On Bayesian Network and Outlier Detection

Sakshi Babbar and Sanjay Chawla
University of Sydney, Sydney

COMAD 2010

Outline

- Introduction
- Motivation and contributions
- Bayesian Network
- Methodology
- Experiments
- Conclusion and future scope of work

Introduction

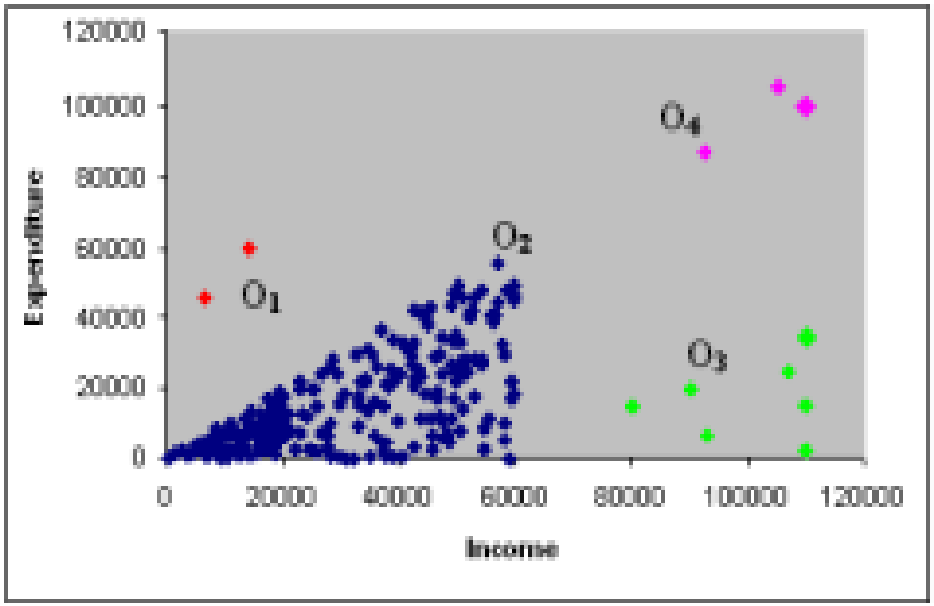
• An outlier is a data instance in a database which is significantly different from the norm.

• A real challenge in outlier detection is to ascertain whether discovered data point is in fact interesting.

• Current outlier mining methods, for example, distance based approach, looks for those data points which are far away from their neighbors.

• Example illustration

Introduction



• As demonstrated, distance based approaches ignore valuable information that is available in the data.

• Challenge here is to overcome mismatch between outliers as entities "which are far away from their neighbors" and "real" outliers

Introduction

• We propose to integrate domain knowledge in discovering process of outlier detection.

• Essential idea is to validate data points discovered as outliers on the grounds of domain.

• We define outliers as :

"unlikely events under the current favored theory of the domain"

• In this present paper, we used Bayesian Network to capture domain knowledge.

• BN captures causal relationship among a set of variables using a graph in which variables are nodes and causations in indicated by arrows.

Introduction

• The strength of relationship among dependent variables is represented in terms of probability.

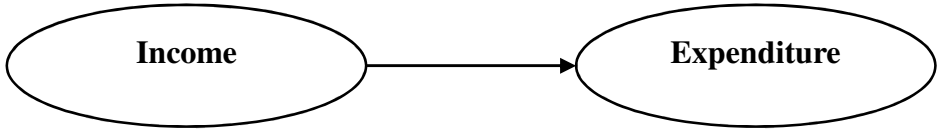
• In our approach, causal relationships encoded in BN were exploited using two quantitative rules to identify anomalous patterns.

• These rules were used to score instances based on joint probability distribution in BN.

• Later, the instances were sorted by their score and top n low probability scored instances were declared as outliers.

• Example illustration :

Introduction



P (I=low)	P (I=high)
90%	10%

I	P (E=low | I)	P (E=high | I)
low	98%	2%
high	70%	30%

- Taking Bayesian structure into account, joint probability of an event :

Pr(Income=high, Expenditure=high) = Pr(Expenditure=high | Income=high) *
Pr(Income=high) = 0.03

Similarly

Pr(Income=low, Expenditure=high) = Pr(Expenditure=high | Income=low) *
Pr(Income=low) = 0.018

Motivation

Example : consider following sequences of coinflip

HHTHTTHTHT
HHHHHHHHHH

Probability will be $(1/2)^{10}$ or 1 in 1024

sequence two is more of a coincidence than the sequence one. Flipping 10 heads in a row all grab our attention because it suggest the existence of hidden causal structure in context where our current understanding would suggest no such structure should exist.

Coincidence are not just unlikely events by events of "unlikely kind/pattern "

HHTHTTHTHT
HHHHHHHHHH

Out of 1024 sequences, 252 are of this kind
Out of 1024 sequences, 2 are of this kind

Contributions

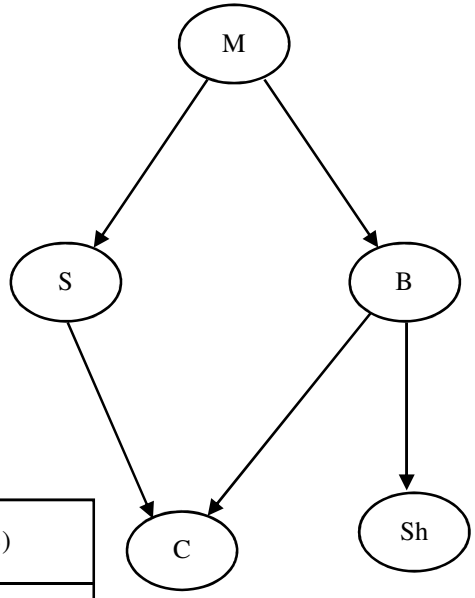
- We present two quantitative rules to help discovering anomalous pattern residing in the dataset in conjunction with the Bayesian network joint probability distribution to sort for those instances where similar anomalous pattern are present to maximum.

- We evaluate the validity of discovered outliers by explaining why identified data points are anomalous which indicates the credibility of our approach.

- Critical analysis of distance based techniques is also presented which highlights why distance based criteria may not be an accurate and effective technique to discover true outliers.

- Our experiments on variety of simulated and real datasets, shows that our overall approach is effective and accurate at the same time.

Bayesian Network



M	P (S=p | M)
p	45%
a	9%

M	P (B=p | M)
p	5%
a	79%

S	B	P(C=p |S,B)
i	p	60%
ni	p	8%
i	a	45%
ni	a	85%

B	P (Sh=p | B)
p	55%
a	76%

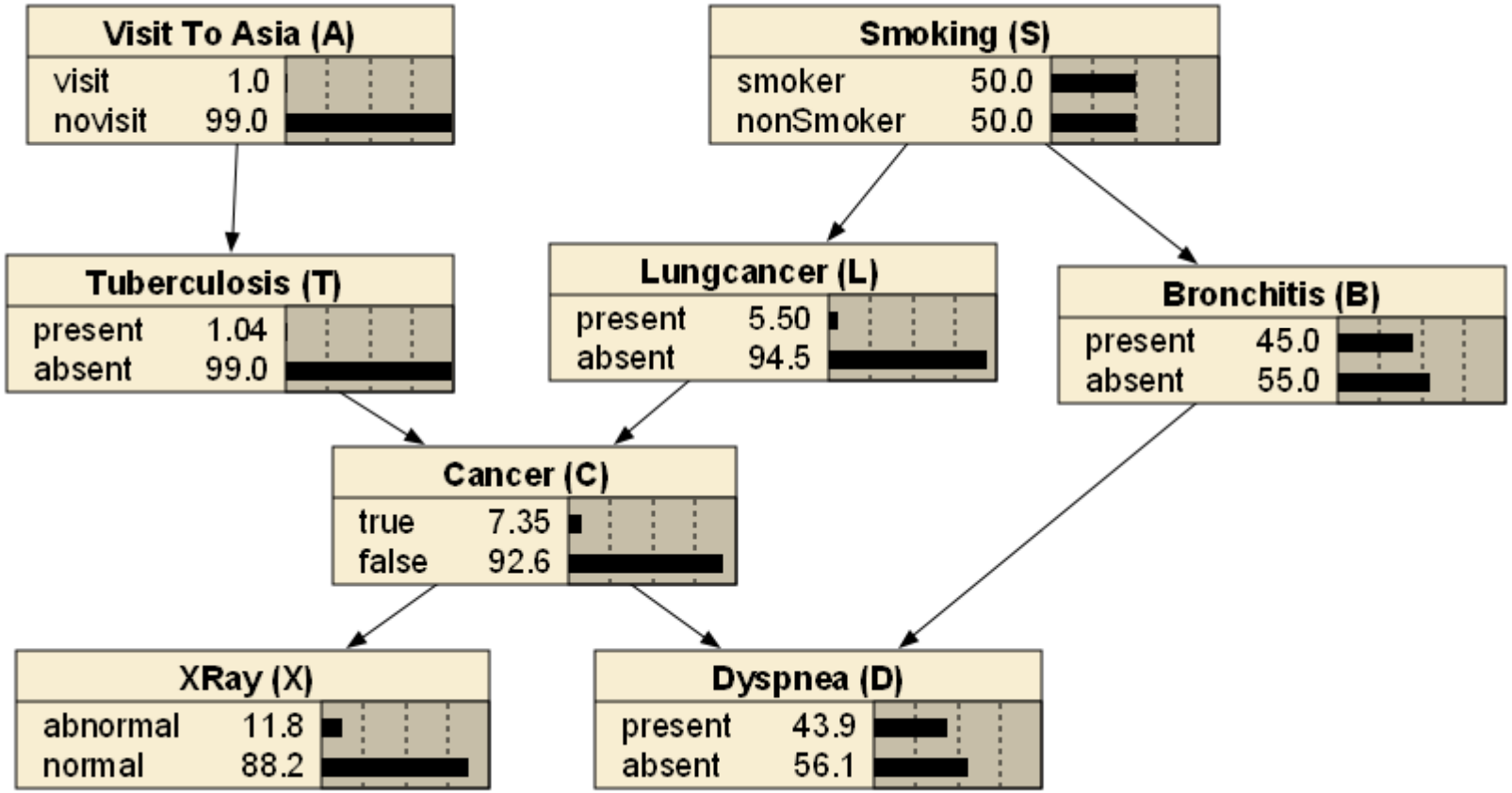
Joint probability in X is given by : $P(X) = \prod_{i=1}^{n} P(x_i|Pa_i)$

Example :

P(M, S, B, C, Sh) = P(S | M) * P(B | M) * P(C | S,B) * P(Sh | B) * P(M)

Bayesian model

ChectClinic



Methodology

- With the need of discovering true and meaningful outliers, we propose to find outliers by finding joint probability using Bayesian network.

- Essential idea is, for a given instance, we find :

  joint probability = prior X conditional probability

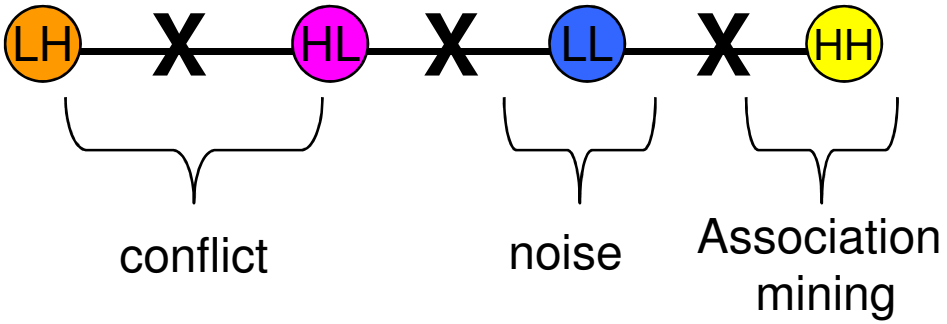
  across each of the variable in a given domain

- The product thus obtained gives us a score for the instance and *low scored instances were treated as potential outliers*.

- An important observation here is, product which constitute a score of a given instance, can give rise to four different situations

Methodology

- Consider



- Joint probability is formed by the combination of above listed situations. However :

  - it is always possible that any situation occur any number of times, while

  - it is not necessary that every situation will be present in the product.

Methodology

- Logically, joint probability of an instance will be low which has maximum of first three situations listed above.

- In order to find true outlying situations from the dataset, our focus is on finding those instances where joint probability is low because of situations one and two only (LH and HL) .

- Keeping this in mind, prior to finding joint probability of an instance, we checked the strength of the relationship.

- If for a given parent variable, prior belief is low and posterior of the direct child of this parent variable is also low then this posterior factor was not considered in finding joint probability

Methodology

- We refer situation one (say R1) and two (say R2) as two quantitative rules which were employed to uncover anomalous patterns in the dataset.

- In line with our definition of outliers, quantitative rules helped in uncovering anomalous patterns and joint probability helped in ranking low for those instances where similar anomalous patterns were found/present to maximum.

- Terminology : minsupp, minconf and maxconf

$$minsupp(X) = \min_i(support(X_i))$$

where X stands for any parent node in the dataset and X_i refers to any state of this node.

Methodology

- Rules

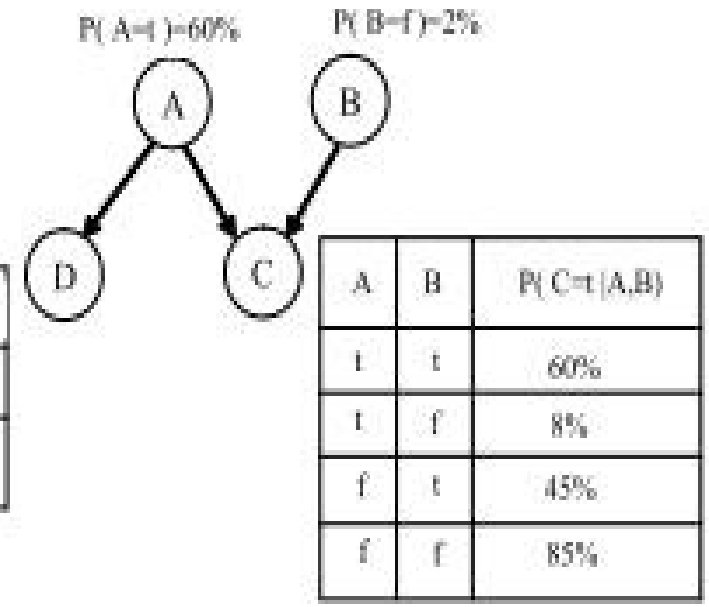
$$R_1 \Rightarrow (P(Pa(C)) = minsupp) \bigwedge (P(C|Pa(C) \geq maxconf))$$

$$R_2 \Rightarrow (P(Pa(C)) > minsupp) \bigwedge (P(C|Pa(C) \leq minconf))$$

Where C stands for any child node and Pa(C) refers to the parent(s) of this child node

Methodology

- Example illustration :



A	P(D=t | A)
t	55%
f	75%

A	B	P(C=t | A,B)
t	t	60%
t	f	8%
f	t	45%
f	f	85%

minsupp(A) = 40% and
minsupp(B) = 2%

Let : minconf = 10 % and
maxconf = 70%

P(A,B,C,D) = P(C | A,B) X
P(D | A) X
P(A) X
P(B)

Let : t_1 = { A= t, B = f, C= f, D= t }
t_2 = { A= f, B = f, C= t, D= t }

- Based on two qualitative rules, score of the instance t_1 and t_2 will be calculated as:

score (t_1) = P(C | A,B)= 0.92 X **P(D | A) = 0.55** X P(B) = 0.02 X P(A) = 0.40 **= 0.011**

score (t_2) = P(C | A,B)= 0.85 X P(D | A) = 0.75 X P(B) = 0.02 X P(A) =0.40 **= 0.005**

Methodology

- Algorithm

Input : Bayesian model (BN(N,E)), parameters minconf and maxconf, and a testset

Output : top *n* low probability data points in a testset

- Compute minsupp for all parents nodes X Є BN(N,E)
- For every testcase in testset, repeat steps(3-4)
- Compute conditional probability in child node given their parent(s), i.e., P(y)=Pr(y | Pa(y)) where Pa(y) Є X
- Apply rules R_1 and R_2 to uncover anomalous patterns and compute joint probability
- Sort joint probability
- Output top *n* low scored data points

Datasets

- Bayesian structure given, simulated dataset examples :

  ChectClinic (256 X 8)

  Busselton (1000 X 15)

- Given dataset, learnt Bayesian model and parameters examples :

  Hepatitis(155 X 19)

  Breast cancer(184 X 9)

Experiments

Testset were used for two different set of experiments :

1. identification of top n outliers and describing why these n data points are outliers.

2. analysis on data points discovered as outliers by our approach and NN approach.

    1. What patterns are observed using our own definition of discovering outliers using Bayesian network ?

    2. What approach does nearest neighbor technique follow to discover outliers?

    3. Why is it that an outlier discovered by Nearest neighbor technique is not an outlier from Bayesian point of view ?

Experiments

- Identification and description

  objective is to **identify outliers** and **present subspaces** which define outliers.

- We applied our methodology on all 10 datasets and explored top *n* outliers.

- We set the thresholds minconf = 10% and maxconf=80%.

- Description follows the annotation $X[x_i](x) \rightarrow Y[y_i](y)$

Experiments

Dataset : ChectClinic

Identification:

absent, abnormal, false, absent, present, absent, visit, smoker

Description: Instance is outlier in

1. **2D space of Visit to Asia[visit](1%) → Tuberculosis[absent](95%)**
2. **2D space of smoke[smoker](50%) → lungcancer[absent](90%)**
3. **2D space of cancer[false](92.6%) → Xray[abnormal](5%)**
4. **3D space of cancer[false](92.6%), bronchitis[absent](55%) → dyspnea[present](10%)**

Experiments

- Analysis on genuine and false outliers

objective to address on why there is a mismatch between outliers as observations **"which are far away from their neighbors"** and **"real"** outliers as identified using Bayesian approach

- we found top *n* outliers using distance based technique and validated these anomalies against the Bayesian model built.

- When any outlier found by distance based technique, stands among top *n* outliers in the domain, then these data points were considered true outliers.

- When experimented on 10 datasets, we found, overall accuracy of NN approach was not more than 40%.

Experiments

- What Bayesian approach follows ?

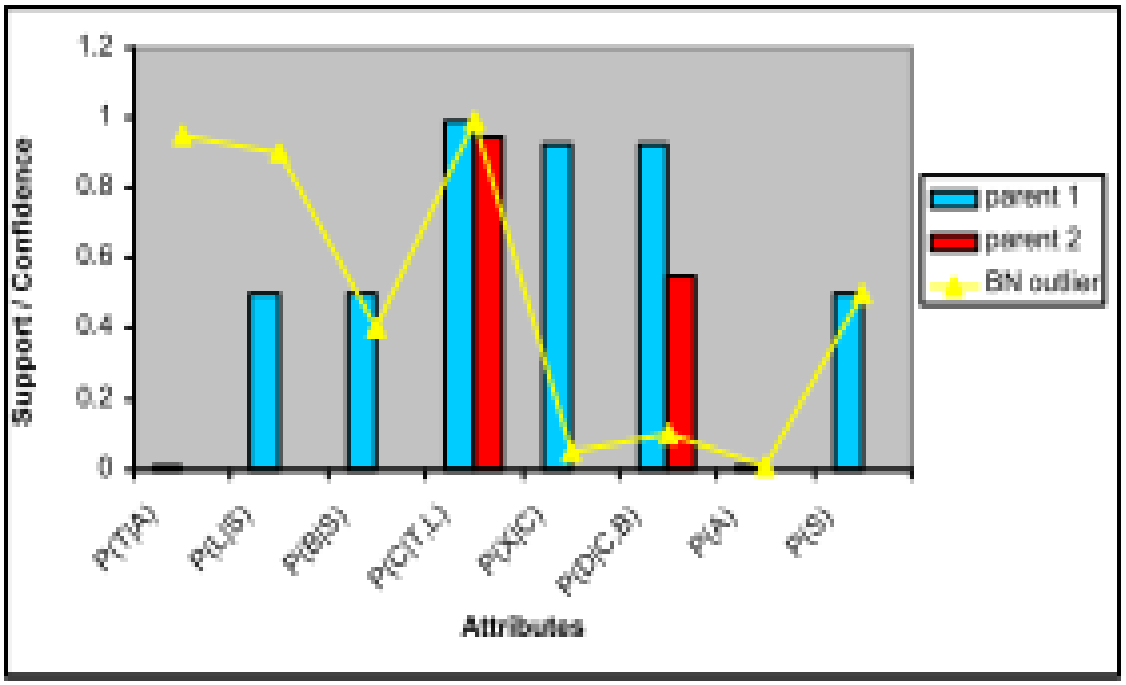
joint probability of several variables can be calculated from the product of individual probabilities of the nodes following chain rule of probability. For eg for ChestClinic BN :

$$P(A,S,T,L,B,C,X,D)=P(X|C) * P(D|C,B) * (C|T,L) * P(T|A) * P(L|S) * P(B|S) * P(S)$$

• Following definition of joint probability in BN, we explored inner structure on similar lines for top *n* outliers discovered by our approach for every dataset.

Experiments

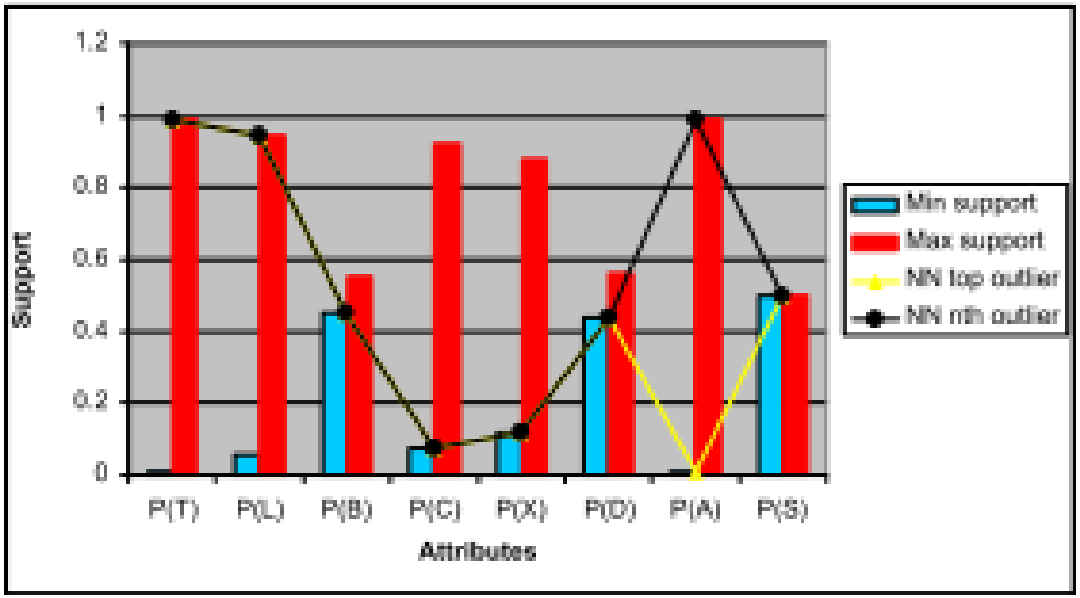
- Graph below represents pattern of top outlier in terms of conditional probability (confidence) and prior (support) which together constitute joint probability in the BN.



Experiments

- What NN technique follows ?

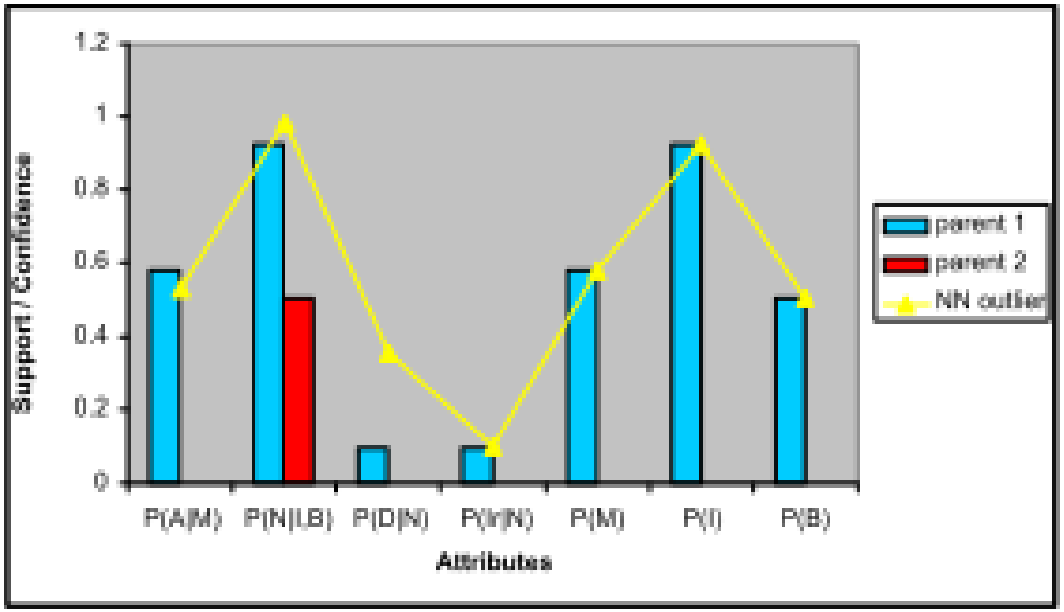
distance based approaches look for those data points where maximum number attributes have low support.



Experiments

- Why distance based outlier is not an outlier in BN ?

To answer this question, we simply took an outlier discovered by distance based technique and analysed pattern of this outlier from Bayesian perspective.



Conclusion

• We have introduced an approach to find meaningful outliers using domain knowledge captured by the Bayesian Network.

• We propose outliers are *unlikely events under the currently favored theory of the domain.*

• By structuring domain knowledge in Bayesian framework, anomalous pattern were uncovered using two quantitative rules.

• We presented the explanation on subspaces which define outliers. Such explanation contributes to a new, vital knowledge for the domain.

• Our approach illustrated why distance based technique fails to discover true outliers in mere support-based mining framework as compared to our approach which in support and confidence based mining framework.

Future scope of work

- We intend to work on high dimensional datasets.

- We also plan to apply our technique to a specific domain and work with specialist in the domain to help uncover potentially useful anomalies.

Questions ?

Thank you