A Spatio-Temporal Bayesian Network Classifier for Understanding Visual Field Deterioration

Allan Tucker [1], Veronica Vinciotti, Xiaohui Liu,

Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK

David Garway-Heath

Glaucoma Unit, Moorfields Eye Hospital, London, UK

Abstract

Progressive loss of the field of vision is characteristic of a number of eye diseases such as glaucoma which is a leading cause of irreversible blindness in the world. Recently, there has been an explosion in the amount of data being stored on patients who suffer from visual deterioration including field test data, retinal image data and patient demographic data. However, there has been relatively little work in modelling the *spatial* and *temporal* relationships common to such data. In this paper we introduce a novel method for classifying Visual Field (VF) data that *explicitly* models these spatial and temporal relationships. We carry out an analysis of this method and compare it to a number of classifiers from the machine learning and statistical communities. Results are very encouraging showing that our classifiers are comparable to existing statistical models whilst also facilitating the understanding of underlying spatial and temporal relationships within VF data. The results reveal the potential of using such models for knowledge discovery within ophthalmic databases, such as networks reflecting the 'nasal step', an early indicator of the onset of glaucoma. The results outlined in this paper pave the way for a substantial program of study involving many other spatial and temporal datasets, including retinal image and clinical data.

Key words: Classification, Multivariate Time Series, Bayesian Networks, Visual Field, Glaucoma

Email address: allan.tucker@brunel.ac.uk (Allan Tucker).

[1] Corresponding author Tel.: +44 (0)1895 816 253; fax: +44 (0)1895 251 686

1 Introduction

Progressive loss of the field of vision is characteristic of a number of eye diseases such as glaucoma, a leading cause of irreversible blindness in the world. Recently, there has been an explosion in the amount of data being stored on patients who suffer from visual deterioration, including visual field (VF) test, retinal image and patient demographic data. The aim now is to extract as much information as possible from these data in order to address fundamental questions still open within the glaucoma community. For example, the diagnosis of glaucoma made by clinicians would be highly improved by the identification of the various causes of VF loss as well as the detection of patterns of VF loss that match with glaucomatous patterns. It would also be very beneficial to be able to integrate different types of clinical data (such as intracocular pressure and visual field data) for both the diagnosis and detection of the disease progression. Furthermore, since the visual field loss is characterised by a slow progression, early detection of glaucoma can be invaluable as early intervention can slow VF deterioration. Statistical and classification models that address all of these issues would therefore be extremely helpful to the glaucoma community.

There has been very little modelling of spatial and temporal relationships which are characteristic of VF data. Swift and Liu [26] have looked into learning statistical time series models of VF data. Other approaches that explore the temporal aspect of VF data include trend analysis [5,14], event analysis [14] and state space models [1]. Ibanez and Simo [17] have investigated spatio-temporal statistical models with the aim of forecasting visual field deterioration, but to date have only looked at visual fields of normal eyes.

Various models have been developed for classifying VF data. Hothorn and Lausen [16] make use of the retinal image data to classify glaucoma with tree classifiers. Goldbaum et al. [8] document a comprehensive comparison of machine learning classifier systems for the classification of glaucomatous visual fields. Many of these classifiers are 'black box' in nature and therefore do not give much insight into the behaviour of the VF. Much research in glaucoma has involved exploring the distribution of point-by-point light sensitivity, at a single point in time, in normal [13,19] and glaucomatous populations [30]. However, much remains unknown about the behaviour of the visual field test, such as the light sensitivity relationship between adjacent and distant visual field test points, the relationship between light sensitivity and other ocular parameters (such as optic nerve appearance and intraocular pressure level), and how stable and deteriorating visual fields behave over time.

It is our intention to make use of the vast amounts of data available in order to build models for classification that fully exploit the spatio-temporal nature of

these data whilst avoiding the inherent problem of black box paradigms. This has led us to investigate the use of Bayesian networks which are transparent in the way that they model data. Here we extend the Bayesian network classifier to explicitly handle the spatial and temporal relationships found within visual field data.

The paper is broken down into the following sections. In Section 2, we describe some relevant background in classification methods. In Section 3, we describe existing Bayesian network classifiers as well as the extension of the Bayesian network to incorporate temporal links. In Section 4, we describe our Spatio-Temporal Bayesian network Classifier (STC), which is a combination of the existing Bayesian classifiers and the temporal Bayesian networks in order to model and classify data with spatial and temporal relationships. This section includes an outline of the architecture of the model, the learning algorithm and the spatial operators used. Section 5 includes a description of the parameters and the datasets used in the experiments, one dataset being non-temporal and the other being a longitudinal time series. In Section 6, the results of the experiments are documented, firstly just applying the non-temporal classifiers on the non-temporal data, and secondly applying our STC to the temporal dataset. Furthermore, we compare our method with two standard statistical classifiers, linear regression and k nearest neighbour, and illustrate the benefits that spatio-temporal Bayesian networks offer over these models. Results include ROC analysis, network structure analysis and inference analysis. Section 7 discusses the implications of our results, whilst in Section 8 conclusions are made and future work is outlined.

2 Overview of Classifcation Problems

In this paper, we look at classification models to predict whether a certain patient has glaucoma or not, given measures of its visual field as well as other variables, like age, gender and intraocular pressure. Denote with $Y = (X, C)$ the vector of variables for this problem, where X is the vector of attributes, and $C = \{0, 1\}$ the corresponding class (with 0=normal, 1=glaucoma). Statistical classifiers provide an estimate of $p(c|x)$, the probability that a patient with observed measurement vector x belongs to class c. Linear and logistic regression, linear and quadratic discriminant analysis, k-nearest neighbours and graphical models are amongst the most popular statistical classifiers.

The estimated $p(c|x)$ provided by one of these models, either directly or indirectly via Bayes theorem, is compared to a threshold t to predict the class of x. This threshold is usually chosen to minimise the expected misclassification

loss [10], resulting in the classification rule: classify x to class 1 if

$$p(1|x) > t = \frac{k_0}{k_0 + k_1}, \tag{1}$$

and otherwise to class 0, where k_i denotes the cost of misclassifying an object from class i.

When the misclassification costs are equal ($k_0 = k_1$), the resulting rule will classify an object to the class with the highest predicted probability. Although this is the default option in many implementations, in real applications the misclassification costs are seldomly the same. In credit scoring applications, for example, it is more costly for the bank to classify a bad customer as a good one than vice versa [29]; in glaucoma applications, like the one that we are considering in this paper, the cost of misclassifying a normal eye as glaucomatous is usually considered much higher than the reverse, since the disease has a low frequency and a slow progression [8]. Hand and Vinciotti [12] discuss the importance of taking the relative misclassification costs into consideration when building and assessing the model. After all, we want the classifier to perform well for the particular choice of costs that we make.

In order to assess the performance of a classifier and to compare different classifiers, it is common practice to use Receiver Operator Characteristic (ROC) curves [10]. An ROC curve allows one to view graphically the performance of a classifier by plotting the *sensitivity*, which in our case is the proportion of glaucomatous eyes correctly classified as glaucomatous, versus (1-*specificity*), the proportion of normal eyes misclassified as glaucomatous, as the threshold t in equation 1 assumes increasing values between 0 and 1. Different points in the curve will correspond to different values of the threshold, i.e. different values of the misclassification costs. The perfect classifier would have an ROC curve that follows the top-left corner of the unit square, whereas the worst situation would be a classifier whose curve follows the diagonal. Real applications will usually show curves between these two extremes.

A global measure of the classifier performance, often used in classification problems, is the Area Under the ROC Curve (AUC). This will be some value between 0.5, associated to the diagonal of the square, and 1, corresponding to the curve that follows the top-left corner. Such a global measure of performance will not be very useful in the situation where curves relative to different classifiers will cross at multiple points, which is actually very common in real-life applications. If curves cross at various points, then one classifier is better than the others on a certain range of values for the threshold, but worse on other values. In situations like this, it is more appropriate to compare the classifiers for the values of the threshold that one is interested in. A common measure of classifier performance relative to a certain threshold is given by

the misclassification cost

$$C = \frac{k_0(\#\text{ misc pts from class 0}) + k_1(\#\text{ misc pts from class 1})}{\text{total }\#\text{ pts}}. \tag{2}$$

In this paper we will compare various statistical classifiers using both the AUC and the misclassification cost measures, in order to gain a better insight into the relative performance of the classifiers.

3 Bayesian Networks

Bayesian Networks (BNs) are probabilistic models that can be used to combine expert knowledge and data. They facilitate the discovery of complex relationships in large datasets and enable non-statisticians to query resultant models. For this reason they are particularly useful in the analysis of VF data when trying to understand underlying relationships between VF points and other clinical variables.

3.1 The Basics

A BN consists of a directed acyclic graph, made up of links between nodes that represent variables in the domain. The links are directed from a parent node to a child node, and with each node there is an associated set of conditional probability distributions. A Bayesian network thus consists of the following: a set of N nodes, $\{Y_1 \ldots, Y_N\}$, representing the N variables in the domain and directed links between the nodes. Associated with each node Y_i with parents π_i, there is a probability table, $p(Y_i|\pi_i)$. The set of these probabilities for all nodes Y_is provides an efficient factorization of the joint probability $p(Y)$ in terms of dependencies between variables [25].

The process of learning a BN from data is made up of two distinct phases. First of all, a network has to be selected amongst the space of all possible models. Then, the probabilities $p(Y_i|\pi_i)$ have to be estimated. Learning the structure of a BN from data [2] is a non-trivial problem due to the large number of candidate network structures. As a result there has been substantial research in developing efficient algorithms within the optimisation communities. Most methods involve scoring candidate network structures and one of the most common metrics is the log-likelihood which is calculated by

$$\log p(D|bn_D) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(F_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} F_{ijk}!, \tag{3}$$

where N is the number of variables in the domain, r_i denotes the number of states that a node Y_i can take, q_i denotes the number of unique instantiations of the parents of node Y_i, F_{ijk} is the number of cases in the database D, where Y_i takes on its kth unique instantiation and the parent set of i takes on its jth unique instantiation, and $F_{ij} = \sum_{k=1}^{r} F_{ijk}$.

3.2 Bayesian Networks for Classification

Bayesian network classifiers have recently shown excellent properties [6]. Because the discovered relationships in the classifiers are explicit, it means that the models can be analysed to understand how classification decisions are made. This is an extremely useful property when trying to understand classification decisions in medical data such as glaucomatous VFs.

There are a number of Bayesian network classifiers, the most common being the naïve Bayes classifier. This architecture assumes that every feature in the classifier is independent given the class, i.e. $p(x|c) = \prod_{i=1}^{N-1} p(x_i|c)$. Despite the fact that the independence assumption in the naïve Bayes classifier is almost always incorrect in real applications, many studies have shown a very good performance of the model, even in comparison with much more sophisticated classifiers(for example[4,6,11,21,22]). These results are of particular interest especially considering the many advantages of the naïve Bayes classifier in practical applications: it is very simple, very efficient and very easy to interpret and implement. Hand and Yu [11] give some mathematical justifications of why such a simple and unrealistic model might perform so well on future observations: simple models, like this, have a lower variance than more complex models. Hence, despite having a larger bias, they might perform better on observations outside the training data.

Other authors have suggested extensions to this classifier that relax the strong assumption of independence. The Tree Augmented Network (TAN) is one of these [6]: in addition to the naïve Bayes structure, this model learns a tree structure amongst the features. Another possible extension to the naïve model is to learn a standard Bayesian network, where the class variable is simply included as one of the nodes.

3.3 Temporal Bayesian Networks

A typical feature of visual field data is their temporal aspect: patients attend the clinic regularly to take VF tests and at each patient visit, a classification is made to decide whether an eye has developed glaucoma or, if glaucoma is already present, whether it has worsened (progressed). For this reason, we

have also looked at classifiers that take the time aspect into consideration. In particular, we looked at Temporal Bayesian Networks (TBNs).

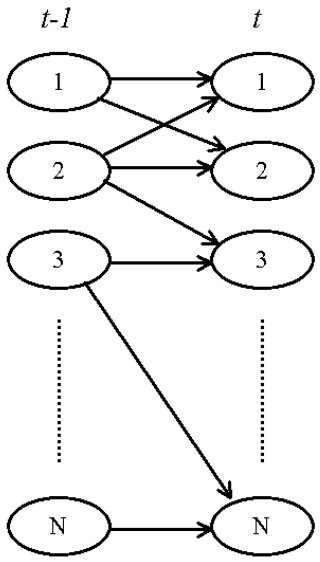

Fig. 1. A Typical TBN with 2 Time Slices. Note the links within one time slice and those spanning from one to the next.

A TBN is a Bayesian network where the N nodes represent variables at differing time slices. Therefore links occur between nodes over time and within the same time lag. Figure 1 shows an example of a TBN where each node represents a variable at a certain time slice and each link represents a conditional dependency between two nodes. Given some evidence about a set of variables at time t, we can infer the most probable explanations for the current observations.

Inference in TBNs is very similar to standard inference in static BNs [3]. In this paper, we use a form of stochastic simulation [15] because of its speed and its intuitive appeal when explaining to clinicians how the prior distributions are calculated. Previously, we have developed methods for learning different specialist TBN structures. In [27] we developed methods for learning TBNs with large time lags and in [28] we developed operators for learning TBNs from spatio-temporal data which we call spatio-temporal Bayesian networks.

To our knowledge, Bayesian network classifiers have not been extended to classify multivariate time series data using TBN models or to classify spatio-temporal data, characteristic of many VF datasets. Within this paper we develop a spatio-temporal Bayesian network model to classify VF data. The method is fully described in the next section.

4 The Spatio-Temporal Bayesian Network Classifier

Previously we have investigated learning temporal Bayesian networks from VF data in order to explore the VF relationships discovered within the network structure. Due to the spatial as well as the temporal nature of VF data we developed spatial operators to efficiently learn spatial-temporal network structures [28]. In this paper, we employ these operators to learn Spatio-Temporal Bayesian network Classifiers (STC) and compare them to standard statistical classifiers.

The STC contains relationships that are both temporal and non-temporal between different variables. When learning these relationships it is assumed that there is a spatial relationship between nodes in the network based upon the cardinal coordinate system. Therefore, nodes that are spatially close to other nodes are deemed more likely to be dependent upon one another. The class node has no spatial relationships but does have temporal properties as we are trying to classify the set of variables at each time point. Figure 2 illustrates the main features of a STC, where C represents the class node and t represents the time slice. Note that the spatial relationships can extend to more than first order cardinal neighbours. Note also that the number and direction of links between the class node and the feature nodes may vary.

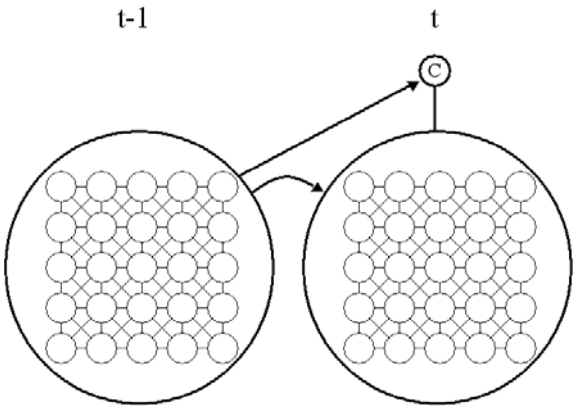

Fig. 2. The Spatio-Temporal Bayesian Network Classifier.

4.1 The Algorithm

In our learning algorithm, candidate structures of a network, bn, given a dataset, D, are scored using the metric in equation 3. In order to increase efficiency of our algorithms, we have developed an evolutionary approach without the necessity of storing a population of candidate solutions. Rather, we con-

sider each point in the spatial dataset to be an individual within the population of points. Therefore, the population, itself, is the candidate solution. We have looked at a similar method before for grouping algorithms [24]. The algorithm also makes use of a simulated annealing type of selection criteria [20], where good operations are always carried forward, but sometimes less good ones are also accepted dependent upon a temperature parameter. A form of elitism [9] is employed to ensure that the final structure is the best discovered. This is to prevent the simulated annealing process from moving away from a better solution when the temperature is still high. We formally define the algorithm below where $maxfc$ is the maximum number of calls to the scoring function, c is the 'cooling parameter', t_0 is the initial temperature, b is the branching factor of a network, and $R(a, b)$ is a uniform random number generator with limits, a and b.

```
Input t_0, b, maxfc, D
fc = 0, t = t_0
Initialise bn to a STC with no links
result = bn
While fc ≤ maxfc do
      score = L(bn)
      For each operator do
          Apply operator to bn
          If bn is valid given b Then newscore = L(bn) fc = fc + 1
                   dscore=newscore-oldscore
                   If newscore > score Then
                       result = bn Else
                       If R(0,1) < e^(dscore/t) Then
                       Undo the operator
                   End If
          End If
      End For
      t = t × c
      End While
Output result
```

4.2 Operators

We now introduce three spatial and three non-spatial operators. All involve manipulating links within the STC. Note that a random link can be either temporal or non-temporal. For the scope of this paper, we only look at first order temporal links.

4.2.1 Non-Spatial Operators

We have chosen three non-spatial operators as these represent common operators used in optimisation techniques such as hill climbing and simulated annealing.

- Add - A link with random parent and child is added to the network.
- Take - Randomly remove a single existing link.
- Mutate - Randomly change the parent of an existing link.

4.2.2 Spatial Operators

For the scope of this paper, we assume that the points in a spatial dataset are located according to cartesian coordinates. Therefore, each point in a dataset with coordinates (x,y) has a first order neighbourhood which includes all nodes with coordinates (i,j) for $i = x \pm 1$ and $j = y \pm 1$. The spatial operators that we have developed exploit the cartesian spatial nature of a dataset.

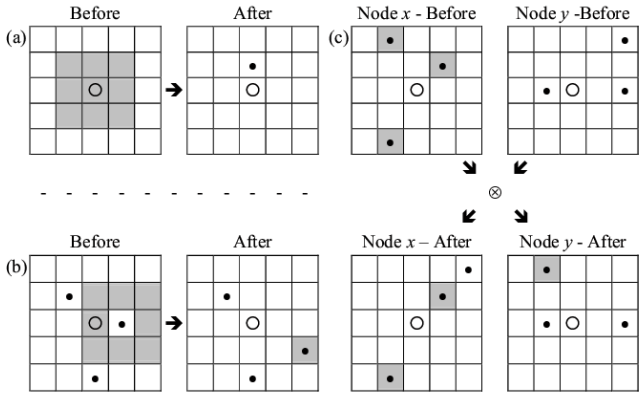

Fig. 3. The Spatial Operators: (a) Add a Link in the 1st Order Neighbourhood (b) Mutate a Parent to Within its 1st Order Neighbourhood (c) Spatial Crossover

Figure 3 shows examples of parent coordinates, relative to the child node, when applying the operators. Unfilled circles represent child nodes, filled circles represent parents of the child.

- Spatial Add (Figure 3a) - Add a link with random child and a parent that is one of the child's first order neighbours.
- Spatial Mutate (Figure 3b) - Randomly change the parent of an existing link by setting it to the first order neighbour of its previous position.

- Spatial Crossover (Figure 3c) - Randomly swap the relative positions of two nodes parents.

5 Experimental Set-up

In this paper, we introduce an extension of Bayesian network classifiers which handles spatio-temporal data and show how the resulting models can be analysed in order to discover new knowledge about visual field deterioration. We compare our method with different classifiers from the statistic and machine learning communities, including the family of Bayesian network classifiers, linear regression and k-nearest neighbour method, in the context of glaucoma detection. We apply these methods to two visual field datasets which we describe in this section, followed by a description of how the data were preprocessed and how the methods were parameterised.

5.1 The Datasets

The Visual Field (VF) test assesses the sensitivity of the retina to light. It is typically measured by automated perimetry, a technique in which the subject views a dim background as brighter spots of light are shone onto the background at various locations in a regular grid pattern. The brightness at which the subject sees the spots of light is related to the retinal sensitivity. See Figure 4 for an example of two VF tests, one from a healthy eye and one from a patient suffering from glaucoma.

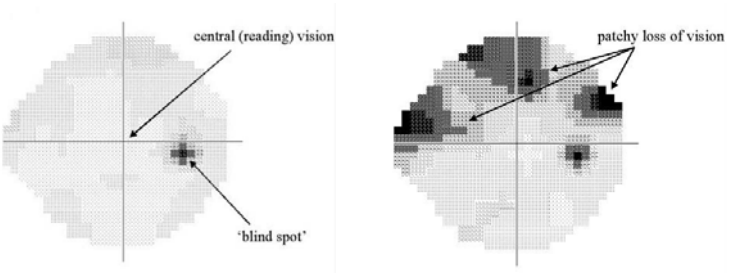

Fig. 4. A Typical VF Test from a Healthy Eye and a Glaucomatous Eye.

All VF testing was performed with the Humphrey Field Analyzer model and the 24-2 full threshold program [32] . We investigate two separate datasets:

(1) Subjects included 78 with established early glaucomatous VF loss and 102 normal volunteers known not to be sufferers. One visual field per subject was used for analysis. Early glaucomatous VF was defined as

an AGIS score between 1 and 5, on three consecutive reproducible and reliable Humphrey 24-2 strategy visual fields, with at least one location consistently below the threshold for normality [32]. Normal subjects had VF tests scoring '0' in the AGIS classification.

(2) The visual fields of 24 subjects attending the Ocular Hypertension Clinic at Moorfield's Eye Hospital were examined at 4-monthly intervals [18]. All subjects initially had normal visual fields and developed reproducible glaucomatous VF damage in a reliable VF during the course of follow-up ('conversion'). Conversion was defined as the development of an AGIS score greater than or equal to 1 from an initial score of 0, on three consecutive reproducible and reliable Humphrey 24-2 strategy visual fields, with at least one location consistently below the threshold for normality. If a patient developed a visual field defect, then the test was repeated within 1 month, and if the same defect was then reproduced on a reliable second field, then a third test was performed 3-4 months after this. Conversion is confirmed if the field defect is present on the three consecutive reliable tests. VF were not re-classified following conversion. For this data, the average number of field tests in each patient's series was 24.08, the maximum was 45, and the minimum was 1.

Both datasets are slightly imbalanced (the number of glaucomatous VF is not equal to the number of non-glaucomatous VFs) and include some other variables such as gender, age, and intraocular pressure. For the scope of this paper, we only look at the right eye of patients. Table 1 provides a summary of each dataset.

INSERT TABLE 1 ABOUT HERE

5.2 Data Preprocessing and Parameterisation

For the Bayesian network (BN) models, all continuous data are discretised into four states using a frequency-based method where bin sizes are determined such that there are equal numbers of each state per variable in the dataset. Discretisation was performed on a point-wise basis.

For our experiments, the following parameters were used. For the BN learning, we set $maxfc = 50000$, $t_0 = 5$ and $c = 0.9999$, with these parameters defined as in Section 4.1. These were chosen as they were found to be the most efficient based upon previous empirical studies. We have found that the individual improvements in score during the early iterations of our algorithm are a good indication for the value of t_0. The maximum number of parents, b, was varied between 1 and 3 depending upon the experiments, and t_0 and c were set so as to finish on a suitably cold temperature to ensure a stable solution.

Inference involved setting the number of stochastic simulations to 10000 so as to compromise the time taken to perform inference with generating an accurate prior distribution.

6 Results

The following results are split into two main sections for each dataset. Firstly, the non-temporal visual field (VF) data are analysed with respect to classification and structural analysis. This involves a comparison of the efficiency of different classifiers in identifying glaucomatous VFs, followed by an analysis of the Bayesian Network (BN) classifier structure. Secondly, a comparison is carried out of the classifiers when applied to the multivariate time series VF data. This includes the results of our spatio-temporal classifier (STC) and an analysis of the BN and STC structures.

6.1 Non-Temporal Classification and Analysis

6.1.1 Comparison of Classifiers

Figure 5 shows the ROC curves of the Bayesian network classifiers when 10-fold cross-validation is applied to each classifier on the non-temporal VF data. Also included is the AUC for each method. It can be seen that of the Bayesian classifiers the best curve is generated when using the Bayesian network with only one parent allowed (AUC is 0.94). The next best is naïve Bayes classifier with an AUC of 0.9 with Tree-Augmented Network (TAN) and the Bayesian network with two parents performing worst. This could well be due to overfitting or lack of data as these models will have a higher number of parameters.

Figure 6 shows a comparison between the best Bayesian classifier found with linear regression and k-nearest neighbour (k-nn), where k was chosen using 10-fold cross validation. The BN with one parent performs generally better than k-nn (which has an AUC of 0.91) and is comparable to linear regression (both have an AUC of 0.94).

6.1.2 Network Analysis

One of the advantages of the BN classifier over the naïve Bayes classifier, k-nn and linear regression is the *transparency* of the model. That is, all of the relationships are made explicit in the BN structure. Figure 7 shows one such structure with respect to the spatial arrangement of the VF. This has

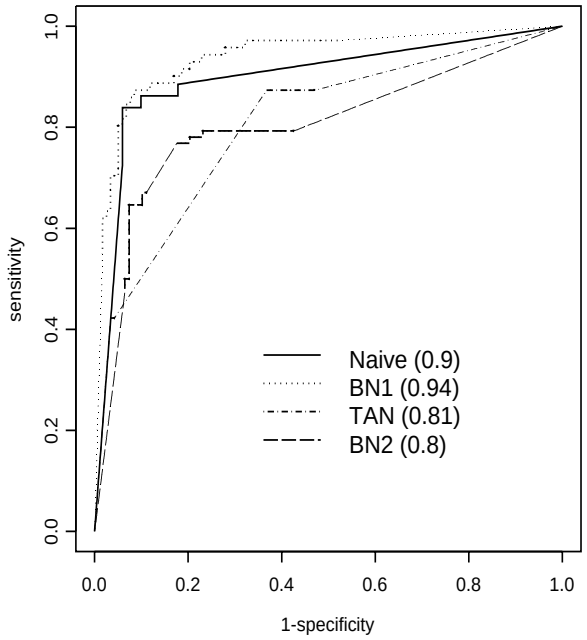

Fig. 5. ROC curves of Bayesian network classifiers on static VF Data

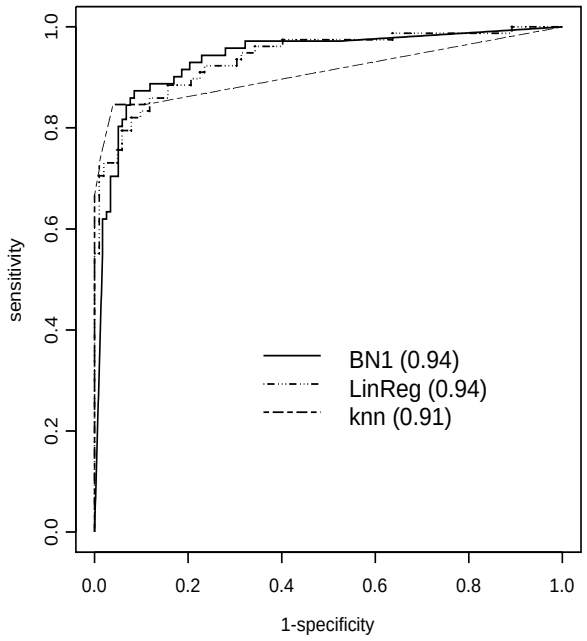

Fig. 6. ROC curves of best Bayesian network and standard statistical classifiers on static VF Data

been learnt from the non-temporal VF dataset. It is obvious that many of the dependencies between VF points are spatial in nature.

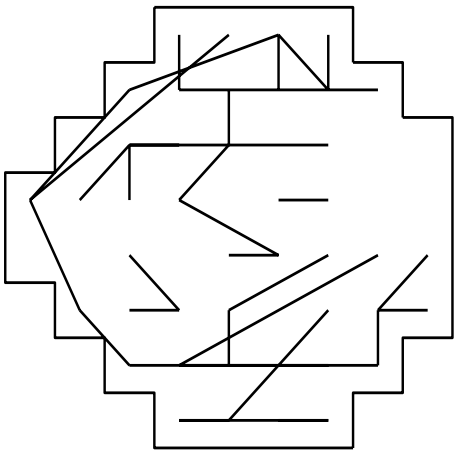

Fig. 7. Structure of BN1 on static data

We now make use of expert knowledge concerning the anatomy of the eye in order to assess the quality of the BN structure. Figure 8 shows the visual field of the right eye with the angle of corresponding nerve fibre bundle entry to the optic-nerve-head of each VF point [7], the optic nerve being where information is carried from the retina to the visual cortex .

			268	262	252	245		
		264	274	281	275	260	246	
	271	285	291	296	298	283	253	229
278	287	291	298	312	329	318		218
83	76	68	55	34	11	13		167
	85	78	66	56	48	60	95	136
		88	81	77	80	93	112	
			93	95	100	108		

Fig. 8. Expert Knowledge: The Angle Between Each Visual Field Point and the Optic-Nerve-Head

It is expected that VF points with similar angles should be more closely related and so we calculate the mean angle difference between the parent of a link and the child of a link. We now introduce a metric to score the mean optic-nerve-head angle distance in equation 4.

$$\sum_{i=1}^{m} \frac{1}{m} |\alpha(par(i)) - \alpha(chd(i))|, \tag{4}$$

where m is the number of links in a network, $\alpha(i)$ returns the optic-nerve-head angle of VF point i, $par(i)$ return the VF point of the parent of link i, and $chd(i)$ returns the VF point of the child of link i.

For the BN with a maximum of one parent it was found to be 15.28 degrees. This is relatively low as the maximum difference between two angles will be 180 degrees.

We are interested in which VF points are most predictive in classifying the VF as glaucomatous or not. Marked in 'x's on Figure 9 are the VF points that are discovered with direct links from the classifier node. Learning these links can be thought of as a form of feature selection for classifying the VF. This is similar, for example, to a selective naïve Bayes classifier [23] which is used to find relevant features. Interestingly, these VF points reflect a common feature of the early stages of glaucoma called the 'nasal step' where particular VF points around the nasal and superior peripheral areas indicate the early onset of glaucoma.

The discovered features also include arcuate paracentral defects which was unexpected. Some of these x's are in the temporal visual field (temporal in the physiological sense), which are not conventionally thought to be important points for classification. This will be followed up in further research and may show the potential of Bayesian network models in learning links that can teach clinicians interesting/informative patterns in clinical data.

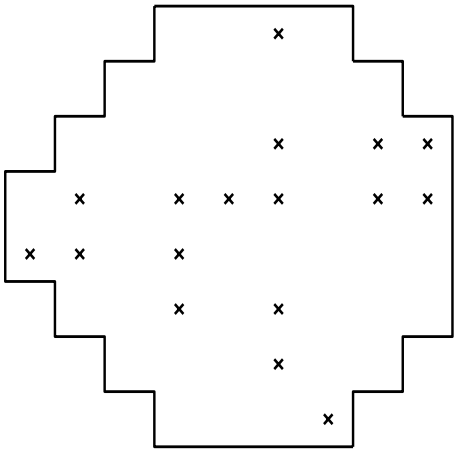

Fig. 9. Features discovered as Direct Descendants of Class Node for BN1 on static data

It is worth pointing out that there is the possibility that our models are learning links that reflect that classification process (in this case, AGIS). The AGIS classification requires clusters of abnormal points. Because of the distribution of VF test points in nerve fibre bundles, clusters are more likely in some regions of the VF than others. The models may, therefore, be more likely to find

'influential' points in some regions of the field than others. This, of course, does not prevent our models finding other 'unrecognised' influential points. The finding of temporal and paracentral influential points may indicate this.

In this section we have shown how Bayesian network classifiers have the potential to perform as well as other statistical classifiers in determining glaucomatous VFs, but also offer a way to help understand the nature of such conditions through the analysis of the network structures. In the next section we extend the BN classifier paradigm to model spatial and temporal VF data.

6.1.3 Comparison of Bayesian Inference with Clinicians' Decisions

Before we train the Bayesian classifiers on the temporal data, we investigate how the non-temporal classifiers, trained on the non-temporal data, perform on unseen time series data and compare how they perform to the classifications made by clinicians. We have done this because it will give us a more accurate idea of how the classifiers perform when tested on new data from completely different distributions. Figure 10 shows some typical results from the experiments: it plots the probability of a VF being glaucomatous according to the BN classifier with a maximum of one parent along with the point of conversion according to clinicians (dotted line). Note that clinicians did not reclassify the field after conversion has been determined.

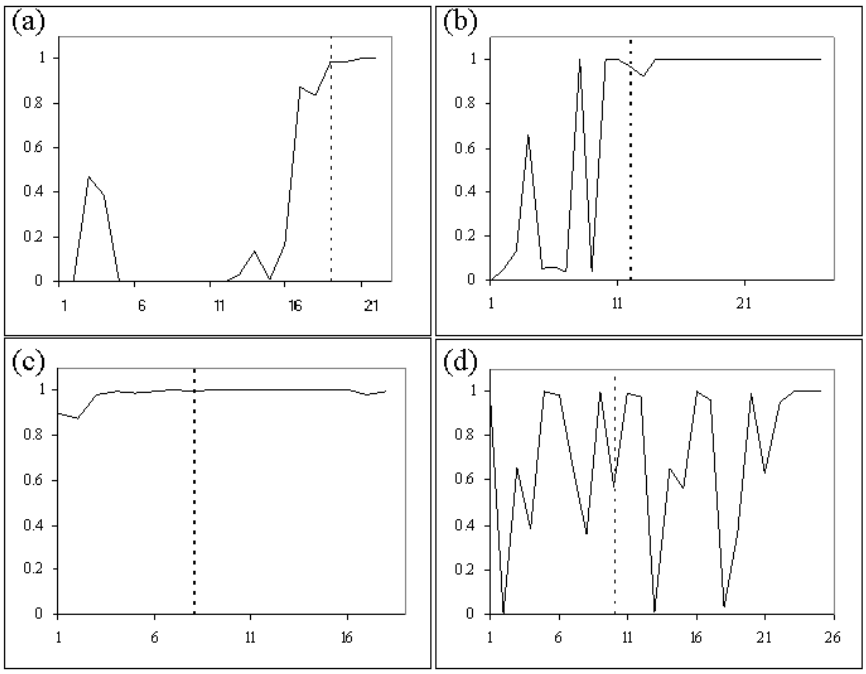

Fig. 10. $p(1|x)$ versus true class for representative set of time series

Many cases occurred within the unseen data where the probability of glaucoma

according to the BN classifier rapidly increases from almost zero to one several time points before the clinician decides the VF has converted to glaucomatous (for example see Figure 10a). This may well be due to the methods employed by clinicians whereby a VF is not considered to be glaucomatous until three VF tests in a row have reached the threshold for classification of glaucoma. Another common feature was where the probability remained low for a while and then began to fluctuate from one time point to the next before settling on a high probability of glaucomatous, shortly before the clinicians classifies the VF as converted (for example, see Figure 10b). This may be useful in that the fluctuation may be an early indicator that the VF is about to convert as it tends to begin some time before the clinician decides. Another interesting result that was observed occasionally was where the BN classifier classified the VF as glaucomatous from the outset (Figure 10c). This could be due to an error or bias in the classifier (such as overfitting or lack of data) or the fact that the classifier has discovered a feature not used by the clinicians and has in fact correctly identified the early onset of glaucoma. Examples were also noticed during the experiments, where the classifications are less clear cut. For example, Figure 10d implies that the BN classifier has really failed to successfully classify the VF before or after conversion, though it must be stressed that these were uncommon (as the ROC results in the previous section indicate). It should also be reiterated that clinicians did not reclassify a VF after it was considered to convert so the classifier may not be performing as badly as it seems (the fluctuations may reflect genuine variance in the VF datasets). We will be investigating these cases in the future.

6.2 Analysis of STC for Temporal Classification

6.2.1 Comparison of Classifiers

We now explore how the different classifiers perform on temporal VF data. This includes the Spatio-Temporal Bayesian Network Classifier (STC) which makes use of our spatio-temporal learning algorithms, described in Section 4. Figure 11 shows the ROC curves for the different Bayesian classifiers including the STC (with a maximum of 2 and 3 parents). It can be seen that the worst performers are naïve Bayes and TAN on this dataset (with respective AUCs of 0.77 and 0.79). This could be due to there being more data than on the static dataset and so the more complex models are less prone to suffering from over-parameterisation. The BN classifier with 2 parents scores an AUC of 0.81. However, the STCs with 2 and 3 parents do better with AUCs of 0.84 and 0.85, respectively.

The misclassification versus cost chart in Figure 12 shows how the classifiers perform when we assign different costs to the misclassifications. This corre-

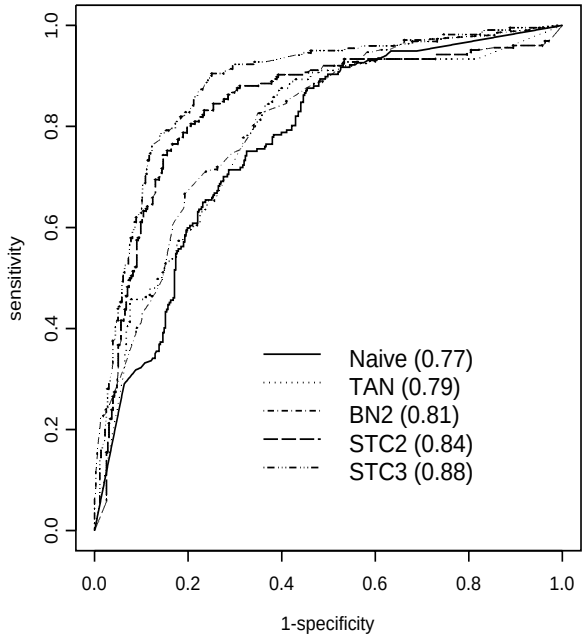

Fig. 11. ROC curves of Bayesian network classifiers on temporal VF Data

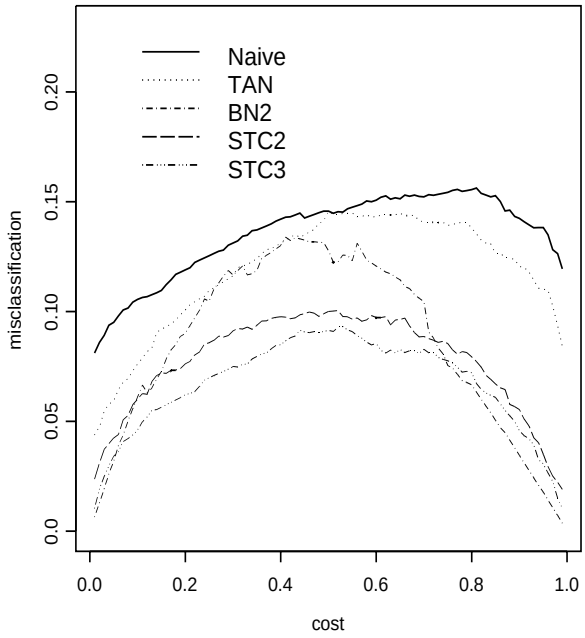

Fig. 12. Misclassification cost (equation 2)versus k_0 for Bayesian network classifiers on temporal VF Data

sponds to considering different values for the threshold t in equation 1. After scaling the costs k_0 and k_1 such that $k_0 + k_1 = 1$, it follows that $t = k_0$, the cost of misclassifying a normal eye as glaucomatous. This is reported on the horizontal axis. As such, higher values will tend to correspond to higher specificity. On the vertical axis the misclassification cost of Equation 2 is plotted. The plot shows that Naïve and TAN perform worse than the other classifiers for all values of the threshold. This is a much stronger result than the one already suggested by the AUC measure, as it implies that the models are not recommended for any value of the threshold that might be chosen. The BN classifier with 2 parents performs worse than the temporal models for an intermediate range of the cost, but is competitive with STC2 and STC3 for low and high values of specificity. The latter was not that evident from the ROC curves. It should be pointed out that results may be more precise than they appear because of inherent imprecisions of clinical classification. Firstly, they require a repeat of 3 abnormal fields. Fields will therefore be glaucomatous earlier in the series. Secondly, fields are not re-classified following conversion (so a field may be labelled as converted but, in reality, may have returned to normal. Our classifiers will pick this up and will suffer as a false negative.

To have a fairer comparison with the STC model, we decided to build the linear regression and k-nn models on the same training data used for STC. The columns of this database are now the VF variables and intraocular pressure at time $t-1$ and t, as well as age and gender. The value of k was chosen using 10-fold cross validation. Figure 13 shows the ROC curve of the STC with a maximum of three parents against linear regression and k-nn, both using the time-shifted variables. It seems that linear regression does somewhat better (AUC $= 0.93$) than STC and k-nn, both of which have similar ROC curves, the AUCs being 0.89 and 0.88, respectively. When we look at the misclassification versus the cost in Figure 14, we can see that whilst linear regression does considerably better when the cost is not extreme (between 0.4 and 0.6), the performances of the three methods are relatively similar beyond these costs, at either extreme.

Generally, because glaucoma prevalence is low and the condition is usually only slowly progressive, false positive diagnoses are regarded as more costly. Cost includes harm to the patient through a false diagnosis as well as the actual cost of treatment. This means that the right-hand side of the plots in Figures 12 and 14 is the most interesting. However, in reality, costs depend also on the individual patient. For instance, in a young patient with advanced disease, the cost of failing to identify progression is greater than in an elderly patient with early disease.

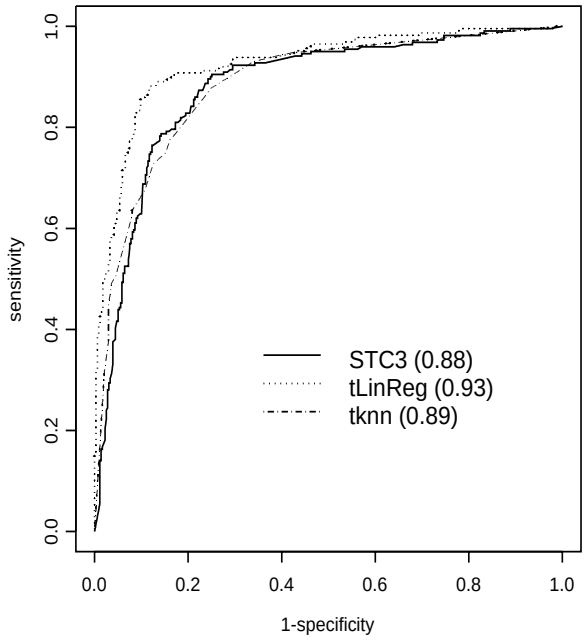

Fig. 13. ROC curves of best temporal Bayesian network, temporal linear regression and temporal k-nn on temporal VF Data

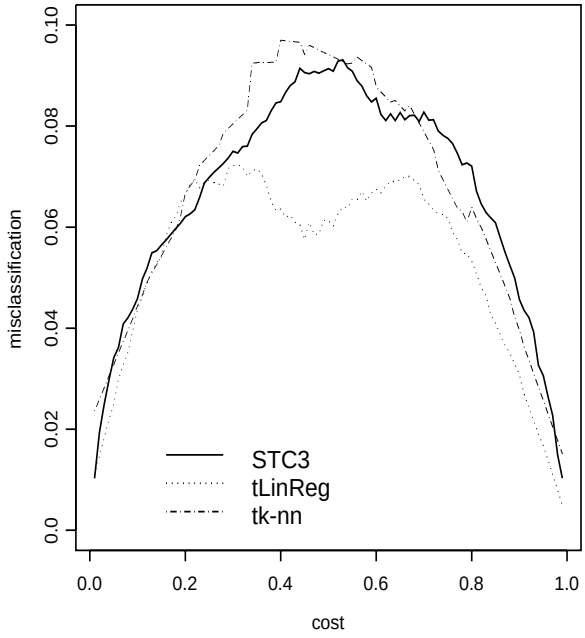

Fig. 14. Misclassification cost (equation 2) versus k_0 for temporal Bayesian network, temporal linear regression and temporal k-nn on temporal VF Data

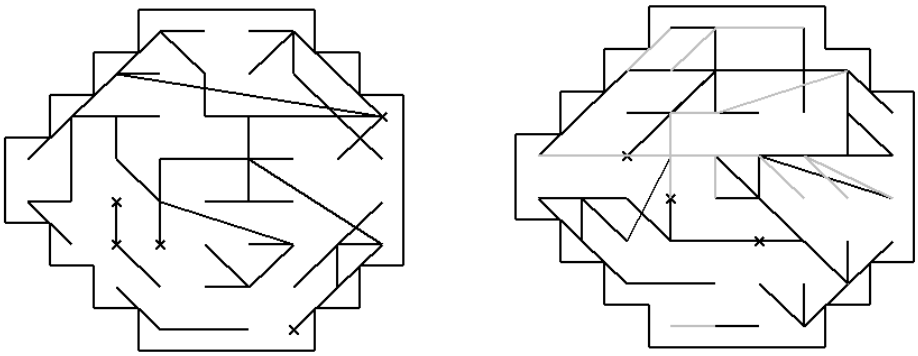

Fig. 15. Network Structures on temporal VF Data using BN2 (left) and STC3 (right). The Grey Lines Represent First Order Temporal Links

6.2.2 Network Analysis

Figure 15 shows the discovered network structures for the BN with a maximum of 2 parents and the STC with a maximum of 3 parents, learnt from the temporal dataset. It is evident primarily that both networks have a distinct spatial nature as would be expected. However, the blind spot, which should be independent of other VF variables, has been linked to other variables. This could be due to noise in the data or due to spurious correlations between variables. This may be made even worse when temporal correlations are allowed. Indeed, the STC does appear to indicate more relationships associated with the blind spot than the non temporal BN (likely due to spurious correlation), though enlargement of the blind spot (peripapillary atrophy) is a feature of glaucoma and may show up.

Looking at the mean optic-nerve-head angle metric that was introduced in equation 4 in Section 6.1.2 for comparing the networks to expected characteristics of the eye, we see a relatively low mean angle difference again, 24.71 degrees for BN2 and 24.74 degrees for STC3. This shows that our networks are reinforced by anatomical knowledge.

The 'x's mark the direct links to the class node. Notice that these are far fewer than in the BN discovered from the non-temporal data (Figure 9). This is likely to be due to the IOP variable that was also included in the temporal data and has been found to be a strong indicator of glaucoma. It was discovered to be linked to the class node in all networks in our experiments. Also, many fields in the temporal series are 'nearly' glaucomatous (but AGIS classified as normal) which may affect the number of direct links to the class node. Links that are grey are temporal. However, many of the temporal links that were discovered were autoregressive links (links from one variable to itself at the subsequent time point) and as such they are not visible in the figures. The improvement in classification when using the STC over the BN, which is evident in the ROC curves and misclassification costs of figures 11 and 12 are likely to be

due to its ability to take into account these temporal relationships in order to classify VFs. In the next section we discuss how this is achieved and possible implications.

7 Discussion

The use of spatio-temporal Bayesian network classifiers (STCs) have allowed us to gain insight into the evolution of glaucomatous damage, by modelling the spatio-temporal nature of the data. The model brings many benefits to the glaucoma community. First of all, it allows one to easily combine different types of data, for example visual field, intraocular pressure (IOP) and optic-nerve-head structures. These measurements are rarely useful when taken in isolation, so a model that combines them into a common framework whilst making explicit the relationships between them is highly useful to clinicians.

Furthermore, the STC allows one to incorporate the temporal aspect of the data in the model. This will help researchers model the disease process and learn about its pathogenesis, which could result in more accurate and precise estimates of the rate of progression and in the identification of risk factors for progression (indeed, the STCs will identify some of these themselves) and responses to therapeutic intervention. What is more, temporal models are capable of modelling more complex interactions.

The obvious improvement in the ROC curves when including the temporal links is likely to be due to the ability of the model to capture the changes within VF points. For example, many links were found such as those illustrated in Figure 16 whereby an autoregressive temporal link was found between a VF variable at time $t-1$ and at time t. VF at time t was also the child of a non-temporal link to the class node. This means that the interaction between VF

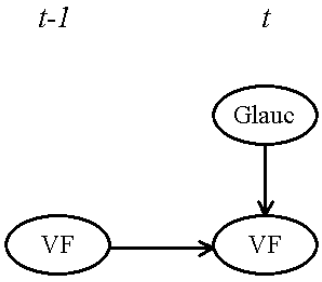

Fig. 16. Temporal Relationships modelling change in VF Values to Improve Classification

at time $t-1$ and VF at time t can affect the probability of the class node (Glauc) due to the 'explaining away effect' [31]. This is where the observation

of one parent of a node (here the VF point at time t-1) will affect the posterior distribution of another parent of that node (here the class node). Obviously, this type of interaction cannot be modelled by non-temporal classifiers.

We have begun experiments on some incoming data, which include more clinical variables such as IOP and medication. We have found interesting results whereby the glaucoma converter class node is regulated by IOP which is in turn regulated by whether a patient receives medication. The medication node, itself is regulated by the glaucoma node and so a cycle exists over time (see Figure 17). If someone exhibits high IOP, it is likely to be related to the risk of

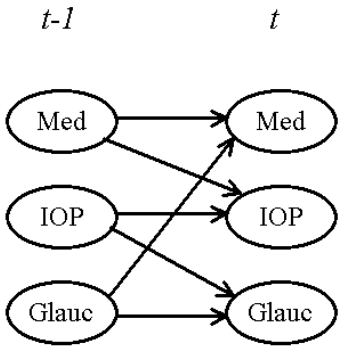

Fig. 17. A Temporal Cycle Whereby IOP Regulates Glaucoma Which in Turn Regulates Medication

converting to glaucomatous VF loss. If someone suffers from glaucoma, they are likely to be given medication (e.g. to lower IOP), resulting in their IOP dropping by their next visit. This means that in many cases a low IOP is observed despite the onset of glaucoma. The temporal nature of this can easily be modelled by temporal BNs. However, such temporal cycles cannot be modelled with static models.

As well as allowing one to combine different types of data and to include temporal links, Bayesian networks have the advantage of explicitly modelling the spatial relationships between the visual field and the clinical variables. Links have been discovered within our networks that were not previously recognized clinically and could lead to the discovery of new characteristics of VF deterioration. Indeed, some known characteristics have been identified in our networks, as well as interesting new ones which we intend to follow up. Furthermore, by querying the network or changing the prior probabilities one can observe the effect of these on the network structure. This could for example inform clinicians about possible sources of risk for progression.

Bayesian network models are similar in nature to the way in which clinicians work - various examinations are made and the result of each is added to 'the clinical picture' to either increase or decrease the probability of disease being present or of disease progressing. As such, the network provides the clinician

with probabilities for abnormality. These are clinically more meaningful than discrete thresholds on scores, which are used to currently determine glaucoma conversion, as described in Section 5.1.

8 Conclusion and Future Work

In this paper we have investigated a number of different classifiers for identifying deterioration in the visual field associated with glaucoma. We have focussed on Bayesian classifiers due to their ability to explicitly model the relationships between variables.

We have introduced a new form of Bayesian classifier which we call the Spatio-Temporal Bayesian network Classifier which has been shown to be an improvement on the other Bayesian classifiers and comparable to other statistical classifiers. We have tested the classifiers on two visual field datasets, one non-temporal and one temporal, and have shown how the resulting Bayesian network classifiers can be used to help understand the nature of visual field deterioration in the form of network structure analysis and inference. We have identified within the structures, various characteristics including the 'nasal step', whereby certain areas of the visual field indicate the onset of glaucoma. Inference has shown that there is potential to understand how the clinicians come to their decisions and possibly use the information to improve upon the current classification algorithms of a visual field.

Since we aim to learn temporal models from visual field data, a natural extension of the work in this paper would be to see how the models can be used to forecast future states of the visual field given previous observations. It would also be extremely valuable if we could use the models to predict future class states given the previous states of a visual field.

BNs facilitate the use of prior knowledge which we intend to explore. For example, we can integrate expert knowledge regarding the structure of nodes as well as the angle from each visual field point to the optic-nerve-head. We also intend to look at other visual field datasets. The problem with the temporal dataset in this paper is that classification required 3 abnormal fields in a row. Our resulting classifiers may well end up 'tied' to the clinical classification and may be reflecting those decisions rather than modelling the true nature of change. Another way around this problem would be to submit only normal patient data and allow our models to identify visual fields that are significantly different from normal. However, this type of model will tell us less about the nature of visual field deterioration as it will be modelling healthy eyes.

Acknowledgements

The work of Allan Tucker on this project was funded by the EPSRC, GR/R35018/01. The work of Veronica Vinciotti on this project was funded by the BBSRC, grant 100/EGM17735. We would like to thank Nick Strouthidis for his help with collating the visual field dataset.

References

[1] K.E. Anderssen and V. Jeppesen, Classifying visual field data, Technical Report, MSc Thesis, Aalborg University, Denmark, 1998.

[2] G.F. Cooper and E. Herskovitz, A Bayesian method for the induction of probabilistic networks from data, *Machine Learning* 9 (1992) 309-347.

[3] P. Dagum, A. Galper and E. Horvitz, Dynamic network models for forecasting, in: *Proceedings of the 8th Annual Conference on Uncertainty in AI* (Morgan Kaufmann, San Mateo,1992) 41-48.

[4] P. Domingos and M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss, *Machine Learning* 9 (1997) 309-347.

[5] F.W. Fitzke, R.A. Hitchings, D. Poinoosawmy, A.I. Mc- Naught and D.P. Crabb, Analysis of visual field progression in glaucoma, *Br J Ophthalmol.* 80(1) (1996) 40-48.

[6] N. Friedman, D. Geiger and M. Goldszmidt, Bayesian network classifiers, *Machine Learning* 29 (1997) 131-163.

[7] D.F. Garway-Heath, F. Fitzke and R.A. Hitchings, Mapping the visual field to the optic disc, *Opthalmology 2000* 107 (2000) 1809-1815.

[8] M.H. Goldbaum, P.A. Sample, K. Chan, J. Williams, T. Lee, E. Blumenthal, C.A. Girkin, L.M. Zangwill, C. Bowd, T. Sejnowski and R. Weinreb, Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry, *Invest Ophthalmol Vis Sci.* 43(1)(2002) 162-169.

[9] J.J. Grefenstette, Optimization of Control Parameters for Genetic Algorithms, *IEEE Transactions on Systems, Man and Cybernetics* 16(1)(1986) 122-128.

[10] D.J. Hand, *Construction and assessment of classification rules* (Wiley, Chichester, 1997).

[11] D.J. Hand and K. Yu, Idiot's Bayes - not so stupid after all?, *International Statistical Review* 69 (2001) 385-398.

[12] D.J. Hand and V. Vinciotti, Local versus global models for classification problems: fitting models where it matters, *The American Statistician* 57(2) (2003) 124:131.

[13] A. Heijl, G. Lindgren and J. Olsson, Normal variability of static perimetric threshold values across the central visual field, *Arch Ophthalmol.* 105 (1987) 1544-1549.

[14] A. Heijl, G. Lindgren and A. Lindgren, Extended empirical statistical package for evaluation of single and multiple fields in glaucoma: Statpac 2, in: *Perimetry Update 1990/1*, Editors R. Mills and A. Heijl A (Amsterdam, Kugler Publications, 1990), 303-15.

[15] M. Henrion, Propagating uncertainty in Bayesian networks by probabilistic logic sampling, in: *Proceedings of the 2nd Annual Conference on Uncertainty in AI* (Elsevier Science, New York, 1998) 149-163.

[16] T. Hothorn and B. Lausen, Bagging tree classifiers for laser scanning images: a data and simulation-based strategy, *Artificial Intelligence in Medicine* 27(1) (2003) 65-79.

[17] M.V. Ibanez and A. Simo, Spatio-temporal modelling of perimetric test data, *Statistics in Medicine*. To appear.

[18] D. Kamal, D. Garway-Heath, S. Ruben, F. O'Sullivan, C. Bunce, A. Viswanathan, W. Franks and R. Hitchings, Results of the betaxolol versus placebo treatment trial in ocular hypertension, *Graefes Arch Clin Exp Ophthalmol.* 241(2003) 196-203.

[19] L. Katz , A. Sommer, Asymmetry and variation in the normal hill of vision, *Arch Ophthalmol.* 104(1986) 65-68.

[20] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, Optimization by Simulated Annealing, *Science* 220(4598) (1983) 671-80.

[21] R. Kohavi, Scaling up the accuracy of naïve-Bayes classifier: a decision-tree hybrid, in: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (AAAI Press, Portland, Oregon, 1996) 202-207.

[22] P. Langley, W. Iba and K. Thompson, An analysis of Bayesian classifiers, in: *Proceedings of the Tenth National Conference on Artificial Intelligence* (AAAI Press, San Mateo, CA, 1992) 223-228.

[23] P. Langley and S. Sage, Induction of selective Bayesian classifiers, in: *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, Seattle, WA, 1994) 399-406.

[24] X. Liu, S. Swift and A. Tucker, Using Evolutionary Algorithms to Tackle Large Scale Grouping Problems, in *Proceedings of the Genetic and Evolutionary Computation Conference* (Morgan Kaufmann, San Francisco, USA, 2001) 7-11.

[25] M. Ramoni and P. Sebastiani, Bayesian Methods, in *Intelligent Data Analysis, An Introduction*, Editors M. Berthold, D. Hand, (Springer, Berlin, 1999) Chapter 4.

[26] S. Swift and X. Liu, Predicting glaucomatous visual field deterioration through short multivariate time series modelling, *Artificial Intelligence in Medicine* 24(1)(2002) 5-24.

[27] A. Tucker, X. Liu and A. Ogden-Swift, Evolutionary learning of dynamic probabilistic models with large time lags, *The International Journal of Intelligent Systems* 16(5)(2001) 621-646.

[28] A. Tucker, D. Garway-Heath and X. Liu, Spatial operators for evolving dynamic probabilistic networks from spatio-temporal data, in: *Proceedings of the Genetic and Evolutionary Computation Conference*, (Springer-Verlag, Chicago, USA, 2003) 12-17.

[29] V. Vinciotti and D.J. Hand, Scorecard construction with unbalanced class sizes, *Journal of the Iranian Statistical Society* 2(2) (2003) 189-205.

[30] J. Weber and S. Rau, The properties of perimetric thresholds in normal and glaucomatous eyes, *Ger J Ophthalmol.* 1(1992) 79-85.

[31] M.P. Wellman and M. Henrion, Explaining explaining away, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(3)(1993)287-307.

[32] Advanced Glaucoma Intervention Study. 2. Visual field test scoring and reliability, *Ophthalmology* 101(1994) 1445-1455.

Table 1
Breakdown of the Datasets

Dataset	VFstatic	VFtemporal
# VF Variables	54	54
# Clinical Variables	2	3
# VF Tests	180	588
# Patients	180	24
Pos/Neg Class Ratio	43:57	39:61

Figure 1: A Typical TBN with 2 Time Slices. Note the links within one time slice and those spanning from one to the next.

Figure 2: The Spatio-Temporal Bayesian Network Classifier.

Figure 3: The Spatial Operators: (a) Add a Link in the 1st Order Neighbourhood (b) Mutate a Parent to Within its 1st Order Neighbourhood (c) Spatial Crossover

Figure 4: A Typical VF Test from a Healthy Eye and a Glaucomatous Eye.

Figure 5: ROC curves of Bayesian network classifiers on static VF Data

Figure 6: ROC curves of best Bayesian network and standard statistical classifiers on static VF Data

Figure 7: Structure of BN1 on static data

Figure 8: Expert Knowledge: The Angle Between Each Visual Field Point and the Optic-Nerve-Head

Figure 9: Features discovered as Direct Descendants of Class Node for BN1 on static data

Figure 10: $p(1|x)$ versus true class for representative set of time series

Figure 11: ROC curves of Bayesian network classifiers on temporal VF Data

Figure 12: Misclassification cost (equation 2)versus k_0 for Bayesian network classifiers on temporal VF Data

Figure 13: ROC curves of best temporal Bayesian network, temporal linear regression and temporal k-nn on temporal VF Data.

Figure 14: Misclassification cost (equation 2) versus k_0 for temporal Bayesian network, temporal linear regression and temporal k-nn on temporal VF Data

Figure 15: Network Structures on temporal VF Data using BN2 (left) and STC3 (right). The Grey Lines Represent First Order Temporal Links

Figure 16: Temporal Relationships modelling change in VF Values to Improve Classification

Figure 17: A Temporal Cycle Whereby IOP Regulates Glaucoma Which in Turn Regulates Medication