On Bayesian Network and Outlier Detection

Sakshi Babbar and Sanjay Chawla

School of Information Technologies, University of Sydney,
Sydney NSW 2006, Australia
sakshi.babbar@gmail.com , sanjay.chawla@sydney.edu.au

Abstract

Existing studies on data mining has largely focused on the design of measures and algorithms to identify outliers in large and high dimensional categorical and numeric databases. However, not much stress has been given on the interestingness of the reported outlier. One way to ascertain interestingness and usefulness of the reported outlier is by making use of a domain knowledge. In this paper, we present a new measure to discover outliers based on background knowledge, represented by a Bayesian network. We define outliers as *"unlikely events under the current favored theory of the domain"*. We introduce two quantitative rules derived from the Bayesian network to uncover outliers. Furthermore, we use these rules to rank the instances based on joint probability distribution in the Bayesian network.

In our approach, we not only identified outliers but also explain why they are likely to be so. A critical analysis on distance based technique is also presented to show why there is a mismatch between outliers as entities "which are far away from their neighbors" and "real" outliers as identified using Bayesian Networks.

1 Introduction

An outlier is a data instance in a database which is significantly different from the norm. The objective in outlier detection, is not only to identify outliers in large and high dimensional databases but also to *correlate* them with actual anomalous events. For example, if the outlier detection techniques are being used for finding anomalies in network traffic, then outliers in network data should correspond to physical anomalies - like denial of service attack or ping flood. Thus if O is a set of discovered outliers from data and A is the set (unknown) anomalies, then an ideal good outlier detection method will have high precision and recall, i.e., both $P(A|O)$ and $P(O|A)$ are high. The challenge in outlier detection is that we rarely, if ever, have access to the anomalous set A. Thus like clustering, outlier detection is an unsupervised learning method.

Current data mining methods identify sparse regions in point cloud data to search for outliers. For example, in distance-based methods, a data point is an outlier if it is effectively far away from its neighbors. Variations on distance-based approaches, like those based on density, incorporate the local density of the region while reporting outliers, though the principle remains the same. However, as we will demonstrate, such approaches ignore valuable information that is available in the data.

Suppose we conceptually place a fine resolution grid on the point cloud space. For example, in an N-dimensional data set we can identify the grid cells with the a lattice Z^n. Now, distance-based outliers are essentially data points which live in sparse cells. In fact we can associate a probability with each cell, which is the percentage of data points which lie in that cell. In the language of pattern mining, cells with low (but non-zero) *support* contain the outliers. A major objective of this paper is to show that when we want to search for outliers and then use them to identify anomalous events, then the focus on *confidence* yields more meaningful results.

To elaborate more on above stated observations, consider a hypothetical dataset belonging to a certain region of the country, highlighting persons income and their expenditures. This example is extended version of one presented in [4]. The sample data in Figure 1 represents relationship between persons income (X-axis) and expenditure(Y-axis). As observed, data points are roughly clustered. We name them as O_1,O_2,O_3 and O_4 respectively. Cluster O_2 which is very dense, indicates that in a given region, persons expenditure is bounded within their income. Unlike cluster O_2, data points forming cluster O_1 indicates that there are very small percentage of people the expenditure of whom are higher than that of their income. Likewise, there are few people in region earning high but choose to spend low as represented by the cluster O_3. Lastly, a small percentage of people have high income and they prefer spending high as indicated by cluster O_4. In the above discussion, if the objective is find to find outliers using existing techniques such as distance based [10] or density based [6] then most likely these approaches will find data points belonging to the clusters O_4 as highly ranked potential outliers. This is because these data points are far away from their *k* nearest neighbor and hence are isolated and easily detected as out-

International Conference on Management of Data
COMAD 2010, Nagpur, India, December 8–10, 2010

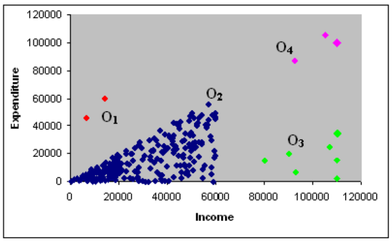
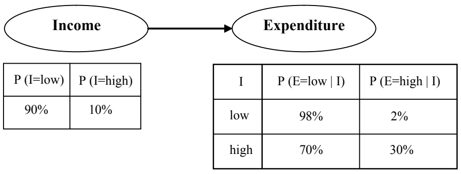


Figure 1: Shows objects in two dimensional space where X-axis represents Income and Y-axis represents Expenditure.

liers. To illustrate more on this, data points forming cluster O_3 has low support from their neighbors in dimension Income and high support in dimension Expenditure. For the data points in cluster O_1 and O_2, there is enough high support from the *k* nearest neighbors of these data points in given two dimensional space. Lastly, three data points shown under cluster O_4 are farthest from their k nearest neighbor, having low support by their neighbors in both the dimensions. Intuitively, high expenditure when income is high as indicated by the data points in cluster O_4 should not be flagged as outliers. Real outliers which "make sense" are the data points belonging to the cluster O_1. Challenge here is to overcome the mismatch between outliers as entities "which are far away from their neighbors" and "real" outliers.

We propose in this present paper a technique through which real outliers can be captured. Based on the above discussion, we propose to use Bayesian network to represent casual knowledge of the domain. Bayesian network capture causal relationships among a set of variables using a graph in which variables are nodes and causation is indicated by arrows. The strength of relationship among dependent nodes is represented in terms of probability. In our approach, casual relationships encoded in the Bayesian network were exploited using two quantitative rules discussed in Section 4 to identify anomalous patterns. These rules were used to score instances based on the joint probability distribution in the Bayesian network. Later, the instances were sorted by their score and top *n* low probability scored instances were declared as outliers.



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

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

Figure 2: A simple example of Bayesian network in causal relationship.

Figure 2 represents a Bayesian network of above taken example. The two variables namely, income and expenditure are represented by the nodes. The arrow from node Income to node Expenditure indicates that persons income influences his spending. Following data distribution as in Figure 1, tables associated with nodes represents prior probabilities for the parent node, i.e., node Income and conditional probabilities for the child node, i.e., node Expenditure. As indicated in the figure, both nodes can take up two states namely, low and high. For example, 90% of the population belonging to the region has low income and rest 10% has high income. As income affects persons spending as indicated in the Bayesian network, 70% of the people has low expenses when their income is high and 30% people spends as high as they earn. Taking Bayesian structure into account, joint probability of an event

$$Pr(Income = high, Expenditure = high) = \\ P(Expenditure = high|Income = High) \times \\ P(Income = high) = 0.03$$

on similar lines, joint probability of an event

$$Pr(Income = low, Expenditure = high) = \\ P(Expenditure = high|Income = low) \times \\ P(Income = low) = 0.018$$

This simple example illustrates how causality can be exploited using joint probability to uncover anomalous patterns in the data. Joint probability of the event Pr(Income=low,Expenditure=high) is low because the conditional probability P(Expenditure=high | Income=low) is very low(2%) in the Bayesian network. Detail explanation on the Bayesian network and our methodology is presented in Sections 3 and 4 respectively.

Bayesian network have also been used for mining outliers in classification settings. However, in the present paper, we worked in an unsupervised environment. We used Bayesian network as a model for a given domain, to justify our objective that it is meaning and relationship among attributes that needs to be explored to discover outliers. The four worthy primitives of the Bayesian network namely, likelihood, conditioning, relevance and causation [13] were extensively utilized to discover true and meaningful outliers. We define outlier as[8]
"unlikely events under the current favored theory of the domain"

We not only identified outliers using our approach but took the identification aspect to explain why identified data point is an outlier. To best of our knowledge, Knorr and Ng.[9] were the first and perhaps only to suggest the usefulness of explaining why discovered data point is an outlier. Though such explanation is vital for the user, their approach on identifying outliers is based on distance based criteria which we proclaim is not an effective approach in discovering true outliers. In addition to explanation aspect, we also present critical analysis on the search methodology of distance based techniques, Bayesian approach and why a data point discovered as an outlier by a distance based technique is not necessarily an outlier from the Bayesian perspective.

We claim following contributions towards mining true and meaningful outliers.

1. 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 anomalous pattern are present to maximum.

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

3. A 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.

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

The rest of the paper proceeds as follows. Section 2, gives a brief overview of common data mining approaches for outlier detection. Section 3, introduces Bayesian networks. Section 4, presents our methodology, experiments and analysis on distance based technique. We conclude paper with summary and direction of future research in Section 5.

2 Related Work

Common outlier detection techniques can be classified as statistical, distance and density based. The effectiveness of these techniques are illustrated using Figure 3. Statistical methods develop statistical models from the given data and then apply a inference test to determine if an instance is likely to have been generated from the model [7]. Statistical techniques are based on the principle that outliers are observations which are far away from the mean. Statistical model can find data points P_1 and possibly point P_3 as outliers, but can not detect data points P_2 and P_4. In this example, since the value of the data point P_2 corresponds to the mean, therefore, it would not be detected by these methods. Moreover, these methods rely on the assumption that data is generated from a particular distribution which may not hold true especially in a high dimensional space.

Distance based methods [5, 9]use metric measure to rank outliers based on the the distance to their nearest neighbor. Outliers are those points for which distance is large. Under their key assumption, they would label data points P_1 and P_2 as outliers because they cannot fulfil the condition of having a required number of neighbors within certain distance threshold. However, as the data points P_3 and P_4 can fulfil this condition, they would not be declared as outliers. The problem arises due to the fact that these methods take into account the global data distribution rather than local isolation with respect to the neighborhood. This makes them perform poorly when the dataset has regions of varying densities. Moreover they have high computational complexity because they need to compute the distance to all data points in the database for finding the k nearest neighbors.

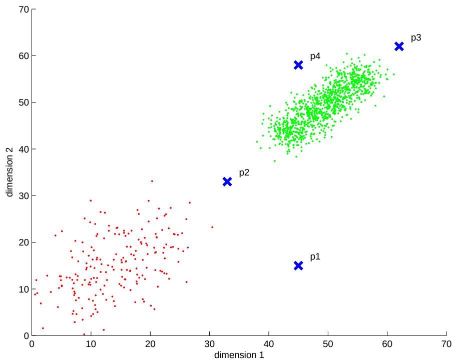


Figure 3: Example showing one dense cluster, one sparse cluster and four outliers.

The density based methods “solve” the problem of finding outliers which are isolated from the entire data objects (global) as well isolated from the local neighborhood (local). Several local outlier detection algorithms have been proposed in the literature including LOF[6] and LOCI[12]. As a consequence such techniques can declare data points P_1, P_2, P_3 and P_4 as outliers. Again, performance of these approaches greatly depends on a distance measure which is very challenging, if the data are complex for e.g. graphs and sequences.

Clustering based approaches make a very simple assumption for finding outliers: normal data points belong to large and dense clusters, while outliers either do not belong to any cluster or form very small clusters[7]. However such techniques are highly dependent upon the effectiveness of clustering algorithms which in turn are dependent on a suitable metric for clustering. This might result in outliers getting assigned to large clusters, therefore, likely to be considered as normal and not outliers.

The use of Bayesian network to find outliers is wide and varied in different applications like video surveillance, intrusion detection in network[15] ,health care[16] and more. This approach is like a classification problem, where a trained Bayesian network on training dataset aggregates information from different variables and provides an estimate on the expectancy of that event to belong to normal/abnormal class for unseen test dataset. The biggest disadvantage of this technique is that they rely on the availability of accurate labels for various classes,which is, most often not possible.

Till date, much focus has been given on discovering point based outliers whereas, studies on finding conditional or contextual outliers are rare. This may be due to the fact that finding contextual outliers needs domain knowledge to understand context of the attributes. Song and Wu [14] were first to propose an approach to discover conditional anomalies. They captured knowledge of the domain in the form of relationships between two sets of attributes. Where one set behaves as a parent and other as a children. By perturbing values of attributes belonging to the child set, i.e., by disturbing the original relationship, they claim to find more anomalous patterns from the set which has been perturbed as compared to the original set where relationships were intact. Nevertheless, this paper highlights importance of discovering outliers based on domain knowledge but key limitation being relationship leant were between two sets of attribute where one set behave as a parent and other as a child. This in one sense, narrows the domain knowledge. The idea of perturbing one set of attributes may not always result in anomalies. It can result in normal data again. Therefore, their approach might fail in this scenario. The focus of the paper in only on highlighting importance of domain knowledge in discovering true outliers. However nothing has been said about how identified anomalies can be useful in updating domain's knowledge.

3 Bayesian Networks

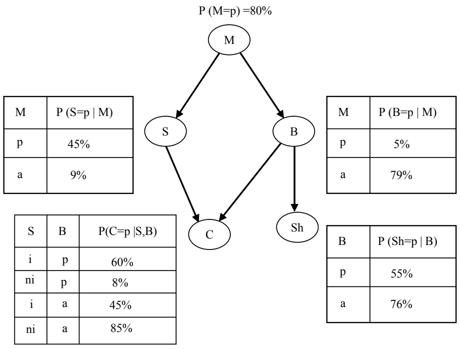
Bayesian network belong to the family of probabilistic graphical models. These graphical models are used to encode knowledge about a domain. In particular, each node in the graph represents a concept(or variable), while links(edges) connect pairs of nodes to represent possible causal relationships. Bayesian network corresponds to graphical models known as directed acyclic graph(DAG), meaning that their edges have direction, and that there is no cycle within the graph. More formally, a Bayesian network over a set of variables X=x_1,x_2...,x_n consists of (1) a network structure ***S*** that encodes plausible relationship among variables in **X** and (2) a set ***P*** of local probability distributions associated with each variable in **X**. An edge from node X_i and X_j in **X** represents a statistical dependence between the corresponding variables. Thus, the arrow indicates that a value taken by variable X_j depends on the value taken by variable X_i,or that X_i "influences" X_j. Node X_i is then referred as parent node and, similarly, X_j is referred to as a child node of X_i. Bayesian network only relates nodes that are probabilistically related by some sort of causal dependency hence the links missing between the variables is because of conditional independence property. In other words, each variable in **X** is independent of its non descendants given the state of the parents[11].

For each variable X_i : $X_i \perp$ nondescendants X_i | Pa^S X_i

Where the symbol **Pa** denotes, parents of variable X_i in network structure ***S*** and symbol $\perp$ denotes conditional independence. A local probability distribution ***P*** associated with each variable in **X** presents prior probability tables for the nodes in the structure ***S*** that have no parents and conditional probability tables(CPTs) for the nodes in ***S*** given their parents. The two components of Bayesian network namely, graphical structure ***S*** and ***P*** of local probability distributions together defines the joint probability distribution for **X**. Given structure ***S*** and parameters ***P***, the joint probability distribution for **X** is given by the product between individual prior probabilities of all parent nodes in **X** and conditional probabilities of all child nodes in **X**. Thus a joint probability distribution in **X** is given by Eq.1.

$$P(X) = \prod_{i=1}^{n} P(x_i|Pa_i) \tag{1}$$

Consider a medical Bayesian network in Figure 4 from [1] on cancer disease as example to illustrate some of the characteristics of Bayesian networks. This Bayesian structure suggests that an event metastatic cancer(denoted by the node M) can cause brain tumor, an event represented by the variable B and serum calcium, an event represented by the variable S. Similarly, severe headache represented by the node Sh is an effect of an event brain tumor. Associated with each nodes are the unconditional and conditional probability tables associated with those. Each node represented in the network can take up two states namely, true(denoted by t) and false(denoted by f) for all variables except variable serum calcium(S)which takes values increased(denoted by i) and not increased(denoted by ni). Only partial probabilities are shown and rest can be inferred from the requirement that the probabilities add up to one. For example, when brain tumor is present and person has increased serum calcium in the body then probability that person will go into coma is 60% and thus the probability that person will not go in coma is 40%. Following the Bayesian network independence assumption, several independence statements can be observed. For example, when metastic cancer is given, variables serum calcium and brain tumor are conditionally independent. Similarly, when brain tumor is given, severe headache is conditionally independent of its ancestor metastatic cancer. The conditional independence characteristics of the Bayesian network provides a compact factorization of the joint probability distribution. Joint probability over five variables represented in Figure 4



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%

Figure 4: Bayesian network representation of the cancer disease. Where M,S,B,C,Sh stands for Metastic Cancer, Serum calcium, Brain tumor, Coma and Severe headache respectively. Each node represented in the network can take up two states namely true(denoted by t) and false(denoted by f) for all variables except variable Serum Calcium which takes values increased(denoted by i) and not increased(denoted by ni).

is given by:

$$P(M,S,B,C,SH) = P(S|M) \times P(B|M) \times P(C|B,S) \times P(SH|B) \times P(M)$$

As all variables are binary, Bayesian network reduces the factors in joint probability from 2^5-1 = 31 to 11 parameters. Such a reduction provides great benefits from inference, learning and computational perspective. Once Bayesian network is built, a process of inferencing is applied which is a task of updating probabilities of outcome based upon relationships in the model and the evidence known about the situation in hand. The updated probabilities reflect the new levels of beliefs in(or probabilities)of all possible outcomes codes in the model. In general, all possible inference queries are evaluated by marginalization, i.e., summing over irrelevant variables. For example, inference query like, what is the probability of metastatic cancer given person is suffering from severe headache?. Such queries in general take exponential time in computation. However, efficient algorithms like message passing[13] can provide approximate answers in polynomial number of steps. The graphical model and associated probabilities can be specified by the domain experts. However, in the absence of domain experts, Bayesian structure and parameters can be learned from the data. Softwares like [1],[3] can be used for the learning task.

4 Methodology and Experiments

4.1 Methodology

With the need of discovering true and meaningful outliers and also motivated by[8], we propose to find outliers by finding joint probability using Bayesian network. We believe outliers are "***low probable, with intrinsic anomalous pattern within***", therefore, by using joint probability distributions and knowledge of the domain, instances can be ranked according to their probability of occurrence . The essential idea is, for a given instance, we find joint probability, which is a product of priors and conditional probabilities across each of the variable in a given domain. The product thus obtained gives us the score for the instance and low scored instances were treated as potential outliers. An important observation here is, product of priors and conditional probabilities, which constitute a score of a given instance, can give rise to four different situations namely,

1. low prior and high conditional probability
2. high prior and low conditional probability
3. low prior and low conditional probability
4. high prior and high conditional probability

In data mining terminology, prior and conditional probability are referred to as support and confidence respectively. A joint probability actually is a product of the above four factors or we can say 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 at the same time it is not also necessary that every situation will be present in the product. This depends upon values taken by attributes and their structure of relationship. Of the four situations, the situations listed at one and two are the only case where there is a conflict between the evidence and event conditional probability provides for a theory and our prior belief about the plausibility of that theory and hence an indication of potential outlying situations. Griffiths and Tenenbaum [8] defines situations one and two above as mere and suspicious coincidence respectively. From outlier mining point of view, low unconditional probability is most likely a "noise event" unless there exists a variable for which there is high conditional probability. Situation three is example of noise. High support and high confidence is example of high correlation and association among attributes. The focus of association rule mining is to discover such patterns from data.

Logically a joint probability of an instance will be low which has maximum number of first three situations listed above. In order to find true outlying situations from the dataset, our focus is on finding those instances where score of joint probability is low because of situations one and two only. Keeping this in mind, prior to finding joint probability of an instance, we checked the strength of the relationship between the two variables. 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. We refer situations one(say R_1) and two(say R_2)as two quantitative rules which can be employed to uncover anomalous patterns in the given data. ***In line with our definition of outliers, quantitative rules helped in uncovering anomalous patterns and joint probability distribution in the Bayesian network ranked low for the instances where such anomalous patterns were present in maximum.*** To apply rules R_1 and R_2 on the dataset, we need to define three parameters namely, low prior and low,high conditional probability. We name these parameters as *minsupp*, *minconf* and *maxconf* respectively. Parameter *minsupp* is computed for every parent node in the Bayesian network. We define *minsupp* of the parent node by Eq.2. Here X stands for any parent node in the dataset and x_i refers to any state of this node.

$$\textit{minsupp(X)} = \min_i(support(X_i)) \tag{2}$$

For example, in Figure 7(a), *minsupp* for the parent node Tuberculosis is 1.104%. Unlike *minsupp* parameter, *minconf* and *maxconf* are user defined thresholds. Following these parameters, we define rules R_1 and R_2 as in by Eq.4 and Eq.5. Where C represents a child node and Pa(C) refers to the parent(s) of the child node C. The factor P(Pa(C)) calculates support of the parent(s) of C in the Bayesian network(BN) defined by Eq.3 and compares it with respective *minsupp* value. Whereas, P(C|Pa(C)) calculates conditional probability of C given parent(s) and compares with *maxconf* and *minconf* thresholds.

$$support(Pa(C)) = P(Pa(C)) \in BN \tag{3}$$

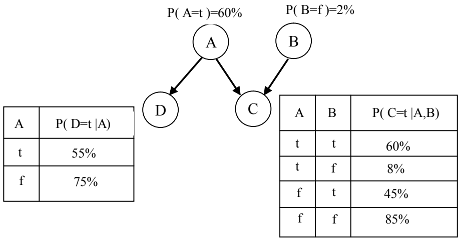
$$R_1 \Rightarrow (P(Pa(C)) = \textit{minsupp}) \bigwedge (P(C|Pa(C) \geq \textit{maxconf})) \tag{4}$$

$$R_2 \Rightarrow (P(Pa(C)) > \textit{minsupp}) \bigwedge (P(C|Pa(C) \leq \textit{minconf})) \tag{5}$$

To help understand our approach, we illustrate how we ranked instances by taking small hypothetical Bayesian network as shown in Figure 5. For this Bayesian network, joint probability distribution over four variables will be represented by Eq.6. Following definition of *minsupp* parameter in Eq.2, *minsupp(A)=0.4(1-0.6)* and *minsupp(B)=0.02*. Let thresholds for parameters *minconf* and *maxconf* are set to 10% and 70%.

$$P(A, B, C, D) = P(C|A, B) \times P(D|A) \times P(A) \times P(B) \tag{6}$$

Let T denotes the testset for the Bayesian network in Figure 5 and let instances follow the attribute order A,B,C,D. Let t$_1$=$\{t, f, f, t\}$ and t$_2$=$\{f, f, t, t\}$ be two testcases such that t$_1$,t$_2$ $\in$ T. With our definition of ranking instances based on two quantitative rules R$_1$ and R$_2$, score of the instance t$_1$=$\{t, f, f, t\}$ will be 0.01104 (P(C|A,B)=0.92 × P(B)= 0.02 × P(A)=0.60) and for the instance t$_2$=$\{f, f, t, t\}$ the score will be 0.005 (P(C|A,B)=0.85 × P(D|A)= 0.75 × P(B)=0.02 × P(A)=0.40). The instance t$_2$ is more anomalous than t$_1$ because it has two patterns satisfying rule R$_1$ whereas, instance t$_1$ has one anomalous pattern which is uncovered by the rule R$_1$. For t$_1$=$\{t, f, f, t\}$, if we calculate joint probability distribution without imposing any rules then it would come out to be 0.006%. The reason why it scored low than the one on which rules where applied is because of the factor P(D|A). The probability P(D|A) in t$_1$=$\{t, f, f, t\}$ is neither an example of R$_1$ nor R$_2$. Thus by sensibly using quantitative rules, instances can be ranked for outlierness. We present pseudo code of the algorithm(OutlierMiner) in 1 below. OutlierMiner takes as input a Bayesian network(BN(N,E)) where N, represents number of nodes and E, set of edges, a testset from which outliers has to be mined and parameters *minconf* and *maxconf*. Algorithm starts by computing *minsupp* for every parent variable(denoted by X)in the Bayesian network. Next, for every testcase in testset, conditional probability in child node(denoted by y) is computed given priors of parents. Rules R$_1$ and R$_2$ are applied on every testcase and joint probability is computed. Finally, top *n* low scored instances are reported as outliers.



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%

Figure 5: A hypothetical Bayesian network of four variables. Each variables takes two states, i.e., true(t) and false(f). Partial probabilities are shown and rest can be calculated by subtracting given probability by one.

The computational complexity of OutlierMiner is governed by factors such as: (1) size of the Bayesian network(i.e. the number of nodes in the net) (2) size of the dataset and (3) probabilistic inference(belief updating)in the Bayesian network. Probabilistic inference in Netica(a commercial Bayesian Network Software) is carried out using join tree algorithm, whose computational complexity is exponential in the worst case.

We experimented on two different sets of Bayesian network: 1) Bayesian structure given, simulated dataset 2) given dataset, learnt Bayesian model and parameters. For the first case, we chose validated Bayesian model from Netica Bayesian net library [1] and simulated dataset using Netica software. For the second category, we learnt Bayesian model from the real datasets taken from UCI repository[2] and B-Course[3]. In order to learn Bayesian model, we divided the dataset into training and testset. Training set was used for the learning task, whereas, testset was used for the experiments. This process is illustrated in step (1), Figure 6. We used B-Course software for learning Bayesian model and parameters. Next, learnt Bayesian structure was used in Netica and an algorithm OutlierMiner was developed using Netica Java API for calculating the probability of a given data point.

Algorithm 1 OutlierMiner

```
Input: Bayesian model(BN(N,E)), paremeters minconf and maxconf, and a testset
Output: top n low probability data points in a testset
1. Compute minsupp for all parents nodes X ∈ BN(N,E)
2. For every testcase in testset, repeat steps(3-4)
3. Compute conditional probability in child node given their parent(s), i.e., P(y)=Pr(y | Pa(y)) where Pa(y)∈ X
4. Apply rules R1 and R2 to uncover anomalous patterns and compute joint probability
5. Sort joint probability
6. Output top n low scored data points
```

Testset was used for two different set of experiments, which infact highlights our contributions. Our first set of experiments were focused on identification of top *n* outliers and describing why these *n* data points are outliers as shown as Task 1 in Figure 6. In another set of experiments, we present analysis on data points discovered as outliers by our approach and nearest neighbor approach. Answers were sought for the following three questions:

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 necessarily an outlier from Bayesian point of view and vice-versa ?

This process is summarized as Task 2,Figure 6

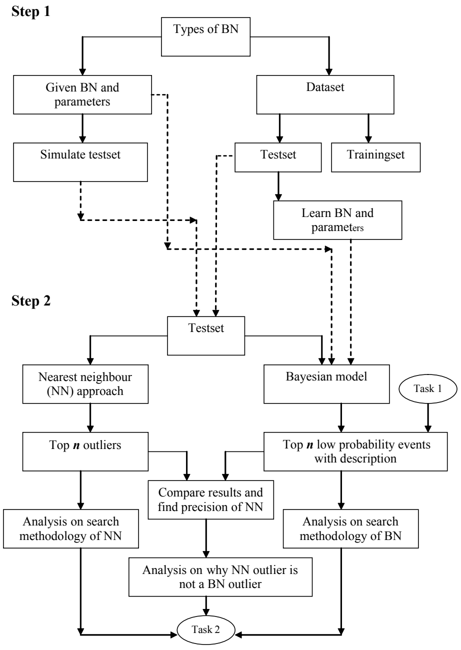

Step 1
Types of BN
Given BN and parameters
Dataset
Simulate testset
Testset
Trainingset
Learn BN and parameters
Step 2
Testset
Nearest neighbour (NN) approach
Bayesian model
Task 1
Top *n* outliers
Top *n* low probability events with description
Compare results and find precision of NN
Analysis on search methodology of NN
Analysis on search methodology of BN
Analysis on why NN outlier is not a BN outlier
Task 2

Figure 6: Methodology

4.2 Assumptions

Our experiments are based on the following key assumptions.

1. As outlier mining is an unsupervised technique, we deleted class labels from the real datasets and tested our approach in an unsupervised way. Though it is difficult to claim relevance of our approach in absence of labels, yet we prove by the quality of our results that our approach is intuitive and meaningful.

2. In order to estimate precision of nearest neighbor technique, we took n, i.e., number of outliers to discover to be 5.

4.3 Datasets

We selected in total ten datasets, out of which five were simulated from validated Bayesian model and rest were real datasets from which Bayesian models were learnt. The outline of each dataset is described below. In addition, we also present the structure of Bayesian models for few datasets. Due to the limitation of space, Bayesian model for few datasets are only shown. For each dataset, the notation (i × j) indicates that the dataset had *i* number of instances and *j* number of attributes.

4.3.1 Given Bayesian network,simulated dataset

Below are the description of five Bayesian model taken from Netica net library through which datasets were simulated.

1. ChestClinic(256 × 8): a simple Bayesian network to diagnose patients arriving at a clinic. All features of this domain were discrete. Figure 7(a) represents the Bayesian network for this domain. Every node encapsulate the attribute name, plausible states and support of every possible state in the dataset. For example, Bronchitis is the attribute name which has two possible states namely, present and absent with 45% and 55% as support, represented by bars next to the states.

2. Busselton(1000 × 15): a Bayesian network to predict risk of Coronary Heart Disease. This domain had a mixture of discrete and continuous features. Detail description of this Bayesian model can be found on Netica website[1].

3. Balpha(1000 × 10): an environmental Bayesian network for the fungus Bondarzewia mesenterica[1]. Features were both discrete and continuous types. Bayesian network is available on Netica website.

4. Diabetes(1000 × 9): a medical Bayesian network for Diabetes. Figure 7(b) represents Bayesian model. All features were discrete as shown in the Figure .

5. System Performance(1000 × 22): a general Bayesian model for troubleshooting. This domain had a mixture of discrete and continuous features. Bayesian network is available on Netica website.

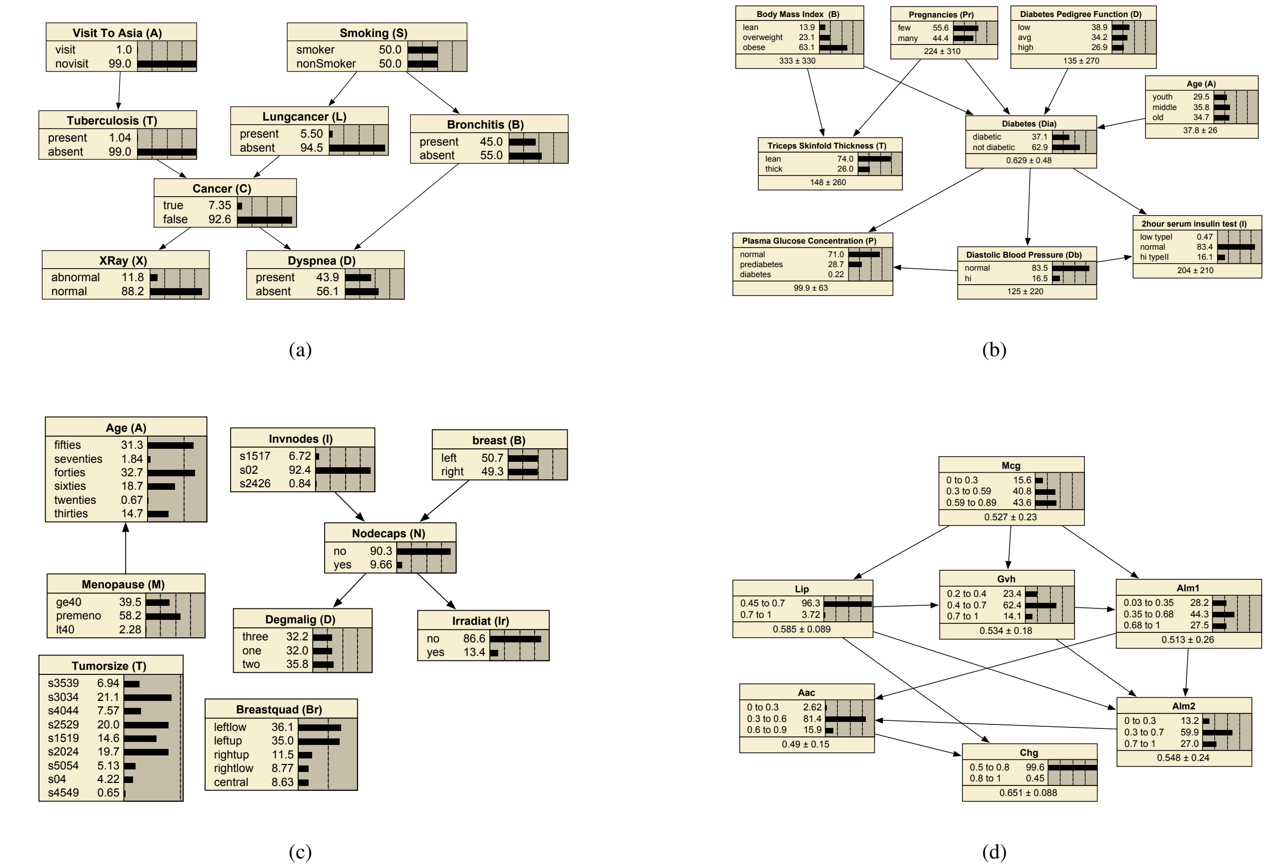
4.3.2 Given dataset, learnt Bayesian structure and parameters

We chose five real datasets, of which four were taken form the UCI archive and other was taken from a web-based data analysis tool for Bayesian modeling called B-Course[3]. During the learning process, attributes which were numerical where automatically discretized by the software. Below are the description of the datasets:

1. Hepatitis(155 × 19): a medical dataset on Hepatitis disease taken from UCI repository. Attributes were mixture of categorical and real data types.

2. Breast cancer(184 × 9): a medical dataset on Breast cancer. For this domain, all attributes were discrete. Bayesian model learnt is represented in Figure 7(c) .This dataset is from UCI repository.

3. Statlog(1000 × 20): a financial dataset which describes important attributes which are accessed before grating credit to the person. Features were mixture of categorical and integer data type. This dataset is from UCI repository. Bayesian model has not been shown due to the limitation of space.



(a)

(b)

(c)

(d)

Figure 7: (a) Bayesian Network of the ChestClinic dataset. Where nodes represents name of the attributes, possible states and support of individual state indicated by bar next to the state name in the Bayesian network. (b) Bayesian Network of the Diabetes dataset.(c) Bayesian network of the Breast cancer dataset.(d) Bayesian network of the Ecoli dataset.

4. Ecoli(336 × 8): a life science dataset taken from UCI repository. All features were real. Bayesian network learnt is shown in Figure 7(d)

5. Boston housing(516 × 14): dataset concerns housing values in suburbs of Boston. All features were numerical. This dataset is from B-Course website[3].

4.4 Experiments and Analysis

As mentioned in the methodology section, we divided experiments in two parts, i.e., "identification and description" and "analysis of false outliers" respectively.

4.5 Identification and description

We applied our methodology on all ten datasets described in above section and explored top *n* outliers. Instances which scored low joint probability in Bayesian network were treated as outliers. In addition to this, we present subspaces which define outliers. We set the thresholds *minconf=10*% and *maxconf*=80%. As our goal is to discover outliers based on causation and correlation, therefore, attributes which were independent like attributes Tumorsize and Breastquad in Breast cancer datatset were not used in the experiments. Due to the limitation of the space, we present top outliers of few datasets only. Top outliers of these datasets are followed by the description on subspace which defines why these data points were outliers. Description follows the annotation X(x) → Y(y). X represents set of parent(s) and Y represents children corresponding to parent(s)in X. Whereas, x stands for support of the parent in the dataset and y, confidence in child node given parent node(s). For example, in the ChestClinic dataset, shown data point is an outlier in three dimensional space of Cancer, Bronchitis and Dyspnea. Support of both the parent nodes are greater than their *minsupp* parameter, i.e., 92.6% and 55% (refer Figure 7(a)) whereas, confidence in child node Dyspnea given support of parent nodes is low(10%). This rule is example of high support and low confidence(rule R_2).

Dataset: ChestClinic
Identification: absent,abnormal,false,absent,present,absent, visit,smoker
Description: Instance is outlier in
1. 2D space of Visit to Asia(1%) → Tuberculosis(95%)
2. 2D space of smoke(50%) → lungcancer(90%)

3. 2D space of cancer(92.6%) → Xray(5%)
4. 3D space of cancer(92.6%), bronchitis(55%)
→ dyspnea(10%)

Dataset: Diabetes
Identification: low,lean,few,youth,lean,low_typeI,normal, normal,not_diabetic
Description: Instance is outlier in
1. 3D space of B(13.86%), Pr(55.62%)
→ T(98.7%)
2. 5D space of B(13.86%), Pr(55.62%),
D(38.9%),Age(29.54%) → Dia(97.7%)
3. 3D space of Dia(62.9%), Db(83.4%)
→ I(0.2%)

Dataset: Ecoli
Identification: 0.52,0.81,0.48,0.5,0.72,0.38,0.38
Description: Instance is outlier in
1. 3D space of Mch(40.8%), Lip(96.3%) → Gvh(0.9%)
2. 3D space of Mch(40.8%), Gvh(14.1%)
→ Alm1(100%)
3. 3D space of Lip(96.3%), Gvh(14.1%)
,Alm1(44.3%)→ Alm2(7.27%)

Dataset: Breast cancer
Identification: twenties,premeno,s3539,s02,no,two,right, rightup,no
Description: Instance is outlier in
1. 2D space of Menopause(58.2%) → Age(1%)
2. 3D space of Invnodes(92.4%), Breast(49.3%)
→ Nodecaps(95%)

4.6 Relevance of our approach

Our emphasis in this section is on the usefulness and relevance of our approach in discovering genuinely anomalous patterns. Any outlier detection technique is novel if it can validate anomalous behavior of the observations and can provide insights into the fact as to why these observations are suspicious. Such insights not only give understanding on data but helps in improving knowledge of the domain. The most authentic way to validate outliers discovered by any outlier detection technique is by evaluating observations using domain knowledge. However, as expertise of the particular domain is not always readily available to disseminate knowledge about the domain and validate outliers; a model representing domain knowledge could be a promising solution in this direction.

We present relevance and quality of our results by discussing an outlier instance discovered by our approach. The idea is, if an explanation of a data point to be anomalous is justified by the domain as an unseen yet interesting knowledge then that observation is a true outlier and therefore an indication of relevance of our approach. We chose dataset ChestClinic for explanation. The reason of choosing this dataset is because the relationship among attributes and the general knowledge of the domain is very easy to understand and hence explaining an outlier instance of this dataset will be effective. Following is the top outlier of ChestClinic dataset with description.

Dataset: ChestClinic
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%),
broncitis[absent](55%) → dyspnea[present](10%)

We amended rule X(x) → Y(y) with additional information which is represented in angular braces. Information in angular braces represents state of the variable. Referring to Figure 7(a), we explain four outlier subspaces identified as follows:

1. Percentage of people who makes visit to Asia(1%) is unlikely to have tuberculosis(95%). This is a suspicious event because we do not have enough evidence(1%, which is very small) to this fact.

2. Referring to second subspace, there is one cause of lungcancer, i.e., smoking. A lay mans opinion says, a person who smokes is mostly likely to get affected by lungcancer. For the given instance, value of the variable smoker is “smoker” and value of lungcancer is “absent”. Which obviously indicates a new dimension to knowledge that there could be other causes leading to lungcancer. The support of smoke is 50%, which is considered as *minsupp* because smoke has only states with same probabilities.

3. In the third subspace, intuitively, a person suffering for cancer should have abnormal xray. Whereas, for this observation, cancer is absent but still xray report is abnormal. It raises question as to why xray is abnormal when cancer is absent. This lead us to a new knowledge that abnormal xray is not only affected by the presence of cancer but there could exist other factors causing abnormal xray.

4. Similarly, for the fourth subspace, two causes of disease dyspnea namely, cancer and bronchitis are absent but still disease dyspnea is present.

4.7 Analysis on genuine and false outliers

In this section, we 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. To start with, 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. We calculated precision of nearest neighbor approach on all ten datasets mentioned above.

Table 1: Precision obtained by Nearest Neighbor technique. Parameter n, number of outliers to discover was set to five.

Dataset name	Precision
ChestClinic	20%
Busselton	80%
Balpha	0%
Diabetes learned	60%
System Performance	0%
Hepatitis	40%
Breast cancer	20%
Statlog	80%
Ecoli	20%
Boston housing	20%

Table 1 summarizes results we achieved. The first column of the tables represents the name of the dataset and second column specifies precision as obtained by the distance based technique. We used algorithm presented in[10] for discovering outliers using distance based criteria. Hamming and Euclidean distance measures were implemented in the algorithm for categorical and numerical data types respectively. As observed from the results, accuracy of distance based technique lies in the range of 0%-80%. For most of the datasets, precision is not more than 40%. Below we present analysis on, search methodology of Bayesian approach, search methodology of distance based technique and finally why distance based outlier is not a genuine outlier from domain perspective.

4.7.1 What Bayesian approach follows ?

Bayesian network tightly integrates relationships among features of the domain and plausibility of an event in probabilistic terms. By exploiting relationships, low and high likely events can be interpreted. Probability that a series of events will happen concurrently can be answered by calculating joint probability. Bayesian networks provides efficient graphical representation of joint probability; by taking advantage of conditional independence, dimensionality of the dataset can be factored into smaller groups indicating dependent attributes and extent of correlation in probability. In other words, joint probability of several variables can be calculated from the product of individual probabilities of the nodes following chain rule of probability. For example, joint probability of all the attributes in the Bayesian model in Figure 7(a) is represented by Equation below.

$$\begin{aligned} P(A,S,T,L,B,C,X,D) &= P(X|C) \times P(D|C,B) \\ &\times (C|T,L) \times P(T|A) \times P(L|S) \\ &\times P(B|S) \times P(S) \end{aligned}$$

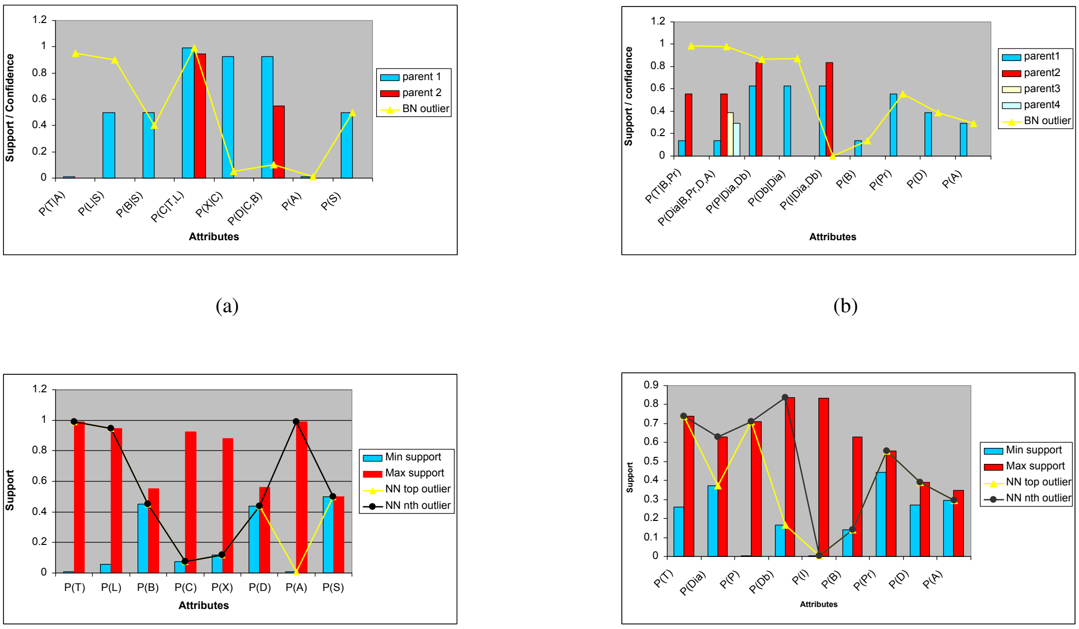
Where A,S,T,L,B,C,X,D represents initials of eight variables in the Bayesian network. Following definition of joint probability in Bayesian network, we explored inner structure on similar lines for top n outliers discovered by our approach for every dataset. It is important to mention here, top n observations where scored low in the Bayesian network because they were having maximum patterns of two qualitative rules we pointed above. However, by structuring these anomalous instances in individual probabilistic terms we can observe these anomalous patterns. Such representation not only indicates search methodology of our technique but also gives understanding on data in general.

Graphs in Figure 8(a) and Figure 8(b) represents pattern of top outlier in terms of conditional probabilities(confidence) and prior(support) which together constitutes joint probability in the Bayesian network. The graphs in Figure 8(a) and 8(b) are for datasets ChestClinic and Diabetes respectively. We chose datasets with minimum number of attributes so that analysis through graph can easily be explained. The attribute names are represented by taking initials as represented in the respective Bayesian network. The X-axis of the graph represents attributes and Y-axis represents support of parent node in bars and confidence in child node through trend line. Here, support of the parent node is defined using Eq.3. For graph in Figure 8(a), first six attributes are child node whereas, rest two are independent nodes. Referring to Figure 7(a) and first bar in graph of the Figure 8(a) indicates, support in parent node Visit to Asia is nearly zero, but confidence in direct child(Tuberculosis) of this parent node (Visit to Asia) is above 95% which is quite high as represented by the point on the trend line just above the bar.

More than one bar at the same position represents number of parents linked with that child. Like, child node Cancer has two parents namely Tuberculosis and Lungcancer and hence shown by two different support bars in fourth term. Trend shown in Figure 8(a) specifies subspaces which define outlier. Terms first, second, fifth and sixth were uncovered by the qualitative rules. Not only outlying subspaces are visible but normal subspaces can also be interpreted by the observing the graph. Subspace Tuberculosis, Lungcancer and Cancer is example of high support and high confidence and hence is normal. Similar explanation can be followed for the graph in Figure 8(b).

4.7.2 What Nearest neighbor technique follows ?

The major difference between Nearest neighbor and Bayesian approach can be summarized as follows: distance based technique treats every attribute of the domain uniformly whereas, for the Bayesian approach, treatment with attributes depends upon relationship among attributes. Any distance based approach will find a pair wise distance between two objects and will declare an object to an outlier which is far away from k nearest neighbors. Intuitively, it implies that an object declared as an outlier does not have enough support by the nearest neighbors, so, is isolated and far away from the dense area. Contrary to this, a dense cluster is formed by those data points which has similar support from the nearest neighbors which is the reason they satisfy condition of k nearest neighbor and hence are normal. Thus distance based approaches look for those



(a)

(b)

(c)

(d)

Figure 8: (a) Pattern of top BN outlier in the ChestClinic dataset. The bars represent support of the parent attribute(s) and conditional probability in a child node is represented by the trend line. Terms first, second, fifth and sixth were uncovered by the qualitative rules.(b) Pattern of top BN outlier in the Diabetes dataset. Terms first, second and fifth were uncovered by the qualitative rules.(c) Pattern of top and n^{th} NN outlier in the ChestClinic dataset. The bars represents minimum and maximum support of the attribute in the Bayesian network and two trend lines represents NN's top and n^{th} outliers respectively. Top outlier(indicated by yellow line) has six attributes with low support whereas, n^{th} outlier(indicated by black line) has five attributes with low support.(d) Pattern of top and n^{th} NN outlier in the Diabetes dataset. Top outlier(indicated by yellow line) has five attributes with low support whereas, n^{th} outlier(indicated by black line) has three attributes with low support.

data points where maximum number of attributes have low support. On the other hand, Bayesian approach considers both conditional probability(confidence) and unconditional probability(support) in order to discern between abnormality and normality.

Analysis on two datasets namely ChestClinic and Diabetes are shown using graphs in Figure 8(c) and Figure 8(d) respectively. Due to the limitation of space we could not present analysis on every dataset. The X-axis represents attributes of the domain and Y-axis represents support of the attributes. Two bars on every attribute of X-axis represents minimum and maximum support attribute has in the Bayesian network. Minimum support of the attribute follows Eq.2 and maximum support of an attribute in the Bayesian network is represented by Eq.7. Where X stands for any parent node in the Bayesian network and x_i refers to any state of this node.

$$maxsupp(X) = max_i(support(X_i)) \qquad (7)$$

In addition, two trend line reveal the pattern of top and n^{th} outlier discovered by distance based technique. Interestingly, top outlier has six attributes with low support(indicated by yellow trend line) whereas, for n^{th} outlier, five attributes have low support(indicated by black trend line) for the ChestClinic dataset as represented by the graph in Figure 8(c). Similar pattern is observed in Figure 8(d). For few datasets we found, distance based outliers chose those data points as outliers where support of few attribute is near to minimum support if not minimum support exactly.

4.7.3 Why distance based outlier is not an outlier in Bayesian network ?

To answer this question, we simply took an outlier discovered by distance based technique and analyzed pattern of this outlier from Bayesian perspective. As discussed above, for Bayesian approach, any data point will be an outlier which satisfies two quantitative rules. If any data point scores high probability in Bayesian network it is simply because that data point does not have any anomalous pattern. Graph in Figure 9 represents a data point which is abnormal by the definition of distance based technique but normal from domain's standpoint. We took example outlier from Breast cancer dataset where precision of distance based technique is very low. Graph in Figure 9, represents patterns which either are example of high support and high confidence or low support and low confidence and hence normal from the Bayesian viewpoint.

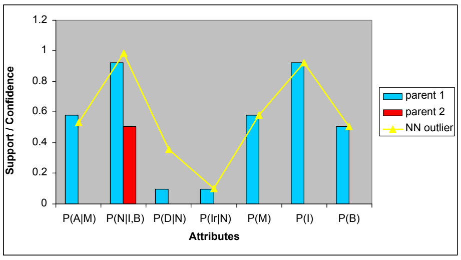


Figure 9: Pattern of NN n^{th} outlier in the Breast cancer dataset which is not an outlier in domains perspective. The bars represent support of the parent attributes and conditional probability in the child attribute is represented by the trend line. All patterns are either example of high support and high confidence or low support and low confidence.

5 Conclusion and future scope of work

In this paper 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 current favored theory of the domain. By structuring domain knowledge in Bayesian framework, anomalous patterns were uncovered using two quantitative rules. Instances were ranked based on the score of the joint probability distribution in the Bayesian network.

We presented the explanation on subspaces which defines outlier. 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 works in support and confidence based mining framework. Netica[1], a powerful application for Bayesian networks was integrated with our algorithm through its Java API(NeticaJ).

As for future work, we intend to work on high dimensional datasets. We also plan to apply our techniques to a specific domain and work with specialists in the domain to help uncover potentially useful anomalies.

References

[1] Bayesian network development software. `http://www.norsys.com/`.

[2] Uci machine learning repository. `http://archive.ics.uci.edu/ml/`.

[3] A web-based data analysis tool for bayesian modeling,. `http://b-course.cs.helsinki.fi/obc/`.

[4] S. Babbar. Integration of domain knowledge for outlier detection in high dimensional space. In *Database Systems for Advanced Applications: DASFAA 2009 International Workshops: BenchmarX, MCIS, WDPP, PPDA, MBC, PhD, Brisbane, Australia, April 20 - 23, 2009*, pages 363–368, Berlin, Heidelberg, 2009. Springer-Verlag.

[5] S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 29–38, New York, NY, USA, 2003. ACM.

[6] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, 2000.

[7] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):1–58, 2009.

[8] T. Griffiths and J. Tenenbaum. From mere coincidences to meaningful discoveries. *Cognition*, 103(2):180–226, 2007.

[9] E. M. Knorr and R. T. Ng. Finding intensional knowledge of distance-based outliers. In *VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases*, pages 211–222, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[10] E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: algorithms and applications. *The VLDB Journal*, 8(3-4):237–253, 2000.

[11] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press., Cambridge, USA, 2009.

[12] S. Papadimitriou, H. Kitagawa, P. Gibbons, and Faloutsos. Loci: Fast outlier detection using the local correlation integral. In *Proceedings. 19th International Conference on In 19th International Conference on Data Engineering*, pages 315–326, 2003.

[13] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[14] X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection. *IEEE Trans. on Knowl. and Data Eng.*, 19(5):631–645, 2007.

[15] J. L. Thames, R. Abler, and A. Saad. Hybrid intelligent systems for network security. In *ACM-SE 44: Proceedings of the 44th annual Southeast regional conference*, pages 286–289, New York, NY, USA, 2006. ACM.

[16] W. Wong, A. Moore, G. Cooper, and M. Wagner. Bayesian network anomaly pattern detection for disease outbreaks. In T. Fawcett and N. Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning*, pages 808–815, Menlo Park, California, August 2003. AAAI Press.