Published in final edited form as:

J Intell Inf Syst. 2010 December 1; 35(3): 465–494. doi:10.1007/s10844-009-0115-6.

A performance evaluation framework for association mining in spatial data

Qiang Wang and Vasileios Megalooikonomou

Data Engineering Laboratory, Department of Computer and Information Sciences, Temple University, 415 Wachman Hall, 1805 N. Broad Str., Philadelphia, PA 19122, USA

Qiang Wang: qwang@lucas.cis.temple.edu; Vasileios Megalooikonomou: vasilis@temple.edu

Abstract

The evaluation of the process of mining associations is an important and challenging problem in database systems and especially those that store critical data and are used for making critical decisions. Within the context of spatial databases we present an evaluation framework in which we use probability distributions to model spatial regions, and Bayesian networks to model the joint probability distribution and the structural relationships among spatial and non-spatial predicates. We demonstrate the applicability of the proposed framework by evaluating representatives from two well-known approaches that are used for learning associations, i.e., dependency analysis (using statistical tests of independence) and Bayesian methods. By controlling the parameters of the framework we provide extensive comparative results of the performance of the two approaches. We obtain measures of recovery of known associations as a function of the number of samples used, the strength, number and type of associations in the model, the number of spatial predicates associated with a particular non-spatial predicate, the prior probabilities of spatial predicates, the conditional probabilities of the non-spatial predicates, the image registration error, and the parameters that control the sensitivity of the methods. In addition to performance we investigate the processing efficiency of the two approaches.

Keywords

Data mining; Association discovery; Spatial databases; Performance evaluation; Simulation and modeling

1 Introduction

Current technology has made enormous amounts of spatial data available from various domains, including weather forecasting, medicine, remote sensing, and environmental monitoring. Spatial data mining (Koperski and Han 1995; Shekhar et al. 2004) refers to the discovery of spatial relationships and relationships between spatial and non-spatial data or to other interesting patterns not explicitly stored in spatial databases. Recent advances in spatial databases (Samet 1990; Gueing 1994) have led to the study of spatial data mining. Statistical methods (Fotheringham and Rogerson 1994) were the first applied to spatial data. Initially these methods did not work well 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 (Piatetsky-Shapiro and Frawley 1991; Agrawal et al. 1993) 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 (Ester et al. 2001; Gorsevski et al. 2000), spatial computation techniques for mining spatial association rules (Koperski and Han 1995; Karasov et al. 2005), proximity techniques for finding characteristics of spatial clusters (Knorr and Ng 1996), spatial clustering (Ester et al. 1996; Son et al. 1998; Han et al. 2001) and spatial co-locations (Morimoto 2001), among others.

Although the term “spatial data” refers mostly to geographic information databases in the computer science literature, due to the numerous applications in this domain, we consider spatial data using the more general definition that includes 2-Dimensional (2-D) and 3-Dimensional (3-D) images. Examples of 3-D images are those obtained from various medical and biological imaging modalities that capture 3-D images that contain structural, functional/physiological, or other information about tissue, internal structures, and anatomy. Mining in 3-D image databases (Smyth et al. 1994; Megalooikonomou et al. 1999, 2000b) can be seen as a special case of spatial data mining. Other examples include 3-D meteorological data collected using advanced sensor and satellite technology, 3-D seismic or other geophysical data, etc. Challenging issues in the mining of spatial databases (Han and Kamber 2000) include: (a) the exploration of statistical dependence that exists between neighboring objects; (b) the large dimensionality of the data involved; and (c) the efficient management of topological or distance information through spatial reasoning, spatial knowledge representation techniques and multidimensional spatial access methods (SAM).

Although much 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 parameters such as sample size has not yet been addressed sufficiently. This problem becomes even more important when dealing with critical data coming from medical studies, environmental studies, navigation, remote sensing, and other important fields. The requirements for a satisfactory way of performing this evaluation are complex: (a) the ground truth (e.g., 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 performance evaluation work has been done on non-spatial data. Megiddo and Srikant (1999), used simulations to determine the threshold for the p-value for associations. They have shown that the support and confidence thresholds prune out most rules that are not statistically significant. In addition, several researchers have systematically studied the problem of sample size corresponding to power of statistical tests (Bennett and Hsu 1960; Gail and Gart 1973) (such as the chi-square and Fisher exact tests of independence), compared with other tests (Tanizaki 1997) and performed simulations to study the power of statistical tests in sample spaces of much higher dimensionality (Osious and Rojek 1992; Tanizaki 1997).

Evaluating mining methods that are applied on spatial 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 region boundary identification) on the performance of the mining methods (Tan et al. 2001) 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 to the evaluation problem does not appear to be possible.

This work is grounded by 3-D medical image database systems (Arya et al. 1996; Letovsky et al. 1998; Armato III et al. 2004) that include a large collection of 3-D images from various studies and different medical imaging modalities and that were built to discover structure-function relationships as well as a set of anatomical templates. Each such study typically involves hundreds or even thousands of samples. The amount of data per sample can easily reach hundreds of megabytes, depending on the kind of study and the spatial and contrast resolutions used. For example, functional-Magnetic Resonance Imaging (fMRI), which shows physiological activity in the brain, requires time-sequences for each sample and each task performed. Adding these numbers together we find that in such environments terabytes of data are not unusual. Megalooikonomou et al. (1999) have presented a data mining process to discover associations between spatial ROIs and clinical conditions. In that paper, several methods were proposed including the use of the chi-square test, Fisher exact test, voxel-based logistic regression, and clustering analysis.

In this paper, we do not propose new spatial data mining methods. Instead, we propose a framework for evaluating spatial data mining methods for the detection of associations between spatial and non-spatial data. The methods are presented in the general framework of spatial databases. Our approach is to design a simulator in order to generate a large number of artificial samples (including spatial data, spatial and non-spatial predicates, and associations among them), apply a spatial association mining method, and compare the true associations with the ones that the method discovered. The simulator creates 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 association mining algorithms.

The remainder of the paper is organized as follows: In Section 2 we present background information. In Section 3 we describe the method used for the evaluation of spatial association mining techniques and we present in detail the various components of a spatial simulator. In Section 4, we demonstrate the use of our evaluation framework by presenting, as a case study, experimental results from the evaluation of two approaches used for learning associations, the Bayesian method and dependency analysis (using statistical tests of independence). In particular, we evaluate a Bayesian network learning technique based on the Minimum Description Length (MDL) score and greedy hill-climbing and a constraint-based technique that uses the Fisher exact test of independence to discover conditional independence properties among data. Finally, the paper concludes with a short discussion in Section 5.

2 Background

Spatial data mining has been categorized based on the possible types of rules one can discover in spatial data (Koperski and Han 1995): (a) spatial discriminant rules, (b) spatial associations and (c) spatial characteristic rules. The first refers to the discovery of spatial features and patterns that are discriminative among classes. The second refers to the elucidation of implications and associations among features. The third refers to the description of general patterns in a set of spatially related data. Here, we focus on spatial association discovery, although the evaluation framework we propose may also be useful for spatial discriminative and characteristic rule discovery.

A spatial association is a rule of the form:

$$s_1 \wedge \dots \wedge s_k \rightarrow c_1 \wedge \dots \wedge c_m$$

where at least one of the predicates $s_1, \dots, s_k, c_1, \dots, c_m$ is a spatial predicate (see Table 1 for a brief description of the notation we use in the paper). A spatial predicate is a statement about a spatial object that attributes a property to it. Various kinds of such predicates are involved in spatial associations representing topological relationships between spatial objects, spatial orientation (or ordering), or contain distance information. Examples of spatial predicates are “close to,” “intersects,” and “inside/outside.” Spatial objects can have other properties as well. For example, they can be of interest or not (e.g., a region in a medical image being abnormal). As defined in Koperski and Han (1995), a spatial association rule can be of two different forms: (a) non-spatial consequent with spatial antecedent(s) and (b) spatial consequent with non-spatial/spatial antecedent(s). In this paper, for the purpose of presentation of the simulator we deal only with associations rules of the first form. The simulator can be extended to generate associations of the second form.

We consider the following example of associations from the medical imaging domain:

$$\text{intersects}(i, \text{thalamus}) \wedge \text{abnormal}(i) \rightarrow \text{present}(k)$$

where i is a spatial region of interest (where abnormality is present) in a medical image (e.g., due to the existence of a lesion in the region), thalamus is a known area of a brain atlas and k is a deficit. The top part of Fig. 5 shows the result of the intersection of a region of interest (i.e., a lesion) with a brain atlas that provides prior information about the locations of brain structures. The presence of a lesion in that location may or may not be associated with one or more deficits.

In order to analyze large sets of spatial data and discover associations and patterns among spatial predicates and among spatial and non-spatial predicates one first needs to make data comparable across samples. Consider the case of satellite images where, for each region, several images from different sensors are available for analysis. In order to consider these images two preprocessing steps need to be performed. The spatial regions that are of interest must be identified (segmented) first (i.e., their boundaries must be delineated). This is done using manual, semiautomatic or automatic techniques. The next step is to perform image registration in order to map homologous regions to the same location in a common spatial standard or template (i.e., a map or an atlas). A map models the exact shapes and positions of regions. This step brings images of the same region in spatial coincidence with each others and with a template. An image does not identify the specific region to which each pixel or voxel (volume element) in 3-D belongs, but a map can provide this information with the accuracy of the registration methods when overlaid on the image. Several linear and nonlinear image registration methods have been developed. In the discussion that follows, we assume that spatial regions that are of interest have been segmented and registered to a map.

Although the evaluation framework we propose in this paper can be applied to the study of various methods, either statistical or non-statistical, in the results section we present as a case study the analysis of two methods used for learning associations. Here, we provide the necessary background for both.

Learning associations from data is a very challenging problem that has received much recent attention. Two kinds of approaches proposed for this purpose are as follows:

1. *Bayesian or scoring-based* methods, where a score function such as the Minimum Description Length (MDL) score is used to evaluate the fitness of a probabilistic network structure to the data, and,
2. *Dependency analysis or constraint-based* methods, where a statistical measure of conditional independence such as the chi-square, Fisher's exact test or mutual information is used to discover conditional independence properties among data.

The *Bayesian Network* (BN) technique for discovering associations is based on algorithms for learning the Bayesian network structure. Basically, there are two different groups of algorithms for structure learning: (a) constraint-based, e.g. the PC algorithm (Spirtes et al. 2000) and (b) score and search based. In this paper, we concentrate on the latter. The score and search approach is performed using a search through the space of possible network structures (directed acyclic graphs (DAGs)). Since the number of DAGs is super-exponential in the number of nodes (Cooper 1999), exhaustively searching the space is not practical and a local search algorithm (e.g., greedy hill climbing perhaps with multiple restarts) or a global search algorithm (e.g., Markov Chain Monte Carlo) is used. In addition to the search procedure, one must specify a scoring function and for that there are two popular choices, the Bayesian score and the Bayesian Information Criterion (BIC).

The Bayesian score integrates out the parameters and is also known as the marginal likelihood of the model. Since the calculation of Bayesian score involves difficult integrals, in practice, another simpler score, BIC is more frequently used. BIC can be derived as a large sample approximation to the marginal likelihood. It is defined as Bogdan et al. (2004)

$$\log P(\mathcal{D}|\hat{\theta}) - \frac{\log N}{2}\xi,$$

where $\mathcal{D}$ is the data set, N is the number of samples in the data set, ξ is the dimension of the model (i.e., the number of parameters) and $\hat{\theta}$ is the maximum likelihood (ML) estimate of the parameters. The second term is a penalty for model complexity. The number of parameters in the model is the sum of the number of parameters in each node. The BIC method has the advantage of not requiring a prior. BIC can be derived as a large sample approximation to the marginal likelihood. It is also equal to the Minimum Description Length of a model.

The *chi-squared test* was proposed in the data mining community (Brin et al. 1997) to measure the significance of associations. This test is used in the analysis as follows: for each spatial–non-spatial predicate pair a contingency table that shows all possible outcomes is constructed and the variables are examined to see whether they are independent of each other. Since computing a statistic for many pairwise tests creates the multiple-comparison problem (Hochberg and Tamhane 1987), i.e., a certain portion of tests will be positive by chance, the Bonferroni correction (Andersen 1997) or heuristic variations such as the sequential bonferroni correction are usually applied. A 2×2 contingency table consisting of four cells for a spatial predicate s_i and a non-spatial predicate 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) of the contingency table. 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 us how rarely we 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 (Brin et al. 1997) the chi-squared test is usually replaced by the Fisher exact test (Fisher 1934) 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. Note that while Fisher's exact test is a conditional test, there are alternatives for an unconditional test, e.g. Barnard's exact test (Barnard 1947). Even though Barnard's test is considered more powerful than Fisher's test, it is not as widely used due to its computational cost. In the rest of the paper, we discuss only Fisher's exact test.

In previous work (Megalooikonomou et al. 2000a) an approach based on a lesion-deficit simulator was used to evaluate a statistical test of independence in the brain imaging domain. The approach was extended in Megalooikonomou (2002) for different types of regions of interest in medical image databases. The error associated with registration was also modeled. Here, we present a general evaluation framework for association mining in spatial databases. We study spatial and non-spatial predicates and focus on employing the proposed framework to systematically compare the performance and efficiency of two types of methods used to detect associations (i.e., Bayesian and dependency analysis methods).

3 Proposed method

We designed a general purpose spatial simulator named Spatial Data Simulator (SDS) to generate a large number of samples consisting of images that include spatial regions, spatial predicates, non-spatial predicates and associations among them, all conforming to predefined distributions. The spatial predicates are produced automatically after overlaying the spatial regions on a map and then calculating the exact positions of the regions as well as intersections with known map areas. Then, we analyze the generated spatial regions, and spatial and non-spatial predicates by applying association 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.

Depending on the partitioning of the space we consider (i.e., coarse to fine partitioning) the number of spatial regions can range from the number of spatially distinct areas of a map to the number of voxels or pixels in a 3-D volume or image. In the first case, a number of neighboring voxels are grouped together and considered as one spatial object. The grouping can be done based on prior knowledge (e.g., taking into account the states of a country in a geographic map). A map-based approach allows the incorporation of more domain knowledge than the voxel-based approach. This approach reduces the dimensionality of the space by grouping neighboring voxels to form a much smaller number of spatial regions that

correspond to the map structures (Megalooikonomou et al. 1999). Thus, it is also less computationally demanding. On the other hand, results from a map-based approach are at best as good as the map itself. In the second case, every individual voxel in a volume is used as a spatial region. Due to the increased complexity introduced in the probabilistic network model when dealing individually with each voxel (each having a prior probability and possibly affecting a spatial or non-spatial predicate), we focus on modeling and analyzing spatial regions instead of individual voxels.

This work extends previous work on using simulators for evaluating the performance of association mining techniques (Megalooikonomou et al. 2000a; Megalooikonomou 2002). Here, we consider general spatial data rather than medical image data. We study spatial and non-spatial predicates. Moreover, we introduce Bayesian network learning and present a study on the selection of the parameters of Bayesian network learning. Although in Megalooikonomou et al. (2000a), Megalooikonomou (2002) the proposed framework was used to evaluate the Fisher's exact test of independence, here, we demonstrate the general applicability of the proposed framework by evaluating two approaches, Bayesian learning and dependency analysis based on statistical tests of independence and by considering two different application domains, biomedical imaging and meteorology. In addition, we provide an extensive comparison of the performance and efficiency of the two approaches including the study of linear and nonlinear associations.

In the remainder of this section, we first describe the spatial simulator SDS in more detail. Then, we describe two association mining techniques, which will be evaluated using the proposed framework and serve as case studies for our methodology.

3.1 Spatial Data Simulator - SDS

We use our framework to characterize the power of spatial association mining methods employed to detect multivariate associations among spatial predicates and among spatial and non-spatial predicates. To accomplish this task, we must be able to produce artificial samples, each of which consists of an image with regions that are of interest (e.g., highly populated areas in geographic maps, or abnormal areas in medical images), a set of non-spatial predicates, and a set of associations among spatial and non-spatial data. The goal of SDS is to generate samples such that both spatial and non-spatial predicates conform to predefined distributions. The predefined distributions can be determined from previously analyzed data sets, empirical distributions based on domain knowledge, or published information. The SDS algorithm shown in Fig. 1 shows the steps used for generating the simulated data.

3.1.1 Extraction of probability distributions from data and generation of simulated regions—For each artificial sample, SDS generates spatial regions whose shape, location, and distribution are given by certain parameters. These regions are the initial regions of interest (ROIs) and their parameters can be either given as fixed inputs (input_stat in Algorithm SDS) or obtained from real studies (extract_statistics). Given a data set $\mathcal{D}$ obtained from a real study (e.g., a set of medical images of a patient group), SDS collects statistics from it on the number of regions per sample, their sizes, and their spatial distribution. From these statistics, the simulator forms probability density functions. Then, for each sample, it performs random draws from these distributions, generating the number of regions, and, for each region, its centroid and size.

To generate artificial samples of spatial ROIs, SDS needs to have the knowledge of the spatial distribution of ROI centroids and the growth model which ultimately decides the shape and size of the ROIs. To obtain the probabilities of each voxel of the volume to be the centroid of an ROI, SDS goes through several steps including smoothing, padding and

normalization to get a proper probability function (see Megalooikonomou 2002 for more details). This function is a smoothed version of the 3D histogram of centroids and has a nonzero value on every point of the 3D volume. Sample spatial distributions of ROIs are presented in Fig. 2.

Once the number of ROIs, their sizes and the locations of their centroids have been calculated from probability density functions previously computed from $\mathcal{D}$, a growth model is required to generate ROIs of different shapes. The model used in SDS is an extension of the 2-D model described in Korn et al. (1998) (Hanson and Tier 1982, Kansal et al. 2000 show complicated growth models). The main idea of this model in two dimensions and a sample 2-D ROI generated using this model are presented in Fig. 3(a), (b) and (c). The growth starts with one active grid cell at time $t=0$; the active grid cell may infect its neighbors with different probabilities.

The boundaries of the regions of interest that are created by SDS are precisely known (even in the case of simulated misregistration). The boundaries of the spatial structures in the map are also well defined by the atlas. However, the exact approach used to define the boundary of a region in real data is domain dependent. This should be performed before acquiring statistics from real data. In our experiments, the regions of interest in real data (i.e., medical images) where the distributions are coming from, have been outlined manually by an expert.

3.1.2 Introduction of image registration error—To reflect the fact that there is no perfect data in a real life applications, SDS is also able to model the error (intro_errors in Algorithm SDS) introduced by image registration. Prior to association mining, the images are preprocessed to convert the regions from all the samples' images to the same coordinate system via a registration method. Although this procedure is very precise, 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 among spatial and among spatial predicates and non-spatial predicates. In order to account for the error that is introduced when images are registered to a map (template image), the spatial regions created in the artificial samples can be displaced according to the distribution of registration error for the spatial-normalization method that is used.

When performing image registration, we need to define a number of landmarks. Landmarks are points that can be easily identified on different images and assist in mapping corresponding points from different images to each other. To obtain the data with noise, with L landmarks defined, an L -dimensional multivariate normal distribution, $\mathcal{N}_L(\mu, \Sigma)$, for the registration error is calculated first (Megalooikonomou 2002). Displacing a set of ROIs for a given sample can then be performed using a displacement produced from $\mathcal{N}_L(\mu, \Sigma)$. Each ROI's centroid is displaced using an inverse distance-weighted markov-random-field equilibration from the displacements of the landmark points. An example of a simulated sample with artificially generated regions along with their displaced counterparts (with Gaussian noise) is presented in Fig. 4.

3.1.3 Generation of synthetic associations—After modeling the spatial ROIs with the extracted statistics (generate_ROIs in Algorithm SDS), SDS is able to produce both spatial and non-spatial predicates. The spatial predicates, such as “*intersects*(i, j)” for ROI i and predefined area j (e.g. a region on a map or a structure on a medical atlas), are produced automatically when the ROIs are overlaid (register in Algorithm SDS) onto a registration map $\mathcal{M}$. In the design of SDS we choose the “intersects” predicate because it is commonly used and is particularly useful in the application domains we consider.

SDS also generates a number of non-spatial predicates (such as the one presented in Section 2). The values of these non-spatial predicates are decided by the values of those spatial predicates generated earlier, through a group of synthetic associations generated with a Bayesian network model in the SDS simulator (line 12 in Algorithm SDS). SDS models the conditional distributions among spatial and non-spatial predicates using the BN model (Pearl 1988), which is a graphical model of potentially casual relationships that represents dependency among the variables.

Because spatial and non-spatial predicate variables are often modeled as categorical (with a discrete, unordered domain of values) and in particular, binary, the BNs that we consider here have two states for each node. For example, the two states of spatial nodes may correspond to a spatial region being of interest or not (e.g., being abnormal (s_i) or not ($\overline{s_i}$)), while the two states of non-spatial predicate nodes correspond to the presence of a medical condition (c_k) or not ($\overline{c_k}$). A more detailed discussion of BN model can be found in Megalooikonomou (2002). Although in this paper we use discrete spatial and non-spatial variables, BNs based on multivariate gaussian distributions (Shachter and Kenley 1989) and mixed discrete continuous distributions (Lauritzen and Wermuth 1989) have also been developed. A non-spatial predicate associated with more than one spatial predicate is represented by a particular model. For example, a noisy-OR model can model probabilistic disjunctive interactions among the causes of an effect (Pearl 1988). In the rest of the paper, we use the term *degree* to refer to the number of incoming arrows (edges) of a certain node in a BN and use *maximum degree* to refer to 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 d* of the BN.

The priors of spatial region being of interest are calculated from the fraction of the volume f_{r_i,d_j} for each spatial region r_i and sample d_j of the real data set; this quantity is defined as the volume of the part of r_i being of interest (satisfying a condition) divided by the total volume of r_i . We model the conditional probability of a spatial region r_i being of interest given f_{r_i,d_j} , $p(r_i|f_{r_i,d_j})$ using a sigmoid function. One way to fit the sigmoid model is to compute $p(c|f_{r_i,d_j} < x)$ for various non-spatial predicates and spatial regions in the data set. The sigmoid function can differ for different spatial regions. For simplicity, a step function can be used instead of a sigmoid function (Megalooikonomou et al. 2000a) and this is what we used in our experiments (see Section 4). The prior probabilities for the spatial regions of interest are computed by calculating for every region of the map, the percentage of samples in the simulated data set with areas of interest in that region. Figure 5 demonstrates, using a flow diagram, how the initial set of ROIs (lesions) and the actual map (brain atlas) are used to calculate the values of the spatial predicates corresponding to the atlas structures. After image registration, the ROIs are overlaid on the atlas and for each atlas structure the fraction that is affected by the ROI is computed and used to decide whether the structure is of interest or not (e.g., has been affected or not).

After identifying a spatial region as interesting or not, we compute the fraction of samples satisfying a condition in this spatial region. This is repeated for all regions (i. e., we calculate for each region the prior probability of its satisfying a condition). Then, for every sample d_j of the simulated data set and spatial region s_k , we sample the prior probability and generate a binary vector of spatial predicates s_j^K of dimension K (where $s_j[k] = 1$ means that the spatial region s_k is of interest for sample d_j).

The spatial correlation or the statistical dependence among neighboring objects is modeled by taking into account the correlation among certain spatial predicates when their values are generated in a sequential manner. Correlation information is provided by the atlas or by any

other form of prior knowledge. In addition, the prior probabilities that are given by real data also contribute into modeling this spatial correlation. However, spatial predicates with the same priors can obviously still produce uncorrelated values if no particular attention is paid to the way the values for each sample are being generated.

By instantiating the state of the spatial nodes of the BN with s_j^K for sample d_j we get a binary vector C_j^M of dimension M for the non-spatial predicates, where, $C_j[i] = 1$ if sample d_j has non-spatial predicate c_i evaluated as true.

3.2 Evaluating two association mining techniques

As a case study, we examine Bayesian network learning and a dependency analysis technique based on the Fisher exact test of independence. Both of these methods are commonly used in detecting associations.

Since the Fisher exact test can only be used for categorical variables, we use the binary vector S_j^K for each spatial region and a sample that defines a regional spatial predicate as being of interest or not based on the fraction of it that satisfies a condition. The test is then applied to each spatial—non-spatial predicate 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.

For the Bayesian network learning, we use the same binary vectors S_j^K and C_j^M . For searching through the space of possible network structures we used a heuristic local search method, the greedy hill-climbing. Hill-climbing starts at a specific point in space, considers all nearest neighbors, and moves to the neighbor that has the highest score; if no neighbors have a higher score than the current point (i.e., we have reached a local maximum), the algorithm stops. One can then restart in another part of the space. A common definition of a neighbor of a graph is a graph that can be generated from the current graph by adding, deleting, or reversing a single arc, depending on the acyclicity constraint. The scoring function we use in the implementation of the BN we evaluated is the Bayesian Information Criterion (BIC) which is equal to the Minimum Description Length (MDL). The Bayesian network learning process continues until there is no improvement in the BIC score. The use of a probabilistic model based on Bayesian networks to evaluate a Bayesian network learning technique does not favor in any way the Bayesian network learning over the dependency analysis technique. This is because the BN as a model is complete (i.e., it can represent any generative process and what is evaluated is actually the learning ability rather than the modeling ability). Evaluating other methods that can be used with continuous variables (e.g., Mann-Whitney test), can be performed using the proposed framework, since the fraction of spatial regions satisfying a condition can then be used directly.

4 Results and discussion

Here, we report and discuss the results of evaluation of two methods used for learning associations (i.e., the Bayesian method and the dependency analysis using the Fisher's exact test of independence). This evaluation was performed using the proposed framework. We measured the performance of the methods in mining associations among spatial and non-spatial predicates as a function of the number of samples needed to discover the existing associations represented by the probabilistic network model; the strengths of associations; the number of associations; the type of associations (linear vs nonlinear); the degree of the

network; the prior probabilities of the spatial predicates; the conditional probabilities of non-spatial predicates; and the parameters that control the sensitivity of the methods. We also studied the effect of the image registration error on the performance of the methods. The datasets and programs used for the experiments (including the simulator, the BN hill-climbing and Fisher's exact test implementations) are available at <http://denlab.temple.edu/repository/spatial-simulator/>. All programs were implemented in C, and compiled with the GNU Compiler Collection (GCC) on a Dell Precision Workstation 530 with dual Intel Xeon processors and 1GB memory running Redhat Linux 9.

To examine the effect of the strength of the spatial–non-spatial predicate associations on the ability of the mining methods to detect them, we considered three cases of conditional probability tables which correspond to strong, moderate, and weak associations (see Table 3). For each case, all non-spatial predicate nodes used the same conditional–probability table.

We used two synthetic data sets for the evaluation: one based on a medical study and one based on weather data. Observe that the priors for generating the synthetic datasets were collected from the real datasets. In this section, we present a comprehensive comparison of the examined association mining methods using the proposed framework on the first data set. To demonstrate the general applicability of the framework we also provide some experimental results on the second dataset.

The parameters we chose for the stochastically generated BN are the following unless otherwise stated. For the first dataset we chose 130 spatial nodes (corresponding to spatial regions), 20 non-spatial predicate nodes, and 69 linear associations (edges) constrained to be from spatial predicates to non-spatial predicates. For the second dataset we chose 36 spatial nodes, 10 non-spatial predicate nodes, and 35 linear associations. In certain experiments, nonlinear associations were also included. The maximum degree (i.e., the maximum number of incoming edges to non-spatial predicate nodes), was restricted to be 4; thus, a non-spatial predicate was related to, at most, 4 different spatial predicates. To generate the first test data, the data set $\mathcal{D}$ that was used as input to SDS to extract the data distributions was taken from a clinical study of traumatic brain lesions (Gerring et al. 1998). The spatial normalization error distribution was determined from this data set by measuring the error on a number of distinct landmarks and interpolating for all other points. To determine this distribution, we collected data from 19 samples, and measured the registration error on 20 distinct landmarks, and interpolated the error at all other points. To generate the second test data set, the input to SDS for extracting the data distributions was the monthly station normals collected at 198 climate stations in Pennsylvania during the years of 1971–2000. This weather data was collected from the National Climatic Data Center (<http://www.ncdc.noaa.gov/oa/ncdc.html>). The climate normals included the maximum and minimum temperatures and precipitation for each month of the year. The raw data was transformed into binary values by setting an approximate threshold, the mean value.

In the case of the brain image dataset, for the prior probability of a spatial region of interest (i.e., being abnormal in this case), a step function was used instead of a sigmoid. Plotting the conditional probability, $p(c|f_{r_i,d_j} < x)$ on an ROC (receiver operating characteristic) graph, with the fraction of the spatial region that is of interest for several non-spatial predicates and considering all spatial regions of the map and samples in the study, showed that a step function with threshold fraction of 0.01 can be used for simplicity in the simulation, instead of a sigmoid function (Megalooikonomou et al. 2000a). This threshold depends on the size of the abnormal regions in a given study. In addition, we set the prior probabilities for spatial regions being abnormal to 0.5 to show the behavior of the evaluated methods for the optimal value of the priors under any value of the conditional probabilities. In the case

where a non-spatial predicate was related to more than one region, a noisy-OR model with leak probability 0.25 was used (unless otherwise specified). Similar operations were performed on the weather data. The climate stations were grouped into 36 regions and the values of spatial predicates were determined by the majority of the binary values within groups. Ten non-spatial predicates were also involved in generating this data.

4.1 Selecting the parameters of the evaluated methods

For the Bayesian network method with a hill-climbing model search algorithm, there is no need for any parameter, the learning process continues until the fractional drop of the BIC score is below a default small threshold. The dependency analysis technique based on the Fisher exact test, however, requires a threshold value to be given so that it can decide the significance between two variables. In Table 4 we present the performance change of the dependency analysis technique based on the Fisher exact test for various values of the threshold of p -value that is used. In addition to the number of detected positives and negatives, the Matthews correlation coefficient (Mathews 1975) is also reported. Defined as

$$\frac{t_p t_n - f_p f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}}$$

Matthews correlation coefficient is a useful performance measure (a value closer to one means better performance).

Although the results presented 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 experiments that follow we choose the threshold 0.001 for the p -value of the Fisher's exact test since this is a good trade-off between the number of true and false positives and true and false negatives regarding the associations that are being discovered. The experiments are repeated for different values of the threshold in Section 4.6. In considering the comparative results we present in the following sections, one should consider that in addition to the selection of the parameters, the particular implementation details may affect the performance of the methods. These experiments are presented mostly to demonstrate the evaluation framework we propose.

4.2 Varying the strength of associations

Having determined the parameters for both methods that we compare, in Fig. 6 we present the methods' performance in elucidating the modeled associations. We consider the case of strong, moderate, and weak associations. We show the total number of associations discovered and the number of existing (correct) associations discovered. The difference between these two numbers is the number of false positives in each case. The graphs illustrate the dramatic effect of the conditional probabilities on the ability of the methods to detect the associations. As expected, the number of samples needed is inversely proportional to the strength of associations.

The BN learning and the dependency analysis are similar in finding associations as the number of samples increases for both strong and moderate associations. For weak associations, the BN is more demanding than the dependency analysis in terms of the number of samples needed. On the other hand though, BN learning demonstrates its advantage with fewer false positives in all three cases (strong, moderate and weak associations). This is due to the heuristic nature of the greedy algorithm used to generate alternative structures for the data. As expected, the number of false positives reduces as the

number of samples increases. Another observation is that the number of false positives in the dependency analysis does not approach zero as the number of samples increases for all the cases we considered. We believe that this is because the dependency analysis, by introducing edges based on the repetitive use of conditional independence tests, becomes unreliable due to the number of tests that need to be performed (in the experiments presented here no correction for the multiple comparison problem is performed).

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. For this experiment we consider models with associations of moderate strength (i.e., the second case of the conditional probability table (see Table 3)). To study the effect of the number of associations, without the loss of generality, we consider a case in which all the nodes have the same degree (defined as degree of the BN) in order to examine the two effects independently of each other. In Fig. 7 we present the performance of the Bayesian and dependency analysis methods for three networks of degree 4 with a different number of associations. It is evident from this figure that both methods perform similarly for probabilistic models with different number of associations but of the same degree. It is interesting to note that the Bayesian network approach requires more samples than the dependency analysis. This may be explained by the heuristic nature of the searching algorithm. For dependency analysis, the deterioration as the number of associations increases is smaller.

We also study the effect of increasing the number of spatial predicates related with a particular non-spatial predicate, or degree of the probabilistic model, (assuming that all the nodes have the same degree), while keeping the number of associations the same (see Fig. 8). As expected, the higher the degree of the model, the more samples are needed to detect the same number of associations. Also, the degree of the network has a much greater effect on the performance of the methods than the number of associations. Moreover, the degree affects the BN learning more than the dependency analysis. This finding was opposed to our expectation that methods that seek multivariate associations may perform better than repeated applications of bivariate tests (such as the Fisher's exact test) in cases where multivariate associations are present. A possible reason is that some prior knowledge is utilized before the actual tests are performed that the associations are only between non-spatial variables and spatial variables. This information helps improve the performance of Fisher's exact test. At the same time, with the increase of degrees, it gets more difficult for the heuristic searching algorithm of the BN to find the optimal network, since it may get stuck at local optima. Again, the Fisher's exact test finds more false positives than the BN does.

4.4 Varying the spatial prior probabilities

In all the experiments presented so far, the prior probabilities for the spatial predicates were set to 0.5 to evaluate the performance of the two methods for the optimal value of priors under any conditional probability table. In this experiment, we obtain the priors for the spatial predicates from the simulated regions data set as was described in Section 3.1. In this case, there are 14 associations from spatial predicates that have zero probability of being true (i.e., these regions do not intersect abnormal areas in the medical images) so the number of associations that can be actually discovered is 52. The smallest nonzero prior is 0.0005 and only 5 out of the 130 structures have priors above 0.2 in this case. Figure 9 demonstrates the performance of the Bayesian and the dependency analysis method for the 3 cases of the conditional probabilities (see Table 3). Comparison of these graphs with those of Fig. 6 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. As in the case of uniform priors, the Bayesian network learning performs well for the model with strong associations but does worse than the dependency analysis in the models with moderate and weak associations. Once again, this is related to the heuristic nature of the BN algorithm used.

4.5 Effect of image registration error

Figure 10 compares the performance of the two methods for strong, moderate, and weak associations showing the effect due to the imperfect image registration method used (use of a particular method is actually a parameter in our model; various image registration methods can be compared). As expected, the registration error (of the particular nonlinear registration method we considered) reduces considerably the power of the methods in detecting the associations, when compared with perfect registration. Observe also that for both methods considered, the reduction on the performance due to registration error increases as the sample size increases.

4.6 Varying the threshold value for dependency analysis

In the previous subsections, we evaluated the dependence analysis based on Fisher's exact test with a threshold $p \leq 0.001$. This value was selected in Section 4.1 as a good compromise between the number of correct associations and the number of false positives discovered. However, in real life applications, it may not be always easy to find an appropriate threshold. If an incorrect threshold value is utilized in the Fisher's exact test, the performance may be deteriorated dramatically (i.e. either by decreasing the number of correctly found associations or by increasing the number of false positive associations).

Figure 11 demonstrates the performance of the dependency analysis using Fisher's exact test with a larger threshold value of 0.01 and a smaller threshold value of 0.0001 for the experiments performed in the previous three subsections. It is obvious that with a larger threshold value (0.01), more associations are discovered together with more false positives. Conversely, with a smaller threshold value (0.0001), more samples are required to find all the correct associations. At the same time, the number of false positives is reduced to a much lower level. These results clearly show the important role of the threshold value in dependency analysis. In contrast, the Bayesian network learning does not require tuning of any parameter.

4.7 Introducing nonlinear associations

In the previous comparisons, we only considered linear associations between spatial and non-spatial variables. A linear association, in this context, is an association only between one spatial variable and one non-spatial variable and is independent of all other variables. When a non-spatial variable is involved in multiple linear associations and has more than one parents, these multiple parents may be independent to each other in which case there is no interaction among these linear associations. However, in some applications, only a specific nonlinear combination of spatial variables may be associated with a non-spatial variable. In this case, the multiple parents of a non-spatial variable are no longer independent. To make the evaluation framework more complete, SDS can also generate nonlinear associations. The ability to detect nonlinear associations becomes another factor in evaluating the performance of an association mining method. Table 5 demonstrates an example of a nonlinear association. From the conditional probability table, we can observe that even though there is no linear association between the non-spatial variable s_2 and the spatial variable c , the combination of s_1 and s_2 has a stronger association with c than s_1 does alone.

In Fig. 12 we present the performance of Bayesian network learning and dependency analysis when nonlinear associations are included. In these experiments there are totally five nonlinear associations. While the Bayesian network learning method is able to detect all the nonlinear associations with as few as 1000 samples, the dependency analysis method fails to detect any of the nonlinear associations. This finding confirms a similar claim made in Herskovits et al. (2004).

There are many different types of nonlinear associations, and the ability of detecting these associations can be affected by the actual implementation of the evaluated methods. The Bayesian network learning method we tested, for example, is based on a hill-climbing algorithm. Since the searching scheme is based on the improvement of adding, deleting, or reversing a single edge in the network structure, it is not able to find the nonlinear associations that are based on the concurrent existence of two or more edges. To test this claim, we repeat experiments on a data set which is generated when five associations between combinations of spatial variables and non-spatial variables are introduced. The values of the non-spatial variables in the five additional associations are decided by the XOR relationship between two spatial variables in the combinations. Figure 13 demonstrates the inability of both methods to detect any of the added nonlinear XOR associations.

4.8 Comparing efficiency

Besides the ability of detecting existing associations, processing speed is another important system evaluation metric. To compare the efficiency of different methods we recorded their running time during the experiments. Figure 14 shows the running time for Bayesian network learning and for the Fisher exact test on two different datasets: (a) data with uniform priors and associations of varying strength; (b) data with moderate associations and varying degrees. From the figures we have several interesting findings:

- The Fisher exact test is much faster than the Bayesian network learning. The running time of the Fisher exact test is approximately linear on the number of samples, while that of the Bayesian network learning has an approximate quadratic relationship with the number of samples;
- The complexity of the data has less effect on the Fisher exact test. The running time of the Fisher exact test stays almost the same when changing association strength or association degree;
- The more complex the data (i.e., with associations of higher degree), the shorter the time Bayesian network learning takes. This is due to the heuristic nature of the algorithm we use. With higher degree, there are more chances for the search algorithm to stop sooner at a local maximum. Early stop of learning also leads to a worse result in terms of the accuracy in detecting associations (see Section 4.3).

Theoretically, the Fisher exact test and BN learning have polynomial time complexity (Bennett and Hsu 1960; Cheng et al. 1997). However, the complexity of the unit calculation for these methods has caused a difference. In our experiments, the Fisher's exact test was implemented using logarithmic conversions and its unit calculation was a simple addition, while the Bayesian network learning was implemented using a hill climbing heuristic search algorithm and each search step (calculating the BIC score) itself had a polynomial complexity.

4.9 Experiments with a weather dataset

To demonstrate the general applicability of the proposed framework, we performed experiments on a weather dataset discussed earlier. The experiments were performed using the various settings that were used in previous subsections. For conciseness, here we only

include figures to demonstrate the results obtained with three settings (varying association strengths with fixed priors). However, please note that all the results observed, including those not included here, were consistent with the results obtained on the brain image dataset.

Figure 15 demonstrates the performance of BN learning and dependency analysis (Fisher exact test with $p \leq 0.001$) on the data generated with uniform (0.5) prior probabilities of the spatial predicates being true and for the three cases of strong, moderate, and weak associations. Similarly to the performance observed on the brain image dataset, BN learning was more demanding than the dependency analysis in terms of the number of samples needed, especially for the case of weak associations. Another interesting finding was that even though we used the same threshold for p -value (0.001) as in Section 4.2, we did not have as many false positives as shown in Fig. 6(b). Once again, this fact indicates the importance of selecting an appropriate threshold for dependency analysis.

In Fig. 16 we present the results from generating the data using non-uniform priors for the spatial predicates. The priors for the spatial predicates were obtained from the raw data downloaded from the National Climatic Data Center website. The figures clearly demonstrate the significant effects of the various priors on the performance of both learning methods. For the case of weak associations, both methods failed to detect some of the existing associations. However, the dependency analysis method had a much better performance for both strong and moderate cases. While the dependency analysis method was able to detect all of the 35 existing associations, the BN learning method discovered 31 associations for the strong case. It is worth noting that the four strong associations undetected by the BN learning method are leading to the same non-spatial predicate. This is an evidence that the hill-climbing BN learning algorithm has its constraints due to its heuristic nature.

Finally, Fig. 17 demonstrates the effects of choosing different thresholds for p -value in the dependency analysis method. Comparing with the results shown in Fig. 15(b), it is clear that with a larger threshold (0.01), the dependency analysis method was able to detect the associations faster, i.e., using fewer subjects. However, at the same time, a larger threshold also brought more false positives. When a smaller threshold value (0.0001) was used, the dependency analysis method required more subjects to discover all associates, but it was also enhanced with the ability to remove all the false positives.

5 Conclusions

We proposed a framework for evaluating the performance of association mining methods for spatial data and presented an analysis of a Bayesian network learning technique and a dependency analysis technique based on the Fisher exact test of independence as a case study. Analyzing simulated data, we demonstrated that the number of samples needed has an inverse relationship to the strength of associations and the prior probability of a spatial predicate being true. 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 predicates associated with a particular non-spatial predicate), has a much greater effect on the performance of the method used, than does the number of associations. While this outcome is intuitively expected, our simulations quantify it. Increasing the image registration error also reduces the power of the methods in detecting the associations; the proposed SDS allows us to take this error into account when calculating the sample size needed for a particular experiment. Both the Bayesian network learning and the dependency analysis are able to detect linear associations if they have thousands of samples derived from a real data set. However, the Bayesian network method requires no parameter adjustment; this is not the case for the dependency analysis where a manually chosen threshold plays a

very important role in detecting associations. When there are nonlinear associations involved, Bayesian network learning is better at detecting the hidden relationship between two variables when the pairwise test of dependency analysis reports nothing. For our implementations, even though it is actually implementation dependent, the Bayesian network method is in general more computationally expensive.

The major contributions of this paper are:

- The adaptation of a framework previously used in medical image analysis to a general form of spatial data mining. It models spatial regions, spatial predicates, non-spatial predicates and associations among them.
- An example use of the proposed framework in evaluating methods currently available for detection of associations; an extensive comparison of performance between a Bayesian network learning technique and dependency analysis based on the Fisher exact test of independence revealing pros and cons of the two methods.
- The study of the recovery of associations as a function of the number of samples needed, the strength, number, and type (linear vs. nonlinear) of associations, the number of spatial ROIs associated with non-spatial data, the prior probabilities of spatial regions being of interest, and the parameters that control the sensitivity of the methods.

Future work includes: (a) the use of SDS in the evaluation of other statistical and non-statistical methods used in spatial data mining, (b) the study of the effect of different preprocessing steps (i.e., registration and segmentation algorithms) in the mining process, (c) the extension of SDS to spatial-temporal data and associated mining methods, and (d) the study of other spatial data mining applications such as spatial clustering, prediction, outlier detection, etc. This work may have wide applicability in the evaluation of 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.

Acknowledgments

The authors wish 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 clinical data set, D. Kontos and D. Pokrajac for providing constructive comments, and D. Margaritis for providing the software we used for Bayesian network learning. This work was supported in part by the National Science Foundation under grants IIS-0237921 and IIS-0705215, and by the National Institutes of Health under grant R01 MH68066-04 (funded by NIMH, NINDS, and NIA). Finally, we would like to thank the anonymous reviewers, whose comments have helped us to significantly improve this paper.

References

- Agrawal, R.; Imielinski, T.; Swami, A. Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD conference*; Washington D.C., USA. 1993. p. 207-216.
- Andersen, E. *Introduction to the statistical analysis of categorical data*. Berlin: Springer; 1997.
- Armato SG III, McLennan G, McNitt-Gray MF, Meyer CR, Yankelevitz D, Aberle DR, et al. Lung image database consortium: Developing a resource for the medical imaging research community. *Radiology* 2004;232:739-748. [PubMed: 15333795]
- Arya M, Cody W, Faloutsos C, Richardson J, Toga A. A 3D medical image database management system. *Int Journal of Computerized Medical Imaging and Graphics, Special issue on Medical Image Databases* 1996;20(4):269-284.
- Babu, S.; Garofalakis, M.; Rastogi, R. *Proceedings of the ACM SIGMOD 2001*. 2001. SPARTAN: A model-based semantic compression system for massive data tables; p. 283-294.

- Barnard GA. Significance tests for 2×2 tables. *Biometrika* 1947;34:139–169.
- Bennett BM, Hsu P. On the power function of the exact test for the 2×2 contingency table. *Biometrika* 1960;47(3,4):393–398. (Correction 48, 1961, p. 475).
- Bogdan M, Ghosh JK, Doerge RW. Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* 2004;167:989–999. [PubMed: 15238547]
- Brin, S.; Motwani, R.; Silverstein, C. Proceedings of the ACM SIGMOD international conference on management of data. *SIGMOD Record*. Vol. 26. New York: ACM; 1997. Beyond market baskets: Generalizing association rules to correlations; p. 265–276.
- Cheng, J.; Bell, DA.; Liu, W. Learning belief networks from data: An information theory based approach. Proceedings of the sixth international conference on information and knowledge management (CIKM'97); Las Vegas, Nevada. 1997.
- Cooper, G. An overview of the representation and discovery of causal relationships using Bayesian networks. In: Glymour, C.; Cooper, G., editors. *Computation, causation & discovery*. Cambridge: MIT; 1999.
- Eden, M. Fourth Berkeley symposium on mathematical statistics and probability. Berkeley, CA: University of California Press; 1961. A two-dimensional growth process.
- Ester, M.; Kriegel, H-P.; Sander, J. Algorithms and applications for spatial data mining. In: Miller, HJ.; Han, J., editors. *Geographic data mining and knowledge discovery*. London: Taylor & Francis; 2001.
- Ester, M.; Kriegel, H-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd international conference on knowledge discovery and data mining (KDD-96); Portland, Oregon. 1996.
- Fayyad, UM.; Smyth, P. Image database exploration: Progress and challenges. Proceedings of the 1993 knowledge discovery in databases workshop; Washington, D.C. 1993.
- Fisher, RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1934.
- Fotheringham, S.; Rogerson, P. Spatial analysis and GIS. London: Taylor and Francis; 1994.
- Fox P. Functional brain mapping with positron emission tomography. *Seminars in Neurology* 1989;9:323–329. [PubMed: 2561931]
- Gail M, Gart JJ. The determination of sample sizes for use with the exact conditional test in 2×2 comparative trials. *Biometrika* 1973;29:441–448.
- Gerring J, Brady K, Chen A, Quinn C, Bandeen-Roche K, Denckla M, et al. Neu-roimaging 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* 1998;37:647–654. [PubMed: 9628085]
- Gorsevski, PV.; Gessler, P.; Foltz, RB. Spatial prediction of landslide hazard using discriminant analysis and GIS. GIS in the Rockies 2000 conference and workshop: Applications for the 21st century; Denver, Colorado. 2000.
- Gueing RH. An introduction to spatial database systems. *VLDB Journal* 1994;3(4):357–400.
- Han J, Cai Y, Cercone N. Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Engineering* 1993;5:29–40.
- Han, J.; Kamber, M. Data mining. San Francisco: Morgan Kaufmann; 2000.
- Han, J.; Kamber, M.; Tung, AKH. Spatial clustering methods in data mining. In: Miller, HJ.; Han, J., editors. *Geographic data mining and knowledge discovery*. London: Taylor & Francis; 2001.
- Hanson F, Tier C. A stochastic model of tumor growth. *Mathematical Biosciences* 1982;61:73–100.
- Herskovits EH, Peng H, Davatzikos C. A Bayesian morphometry algorithm. *IEEE Transactions on Medical Imaging* 2004;23(6):723–737. [PubMed: 15191147]
- Hochberg, Y.; Tamhane, A. Multiple comparison procedures. New York: Wiley; 1987.
- Kansal AR, Torquato S, Harsh GR, Chiocca EA, Deisboeck TS. Simulated brain tumor growth using a three-dimensional cellular automaton. *Journal of Theoretical Biology* 2000;203(4):367–382. [PubMed: 10736214]
- Karasov, V.; Krisp, JM.; Virrantaus, K. Application of spatial association rules for improvement of a risk model for fire and rescue services. Proceedings of ScanGIS2005; Stockholm. 2005.

- Knorr E, Ng R. Finding aggregate proximity relationships and commonalities in spatial data mining. *IEEE Transactions on Knowledge and Data Engineering* 1996;8(6):884–897.
- Koperski, K.; Han, J. Discovery of spatial association rules in geographic information databases. In: Egenhofer, MJ.; Herring, JR., editors. *Proceedings of the 4th international symposium on advances in spatial databases, (SSD)*, Portland, Maine. Vol. 951. New York: Springer; 1995. p. 47–66.
- Korn F, Sidiropoulos N, Faloutsos C, Siegel E, Protopapas Z. Fast and effective retrieval of medical tumor shapes. *IEEE Transactions on Knowledge and Data Engineering* 1998;10(6):889–904.
- Lauritzen SL, Wermuth N. Graphical models for associations between variables, some of which are qualitative and some of which are quantitative. *The Annals of Statistics* 1989;17:31–57.
- Letovsky S, Whitehead S, Paik C, Miller G, Gerber J, Herskovits E, et al. A brain-image database for structure-function analysis. *American Journal of Neuroradiology* 1998;19(10):1869–1877. [PubMed: 9874539]
- Margaritis, D.; Faloutsos, C.; Thrun, S. NetCube: A scalable tool for fast data mining and compression. *Proceedings of the 27th int. conference on very large data bases (VLDB)*; 2001. p. 311–320.
- Mathews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochemical and Biophysical Acta* 1975;405:442–451.
- Megalooikonomou, V. Evaluating the performance of association mining methods in 3-D medical image databases. *Proceedings of the 2nd SIAM international conference on data mining*; Arlington, VA. 2002. p. 474–494.
- Megalooikonomou, V.; Davatzikos, C.; Herskovits, EH. Mining lesion-deficit associations in a brain-image database. *Proceedings of the 5th international conference on knowledge discovery and data mining (KDD-99)*; San Diego, CA. 1999. p. 347–351.
- Megalooikonomou V, Davatzikos C, Herskovits E. A simulator for evaluation of methods for the detection of lesion-deficit associations. *Human Brain Mapping* 2000a;10(2):61–73. [PubMed: 10864230]
- Megalooikonomou V, Ford J, Shen L, Makedon F, Saykin A. Data mining in brain imaging. *Statistical Methods in Medical Research* 2000b;9(4):359–394. [PubMed: 11084713]
- Megiddo, N.; Srikant, R. Discovering predictive association rules. *Proceedings of the 4th international conference on knowledge discovery and data mining (KDD-98)*; New York City, NY. 1999. p. 274–278.
- Morimoto, Y. Mining frequent neighboring class sets. *Proceedings of the 7th international conference on data mining (KDD-01)*; San Francisco, CA. 2001. p. 353–358.
- Ng, R.; Han, J. Efficient and effective clustering method for spatial data mining. *Proceedings of the 1994 int. conf. very large data bases (VLDB)*; Santiago, Chile. 1994. p. 144–155.
- Osius G, Rojek D. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association* 1992;87(420):1145–1152.
- Pearl, J. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo: Morgan Kaufmann; 1988.
- Piatetsky-Shapiro, G.; Frawley, WJ., editors. *Knowledge discovery in databases*. Menlo Park: AAAI/MIT; 1991.
- Samet, H. *The design and analysis of spatial data structures*. Reading: Addison-Wesley; 1990.
- Shachter RD, Kenley CR. Gaussian influence diagrams. *Management Science* 1989;35:527–550.
- Shekhar, S.; Zhang, P.; Huang, Y.; Vatsavai, R. Trends in spatial data mining. In: Kargupta, H.; Joshi, A.; Sivakumar, K.; Yesha, Y., editors. *Data mining: Next generation challenges and future directions*. Cambridge: MIT; 2004.
- Smyth, P.; Burl, MC.; Fayyad, UM.; Perona, P. Knowledge discovery in large image databases: Dealing with uncertainties in ground truth. *Proceedings of the AAAI-94 workshop on KDD*; Seattle, WA. 1994.
- Son, E-J.; Kang, I-S.; Kim, T-W.; Li, K-J. A spatial data mining method by clustering analysis. *Proceedings of the sixth international symposium on advances in geographic information systems, GIS'98*; 1998. p. 157–158.
- Spirtes, P.; Glymour, C.; Scheines, R. *Causation, prediction and search*. Cambridge: MIT; 2000.

- Tan, P.; Steinbach, M.; Kumar, V.; Potter, C.; Klooster, S.; Torregrosa, A. Finding spatio-temporal patterns in earth science data. Proceedings of KDD workshop on temporal data mining; San Francisco, CA. 2001.
- Tanizaki H. Power comparison of non-parametric tests: Small-sample properties from Monte Carlo experiments. *Journal of Applied Statistics* 1997;24(5):603–632.
- Tong, YL. The multivariate normal distribution. New York: Springer; 1990.

J Intell Inf Syst. Author manuscript; available in PMC 2011 December 1.

Fig. 2.
Sample distributions of spatial ROIs: **a** brain lesions, **b** brain activations, **c** areas of high precipitation

Fig. 3.
a Cell probabilities at time t=5 (2-D grid). Probability thresholds for E, W, N, S neighbors are [0.2 0.4 0.1 0.8] **b** Cell infection epoch **c** A 2-D region generated using the growth model

Fig. 4. The artificial regions (in *white*) of a simulated sample **a** before and **b** after displacement due to image registration error (3 consecutive slices of the 3-D volumes are shown in each case)

Brain structure	Spatial predicate	Fraction with lesion	>0.05 (threshold)?	Predicate value
<i>Sturct1</i>	S_1	10%	Yes	1
<i>Sturct2</i>	S_2	2%	No	0
<i>Sturct3</i>	S_3	0%	No	0
<i>Sturct4</i>	S_4	20%	Yes	1
...	...	...	...	...
<i>Sturct_n</i>	S_5	0%	No	0

Fig. 5. Illustration of the process of generating the values of the spatial predicates. Different structures of the atlas are identified by different colors. (Intended for color reproduction)

Fig. 6. Performance of **a** Bayesian and **b** dependency analysis (Fisher exact test with $p \leq 0.001$) methods for *uniform* (0.5) *prior probabilities* of the spatial predicates being true and for the 3 cases of simulated model of Table 3 that correspond to strong, moderate and weak associations

Fig. 7. Varying the number of associations: Performance of **a** Bayesian and **b** dependency analysis (Fisher exact test with $p \leq 0.001$) methods for three different simulated models (of degree 4) with 20 (g20), 40 (g40), and 80 (g80) associations respectively

Fig. 8. Varying the degree of associations: Performance of **a** Bayesian and **b** dependency analysis (Fisher exact test with $p \leq 0.001$) methods for three different BNs with 48 associations and of degree 4 (d4), 6 (d6), and 8 (d8) respectively

Fig. 9. Performance of **a** Bayesian and **b** dependency analysis (Fisher exact test with $p \leq 0.001$) methods for *non-uniform prior probabilities* of the spatial predicates. The probabilities are calculated from a simulated data set and for the 3 cases of simulated model (strong, moderate and weak associations)

Fig. 10. Performance of **a** Bayesian and **b** dependency analysis (Fisher exact test with $p \leq 0.001$) methods with and without registration error, for strong, moderate, and weak associations (a maximum of 52 out of a total of 69 associations can be discovered)

Fig. 11.

Repeating the performance calculations of the dependency analysis using the Fisher's exact test with threshold $p \leq 0.01$ and $p \leq 0.0001$. **a, c, e, g** show the four cases for threshold $p \leq 0.01$: uniform priors, varying numbers of associations, varying degrees of associations and non-uniform priors. **b, d, f, h** show the four cases for threshold $p \leq 0.0001$

Fig. 12.
Performance of BN and Fisher's exact test when nonlinear associations are involved

Fig. 13.
Performance of BN and Fisher's exact test when nonlinear XOR associations are involved

Fig. 14. Running time for the two methods on **a** data with uniform priors and associations of varying strength and **b** data with moderate associations and varying degrees

Fig. 15.

Performance of **a** Bayesian and **b** dependency analysis (Fisher exact test with $p \leq 0.001$) methods on the weather dataset, for *uniform (0.5) prior probabilities* of the spatial predicates being true and for the 3 cases of simulated model of Table 3 that correspond to strong, moderate and weak associations

Fig. 16. Performance of **a** Bayesian and **b** dependency analysis (Fisher exact test with $p \leq 0.001$) methods on the weather dataset, for *non-uniform prior probabilities* of the spatial predicates

Fig. 17. Performance of dependency analysis (Fisher exact test) method on the weather dataset with uniform priors and various p -value thresholds. **a** $p \leq 0.01$; **b** $p \leq 0.0001$

Table 1

Symbol table

Symbol	Definition
$\mathcal{D}$	Real data set $\{d_1, d_2, \dots, d_N\}$ of size N
$\mathcal{D}'$	Simulated data set $\{d'_1, d'_2, \dots, d'_{N'}\}$ of size N'
$\mathcal{M}$	Registration map
R	Set of spatial regions $\{r_1, r_2, \dots, r_G\}$
$\mathcal{S}$	Set of spatial predicates $\{s_1, s_2, \dots, s_K\}$
$\mathcal{C}$	Set of non-spatial predicates $\{c_1, c_2, \dots, c_M\}$
w	Strength of associations
d	Number of spatial regions affecting a non-spatial predicate
f_{r_i, d_j}	Fraction of volume of interest for spatial region r_i and sample d_j
ζ	Number of parameters of the probabilistic network model
t_a	% of existing (correct) associations discovered
f_a	% of false positives in the associations discovered

Table 2

A contingency table for a spatial predicate s_i and a non-spatial predicate c_k

		s_i	
		False	True
c_k	False	a	b
	True	c	d

Table 3

The 3 cases of probabilistic models considered

Case	Association	$p(c s)$
1	Strong	0.99
2	Moderate	0.75
3	Weak	0.51

Table 4

Number of true and false positives (t_p, f_p), true and false negatives (t_n, f_n) discovered by dependency analysis (Fisher exact test) and the corresponding correlation coefficients ($coef$) for three values of the p -value threshold and for moderate strength of associations

# samples	$p \leq 0.01$					$p \leq 0.001$					$p \leq 0.0001$				
	t_p	f_p	t_n	f_n	$coef$	t_p	f_p	t_n	f_n	$coef$	t_p	f_p	t_n	f_n	$coef$
500	57	24	2507	12	0.76	49	32	2528	20	0.81	38	0	2531	31	0.74
1000	69	30	2501	0	0.83	68	1	2530	1	0.99	61	0	2531	8	0.94
1500	69	24	2507	0	0.86	69	2	2529	20	0.99	67	0	2531	2	0.99
2000	69	21	2510	0	0.87	69	1	2530	0	0.99	69	0	2531	0	1.00

Table 5

Conditional probability table of a nonlinear association

Node	Conditional-probability table
s_1	$p(s_1) = 0.5$
s_2	$p(s_2) = 0.5$
c	$p(c s_1) = 0.63, p(c \overline{s_1})=0.38,$ $p(c s_2) = 0.5, p(c \overline{s_2})=0.5,$ $p(c s_1, s_2) = 0.83, p(c s_1, \overline{s_2})=0.42,$ $p(c \overline{s_1}, s_2)=0.16, p(c \overline{s_1}, \overline{s_2})=0.59$