3-D spatial distributions of ROI centroids are presented in Figure 1.

ROI growth model: Each ROI is generated, given its centroid and size, using a discrete-time 3-D version of the growth model by Eden [10]. Our model is an extension of the 2-D model described in [32] (the interested reader can check [26, 25, 15, 29] for more complicated growth models). The main idea of this model in two dimensions and a sample 3-D ROI generated using this model are presented in Figure 2(a) and (b) respectively. The “infection” process starts with one infected cell at time $t=0$ and progresses using the following rule: each infected grid cell may infect its six non-diagonal neighbors with probabilities p_N , p_S , p_E , p_W , p_U , and p_D , respectively, where the probabilities are not necessary equal. This rule is general enough to capture the possible anisotropy in growth due to properties of surrounding tissue. Each ROI is grown until it reaches a certain volume since its number of voxels is prespecified in our model. An example of simulated patient with artificially generated ROIs is presented in Figure 3.

Modeling of spatial normalization error

Prior to association mining, the images are being preprocessed; the ROIs from all patients’ images are placed in the same coordinate system via an elastic registration method. Although this procedure is very accurate, it is imperfect, i.e., it does not necessarily map corresponding regions to the same location in the space of the common spatial standard. Misregistration introduces noise, in the form of false-negative and false-positive associations. Here, we show how we model this important source of error so that we can analyze its effect on the data mining process.

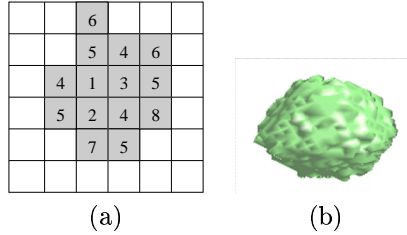


Figure 2. (a) Cell infection at time $t=8$ (2-D grid). (b) A sample 3-D ROI generated using the growth model.

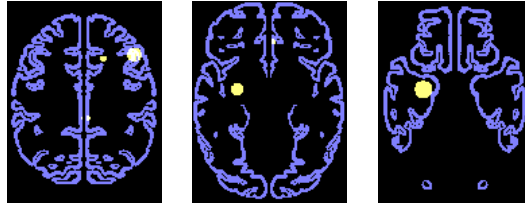


Figure 3. The artificial ROIs (in white) of a simulated patient (3 representative slices of the anatomical atlas are presented).

After identifying manually a number, N , of landmarks ⁶ to use, we calculate their coordinates in spatially normalized images from M patients as well as in the atlas itself. To reduce the variability of measurements obtained by different experts, we use the mean of a number of independent measurements for each landmark. Let $E_i^N, i : 1, \dots, M$ be a vector of dimension N with elements the Euclidean distances $e_{i,j}, j : 1, \dots, N$ between the landmarks and the corresponding displaced landmarks for patient i . The mean vector, μ , and the covariance matrix, Σ , are calculated from E_i^N . The displacement errors for the landmarks follow a normal distribution (this was verified with the Kolmogorov-Smirnov and Shapiro-Wilk tests of normality). If Σ is positive definite, a method that uses the Cholesky decomposition of Σ and N univariate normal variates can be used to produce an N -dimensional multivariate normal distribution [52], $\mathcal{N}_N(\mu, \Sigma)$ for the displacement error. Displacing a set of ROIs for a given patient can now be performed using a displacement produced from $\mathcal{N}_N(\mu, \Sigma)$. Each ROI's centroid is then displaced using an inverse distance-weighted markov-random-field equilibration from the displacements of the landmark points.

Generation of synthetic associations

We chose to model the conditional distributions among spatial regions and deficit variables by Bayesian Networks (BNs) [41]. This is a mature and well respected approach for modeling data in the machine learning community although only very

⁶The main criterion for choosing the landmarks is how reliably and accurately they can be identified and reproduced in an image, and how representative they are of the performance of the registration algorithm.

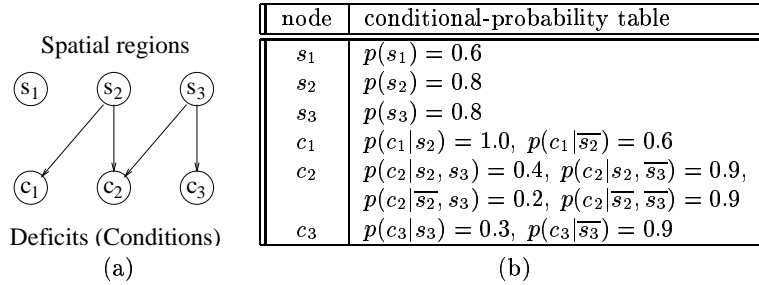


Figure 4. An example of a Bayesian network with 6 nodes (3 spatial regions and 3 deficits) and 4 edges from spatial regions to deficits, and the conditional-probability table for each node.

recently, has been used in the database/data mining community [4, 35]. In the BN model, as shown schematically in Figure 4, nodes represent variables of interest such as spatial regions or deficits, and edges represent associations among these variables. Each node has a conditional-probability table that quantifies the strength of the associations among that node and its parents. Providing the prior probabilities for the root nodes and conditional probabilities for every other node, we can derive all joint probabilities [41].

Because structure and deficit variables are categorical (in particular, binary) the BNs that we consider have 2 states for each node. The two states of spatial nodes correspond to a spatial region being of interest or not (i.e., being abnormal (s_i) or not ($\overline{s_i}$)), while, the two states of deficit nodes correspond to presence (c_k) or absence ($\overline{c_k}$) of a deficit. A BN with 6 nodes, 4 edges, and 2 states for each node is shown in Figure 4. Note that although in this paper we use discrete structure and function variables, BNs based on multivariate gaussian distributions [44] and mixed discrete continuous distributions [33] have been developed.

Parameters of the network are the number of spatial region and deficit nodes, the number of edges (associations), the prior probabilities of spatial regions being of interest, i.e., being abnormal, the conditional probabilities of deficits being present which denote the strength of associations between ROIs and deficits, and the degree of the network defined next. We use the term *degree* to refer to the number of ROIs affecting a particular deficit. Below, we refer to the *maximum degree of a BN*, which is the maximum of the degrees of its nodes. Moreover, in the case where all the nodes have the same degree, we denote this number as *degree of the BN*.

Using this model we can manipulate the numbers of region and deficit variables, the number and strengths of associations, the deficit(s) or abnormal condition(s) caused by abnormal regions, and a function mapping the abnormal fraction of a spatial region to the probability that this region will function abnormally.

A deficit associated with more than one spatial regions is represented by a noisy-OR model, which can model probabilistic disjunctive interactions among the causes of an effect [41]. The noisy-OR model is a boolean OR gate with a failure function associated with each input line – there is a *leak* probability q_i that line i will fail. When no failure occurs, each line’s input is passed to a boolean OR gate.

This overall structure induces a probability distribution which is easily computed; the probability of no failure occurring is denoted by p^{nf} : $p^{nf} = 1 - \sum_{i \in \mathcal{M}} q_i$, where \mathcal{M} is the subset of lines with activated input.

The priors of spatial region abnormality are calculated from the fraction of abnormal volume f_{s_i, ρ_j} for each spatial region s_i and patient ρ_j ; this quantity is defined as the volume of the abnormal part of s_i divided by the volume of s_i . The conditional probability of abnormality of a spatial region s_i given f_{s_i, ρ_j} , $p(s_i | f_{s_i, \rho_j})$, is expected to be a sigmoid function. One way to fit the sigmoid model is to compute $p(c | f_{s_i, \rho_j} < x)$ for various deficits and spatial regions in our data set. The sigmoid function can differ for different spatial regions. For simplicity a step function can often be used instead of a sigmoid function [36].

The prior probabilities for the abnormality of the spatial regions were computed by calculating for every region of the anatomical template, the percentage of patients in the simulated data set with abnormal areas in that region.

After applying the stepwise function of the fraction of abnormal volume for a spatial region to identify it as abnormal or not, we compute the fraction of patients with abnormality in this spatial region. This is repeated for all regions, i. e., we calculate for each region the prior probability of its being abnormal. Then, for every patient ρ_j and spatial region s_k we sample the prior probability and generate a binary vector S_j^K of dimension K (where $S_j[k] = 1$ means that the spatial region s_k is abnormal for patient ρ_j). By instantiating the state of the spatial nodes of the BN with S_j^K for patient ρ_j we get a binary vector C_j^M of dimension M for the deficits, where, $C_j[i] = 0$ if patient ρ_j has deficit c_i present.

3.2 Evaluating the Fisher exact test

As a case study, we examine the Fisher exact test of independence which is commonly used in detecting associations. Since this test can only be used for categorical variables we do not use directly the abnormal fraction of a spatial region in the analysis but rather we use the binary vector S_j^K for each spatial region and patient defining a spatial region as abnormal or not based on the fraction of it that is abnormal. The test is then applied to each spatial region–deficit variable pair (i.e., the binary vectors S_j^K and C_j^M). We use a threshold for the p -value and we report all the associations found with a p -value smaller than the threshold. Finally we compare the true associations with the ones that we discovered. Evaluating other methods that can be used with continuous variables (e.g., Mann-Whitney test), can be performed using the proposed framework, since the abnormal fraction of spatial regions can then be used directly.

4 Results and discussion

Here, we report and discuss the results of the evaluation of the the Fisher exact test of independence using our proposed evaluation framework. We measure the performance of this test in mining spatial region–deficit associations as a function of the number of samples (patients) needed to discover the existing associations represented by a BN; the strengths of associations; the number of associations; the

Table 3. *The 3 cases of BNs considered.*

Case	Association	$p(c s)$
1	Strong	1
2	Moderate	0.75
3	Weak	0.51

degree of the BN, i. e., the number of spatial regions related to a particular deficit; the prior probabilities of spatial regions being of interest, i.e., being abnormal. We also examine the effects of registration error.

To examine the effects of strengths of the spatial region-deficit associations on the ability of the mining methods to detect them, we consider three cases (presented in Table 3) which correspond to strong, moderate, and weak associations. For each case, all deficit nodes use the same conditional-probability table. The parameters we chose for the stochastically generated BN are the following unless otherwise stated: 132 spatial nodes (corresponding to the atlas structures), 20 deficit nodes, and 69 associations (edges) constrained to be from spatial regions to deficits. The maximum degree, i. e., the maximum number of incoming edges to deficit nodes, is restricted to be 4; thus, a deficit is related, at most, to 4 different spatial regions. The data set \mathcal{D} that was used as input to our simulator to extract the data distributions is taken from a study of traumatic brain lesions [20]. The normalization error distribution was determined from this data set by measuring the error on a number of distinct anatomical landmarks and interpolating for all other points. For the conditional probability of the abnormality of a spatial region a step function was used instead of a sigmoid. Plotting this conditional probability, i.e., $p(c | f_{s_i, \rho_j} < x)$, with the abnormal fraction of spatial region (in a ROC graph) for several deficits, and considering all spatial regions of the atlas and patients in the study, showed empirically that a step function with threshold fraction of 0.01 can be used for simplicity in the simulations, instead of a sigmoid function [36]. Note also that this threshold depends on the size of the abnormal regions in a given study. The threshold value 0.01 is also the mean of the optimal thresholds with respect to p -value, i. e., the mean of the thresholds that give the smallest p -value for all structures and deficits we considered. The prior probabilities for spatial regions being abnormal were set to 0.5 to show the behavior of the Fisher’s exact test for the optimal value of the priors under any value of the conditional probabilities. In the case where a deficit was related to more than one region, a noisy-OR model with leak probability 0.25 was used.

4.1 Choosing the p -value threshold

In Table 4 we present the performance change of the Fisher exact test with the threshold of p -value that is used. Generally, the lower the threshold for the p -value, the smaller the number of false positives and number of existing associations discovered (i. e., more conservative the method). Although these results are for the case of moderate strength of associations (case 2 of Table 3), similar results were observed for the cases of strong and weak associations. In the following experiments

Table 4. Percentage of existing associations (t_a) and false positives (f_a) discovered by the Fisher exact test for three values of the p -value threshold and for moderate strength of spatial region-deficit associations.

# samples	$p \leq 0.01$		$p \leq 0.001$		$p \leq 0.0001$	
	t_a	f_a	t_a	f_a	t_a	f_a
500	83	35	71	4	55	0
1000	100	43	99	1	88	0
1500	100	35	100	3	97	0
2000	100	30	100	1	100	0

we use the threshold 0.001 for the p -value, since this is a good trade-off between the number of existing associations and the number of false positives discovered.

4.2 Effect of conditional probabilities

Having determine the threshold (0.001) of p -value to use, in Figure 5(a) we present the performance of the Fisher exact test ($p \leq 0.001$) for the case of strong, moderate, and weak associations. The graphs show the dramatic effect of the conditional probabilities on the ability of the test to detect the causal relationships. In order to discover 70% of the total number of existing associations, we need approximately 180, 500, and 2000 samples for the strong, moderate, and weak associations, respectively. As expected, the number of samples needed is inversely proportional to the strength of associations.

4.3 Varying the number and degree of associations

Here we investigate the effect of (a) the number of associations and (b) their degree in detecting them. We select the moderate case for the conditional probabilities (i. e., case 2). To study the effect of the number of associations, without loss of generality, we consider the case where all the nodes have the same degree (defined as degree of the BN). In Figure 6(a) we present the performance of the Fisher exact test for three BNs of degree 4 with a different number of associations. It is evident from this figure that the Fisher exact test performs similarly for BNs with different number of associations but of the same degree. The deterioration as the number of associations increases is small.

We also study the effect of the number of spatial regions affecting a particular deficit, or degree of the BN, assuming that all the nodes have the same degree. Figure 6(b) shows the effect of increasing the degree of the BN while keeping the number of associations the same. As expected, the higher the degree of the BN, the more samples are needed to detect the same number of associations. Also, the degree of the BN has much greater effect on the performance of the test than the number of associations. Since multivariate associations may not be detected by repeated application of bivariate tests, such as the Fisher exact test, development of different methods which will seek multivariate associations [7] is necessary.

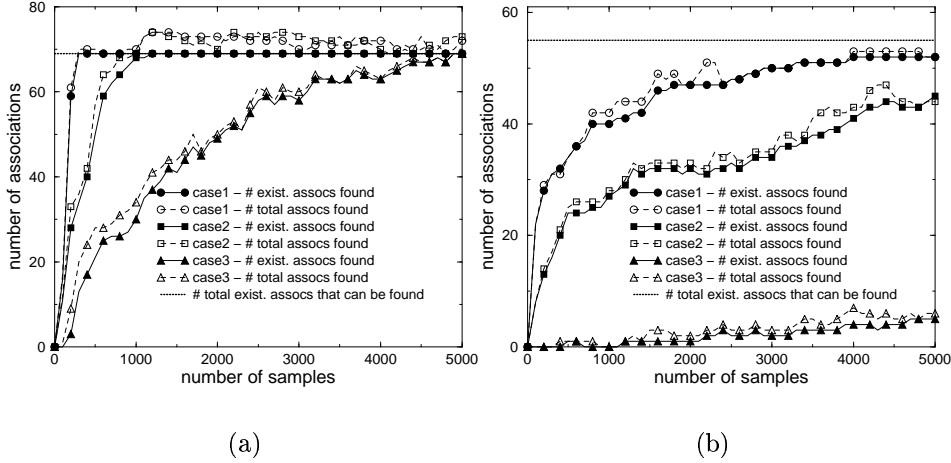


Figure 5. Performance of the Fisher exact test ($p \leq 0.001$) for (a) uniform (0.5) and (b) non-uniform prior probabilities of spatial region abnormality calculated from a simulated data set, and for the 3 cases of BNs of Table 3 that correspond to strong (case 1), moderate (case 2) and weak (case 3) associations. The difference between the total number of associations and the number of existing associations discovered is the number of false positives in each case.

4.4 Varying the spatial prior probabilities

In all the experiments presented so far the prior probabilities for the spatial regions being abnormal were set to 0.5 to evaluate the performance of the Fisher exact test for the optimal value of priors under any conditional probability table. In this experiment we obtain the priors of spatial region abnormality from the simulated regions data set as was described in Section 3.1. Since there are 14 associations from spatial regions (structures) that do not intersect any abnormal region, the number of associations that can be actually discovered is 55. The smallest nonzero prior is 0.0004 and only 5 out of the 132 structures have priors above 0.2 in this case. Figure 5(b) demonstrates the performance of the Fisher exact test for the 3 cases of BN conditional probabilities (see Table 3). Comparing this figure with Figure 5(a) shows that the number of samples required to detect all associations is much larger than in the simplified case of uniform 0.5 priors. This is due to the fact that some prior probabilities are very small. As expected, the number of samples needed is inversely proportional to the smallest prior probability. The detection of false-positive associations is due to the existence of associations between (neighboring) spatial regions. These associations are due to abnormal regions that intersect with more than one structure. Additional false positives can be observed in the case where there are associations between the deficit variables. Finally, the limitation imposed by the prior probabilities can be ameliorated, since for a given collection of images, we know that results associating spatial regions with prior probabilities closer to 0.5 are more likely to be true associations.

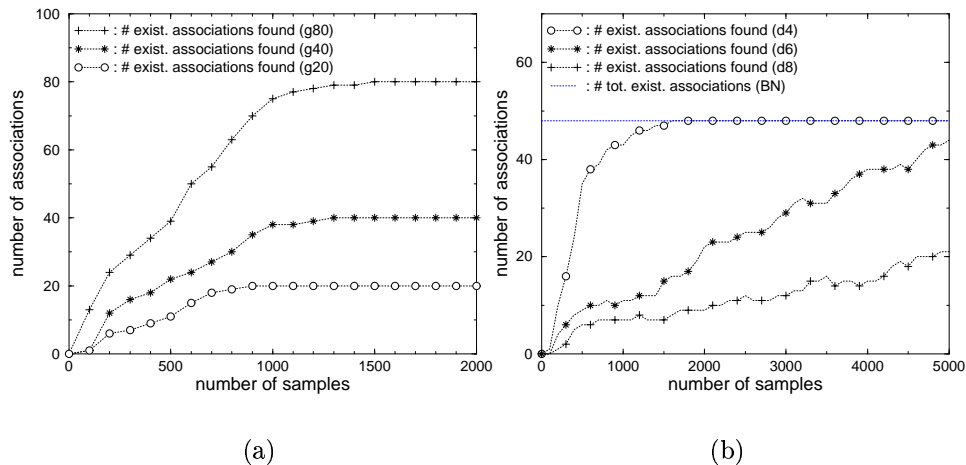


Figure 6. Varying (a) the number and (b) the degree of associations: Performance of the Fisher exact test for (a) three different BNs (of degree 4) with 20 (g20), 40 (g40), and 80 (g80) associations, and (b) three different BNs with 48 associations and of degree 4 (d4), 6 (d6), and 8 (d8) respectively.

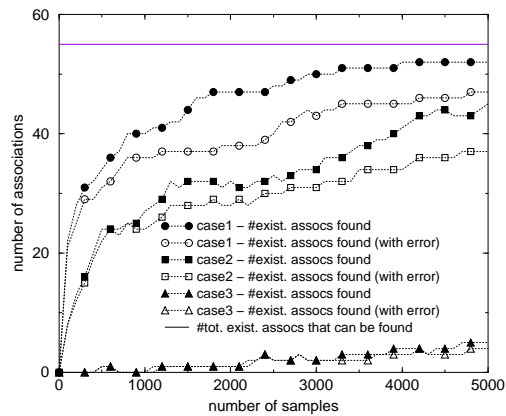


Figure 7. Performance of the Fisher exact test ($p \leq 0.001$) with and without registration error, for strong (case 1), moderate (case 2), and weak (case 3) associations (a maximum of 55 out of a total of 69 associations can be discovered).

4.5 Effect of registration error

Figure 7 compares the performance of the Fisher exact test for strong (case 1), moderate (case 2), and weak (case 3) associations showing the effect due to the imperfect normalization method used (this is actually a parameter in our model; various normalization methods can be compared). As expected, the registration error (of the particular nonlinear registration method we considered) reduces con-

siderably the power of the Fisher exact test in detecting the associations, when compared with perfect registration.

5 Conclusions

We have proposed a method to evaluate the performance of mining methods for 3-D medical image data and presented as a case study an analysis of the Fisher exact test of independence. Analyzing simulated data, we demonstrated that the number of samples needed to detect all the associations while reducing false positives has an inverse relationship to the strength of associations and the smallest prior probability of a spatial region being of interest. The more we descend from the 0.5 level for prior probabilities, the more difficult it becomes to discover associations. The degree of associations, i. e., the number of spatial regions associated with a particular deficit or medical condition, has much greater effect on the performance of the statistical test used here, than does the number of associations to discover. This is intuitively expected, although our simulations quantify it. Registration error also reduces considerably the power of the statistical test in detecting the associations; this simulator allows us to take this error into account when calculating the sample size needed for a particular experiment.

The major contributions of this paper are:

- The design of a spatial simulator that models spatial regions of interest, deficits (or conditions) and associations between them, and the effects of registration, noise, and segmentation of ROIs.
- An example use of the spatial simulator in evaluating the Fisher exact test of independence, one of the methods available for the detection of associations between spatial and non-spatial data.
- The study of recovery of associations as a function of the number of samples needed, the strength and number of associations, the number of spatial ROIs associated with non-spatial data, the prior probabilities of spatial regions being of interest, and the registration error.
- The use of Bayesian networks in an “inverse mode” i.e., in generating data with associations between them that are precisely modeled so that to evaluate other association mining methods.
- The introduction of interesting problems from the brain imaging domain to data mining researchers.

Future work includes (a) the use of this simulator in the evaluation of other statistical and non-statistical methods used in 3-D image data mining, (b) the study of the effect of different spatial normalization and segmentation algorithms in the mining process, and (c) the extension of the simulator to other spatial and spatial-temporal data and associated mining methods. This work is expected to have wide applications in evaluating knowledge that is being discovered in geographic information systems (GIS), environmental studies, remote sensing, CAD, navigation, medical imaging and many other areas where spatial data are used.

Bibliography

- [1] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD Conference*, pages 207–216, Washington D.C., USA, May 1993.
- [2] E. Andersen. *Introduction to the Statistical Analysis of Categorical Data*. Springer Verlag, Berlin, 1997.
- [3] M. Arya, W. Cody, C. Faloutsos, J. Richardson, and A. Toga. A 3D Medical Image Database Management System. *Int. Journal of Computerized Medical Imaging and Graphics, Special issue on Medical Image Databases*, 20(4):269–284, Apr. 1996.
- [4] S. Babu, M. Garofalakis, and R. Rastogi. SPARTAN: A Model-Based Semantic Compression System for Massive Data Tables. In *Proceedings of the ACM SIGMOD 2001*, pages 283–294, May 2001.
- [5] B. M. Bennett and P. Hsu. On the power function of the exact test for the 2x2 contingency table. *Biometrika*, 47(3,4):393–398, 1960 [Correction 48 (1961), p.475].
- [6] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, volume 26,2 of *SIGMOD Record*, pages 265–276, New York, May13–15 1997. ACM Press.
- [7] G. F. Cooper and E. H. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.
- [8] C. Davatzikos. Spatial transformation and registration of brain images using elastically deformable models. *Comp. Vision and Image Understand.*, 66(2):207–222, 1997.
- [9] S. Dutta. Qualitative Spatial Reasoning: A Semi-quantitative Approach Using Fuzzy Logic. In *Proceedings of the 1st Symp. SSD'89, Santa Barbara, CA*, pages 345–364, Jul. 1989.
- [10] M. Eden. A two-dimensional growth process. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, CA*, 1961.

- [11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, 1996*.
- [12] A. Evans, S. Marrett, J. Torrescorzo, S. Ku, and L. Collins. MRI-PET correlation in three dimensions using a volume-of-interest (VOI) atlas. *Journal of Cerebral Blood Flow Metabolism*, 11:A69–78, 1991.
- [13] U. M. Fayyad and P. Smyth. Image Database Exploration: Progress and Challenges. In *Proceedings of the 1993 Knowledge Discovery in Databases Workshop, Washington, D.C., July 1993*.
- [14] R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1934.
- [15] W. F. Forbes and R. W. Gibberd. Mathematical models of carcinogenesis: A review. *The Mathematical Scientist*, 9:95–110, 1984.
- [16] S. Fotheringham and P. Rogerson. *Spatial Analysis and GIS*. Taylor and Francis, 1994.
- [17] P. Fox. Functional brain mapping with positron emission tomography. *Seminars in Neurology*, 9:323–329, 1989.
- [18] Y. X. Fu and J. Arnold. A Table of Exact Sample Sizes for Use with Fisher's Exact Test for 2x2 Tables. *Biometrics*, 48(4):1103–1112, Dec. 1992.
- [19] M. Gail and J. J. Gart. The determination of sample sizes for use with the exact conditional test in 2x2 comparative trials. *Biometrika*, 29:441–448, Sept. 1973.
- [20] J. Gerring, K. Brady, A. Chen, C. Quinn, K. Bandeen-Roche, M. Denckla, and R. Bryan. Neuroimaging Variables Related to the Development of Secondary Attention Deficit Hyperactivity Disorder in Children who have Moderate and Severe Closed Head Injury. *Journal of the American Academy of Child and Adolescent Psychiatry*, 37:647–654, 1998.
- [21] R. H. Gting. A dynamic index structure for spatial searching. In *Proceedings of the ACM SIGMOD Int. Conf. on Management of Data, Boston, MA*, pages 47–57, 1984.
- [22] R. H. Gting. An introduction to spatial database systems. *VLDB Journal*, 3(4):357–400, Oct. 1994.
- [23] J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. *IEEE Trans. Knowledge and Data Eng.*, 5:29–40, 1993.
- [24] J. Han and M. Kamber. *Data Mining*. Morgan Kaufmann, 2000.

- [25] F. Hanson and C. Tier. A stochastic model of tumor growth. *Mathematical Biosciences*, 61:73–100, 1982.
- [26] T. E. Harris. Contact interactions on a lattice. *Annals of Probability*, 2:969–988, 1974.
- [27] M. Heath, K. Bowyer, D. Kopans, and et al. Current status of the digital database for screening mammography. In *Digital Mammography*, pages 457–460. Kluwer Academic Publishers, 1998.
- [28] Y. Hochberg and A. Tamhane. *Multiple Comparison Procedures*. John Wiley and Sons, New York, 1987.
- [29] A. R. Kansal, S. Torquato, G. R. Harsh, E. A. Chiocca, and T. S. Deisboeck. Simulated Brain Tumor Growth using a Three-Dimensional Cellular Automaton. *Journal of Theoretical Biology*, 203(367), 2000.
- [30] E. Knorr and R. Ng. Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining. *IEEE Trans. Knowledge and Data Engineering*, 8(6):884–897, Dec. 1996.
- [31] K. Koperski and J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. In *Proceedings of the 4th International Symposium on Large Spatial Databases (SSD'95), Portland, Maine*, Aug. 1995.
- [32] F. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas. Fast and Effective Retrieval of Medical Tumor Shapes. *IEEE Trans. on Knowledge and Data Engineering*, 10(6):889–904, 1998.
- [33] S. L. Lauritzen and N. Wermuth. Graphical Models for associations between variables, some of which are qualitative and some of which are quantitative. *The Annals of Statistics*, 17:31–57, 1989.
- [34] S. Letovsky, S. Whitehead, C. Paik, G. Miller, J. Gerber, E. Herskovits, T. Fulton, and R. Bryan. A brain-image database for structure-function analysis. *American Journal of Neuroradiology*, 19(10):1869–1877, 1998.
- [35] D. Margaritis, C. Faloutsos, and S. Thrun. NetCube: A Scalable Tool for Fast Data Mining and Compression. In *Proceedings of the 27th Int. Conference on Very Large Data Bases (VLDB)*, pages 311–320, Sept. 2001.
- [36] V. Megalooikonomou, C. Davatzikos, and E. Herskovits. A Simulator for Evaluation of Methods for the Detection of Lesion-Deficit Associations. *Human Brain Mapping*, 10(2):61–73, 2000.
- [37] V. Megalooikonomou, C. Davatzikos, and E. H. Herskovits. Mining lesion-deficit associations in a brain-image database. In *Proceedings of the the 5th International Conference on Knowledge Discovery and Data Mining (KDD-99), San Diego, CA*, pages 347–351, 1999.

- [38] N. Megiddo and R. Srikant. Discovering Predictive Association Rules. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pages 274–278, New York City, NY, Aug. 1999.
- [39] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. In *Proceedings of the 1994 Int. Conf. Very Large Data Bases (VLDB)*, pages 144–155, Santiago, Chile, Sept. 1994.
- [40] G. Osius and D. Rojek. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association*, 87(420):1145–1152, Dec. 1992.
- [41] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.
- [42] G. Piatetsky-Shapiro and e. W. J. Frawley. *Knowledge Discovery in Databases*. AAAI/MIT Press, Menlo Park, CA, 1991.
- [43] H. Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, 1990.
- [44] R. D. Shachter and C. R. Kenley. Gaussian influence diagrams. *Management Science*, 35:527–550, 1989.
- [45] E. C. Shek, R. R. Muntz, E. Mesrobian, and K. Ng. Scalable Exploratory Data Mining of Distributed Geoscientific Data. In *Proceedings of the 2nd International Conference on Data Mining (KDD-96), Portland, Oregon*, pages 32–37, Aug. 1996.
- [46] P. Smyth, M. C. Burl, U. M. Fayyad, and P. Perona. Knowledge Discovery in Large Image Databases: Dealing with Uncertainties in Ground Truth. In *Proceedings of the AAAI-94 workshop on KDD, Seattle, WA, July 1994*.
- [47] E.-J. Son, I.-S. Kang, T.-W. Kim, and K.-J. Li. A spatial data mining method by clustering analysis. In *Proceedings of the Sixth International Symposium on Advances in Geographic Information Systems, GIS'98*, pages 157–158, 1998.
- [48] P. Stolorz and C. Dean. Quakefinder: A Scalable Data Mining System for Detecting Earthquakes from Space. In *Proceedings of the 2nd International Conference on Data Mining (KDD-96), Portland, Oregon*, pages 208–213, Aug. 1996.
- [49] J. Talairach and P. Tournoux. *Co-planar Stereotaxic Atlas of the Human Brain*. Thieme, Stuttgart, 1988.
- [50] H. Tanizaki. Power comparison of non-parametric tests: Small-sample properties from Monte Carlo experiments. *Journal of applied statistics*, 24(5):603–632, Oct. 1997.

- [51] R. Thomas and M. Conlon. Sample-size determination based on Fisher exact test for use in 2x2 comparative trials with low event rates. *Controlled clinical trials*, 13(2):134–147, 1992.
- [52] Y. L. Tong. *The Multivariate Normal Distribution*. Springer-Verlag, 1990.