Summer 8-16-2018

A Semi-Supervised Feature Engineering Method for Effective Outlier Detection in Mixed Attribute Data Sets

Girish Srivatsa Rentala

A SEMI-SUPERVISED FEATURE ENGINEERING METHOD FOR EFFECTIVE OUTLIER DETECTION IN MIXED ATTRIBUTE DATA SETS

by

Girish Srivatsa Rentala, Bachelor of Technology

A Thesis Presented in Partial Fulfillment
of the Requirements of the Degree
Master of Science

August 2018

COLLEGE OF ENGINEERING AND SCIENCE
LOUISIANA TECH UNIVERSITY

LOUISIANA TECH UNIVERSITY

THE GRADUATE SCHOOL

JUNE 22, 2018
Date

We hereby recommend that the thesis prepared under our supervision by

Girish Srivatsa Rentala, Bachelor of Technology

entitled **A Semi-Supervised Feature Engineering Method for Effective Outlier Detection in Mixed Attribute Data Sets**

be accepted in partial fulfillment of the requirements for the Degree of

Master of Science in Computer Science

Supervisor of Thesis Research

Head of Department
Computer Science
Department

Recommendation concurred in:

Advisory Committee

Approved:

Director of Graduate Studies

Dean of the College

Approved:

Dean of the Graduate School

ABSTRACT

Outlier detection is one of the crucial tasks in data mining which can lead to the finding of valuable and meaningful information within the data. An outlier is a data point that is notably dissimilar from other data points in the data set. As such, the methods for outlier detection play an important role in identifying and removing the outliers, thereby increasing the performance and accuracy of the prediction systems. Outlier detection is used in many areas like financial fraud detection, disease prediction, and network intrusion detection.

Traditional outlier detection methods are founded on the use of different distance measures to estimate the similarity between the points and are confined to data sets that are purely continuous or categorical. These methods, though effective, lack in elucidating the relationship between outliers and known clusters/classes in the data set. We refer to this relationship as the context for any reported outlier. Alternate outlier detection methods establish the context of a reported outlier using underlying contextual beliefs of the data. Contextual beliefs are the established relationships between the attributes of the data set. Various studies have been recently conducted where they explore the contextual beliefs to determine outlier behavior. However, these methods do not scale in the situations where the data points and their respective contexts are sparse. Thus, the outliers reported by these methods tend to lose meaning. Another limitation of these methods is that they assume all features are equally important and do not consider nor determine subspaces among the features for identifying the outliers. Furthermore, determining subspaces is computationally exacerbated, as the number of possible subspaces increases with increasing dimensionality. This makes searching through all the possible subspaces impractical.

In this thesis, we propose a Hybrid Bayesian Network approach to capture the underlying contextual beliefs to detect meaningful outliers in mixed attribute data sets. Hybrid Bayesian Networks utilize their probability distributions to encode the information of the data and outliers are those points which violate this information. To deal with the sparse contexts, we use an angle-based similarity method which is then combined with the joint probability distributions of the Hybrid Bayesian Network in a robust manner. With regards to the subspace selection, we employ a feature engineering method that consists of two-stage feature selection using Maximal Information Coefficient and Markov blankets of Hybrid Bayesian Networks to select highly correlated feature subspaces.

This proposed method was tested on a real world medical record data set. The results indicate that the algorithm was able to identify meaningful outliers successfully. Moreover, we compare the performance of our algorithm with the existing baseline outlier detection algorithms. We also present a detailed analysis of the reported outliers using our method and demonstrate its efficiency when handling data points with sparse contexts.

APPROVAL FOR SCHOLARLY DISSEMINATION

The author grants to the Prescott Memorial Library of Louisiana Tech University the right to reproduce, by appropriate methods, upon request, any or all portions of this Thesis. It is understood that "proper request" consists of the agreement, on the part of the requesting party, that said reproduction is for his personal use and that subsequent reproduction will not occur without written approval of the author of this Thesis. Further, any portions of the Thesis used in books, papers, and other works must be appropriately referenced to this Thesis.

Finally, the author of this Thesis reserves the right to publish freely, in the literature, at any time, any or all portions of this Thesis.

Author ______________________________

Date ______________________________

DEDICATION

This work is dedicated to my parents and family who have always been a constant source of support and encouragement throughout my life.

TABLE OF CONTENTS

ABSTRACT ... iii

APPROVAL FOR SCHOLARLY DISSEMINATION ... v

DEDICATION ... vi

LIST OF FIGURES ... x

LIST OF TABLES ... x

ACKNOWLEDGMENTS ... xiii

CHAPTER 1 INTRODUCTION ... 1

1.1 Data Mining ... 1

1.2 Outliers ... 3

1.3 Problem Definition ... 4

1.4 Contributions ... 8

1.5 Organization ... 9

CHAPTER 2 BACKGROUND ... 11

2.1 Density Based Methods ... 12

2.1.1 Benefits and drawbacks of Density Based Methods ... 13

2.2 Distance Based Methods ... 14

2.2.1 Benefits and drawbacks of Distance Based Methods ... 15

2.3 Contextual Based Methods ... 15

2.3.1 Benefits and drawbacks of Contextual Based Methods ... 17

2.4 Bayesian Network Based Methods ... 17

2.4.1 Benefits and drawbacks of Bayesian Network Based Methods ... 18

2.5 Subspace Based Methods 19
2.5.1 Benefits and drawbacks of Subspace Based Methods 20
2.6 Summary 21
CHAPTER 3 BAYESIAN NETWORK MODELS 22
3.1 Bayesian Networks 23
3.2 Bayesian Network Independencies 26
3.3 Joint Probabilities in Bayesian Network 27
3.4 Markov Blanket of Bayesian Networks 28
3.5 Bayesian Network Structural Learning 30
3.6 Hybrid Bayesian Networks 32
CHAPTER 4 METHODOLOGY 34
4.1 Specific Aims 34
4.2 Feature Selection using Maximal Information Coefficient 35
4.3 Subspace Discovery using Markov Blankets 36
4.4 Angle Based Similarity Measure 39
4.5 Mining Outliers in Markov Blanket Subspaces of Hybrid Bayesian Networks 39
4.5.1 Hybrid Bayesian Networks Learning 41
4.5.1.1 Structural Learning 42
4.5.1.2 Parameter Learning 43
4.5.1.3 Learning Local Hybrid Bayesian Networks in MB Subspaces 43
4.5.2 Mining Outliers 44
4.5.2.1 Algorithm 47
4.6 Complexity Analysis 48
4.7 Data Sets 49
4.7.1 KSL Danish Elderly Data Set 49

CHAPTER 5 RESULTS & DISCUSSION 50

5.1 Evaluation Metrics Used 50

5.1.1 Precision 50

5.1.2 Recall 51

5.1.3 F-Measure 51

5.1.4 ROC Curves 51

5.2 Experimental Results 52

5.2.1 Feature Selection using Maximal Information Coefficient 52

5.2.2 Hybrid Bayesian Network on Full Attribute Space 53

5.2.3 Hybrid Bayesian Networks in Markov Blanket Subspaces 54

5.2.4 Analysis of the Results 58

5.2.5 Evaluation of the Proposed Model 62

5.2.6 Visualization of the Reported Outliers 64

CHAPTER 6 CONCLUSIONS AND FUTURE WORK 67

6.1 Conclusions 67

6.2 Future Work 67

APPENDIX A 68

A.1 KSL data set description 68

A.2 Summary of notations 69

Bibliography 70

LIST OF TABLES

Table 4-1: Description of dimensions of the data sets used. 49

Table 5-1: Outliers discovered in the KSL dataset. 59

Table 5-2: Outlier analysis on the KSL dataset using proposed method. 60

Table 5-3: Outlier analysis on the KSL dataset using LOF. 61

Table 5-4: Outlier analysis on the KSL dataset using SOD. 61

Table 5-5: Summary of the results obtained for the KSL dataset. 64

Table A-1: Description of attribtes in the KSL data set. 68

Table A-2: Notations. 69

LIST OF FIGURES

Figure 1-1: Taxonomy of Data Mining models. 2

Figure 1-2: Scatterplots showing outliers in 2-D. 3

Figure 1-3: A hypothetical example of health characteristics of a person. 6

Figure 3-1: Illustration of a simple Bayesian network. 23

Figure 3-2: Bayesian network representation of the cancer disease. 25

Figure 3-3: Relational dependency between attributes metastatic cancer M and brain tumor B. 26

Figure 3-4: Conditional independency between attributes metastatic cancer and coma given the brain tumor. 27

Figure 3-5: Example of a markov blanket for node Xi with Yi as its parent nodes, child nodes and spouses. 29

Figure 3-6: Example of Hybrid Bayesian Network. 33

Figure 4-1: A Bayesian network for diabetes. The darker nodes indicate the Markov blanket of attribute Diastolic blood pressure. 38

Figure 4-2: Outline of the proposed methodology for outlier detection. 41

Figure 4-3: Bayesian networks in the Markov blanket subspaces. 44

Figure 5-1: Maximal Information Coefficient of KSL data set 53

Figure 5-2: Hybrid Bayesian Network on full attribute space of KSL data set 54

Figure 5-3: Hybrid Bayesian Network on Markov blanket subspace of FEV. 54

Figure 5-4: Hybrid Bayesian Network on Markov blanket subspace of Kol. 55

Figure 5-5: Hybrid Bayesian Network on Markov blanket subspace of BMI. 55

Figure 5-6: Hybrid Bayesian Network on Markov blanket subspace of Smok. 56

Figure 5-7: Hybrid Bayesian Network on Markov blanket subspace of Alc. 56

Figure 5-8: Hybrid Bayesian Network on Markov blanket subspace of Work. 57

Figure 5-9: Hybrid Bayesian Network on Markov blanket subspace of Sex. 57

Figure 5-10: Hybrid Bayesian Network on Markov blanket subspace of Hyp. 58

Figure 5-11: ROC curve of proposed algorithm against LOF. 63

Figure 5-12: ROC curve of proposed algorithm against SOD. 63

Figure 5-13: 2D visualization of causal subspace of FEV and Kol in the KSL data set. 65

Figure 5-14: 2D visualization of causal subspace of Kol and BMI in the KSL data set. 66

ACKNOWLEDGMENTS

In the name of the almighty the most beneficent and merciful, first I want to praise the Lord for helping me in all the stages of this dissertation work.

Foremost, I would like to express my sincere gratitude to my advisor Dr. Pradeep Chowriappa for the continuous support of my master’s study and research, for his patience, motivation, enthusiasm, and immense knowledge.

To the committee members Dr. Kevin Cherry and Dr. Ankunda Kiremire, I express my heartfelt gratitude for accepting to serve on my advisory committee and for being generous with their time.

Finally, I am grateful to my parents and friends for their support and encouragement directly or indirectly in completing this thesis.

CHAPTER 1

INTRODUCTION

1.1 Data Mining

Data mining is defined as the “extraction of non-trivial, implicit, previously unknown and potentially useful information from the data” [1]. It depends upon the different fields like machine learning, pattern recognition, artificial intelligence, statistics, database systems. Data mining is used as a tool by businesses to make improved decisions for solving a problem by providing important information.

The idea of a testable hypothesis drives data mining techniques. They can extract implicit patterns of the data. Several techniques both supervised and unsupervised have been used so far for analyzing data to better understand the underlying patterns which will help make informed decisions.

Largely, data mining models are divided into two types; i.e. predictive models and descriptive models as shown in **Figure 1-1**. Predictive models deal with the prediction or forecast of the explicit value of a particular attribute and are classified into two types, namely classification models and regression models. Classification models predict according to the class labels. However, a regression model analyzes the dependencies among the attributes and class labels. Descriptive models analyze the hidden patterns in the data and categorizes them into relevant subgroups. These are classified into two types, namely clustering models and association models. Cluster models group together similar things, events, or people to reduce data complexity. However, association models determine the frequency of relevant associations and provide interesting insights [1].

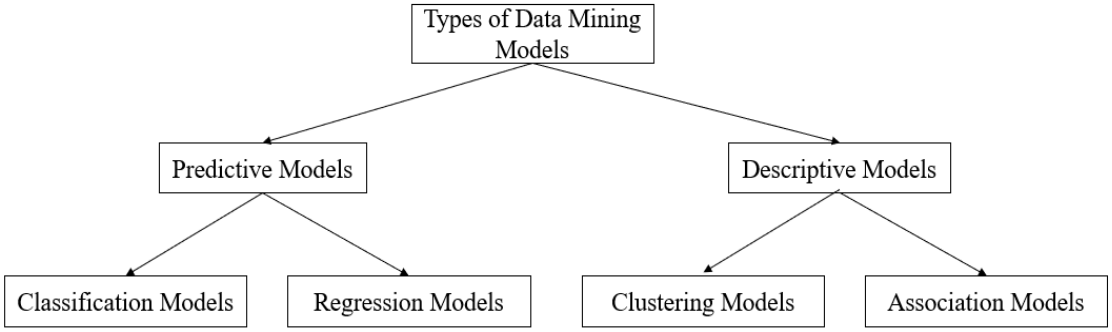


Figure 1-1: Taxonomy of Data Mining models.

Applications of data mining include marketing analysis, predicting subsets of customers likely to respond to a given promotion based on income levels or the amount of previous purchases. Banks and insurance companies use data mining to analyze claim patterns and to predict credit fraud. Other applications include stock market analysis, modelling proteins, and genes in DNA sequences. Moreover, the fact that data mining has become so successful is due to faster and cheaper computer hardware which have led to the development of specialized algorithms to analyze large volumes of data efficiently.

Although data mining is mostly used for discovering relationships or hidden patterns in the data, an often overlooked but important task is the ability to detect outliers or anomalies in the data. Certainly, the patterns may be well established for some applications such as health insurance and credit card fraud, but it is often the exceptions to those patterns that require special attention. Outliers or anomalies may be the result of recording or measurement errors, but they may also be genuine data which may point out surprising, suspicious, and fraudulent activities.

1.2 Outliers

An outlier or anomaly is an observation in the data that is significantly different from other observations in the data. These are also stated to as exceptions, irregularities, deviations, or aberrations. Hawkins [2] defined an outlier as “an instance that is remarkably different from other instances in the sample”. The two-dimensional scatterplot in **Figure 1-2** shows an example of an outlier. In the left side of **Figure 1-2**, the lower left observation is an outlier that is far away from the dense cluster of points. In the right side of **Figure 1-2**, the observation which is noticeably separated from the dense cluster of points is an outlier.

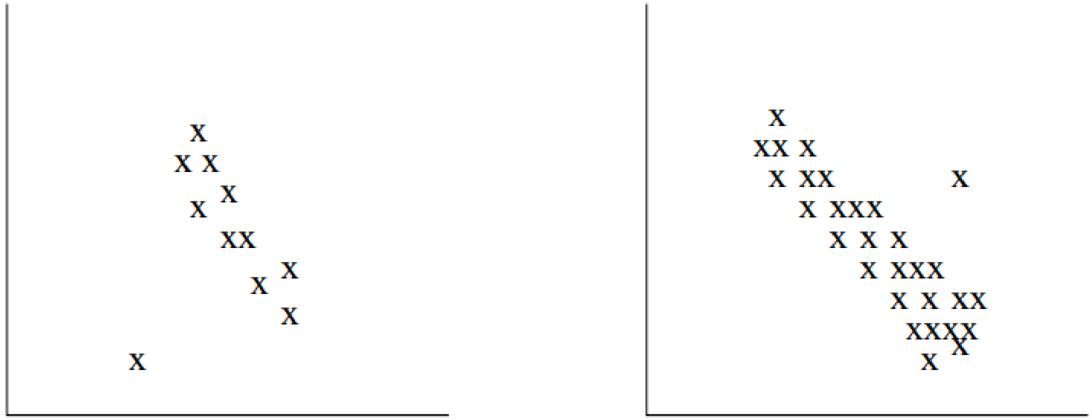


Figure 1-2: Scatterplots showing outliers in 2-D.

The outlier phenomenon is often mysterious to data analysts. Occasionally, outliers may appear in the data sets of poor quality where no relative evidence is displayed by the outlier. Outliers of this kind can be removed from the dataset. However, in some cases, outliers display interesting and truthful information. These types of outliers should not necessarily be removed from the underlying dataset as they may probably provide new evidence. Thus, to improve the quality and mine new evidence from the data, outlier identification is important. Examples include credit card activity monitoring to identify fraudulent transactions and predict misuse [3]. Likewise, sensor monitoring of instruments and devices in industries help to identify system defects [4]. Furthermore, outlier detection methods help to detect new disease outbreaks in public health monitoring systems [5] and are also useful in detecting network intrusions [6].

The task of anomaly or outlier detection is very challenging due to several factors. First is declaring an observation as an outlier when it does not fall in a region representing normal behavior. However, defining a normal behavior for each and every data point is a challenging task. The second factor is the limited availability of labeled data for classifying outliers. Thus, unsupervised techniques are best suited in these cases, where only the normal behavior is modeled to determine the outliers. The third factor is that we cannot apply outlier detection algorithm designed for one application to another application due to the specific nature of the outliers. Finally, the quality of the outliers identified by outlier detection techniques is difficult to examine.

1.3 Problem Definition

Previous research on outlier detection has mainly focused on developing algorithms to categorize outliers in the data sets using either distance/similarity-based or density-based approaches. These techniques will calculate the pair-wise distance among all the points in the data set and will declare a point far away from its nearest neighbors as an outlier. Naturally, these approaches suggest that the point, which is isolated, will not have enough support from its neighborhood to classify it as normal. Also, the existing algorithms are only applicable to the data sets having a specific attribute type which is either categorical or continuous. The algorithms designed for continuous data sets cannot be directly applied to categorical data sets and vice-versa. Moreover, most of the current approaches do not consider the quality of reported outliers; i.e. they ignore valuable information available in the data and fail to tell us why a particular point has been labelled exceptional.

For example, assume a data set belongs to a particular region of a country representing health characteristics of people such as cholesterol levels and blood pressure as shown in **Figure 1-3**. The cholesterol levels are plotted on the X-axis, while the blood pressure is plotted on the Y-axis. The data points in the data set are clustered into four clusters denoted by C_1, C_2, C_3, and C_4 respectively. The cluster C_1, which is dense, represents a region in which people with high cholesterol have high blood pressure and vice versa. Unlike C_1, cluster C_2 contains a small percentage of people with high cholesterol levels and high blood pressure. Finally, cluster $C_3$and C_4 specify a situation where blood pressure is more than cholesterol. If the goal is to detect outliers from this data, using distance [7] and density [8] based approaches, then most probably the clusters C_2, C_3 and C_4 will be flagged as outliers, because they are far-off from their neighbors. However, if we investigate the data points in cluster C_2, we find that it contains information about the people who have high cholesterol with high blood pressure, and as such, these points should not be treated as outliers. Furthermore, the data points in clusters C_3 and C_4 appear interesting as they represent a situation where the blood pressure of a person is more than the cholesterol level. Therefore, identifying these types of outliers which represent valuable information help the stakeholders to improve the understanding of the data and make the appropriate decisions.

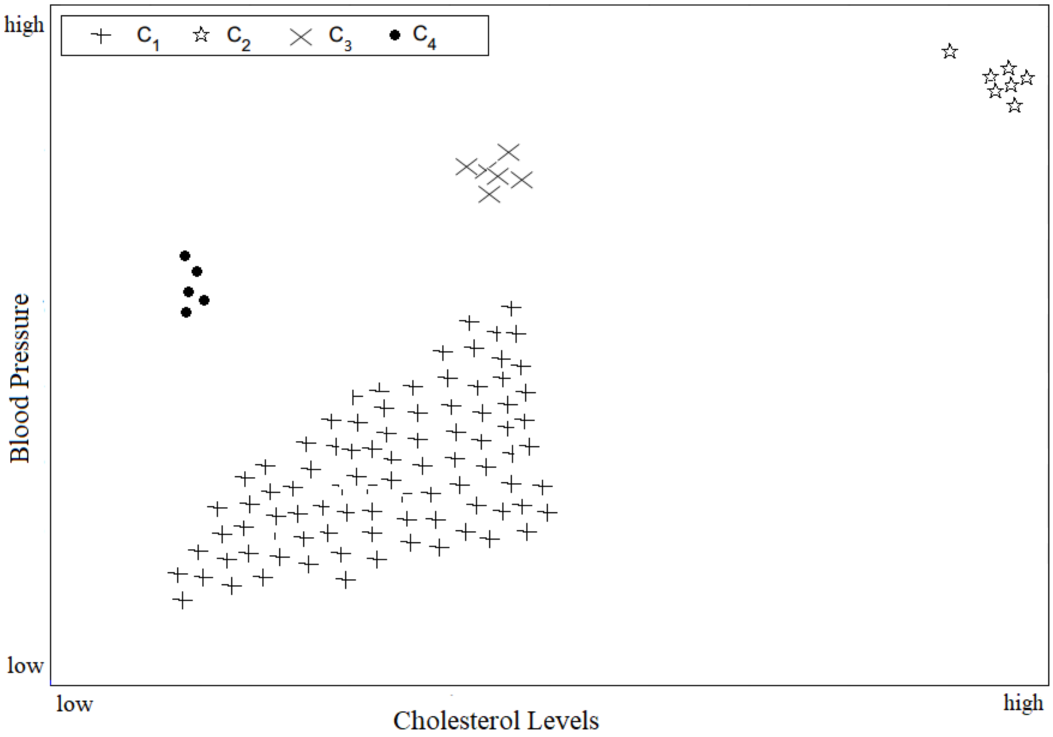


Figure 1-3: A hypothetical example of health characteristics of a person.

The main task in identifying the above-mentioned points in clusters C_3 and C_4 is that they require contextual beliefs of the data. Basically, the meaning of the context decides the interesting aspects of the points. For example, a data point may behave differently in one context, but it appears normal in others. However, in high dimensional space, the data points become sparse; i.e. all the points appear the same and can be regarded as outliers. Consequently, the contexts associated with the points in high dimensions will also be the same leading to a difficulty in defining proper contexts. To explain this, consider the same example shown in **Figure 1-3**. Using state-of-the-art contextual outlier detection algorithms [9, 10], there might be a chance that cluster C_2 will be flagged as an outlier instead of C_3 and C_4 as the cluster C_2 is sparse and lack support from the reference groups. The points in C_2 are not outliers as they follow the normal pattern between the two attributes. Hence, the outliers detected in this case have no meaning. Therefore, a more robust approach is needed to distinguish clusters C_3 and C_4 with cluster C_2 in high dimensions and identify data points from C_3 and C_4 as true outliers. Moreover, due to high dimensionality, the model will be computationally expensive to learn.

Another drawback of the traditional outlier detection techniques is that they detect outliers on full attribute space but do not consider the interrelationships among subspaces of relevant attributes to detect the outliers. A data analyst may find vital information about the underlying processes that lead to the outliers by analyzing the subspaces. For instance, a person suffering from a cardiac disease is typically different from a normal person in features associated with heart such as heart rate, angina, arterial plague, and atrial fibrillation. Other features like skin and hair type may be irrelevant for identifying this type of person. Therefore, having such knowledge about the attributes leads to the effective identification of the outliers. Also, the number of subspaces increase exponentially when the dimensionality increases leading to complexity issues [11]. Therefore, the key challenge here is to select relevant or meaningful subspaces to detect the outliers while avoiding a complete search over all possible subspaces.

In this thesis, we propose a new outlier detection approach for data sets that contain both categorical and continuous attributes. This approach uses an effective subspace sampling method that picks relevant subspaces in preference over full attribute subspace and uses the contextual information of the data as well as a similarity measure to identify the outliers. We base this heuristic on the hypothesis that true and meaningful outliers are likely to be identified in highly correlated feature subspaces by considering both contextual information and similarity of the data points. This hypothesis is based on the challenges with sparse contexts in high dimensional data, wherein the contexts are not informative enough in high dimensional space and are not very useful for outlier detection.

1.4 Contributions

We make the following contributions:

1. The primary contribution of this thesis is to present a framework for identifying true and meaningful outliers in mixed data sets consisting of categorical and continuous attributes. As mentioned earlier, to detect meaningful outliers we require contextual belief of the domain; thus, we propose to use Bayesian networks to define contexts in the data sets. Bayesian networks capture causal relationships among attributes that exist in the data sets, thereby allowing us to explore these relationships to mine interesting outliers. Specifically, in this thesis, we use a special type of Bayesian network called Hybrid Bayesian Network, which provides an ideal representation for capturing contextual knowledge consisting of both categorical and continuous attributes. Additionally, we also describe why a reported point is an outlier.
2. To overcome the difficulty of sparse contexts in high dimensional data, we use a unique characteristic of the data called similarity; i.e. points which are closer to each other have similar properties than the points which are far from each other.

In this thesis, we propose to use angle-based similarity measure to compare the data points since the distance or nearest neighbor concepts become less meaningful due to sparseness in high dimensions. The angle-based similarity measures the degree of outlierness of each point on the evaluation of the broadness of its angle spectrum. The smaller the angle spectrum of a point to other pairs of points, the more likely it is an outlier. Therefore, this approach does not significantly worsen in high-dimensional data because angles are more stable than distances in high dimensions.

3. For subspace selection, we propose a two-stage process to select relevant attributes for forming a subspace. In the first stage, we apply Maximal Information Coefficient to select a subset of attributes that are highly correlated with class labels. In the second stage, the selected attributes are used to learn a Hybrid Bayesian Network and then extract the Markov blankets of each attribute from the learned Hybrid Bayesian Network to form a highly correlated subspace. Then for each subspace, we learn a local Hybrid Bayesian Network. Finally, we derive an outlier score for each point by considering both the joint probability distributions and angle-based similarity measure to identify outliers which violate both the underlying contextual beliefs and neighborhood criteria. The data points with the lowest scores are reported as outliers.

1.5 Organization

This thesis is organized as follows. Chapter 2 describes the literature review of the current outlier detection techniques along with their main strengths and weaknesses. The fundamental concepts of Bayesian networks such as conditional independence, joint probability distributions, Markov blankets and Bayesian structural learning are discussed in Chapter 3. In Chapter 4, we present a methodology to identify meaningful outliers in subspaces using probability distributions of Hybrid Bayesian Network and angle-based similarity measure. In Chapter 5, we show experimental evaluations of our proposed approach with real-world data sets. Chapter 6 concludes the thesis with a summary and directions for future work.

CHAPTER 2

BACKGROUND

In this chapter, we present the literature review of existing outlier detection techniques under four main categories, namely density based, distance based, contextual-based, and Bayesian network-based approaches. Distance and density-based techniques are the oldest and widely used technique for outlier detection. These approaches use a similarity or distance measure to compute the distances between the points, and points far away from their neighborhoods are flagged as outliers. In contrast to distance and density-based techniques, contextual outlier detection techniques use background knowledge or contextual information of the domain to find outlier patterns. Similarly, Bayesian network-based techniques consider the underlying probability distributions of the data set to identify outliers. Outliers are those observations which have low probability.

In addition to this categorization, we explore subspace-based techniques. These approaches mine outliers from the subset of relevant attributes selected from the high dimensional dataset. Also, there are several alternate techniques which focus on discovering outliers using information and spectral theory. Dutta $et\ al.$ [12] and Wenke $et\ al.$ [13] proposed spectral anomaly detection and information theoretic techniques to mine anomalies in astronomical data and network data, respectively.

2.1 Density Based Methods

Density-based methods consider the underlying distribution of the input data and divide the data into high-density and low-density regions. The points which are lying in regions of low density are identified as outliers while the points that lie in the dense neighborhood are normal. These techniques estimate the density of the neighborhood of each point in the data.

The popular and widely used density-based outlier detection algorithm is the Local Outlier Factor (LOF) introduced by Breunig $et\ al.$ [8]. The main idea of LOF is based on the estimation of local density of the points. The local density of a point is computed using the reachability distance method. The reachability distance between two points x and y is interpreted as the maximum k-nearest neighbor distance from y to its outermost point in y's region and the distance from y to x. The computed local density of the point is compared with the local densities of its neighbors. The data points which have lower local density than their neighbors are considered to be outliers.

The local reachability density of x is defined as

$$lrd(x) = 1\Bigg/\left(\frac{\sum_{y \in N_k(x)} reachdist_k(x,y)}{|N_k(x)|}\right) \qquad \textbf{Eq. 2-1}$$

where $|N_k(x)|$ is the number of data points in x's k-nearest neighbor regions and the Local outlier factor of x is defined as

$$LOF_k(x) = \frac{\sum_{y \in N_k(x)} \frac{lrd_{(y)}}{lrd_{(x)}}}{|N_k(x)|} \qquad \textbf{Eq. 2-2}$$

The data point which has a LOF score greater than 1 will be treated as an outlier.

Another widely used density-based algorithm is DBSCAN initially developed to cluster spatial systems with unrestricted cluster shapes [14]. DBSCAN algorithm takes into consideration the minimum number of points and a distance measurement to group together points that are close to each other. Usually, Euclidean distance is used as the distance measure. This algorithm requires two user specified parameters called eps and $minPoints$. If the distance between two points is less than or equal to the eps value, then these points are added to the neighborhood. $MinPoints$ define the number of points to form a dense region. As this method depends on user specified parameters, it suffers from accuracy problems. For example, if eps value is too small, even the normal points are flagged as outliers, and if the eps value is too large, the outliers will be considered as normal points. On the other hand, choosing $minPoints$ also play important role in improving accuracy of the model.

Additionally, many density-based techniques were developed as an extension to the LOF algorithm such as GridLOF algorithm [15], and Connectivity-based Outlier Factor (COF) [16].

2.1.1 Benefits and drawbacks of Density Based Methods

The following are the benefits of density based [17] methods:

1. Density-based methods are unsupervised in nature and can be readily used in a wide variety of applications.
2. Specifically, the data sets which carry varying densities of data points are benefitted from identifying local outliers.

The following are the drawbacks of density based [17] methods:

1. Determining the quality of reported outliers is a problem because these techniques do not consider the background knowledge of the domain.
2. Additionally, these methods have high computational complexity, since these methods estimate the density of each point in their neighborhood.

2.2 Distance Based Methods

The distance-based methods are the oldest and most widely used methods for anomaly detection. In these techniques, each data point is analyzed with respect to its nearest neighbor. These techniques assume that outliers are the points with fewer than k nearest neighbors in the data, where a neighbor is an object that is within a distance. A distance or a similarity measure is required by these techniques to measure the distance between two data points and can be computed in different ways. Euclidean distance is a general choice for continuous variables, but other measures such as Manhattan distance, Mahalanobis distance, Minkowski distance and Cosine similarity can be used as suggested by Tan $et\ al.$ [18]. For categorical attributes, simple matching coefficient is used as described in Boriah $et\ al.$ [19].

Knorr $et\ al.$ [20] firstly introduced the concept of distance-based outlier. According to the authors, a data point p is an outlier if at least a fraction of the data points in the data set lie greater than some distance d from p. This definition was later extended to include a predefined number of points in a neighborhood and measure the distance of a point to its kth nearest neighbor, called as the k-nearest neighbor method. Dang $et\ al.$ [21] presented a k-nearest neighbor approach to detect outliers in large scale traffic data collected from some cities. They considered for any data point p which satisfies the condition $D^K\ (p)\ >\ t,$ where t is some threshold as an outlier. The authors use Euclidean distance to compute the distance between the points. Shirazi $et\ al.$ [22] proposed a combination of two outlier detection techniques called SF-KNN and SUS-KNN based on the best selected features and k-nearest neighbor algorithm to detect network intrusions like U2R and R2L.

Furthermore, the performance of distance-based methods highly depends on the distance or similarity measure adopted.

2.2.1 Benefits and drawbacks of Distance Based Methods

The following are the benefits of distance-based methods [17]:

1. As in the case with density-based methods, the distance-based methods are unsupervised in nature, purely data driven and do not make any background assumptions of the data.
2. By employing an appropriate distance metric, these techniques can be applied on data sets containing mixed attribute types.

The following are the drawbacks of distance-based methods [17]:

1. Like the density-based methods, determining the quality of reported outliers is challenging.
2. These techniques will fail in detecting outliers if there is not enough similar data.
3. Defining a distance metric for complex data sets containing unstructured, semi-structured and structured data could be challenging.

2.3 Contextual Based Methods

Contextual outliers also known as conditional outliers are outliers in a specific context but not otherwise [23]. The context is defined by the domain knowledge or by the structure in the data set and must be specified before applying. Two different attributes; i.e. contextual attributes and behavioral attributes, are used for defining a data point. To define a context for a particular data point, the contextual attributes are used. The best examples are the latitude and longitude of a location in spatial data and time in temporal data. The non-contextual characteristics of a data point are defined by the behavioral attributes. For example, drinking water conditions at any location is a behavioral attribute in spatial data.

The behavioral attributes are used within a specific context to detect any outlier behavior. In a specific context, a point may be a contextual outlier, but if we consider behavioral attributes, the same point might be normal in a different context. This property is key in classifying contextual and behavioral attributes for finding contextual outliers. One such example of a contextual outlier is the different temperatures recorded in summer and winter seasons at a specific place. A temperature of $30°F$ in the summer would be an outlier at that place, but the same temperature would be normal in winter. Here, contextual attributes are winter, and place and the behavioral attribute is temperature.

Contextual outliers are usually explored in streaming data [24], spatial data [25] and image recognition [26]. Wei $et\ al.$ [27] identified contextual outliers using a hypergraph based on the frequent item sets in the data. They grouped together objects containing frequent item sets denoted by hyperedge and designed a deviation score to compute the outlier aspect of the data in a particular attribute with respect to hyperedge. A data point is an outlier if the deviation score of that point is below some threshold θ.

The other contextual outlier detection studies include the one proposed by Valko *et al.* [28]. They used soft harmonic solutions to detect contextual mislabeled anomalies. Regularization methods were used to avoid detection of isolated and distribution boundary instances. Wang *et al.* [29] introduced a method using random walks without a priori contextual information to mine the context and outliers automatically.

The significance of the contextual outliers in the target application domain determines the choice of applying these techniques.

2.3.1 Benefits and drawbacks of Contextual Based Methods

The following are the benefits of contextual based methods [17]:

1. Contextual based approaches are applicable to real-world problems where the data tends to be similar within a context.

The following are the drawbacks of contextual based methods [17]:

1. These techniques depend on the contextual attributes, and it may be challenging to define the context for every application area.
2. In high dimensional spaces, the contexts become sparse.

2.4 Bayesian Network Based Methods

Bayesian networks are frequently used in multi-class outlier detection problems. Bayesian networks are directed acyclic graphical model for depicting probabilistic relationships among a set of variables and come under classification-based methods. For a given data point, the Bayesian networks estimate the probability of observing a class label from a set of normal class labels and the outlier class label. The data point with the largest posterior probability is chosen as the predicted class. Due to its graphical representation of the relationships and strong inference mechanism, Bayesian networks attract a great number of researchers for various applications.

Babbar $et\ al.$ [30] proposed an anomaly detection method using two probabilistic association rules derived from Bayesian networks. They based these rules on two different situations occurring in joint probability distribution; i.e. low prior - high posterior probability and high prior - low posterior probability. Nicholas $et\ al.$ [31] presented a two-stage Bayesian model to detect outliers in social networks. Conjugated Bayesian models are used in the first stage to judge normality of behavior by tracking the pairwise links of all the nodes in the graph. Standard network inference tools are applied in the second stage on a reduced subset of potentially outlier nodes. Rashidi $et\ al.$ [32] presented a technique to identify outliers in categorical data sets using Bayesian networks and attribute value combinations. An AD Tree structure was used to store attribute value combinations.

Wong $et\ al.$ [5] developed a method for detecting disease outbreaks using Bayesian networks. In this approach, the authors form a probabilistic relation between attributes of environmental set which consists of disease trends to attributes in the indicator set which contain all other attributes. The outlier patterns which cause disease outbreaks are identified by comparing the test data against the already established disease patterns. Masood $et\ al.$ [33] and Malhas $et\ al.$ [34] proposed an iterative model to use multiple probabilistic interesting aspect measures to mine anomalous patterns from Bayesian networks. Specifically, they used contingency tables to calculate probabilities.

2.4.1 Benefits and drawbacks of Bayesian Network Based Methods

The following are the benefits of Bayesian network-based methods:

1. Mining genuine outliers are possible because Bayesian networks encapsulate the background knowledge of the domain.
2. The testing phase is fast and is a powerful tool to differentiate between instances of different classes.
3. The conditional independence properties and joint probability distributions provide an easy explanation of why an identified data instance is an outlier.

The following are the drawbacks of Bayesian network-based methods:

1. The joint probabilities between data points will become increasingly similar as the dimensionality increases, which could lead to high false positive rates.

2.5 Subspace Based Methods

New challenges have been introduced due to ever increasing volume and dimensionality of the data sets. Due to the curse of dimensionality as described by Aggarwal $et\ al.$ [35], the existing outlier detection methods fail to perform well when directly applied to the full attribute space in high dimensional data. Thus, identifying the outliers in selected features in low dimensions can be used as an alternate approach to solve high dimensionality problems. The subspace outlier detection techniques aim to discover outliers deviating from the majority in some selected attribute feature space.

The subspace selection and outlying measurement design are two major components in subspace outlier detection. Based on the challenges encountered in high dimensional data, many researchers proposed subspace selection methods to select meaningful subspaces. Aggarwal $et\ al.$ [35] discovered that all the points to be equidistant in high dimensional data with no distinction among them proposed a method to search lower dimensional subspaces for outliers. Knorr $et\ al.$ [36] introduced the concept of determining strong and weak outliers in minimal subspaces and also explain why an identified point is an outlier.

Additionally, the number of subspaces available is directly proportional to the number of attributes in the data set; i.e. as the number of attributes increases, the number of subspaces also increases. Therefore, to solve this problem, several methods have been proposed. Keller $et\ al.$ [37] proposed to select subspaces with high contrast using statistical approaches. Lazarevic $et\ al.$ [38] proposed the idea to mine outliers by randomly sampling the feature set to obtain subspaces of varying sizes. Furthermore, Aggarwal $et\ al.$ [35] developed a method to determine the outliers in subspaces by identifying the exact number of attributes required to form a subspace. Another example is that Kriegel $et\ al.$ [39] designed a method to select subspaces according to the nearest neighbors to mine the outliers. Cherbrolu $et\ al.$ [40] and Wang $et\ al.$ [41] built an intrusion detection system by selecting important features using Markov blanket model and decision trees. Bayesian networks and regression trees are used to create an intrusion detection model. Duan $et\ al.$ [42] introduced a method to find high contrast subspaces to mine meaningful outliers. Joshi $et\ al.$ [43] developed a method to determine a set of contiguous subspaces to search for similar outliers.

2.5.1 Benefits and drawbacks of Subspace Based Methods

The following are the benefits of subspace-based methods:

1. Subspace-based techniques automatically reduce high dimensional space to low dimensional space by selecting a subset of relevant features.

2. These techniques can be used in both supervised and unsupervised models.

The following are the drawbacks of subspace-based methods:

1. Subspace-based techniques typically have high computational complexity.

2.6 Summary

This chapter highlighted the existing research in outlier detection domain that is related to this thesis. Following the introduction, this chapter discussed the four main outlier detection techniques that are available in the literature: density-based, distance-based, contextual-based and Bayesian network-based. Additionally, we also discuss a special outlier detection technique known as subspace outlier detection. This was followed by describing the strengths and weaknesses of each technique and their applications in the literature.

CHAPTER 3

BAYESIAN NETWORK MODELS

Bayesian networks are directed acyclic graphs which provide a graphical schematic to represent the underlying probabilities that define the associations between attributes of some data. Bayesian networks are a universal tool for modeling and reasoning under uncertainty in machine learning research. Since the domain information is not always available in many real-world problems leading to uncertain and inappropriate conclusions, Bayesian networks use their inference scheme in such cases to find solutions which are possible. Additionally, Bayesian networks support structural and parameter learning as well as incorporate new evidences into the model.

The rest of the Chapter is organized as follows. Section 3.1 briefly describes the concept of Bayesian networks and Bayes theorem along with an example. Key concepts such as dependency, independency and conditional independence in Bayesian networks are discussed in Section 3.2. Section 3.3 deals with joint probability distributions in Bayesian networks while Section 3.4 is focused on Markov blankets. Section 3.5 describes Bayesian structure learning and scoring. Finally, Section 3.6 highlights the concept of hybrid Bayesian networks.

3.1 Bayesian Networks

Bayesian networks are characterized by their use of probability distributions for handling the interdependencies between attributes. They have a directed acyclic graph consisting of nodes and arcs, with nodes representing the attributes and arcs representing the relation between the attributes. Two attributes connected using an arc are said to influence each other. The direction of the arc characterizes the parent and child nodes, which are interpreted as two attributes being dependent on each other. The two important components of a Bayesian network are structural representation of the model, and its underlying probability distribution. Probability allows to deal with uncertainty in real world problems, whereas structural models help in representing the real-world situations in a diagrammatic form making it simpler for the user to understand [44].

Figure 3-1 illustrates a simple Bayesian network. The nodes A and C are not connected to each other, indicating that they are mutually independent. Nodes C and D are connected to each other with an arc directed from C to D. This denotes that the node C is the parent and D is the child. Similarly, E is the child node of D. Additionally, D is the child node with two parent nodes A and C.

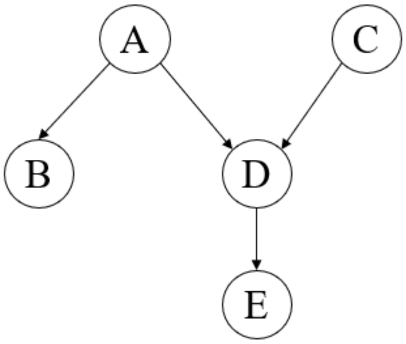


Figure 3-1: Illustration of a simple Bayesian network.

The most important components of the Bayesian networks are the prior, posterior and the likelihood. Let us assume A represents known attribute values while Y represents the class labels. The basic formula of the Bayes theorem is given as

$$P(A|Y) = \frac{P(Y|A)P(A)}{P(Y)} \qquad \textbf{Eq. 3-1}$$

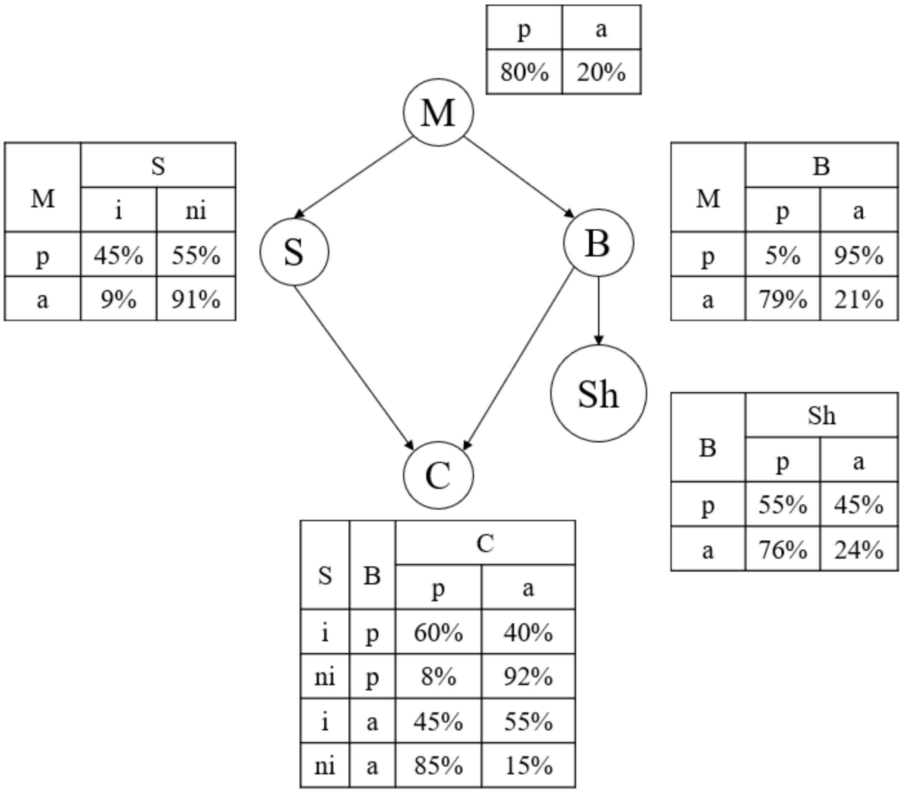
where, $P(A)$ and $P(Y)$ are the probabilities of A and Y irrespective of each other, $P(Y|A)$ is the probability of event Y given A is true and $P(A|Y)$ is the vice versa. In terms of the cause and the effect, it can be restated as

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)} \qquad \textbf{Eq. 3-2}$$

Here, $P(Cause)$ is the prior probability, $P(Cause|Effect)$ is the posterior probability, and $\frac{P(Effect|Cause)}{P(Effect)}$ is the likelihood. Therefore,

$$Posterior\ probaility\ = Likelihood\ \times\ Prior\ Probability \qquad \textbf{Eq. 3-3}$$

For concrete understanding of Bayesian networks, consider the following medical Bayesian network on cancer disease as shown in **Figure 3-2**. This Bayesian network has five binary attributes denoted by five circles with their respective names. The relational dependency between the attributes is represented by the directed arrows or arcs. According to this Bayesian network, the events of brain tumor and serum calcium, denoted by the nodes B and S respectively are caused by metastatic cancer, which is denoted by the node M. Here, M is the parent node and B, S are child nodes. Similarly, brain tumor causes severe headache and coma represented by the nodes Sh and C, respectively. Moreover, the event serum calcium also affects the event coma. Each node is associated with the unconditional (prior) and conditional probability (posterior) table. Each node represented in the network can take up two states, namely present, which is denoted by p, and absent, which is denoted by a except the node serum calcium which takes values increased denoted by i, and not increased denoted by ni. The table for node M contains unconditional probability distributions because it does not have any parent nodes. For example, probability of metastatic cancer is 80% and absence of metastatic cancer is 20%. Since node B is dependent on node M, it contains conditional probabilities indicating the probability of a brain tumor in the presence or absence of metastatic cancer. For example, the presence of metastatic cancer in the body causes a brain tumor with 5% chance. Likewise, the tables of other attributes S, Sh and C contain similar information.



p	a
80%	20%

M	S	
	i	ni
p	45%	55%
a	9%	91%

M	B	
	p	a
p	5%	95%
a	79%	21%

B	Sh	
	p	a
p	55%	45%
a	76%	24%

S	B	C	
		p	a
i	p	60%	40%
ni	p	8%	92%
i	a	45%	55%
ni	a	85%	15%

Figure 3-2: Bayesian network representation of the cancer disease.

3.2 Bayesian Networks Independencies

Key concepts such as dependency, independency and conditional independence among variables can be observed in Bayesian networks. Dependency is said to exist between two variables if one variable provides the predictive value for another variable. This is represented by an arc joining two nodes in the Bayesian network. For example, in **Figure 3-2**, we can predict the probability of brain tumor by knowing the state of metastatic cancer because a brain tumor is dependent on metastatic cancer. These are called dependent nodes. However, there are situations in the graph where the information does not flow directly between two nodes as they are not connected to each other. For example, the node age provides no information about the state of metastatic cancer in a person. This property is called independence [44], which is defined below.

Definition 1- Independence: An attribute X is said to be independent of another attribute Y corresponding to a probability distribution P if and only if

$$P(X|Y) = P(X) \ or \ if \ P(Y) = 0 \qquad \textbf{Eq. 3-4}$$

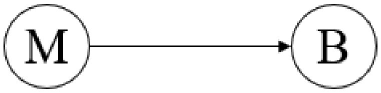


Figure 3-3: Relational dependency between attributes metastatic cancer M and brain tumor B.

Apart from dependent and independent events, there exist situations in the graph where predictive information flows between two unconnected nodes through a third node.

These are called conditionally independent nodes [44]. For example, the knowledge of metastatic cancer determines the predictive value of a coma through a node brain tumor. Therefore, the two attributes metastatic cancer and coma are conditionally independent of each other given the knowledge of the brain tumor. The formal definition of conditional independency is given below.

Definition 2-Conditional Independence: An attribute X is said to be conditionally independent of attribute Y given another attribute Z corresponding to a probability distribution P if and only if

$$P(X,Y|Z) = P(X|Z) \times P(Y|Z) \qquad \textbf{Eq. 3-5}$$

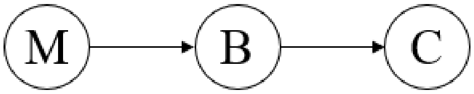


Figure 3-4: Conditional independency between attributes metastatic cancer and coma given the brain tumor.

3.3 Joint Probabilities in Bayesian Network

Bayesian networks are the concise and compact graphical representations of joint probability distributions. If there are d nodes in a Bayesian network denoted by X_1 to X_d, then the joint probability distribution is given as $P(X_1 = x_1,\ X_2 = x_2, \dots, X_d = x_d)$. Since a node in the Bayesian network is conditioned only on its parent node, the joint probability distributions can be broken down using chain rule of probability in the following way [44]:

$$P(X_1 = x_1, X_2 = x_2, \dots, X_d = x_d)$$
$$= P(X_1 = x_1) \times P(X_2 = x_2 | X_1 = x_1) \times \dots$$
$$\times P(X_d = x_d | X_1 = x_1, \dots, X_{d-1} = x_{d-1}) \qquad \textbf{Eq. 3-6}$$

$$P(X_1 = x_1, X_2 = x_2, \dots, X_d = x_d) = \prod_{k=1}^{d} P(X_k = x_k | Pa(X_k)) \qquad \textbf{Eq. 3-7}$$

where $Pa(X_k)$ is the parent of X_k.

For example, by using **Eq. 3-7,** we can compute the joint probability distribution of a situation $P(M = p, S = ni, B = a, Sh = p, C = p)$ in **Figure 3-2** as shown in **Eq. 3-8**. This computation results in 22% probability of occurrence:

$$P(M = p, S = ni, B = a, Sh = p, C = p)$$
$$= P(S = i|M = p) \times P(B = a|M = p)$$
$$\times P(C = p|B = a, S = ni) \times P(Sh = p|B = a)$$
$$\times P(M = p) = 22\% \qquad \textbf{Eq. 3-8}$$

3.4 Markov Blanket of Bayesian Networks

Pearl [45] first introduced the concept of Markov blankets. In a Bayesian network, the Markov Blanket for node X_i which we denote by $MB(X_i)$ is a set of nodes composed of X_i's parents, its children and its children's other parents (spouses) as shown in **Figure 3-5** [46]. Formally, the definition of Markov blanket in a Bayesian network, or more general in a graph, is as follows:

$$MB(X_i) = Pa(X_i) \bigcup Ch(X_i) \bigcup_{Y \in Ch(X_i)} Pa(Y)$$ **Eq. 3-9**

where $Pa(X_i)$ is the parent node of X_i, $Ch(X_i)$ is the child node of X_i and $Pa(Y)$ denotes the other parents (spouses) of X_i's child node.

From **Eq. 3-9**, we can observe that the Markov blanket of attribute X_i consists of just its parents, children and spouses and is independent of all the other attributes in the Bayesian network. Thus, these attributes are highly correlated and are sufficient to provide information about the attribute X_i. The other attributes in the network are unrelated to X_i. Thus, this property of the Markov blanket is helpful for causal discovery; i.e. to reduce the number of variables, an experimentalist must consider in order to discover the direct causes of X_i.

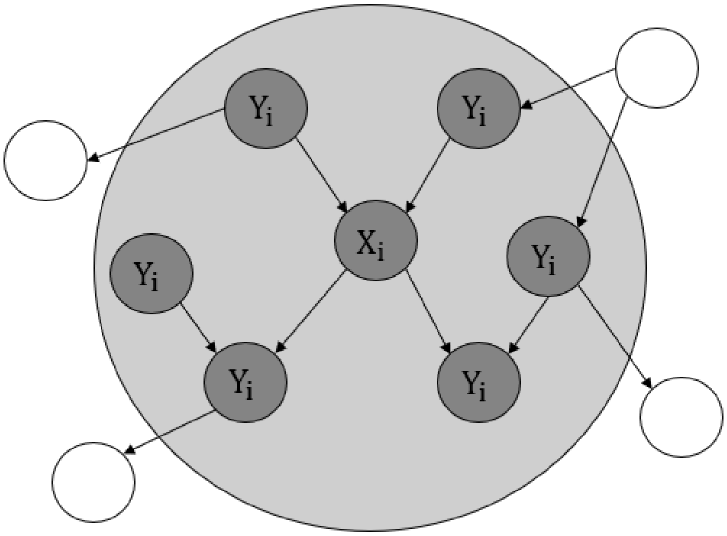


Figure 3-5: Example of a markov blanket for node X_i with Y_i as its parent nodes, child nodes and spouses.

3.5 Bayesian Network Structural Learning

For learning the structure of the Bayesian network, two main approaches are used: constraint-based methods and score-based methods.

Constraint-based algorithms first try to estimate whether certain conditional independencies between data hold true and then try to discover the network structure that best fit these constraints [47]. The estimations are performed using statistical or information theory measures. Moreover, in these approaches, the independence properties are separated from structural findings resulting in a single graphical output with clear semantics. However, it is difficult to optimize the network structure and find reliable conditional independence properties. The Incremental Association Markov Blanket (IAMB) algorithm is one of the examples of the constraint-based algorithms. It uses a forward selection scheme to discover the Markov blanket of the class label followed by an attempt to remove false positives [48], where the Markov blanket of a node is defined as the knowledge needed to predict the behavior of that node.

Score-based approaches uses a scoring function to find the best value for the graph structure by searching through the space all possible structures [47]. Examples of score-based algorithms include Hill Climbing (HC) and Tabu search. Furthermore, the score-based approaches require a scoring function which gives a good score when the graph best fits the data. Several scoring functions have been developed to fit a wide variety of data such as Bayesian Dirichlet (BD) criterion, Bayesian information criterion (BIC), Akaike information criterion (AIC), K2 and Loglik scoring function.

Bayesian Dirchlet (BDe) scoring function: Heckerman $et\ al.$ [49] proposed this scoring function. Given a directed acyclic graph G, it makes four assumptions on parameter independence, parameter modularity, uniformity of prior distributions, and lack of missing values. The equation below represents the BD score function, where D denotes the data, τ denotes a gamma function, $P(G)$ the prior probability of the network, and N'_{ij} denote the hyperparameters of the network.

$$BD(G,D) = \log\big(P(G)\big) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{\tau\big(N'_{ij}\big)}{\tau\big(N_{ij}+N'_{ij}\big)}\right) + \sum_{k=1}^{r_i}\log\left(\frac{\tau\big(N_{ijk}+N'_{ijk}\big)}{\tau\big(N'_{ijk}\big)}\right)\right) \quad \textbf{Eq. 3-10}$$

Since the N'_{ij} are quite difficult to compute, an additional assumption of likelihood equivalence is considered resulting in the BDe scoring function given by

$$P(G,D) = \log\big(P(G)\big) \times \prod_{i=1}^{n}\prod_{j=1}^{q_i}\left(\frac{\tau\big(N'_{ij}\big)}{\tau\big(N_{ij}+N'_{ij}\big)} \times \prod_{k=1}^{r_i}\frac{\tau\big(N_{ijk}+N'_{ijk}\big)}{\tau\big(N'_{ijk}\big)}\right) \quad \textbf{Eq. 3-11}$$

where $N'_{ijk} = \; N' \times P\big(X_i = x_{ik}, \prod_{X_i} = w_{ij}|G\big)$ [49].

K2 scoring function: This is one of the first Bayesian scoring functions proposed by Cooper and Herskovits [50]. It is a particular case of Bayesian Dirichlet with the uninformative assignment $N'_{ijk} = 1$ which corresponds to the zero pseudo-counts. Since $\tau(c) = (c-1)!$ with c being an integer, and τ denotes a gamma function. The K2 score can be expressed as follows:

$$K2(G,D) = \log\big(P(G)\big) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{(r_i-1)!}{\big(N_{ij}+r_i-1\big)!}\right) + \sum_{k=1}^{r_i}\log\big(N_{ijk}!\big)\right) \quad \textbf{Eq. 3-12}$$

Loglik scoring function: This score is the logarithm of the likelihood of data D given the network G. It is obtained by $\log(P\big(G(D)\big)) = -L(D|G)$.

The Loglik score is computed using the following equation:

$$LL(D|G) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right) \qquad \textbf{Eq. 3-13}$$

From the above equation, we can see that maximizing Loglik function of the network minimizes the information content of the data.

3.6 Hybrid Bayesian Networks

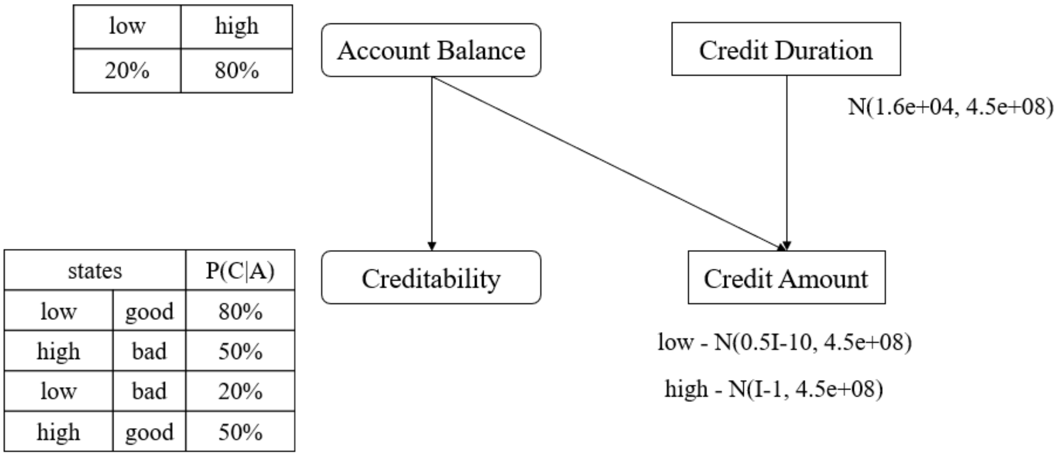
In general, a Bayesian network learns from the data which is either fully discrete (categorical) or fully continuous. However, a Hybrid Bayesian Network can learn a network on both discrete and continuous variables [51]. Compared to standard Bayesian networks, Hybrid Bayesian networks are useful in wider applications consisting of attributes of different types and they can model the true distribution of the data without discretization.

Consider a directed acyclic graph G and its probability distribution $P_k = P(s_k|pa_k)$, where pa_k is the set of parent nodes of s_k. Assume C is the set of attributes partitioned into discrete attributes denoted by ◆ and continuous attributes denoted by τ. Therefore, a Hybrid Bayesian Network $H = (C, P)$ is defined over this graph G by the conditional distribution of continuous attributes from the Gaussian model which is represented as

$$P(s_k|K = k, L = l) = N(\alpha(k) \ + \ \beta(k) \times \ z, \gamma(k))\ s_k \in \tau \qquad \textbf{Eq. 3-14}$$

where K and L are the set of discrete and continuous parents of s_k, respectively, and N represents multi-variate normal distribution with mean μ and standard deviation σ.

An example of Hybrid Bayesian Network is shown in **Figure 3-6**. The attributes Account Balance and Creditability are discrete or categorical attributes represented by round boxes, whereas Credit Amount and Duration of Credit are continuous attributes represented by square boxes. The attribute Account Balance contains prior probabilities of two states, namely low, and high, whereas attribute Creditability contains conditional probabilities in its states, good and bad conditioned on its parent attribute Account Balance. Conversely, for the continuous attribute Credit Duration, the information is denoted with mean µ and standard deviation σ. For attribute Credit Amount, the probability density function is given by **Eq. 3-14**.



low	high
20%	80%

states		P(C|A)
low	good	80%
high	bad	50%
low	bad	20%
high	good	50%

Figure 3-6: Example of Hybrid Bayesian Network.

CHAPTER 4

METHODOLOGY

This chapter describes the procedure to determine the meaningful outliers in the set of subspaces. We analyze such a set of subspaces to provide hints about why the outliers might be occurring. Our methodology uses Maximal Information Coefficient and Markov blankets of Hybrid Bayesian Networks (HBN) to determine subspaces. Hybrid Bayesian Network allows us to define contextual information using joint probabilities. We then combine these joint probability distributions and angle-based similarity measure to determine and explain the outliers in the data set containing both categorical and continuous attributes. We base this approach on the hypothesis that true and meaningful outliers are likely to be identified in highly correlated feature subspaces by considering both contextual information and similarity of the data points.

This chapter provides a detailed overview of the proposed methodology, and the specific aims of this work. Beginning with a discussion of the concepts relating to Markov blankets and angle-based similarity measure, we then provide a detailed explanation of the developed algorithm followed by a description of the data set used.

4.1 Specific Aims

The specific aims of this thesis are as follows:

1. Determining a set of highly correlated subspaces from full attribute space to analyze the outliers.
2. Derive an outlier score to identify the data points that violate both the contextual beliefs and neighborhood criteria.
3. Summarize the collected information for each detected outlier and identify the corresponding subspaces for those outliers.
4. Lastly, examine the subspaces computed in the previous steps and provide insights on why a given point is labeled an outlier.

4.2 Feature Selection using Maximal Information Coefficient

Maximal Information Coefficient (MIC) is a feature selection method which is used to determine the correlation between two attributes in the data set [52]. It is capable of measuring both linear and non-linear relationships between the attributes. The basic idea behind MIC is that it depends on the mutual information between two attributes to measure the degree of their relationship. MIC reveals the dependency between attributes and evaluate their statistical importance and rank them according to the strength of the relationship by representing the mutual information scores in the range of 0 and 1. Additionally, MIC searches for an ideal number of bins in such a way that mutual information between attributes is maximized, thereby avoiding user-specified bins. Also, the values of MIC are not influenced by presence of outliers in the data set.

The Maximal Information Coefficient between two attributes F and L denoted by $MIC(\mathrm{F}, \mathrm{L})$ in a data set D is determined by computing the mutual information $I(F, L)$ normalized by the minimum entropy of F and L. This is represented as

$$MIC(F,L) = \frac{I(F,L)}{min\{\mathrm{H(F)},\mathrm{H(L)}\}} \qquad \textbf{Eq. 4-1}$$

The above **Eq. 4-1** can be further simplified into the following:

$$MIC(F,L) = \frac{H(F) - H(F|L)}{\min\{\mathrm{H(F)},\mathrm{H(L)}\}} = \frac{H(L) - H(L|F)}{\min\{\mathrm{H(F)},\mathrm{H(L)}\}} \qquad \textbf{Eq. 4-2}$$

where $H(\mathrm{F|L})$ is the conditional entropy.

Conditional entropy describes the amount of evidence required to estimate the outcome of F given the value of L. If $MIC(F,L) = 0$, then F and L are statistically independent, and if $MIC(F,L) = 1$, then F and L are statistically dependent.

4.3 Subspace Discovery using Markov Blankets

As described in Chapter 3, Markov blankets were initially presented by Pearl [45]. In this section, we provide an explanation for Markov blankets and its characteristics, followed by the reasoning for choosing Markov blankets for subspace selection.

Assume a training data set that contains n samples and D attributes. Let G be a directed acyclic graph through which joint probability distributions P are learned. Then, a Bayesian network B satisfies the Markov condition if every attribute in the Bayesian network is conditionally independent of its non-descendant attributes conditioned on its parents [45]. Therefore, if $Pa(D_i)$ is the set of parents of D_i in B, then the joint probability P is represented as

$$P(D) = \prod_{i=1}^{d} P(D_i|\, Pa(D_i)) \qquad \textbf{Eq. 4-3}$$

Definition 1-Bayesian Faithfulness: A Bayesian network B is said to be faithful to its probability distribution P if and only if every conditional independency present in P is also present in B.

Definition 2-Markov Blanket (Graphical view point): From the faithfulness definition, the Markov blanket of an attribute D in the Bayesian network B is the set of D's parents, children, and its children's other parents (spouses).

$$MB(D) = Pa(D) \bigcup Ch(D) \bigcup Sp(D) \qquad \textbf{Eq. 4-4}$$

Definition 3-Markov Blanket (Probability view point): From the faithfulness definition, the Markov blanket of an attribute D in the Bayesian network B is a minimal set of attributes conditioned on D that make D statistically independent from all the remaining attributes.

$$P(D|MB(D)) = P(D|Pa(D)) \prod_{Z_j \in Ch(D)} P(Z_j|Pa(Z_j)) \qquad \textbf{Eq. 4-5}$$

From the above discussion, we conclude that the Markov blanket of each attribute will be unique when a Bayesian network satisfies the faithfulness condition and is thus suitable for forming a subspace.

Definition 4-Markov Blanket Subspace: A subspace that consists of an attribute and its parents, children, and spouses.

Figure 4-1 shows an example of Bayesian network consisting of nine attributes of a person having diabetes [9]. If Diastolic blood pressure is the attribute of interest, then one might wish to determine the value of this attribute given some assignment of values to the other attribute in the domain. However, using the Markov blanket concepts, we can eliminate all irrelevant attributes and consider attributes that are highly correlated with

Diastolic blood pressure. The parent of Diastolic blood pressure is {Diabetes} and the set of children is {Plasma Glucose Concentration, Serum Insulin}. The spouse of the children is {Diabetes}. Therefore, the Markov blanket of the attribute Diastolic blood pressure is the set {Diabetes, Plasma Glucose Concentration, Serum Insulin}. This blanket is depicted in **Figure 4-1**. For the attribute Diastolic blood pressure, knowledge about other attributes become irrelevant if we know {Diabetes, Plasma Glucose Concentration, Serum Insulin} because the blanket shields Diastolic blood pressure from the effects of those attributes outside it.

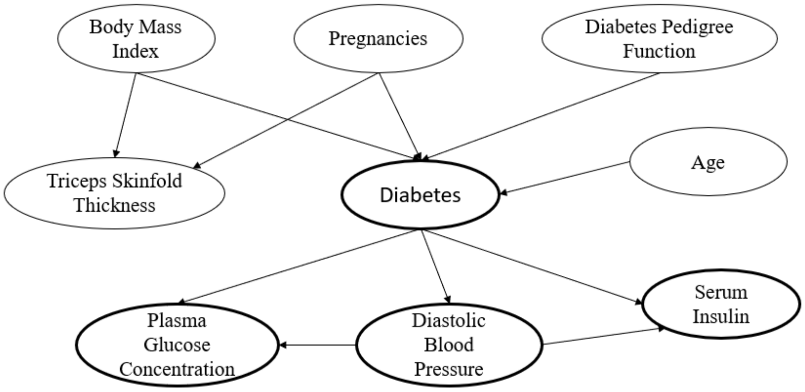

Body Mass Index
Pregnancies
Diabetes Pedigree Function
Age
Triceps Skinfold Thickness
Diabetes
Serum Insulin
Plasma Glucose Concentration
Diastolic Blood Pressure

Figure 4-1: A Bayesian network for diabetes. The darker nodes indicate the Markov blanket of attribute Diastolic blood pressure.

In general, the number of subspaces increases exponentially with an increase in dimensionality; i.e. given d attributes, there are $2^d - 1$ subspaces. Since a Markov blanket provides meaningful information for any attribute, we can directly form a highly correlated subspace using the Markov blanket instead of searching through $2^d - 1$ possible subspaces.

4.4 Angle Based Similarity Measure

One of the problem we try to address in this thesis is the problem of sparse contexts in high dimensional data. For this purpose, we use the angle-based similarity measure to compare the data points in high dimensional space because standard distance metrics such as Euclidean distance and Manhattan distance become meaningless with increasing high-dimensional space causing the methods to lose their accuracy. Therefore, in high dimensions, the angles are more stable than the distances.

In this section, we present the general idea of the angle-based similarity measure [53] to score data points. Consider a point A in the data set for which we must determine the similarity measure. For that given point A, examine the cosine angle between the vectors $\overrightarrow{\mathrm{AX}}$ and $\overrightarrow{\mathrm{AY}}$ for every pair of points X, Y in the data. Then, this cosine angle is inversely weighted by the distance between the points to obtain a spectrum of angles. Then the variance in the spectrum of this angle is measured by varying the data points X and Y, while keeping the value of A fixed. The data points with the smaller variance of angles are considered as outliers.

$$ABSM(A) = VAR_{X,Y \in D}\left(\frac{(\overline{AX},\ \ \overline{AY})}{\|\overline{AX}\|^2.\ \ \|\overline{AY}\|^2}\right) \qquad \textbf{Eq. 4-6}$$

4.5 Mining Outliers in Markov Blanket Subspaces of Hybrid Bayesian Networks

Figure 4-2 shows an outline of the proposed methodology for mining outliers with the following steps:

1. Perform feature selection using Maximal Information Coefficient.
2. Construct a Hybrid Bayesian Network on the complete attribute space of the selected features.

3. Now for each attribute in the Hybrid Bayesian network, identify its Markov blanket subspaces.
4. Build a Hybrid Bayesian network for each Markov blanket subspace.
5. For pure categorical case in each Markov blanket subspace, compute the score using joint probability distributions for each instance.
6. For pure continuous case in each Markov blanket subspace, compute the outlier score for each instance using angle-based similarity measure.
7. Compute the final score of a data point by adding the scores obtained from joint probability distributions and angle-based similarity measure.
8. Identify the data points with the lowest scores and report them as outliers.

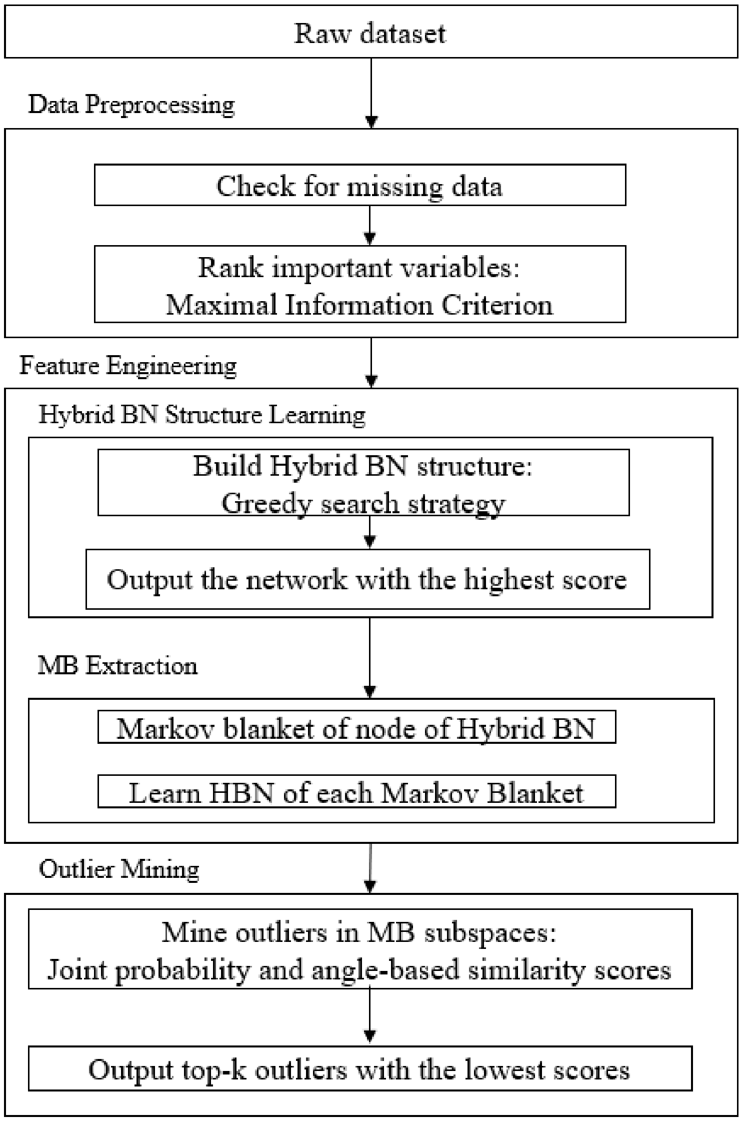

Raw dataset
Data Preprocessing
Check for missing data
Rank important variables:
Maximal Information Criterion
Feature Engineering
Hybrid BN Structure Learning
Build Hybrid BN structure:
Greedy search strategy
Output the network with the highest score
MB Extraction
Markov blanket of node of Hybrid BN
Learn HBN of each Markov Blanket
Outlier Mining
Mine outliers in MB subspaces:
Joint probability and angle-based similarity scores
Output top-k outliers with the lowest scores

Figure 4-2: Outline of the proposed methodology for outlier detection.

4.5.1 Hybrid Bayesian Networks Learning

This section presents the approaches to learn Hybrid Bayesian Networks for a Markov blanket subspace. From the discussion in Section 4.3, we know that the Markov blanket of an attribute consists of its parents, children, and its children’s other parents (spouses). Therefore, we use this definition to construct local Hybrid Bayesian networks in a Markov blanket subspace. First, we learn a Hybrid Bayesian network on full attribute space and then from that structure, we identify subspaces using Markov blankets and build a Hybrid Bayesian Network for that subspace.

4.5.1.1 *Structural Learning*

In order to learn the Hybrid Bayesian Network, we use the Deal package in R [54]. Deal can learn the graph from a data set containing both categorical and continuous attributes. To learn the network, we apply greedy search with random restarts.

Greedy search works as follows:

1. The search is started by selecting an initial DAG G_0.
2. The Bayes factors are calculated between G_0 and all the likely networks by varying one arrow at a time, that is
    a. One arrow is added to G_0.
    b. One arrow is deleted in G_0.
    c. One arrow is reversed in G_0.
3. The network with the highest Bayes factor among all networks is selected.
4. The search is stopped when the Bayes factor does not increase. Otherwise, the network G_0 is chosen and the procedure is repeated from Step 2.

We use the ratio of posterior odds for comparing the network scores of two different DAG's, G_0 and g given data D, where $P(G_0)/P(g)$ is the prior odds and $P(g|D)/P(G_0|D)$ is the Bayes factor.

$$\frac{P(D|G_0)}{P(D|g)} = \frac{P(G_0)}{P(g)} \times \frac{P(g|D)}{P(G_0|D)} \qquad \textbf{Eq. 4-7}$$

Restarts can also be used with the search algorithm by disturbing the initial network according to the parameters and then starting the search with the disturbed network. This procedure can be restarted multiple times given the restart option. Finally, after searching, a group of all the visited networks is returned. In this way, we obtain a Bayesian Network with the highest network score for the mixed data types.

4.5.1.2 *Parameter Learning*

In our work, we use maximum likelihood estimation for parameter learning in Hybrid Bayesian Networks [43]. For a data set D and a Bayesian network B, the goal of maximum likelihood estimation is to select parameters θ that satisfy the following equation:

$$L(\theta^* : D|B) = max_{\theta \in \Theta} L(\theta\colon D|B) \qquad \textbf{Eq. 4-8}$$

The parameter Θ is in the range of 0 and 1. Through the Markov condition of Bayesian networks, the likelihood $L(\Theta\colon D)$ can be stated as follows:

$$L(\theta\colon D|B) \ = \prod_i L_i(\theta_{O_i|Pa_{O_i}}\colon D|B) \qquad \textbf{Eq. 4-9}$$

where O_i is the local likelihood function which is given as

$$L(\theta_{O_i|Pa_{O_i}}\colon D) = \prod_j P(O_i^j|Pa_{O_i}^j\colon \theta_{O_i|Pa_{O_i}}) \qquad \textbf{Eq. 4-10}$$

From the data set D and Bayesian network structure B, $L(\theta\colon D|G)$ is reduced to approximating $\theta_{ijk} = P(O_i = j|Pa(O_i) = k)$, that is, the maximum likelihood estimates are simply the observed frequency estimates $\hat{\theta}_{ijk} = n_{ijk}/n_{ij}$,where n_{ijk} is the number of occurrences in the training set of the k^{th} state of O_i with the j^{th} state of its parents, and n_{ij} is the sum of n_{ijk} over all k.

4.5.1.3 *Learning Local Hybrid Bayesian Networks in MB Subspaces*

Assume a directed acyclic graph $DAG(G)$ is derived from the Markov blanket subspace of attribute G. Similarly, $DAG(Y)$ is derived from the Markov blanket subspace of attribute Y. If $Y \in PC(G)$, $G \in PC(Y)$, and an arc $Y \rightarrow G$ is in $DAG(G)$, then the arc $Y \rightarrow G$ must be in $DAG(Y)$.

Therefore, by using this strategy, the direction of the arcs between the attributes will be consistent in each local Bayesian network. This leads to a consistent joint probability distribution for each attribute.

For understanding, consider the example in **Figure 4-1**. We could generate two different Bayesian networks for attributes Diastolic Blood Pressure and Triceps Skinfold Thickness as shown in **Figure 4-3**. The direction of the arcs between Diastolic Blood Pressure and Skinfold Thickness should be constant in both the Bayesian networks.

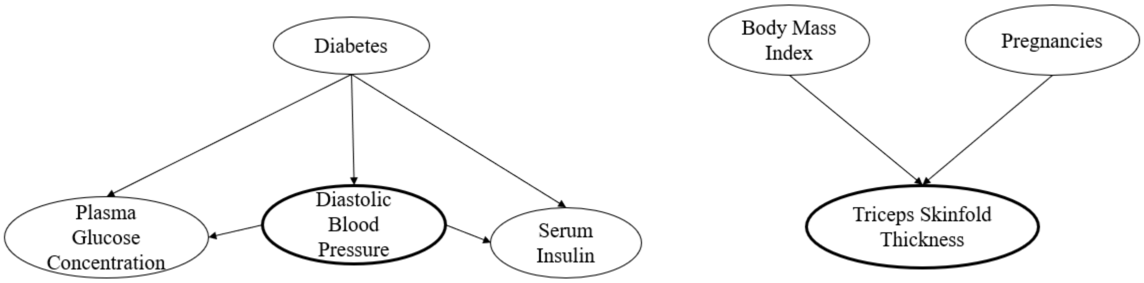


Figure 4-3: Bayesian networks in the Markov blanket subspaces.

4.5.2 Mining Outliers

In this section, a measure to discover the outliers in the Markov blanket subspaces of a Hybrid Bayesian Network is presented. Since the Hybrid Bayesian Network captures the contextual beliefs of the data in a probabilistic manner, it is natural to use joint probability distributions as a measure to detect the outliers. Furthermore, to overcome the problem of sparse contexts, we employ an angle-based similarity measure. Therefore, the overall score for a data point is formed by combining its joint probabilities and angle-based similarity measure.

As the method utilizes a Hybrid Bayesian Network to capture causal relations in a mixed attribute data set, there exist three types of relationships among the attributes which are described below.

Pure categorical case: A subspace where categorical parent nodes are conditioned on categorical child nodes. The symbol $CS_{\blacklozenge}$ is used to represent a set of causal subspaces involving pure categorical case. The notation $|CS_{\blacklozenge}|$ represents the overall number of categorical subspace.

Pure continuous case: A subspace where continuous parent nodes are conditioned on continuous child nodes. The symbol CS_{τ} is used to represent a set of causal subspaces involving pure continuous case. The notation $|CS_{\tau}|$ represents the overall number of continuous subspaces.

Mix of categorical and continuous case: A subspace where categorical and continuous parents are conditioned on a continuous child node. The symbol CS_{λ} is used to represent a set of causal subspaces involving a mixed case. The notation $|CS_{\lambda}|$ represents the overall number of mixed attribute subspaces.

In a Hybrid Bayesian Network, for each Markov blanket subspace with pure categorical case $CS_i \in CS_{\blacklozenge}$, the score of CS_i is formed using **Eq. 4-11**. This score is calculated by multiplying posterior probability with the prior probability:

$$Score_{\blacklozenge}(CS_i)_{(i \in CS_{\blacklozenge})} = P(C|Pa(C)) \times P(Pa(C)) \qquad \textbf{Eq. 4-11}$$

Therefore, the final score of each point n in all the categorical subspaces is calculated by the following **Eq. 4-12**.

$$Score_{\blacklozenge}(n_{CS_{\blacklozenge}}) = \sum_{i=1}^{|CS_{\blacklozenge}|} Score_{\blacklozenge}(CS_i) \qquad \textbf{Eq. 4-12}$$

For causal subspace with pure continuous case $CS_j \in CS_\tau$, the concept of angle-based similarity is used which is represented by **Eq. 4-13**. This method works by taking each observation and computing cosine similarities between all pairs of points. Observations with the smallest variance of these similarities are the outliers.

$$Score_\tau(CS_j)_{(j\in CS_\tau)} = VAR_{x,y\,\in\,D}\left(\frac{(\overline{n_j x},\ \overline{n_j y})}{\left\|\overline{n_j x}\right\|^2\left\|\overline{n_j y}\right\|^2}\right) \qquad \textbf{Eq. 4-13}$$

Therefore, the final score of each point n in all the continuous subspaces is calculated by the following **Eq. 4-14**:

$$Score_\tau(n_{CS_\tau}) = \sum_{j=1}^{|CS_\tau|} Score_\tau(CS_j) \qquad \textbf{Eq. 4-14}$$

For causal subspaces with mixed attribute case, $CS_k \in CS_\lambda$, we proceed to use the angle-based similarity measure to compute the score for each data point. For this purpose, the categorical attributes are converted to binary values using one-hot encoding and the continuous attributes are normalized in the range of 0 and 1. The score of mixed attribute case is represented in **Eq. 4-15**:

$$Score_\lambda(n_{CS_\lambda}) = \sum_{k=1}^{|CS_\lambda|} Score_\lambda(CS_k) \qquad \textbf{Eq. 4-15}$$

Therefore, the complete score of a point n in a data set is calculated by adding scores from **Eq. 4-12**, **Eq. 4-12** and **Eq. 4-15** as represented by **Eq. 4-16**.

$$Score(n) = Score_{\blacklozenge}(n_{CS_{\blacklozenge}}) + Score_{\tau}(n_{CS_{\tau}}) + Score_{\lambda}(n_{CS_{\lambda}}) \qquad \textbf{Eq. 4-16}$$

With $Score(n)$, we sort the top-k data points with the lowest scores as possible outliers.

4.5.2.1 *Algorithm*

Input: A data set D

Output: Top-n low scoring data points

```
Feature selection using Maximal Information Coefficient
Learn a Hybrid Bayesian Network on full attribute space of selected features
// Identify Markov blanket (MB) subspaces from the full HBN
for $i = 1$ to h do
    a. $\mathrm{MB(i)} = \mathrm{MB}(\mathrm{D_i})$
// Build a Hybrid Bayesian Network in each Markov blanket subspace
    for $i = 1$ to h do
        a. Learn the structure of $\mathrm{HBN_i}$ on MB(i) using the greedy search
           strategy
        b. Learn parameters for $\mathrm{HBN_i}$
    end for
end for
// Outlier detection over various Hybrid Bayesian Networks
// Assume n data points
for $i = 1$ to n do
    for $q = 1$ to h do
        if $(CS_i \in CS_{\blacklozenge}\ in\ HBN_q)$ then
            Compute $Score_{\blacklozenge}(n_{CS_{\blacklozenge}})$ using Eq. 4-12
        **else**
            **if** $(CS_i \in CS_\tau\ in\ HBN_q)$ **then**
                Compute $Score_\tau(n_{CS_\tau})$ using **Eq. 4-15**
            **else**
                **if** $(CS_i \in CS_\lambda\ in\ HBN_q)$ **then**
                    Compute $Score_\lambda(n_{CS_\lambda})$ using **Eq. 4-16**
                **end if**
            **end if**
        **end if**
    **end for**
    Compute $Score(n)$ using **Eq. 4-16**
**end for**
Report the data points with the lowest scores as the outliers
```

4.6 Complexity Analysis

The computational complexity of the algorithm is dependent on four aspects, i.e. size of the dataset, subspace feature selection, inference in Hybrid Bayesian Network, and angle-based similarity score. The exact inference in Bayesian Network requires exponential time in the worst case since it is an NP-hard problem. In the case of subspace selection, the average time complexity is $\Theta(d \times n^2)$, where d is the number of subspaces and n is number of data points. This is due to traversing each subspace and calculating the outlier scores for n points. The time complexity for probabilistic inferring using maximum likelihood estimation in d subspaces is $\Theta(c \times d \times n)$ where c is the number of classes. The time complexity for angle-based similarity measure is $\Theta(n^2)$.

Since we repeat this calculation for d subspaces, the time complexity is $\Theta(d \times n^2)$. Therefore, the overall time complexity is $\Theta(c \times d \times n + d \times n^2) \cong \Theta(n^2)$.

4.7 Data set

To evaluate our algorithm, we use real world mixed data set with continuous and categorical attributes. For this purpose, we chose KSL data set which is described below.

4.7.1 KSL Danish Elderly Data Set

The KSL data set, taken from Deal package [54], is from a study measuring health and social characteristics of representative samples of Danish 70-year old people, taken in 1967 and 1984. The data has 300 observations, and each observation has 9 attributes. Description of the variables of the data set has been provided in **Table A-1**. The variables FEV, Kol and BMI are continuous attributes and the rest are categorical attributes. The attribute hypertension is the class label with two possible outcomes--yes or no. The number of people without hypertension are 136 while people with hypertension are 164. This data set is called KSL data for the rest of our work.

Table 4-1: Description of dimensions of the data sets used.

Data Set Name	Number of observations	Number of attributes
KSL Data set	300	9

CHAPTER 5

RESULTS & DISCUSSION

In this chapter, we describe the experimental environment used to evaluate our algorithm and the results obtained.

We use a baseline subspace outlier mining algorithm called Subspace Outlier Detection (SOD) [55] and a full attribute space density-based algorithm called Local Outlier Factor (LOF) [8] to compare the performance of our algorithm. SOD uses the shared nearest neighbors to evaluate the similarity among observations and a subspace set is selected based on similarity measures. LOF determines the outlier score by calculating the ratio between the density of a point to the density of its k nearest neighbors.

5.1 Evaluation Metrics Used

The following performance evaluation metrics are used to compare the performance of the proposed algorithm. These metrics are precision or positive predicted values, recall or sensitivity, F-measure, and ROC curves, and are defined below.

5.1.1 Precision

Precision is defined as the fraction of true positives to the sum of true positives and false positives [56].

$$Precision = \frac{TP}{TP + FP} \qquad \textbf{Eq. 5-1}$$

True positive (TP) denotes the subjects with positive class label correctly identified as positive.

True negative (TN) denotes the subjects with negative class label correctly identified as negative.

False Positive (FP) denotes the subjects with a negative class label falsely identified as positive.

False Negative (FN) denotes the subjects with a positive class label falsely identified as negative.

5.1.2 Recall

It is also called as the true positive rate or the sensitivity. It denotes the proportion of positive class labels identified as positive:

$$Recall = \frac{TP}{TP + FN} \qquad \textbf{Eq. 5-2}$$

5.1.3 F-Measure

F-measure is defined as the weighted average of precision and recall. Therefore, this score takes both the precision and recall into significance.

$$F - measure = \frac{2\ \times (Precision \times Recall)}{(Precision + Recall)} \qquad \textbf{Eq. 5-3}$$

5.1.4 ROC Curves

The area under the ROC curve helps us to evaluate the discriminative power of a test. It is the representation of the graph between sensitivity and specificity. The model will have better accuracy if the area under the curve is larger. The range of AUC lie between 0 and 1 and represents the quality of the test. Consequently, the higher AUC value indicates the better accuracy of the model.

5.2 Experimental Results

This section presents the results obtained by applying our algorithm on the KSL data set obtained from the Deal package.

The KSL data has three continuous attributes and six categorical attributes. The continuous attributes follow normal distribution and any missing records are removed from the data set. Furthermore, we do not discretize continuous data due to loss of information.

5.2.1 Feature Selection using Maximal Information Coefficient

We apply the Maximal Information Coefficient on full attribute space in the KSL data set. **Figure 5-1** shows the comparison of total MIC scores between the target attribute and other attributes in the data set. We can see that the attributes BMI, Kol, and FEV are highly correlated with target attribute Hyp. Therefore, these attributes are selected for outlier detection. The other attributes Work, Smok, Sex and Alc are slightly less correlated with Hyp, but we consider them as they may provide additional knowledge on the outlier behavior. The attribute Year is least correlated with Hyp, and is thus not considered for outlier detection.

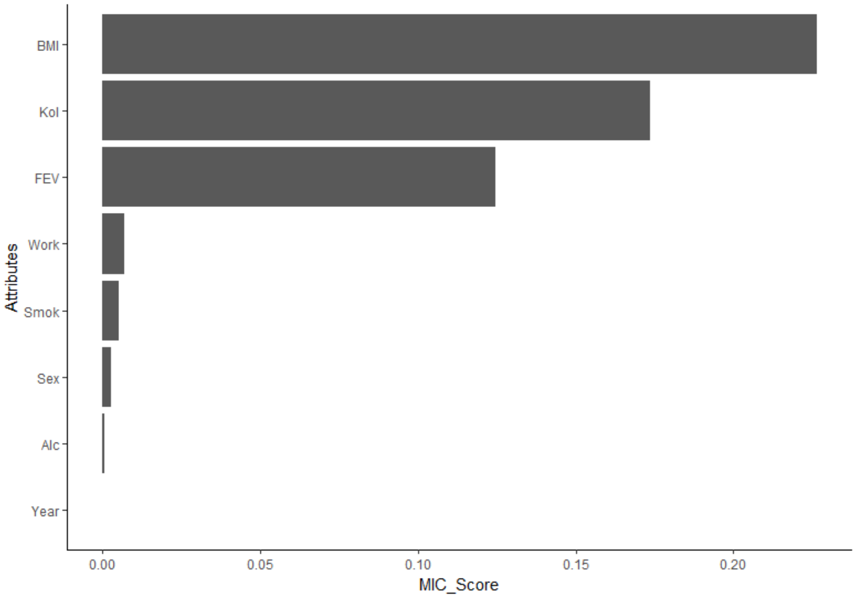


Figure 5-1: Maximal Information Coefficient of KSL data set

5.2.2 Hybrid Bayesian Network on Full Attribute Space

In **Figure 5-2**, we show Hybrid Bayesian Network learned over the KSL data set by taking all the attributes. The names of the attributes represented in the Hybrid Bayesian Networks are the same as the ones given in the data set. In **Table A-1**, we present the description of these attributes.

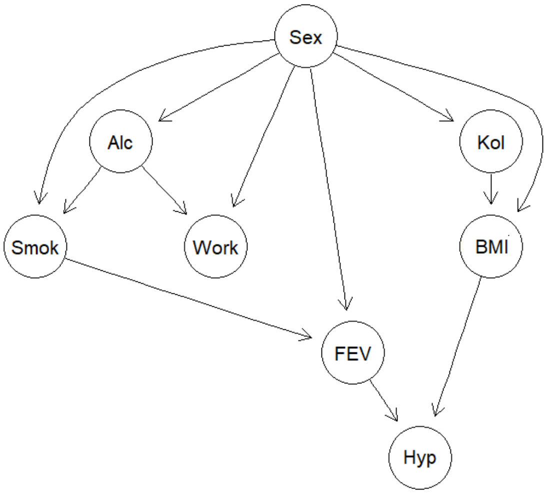


Figure 5-2: Hybrid Bayesian Network on full attribute space of KSL data set

5.2.3 Hybrid Bayesian Networks in Markov Blanket Subspaces

Figure 5-3 to **Figure 5-10** represents the Hybrid Bayesian Networks learned for each of the Markov blanket subspaces obtained from the KSL data set.

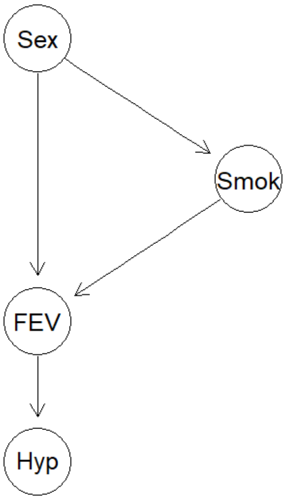


Figure 5-3: Hybrid Bayesian Network on Markov blanket subspace of FEV.

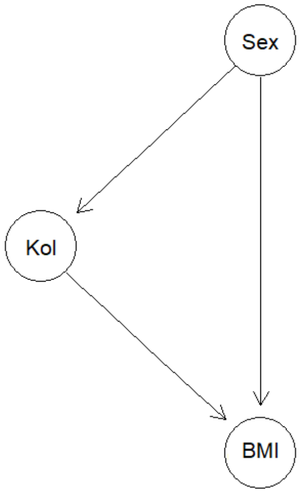


Figure 5-4: Hybrid Bayesian Network on Markov blanket subspace of Kol.

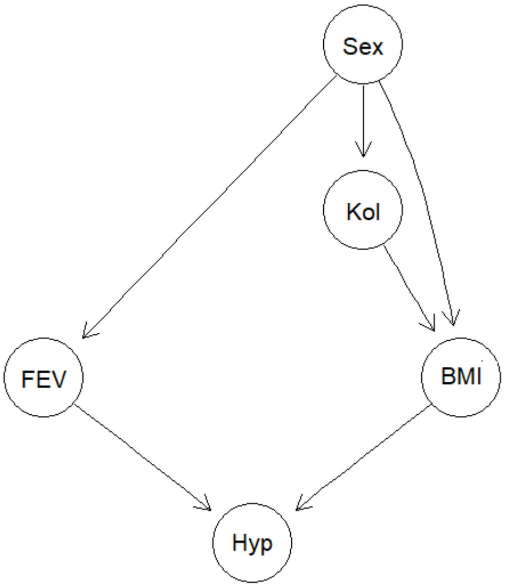


Figure 5-5: Hybrid Bayesian Network on Markov blanket subspace of BMI.

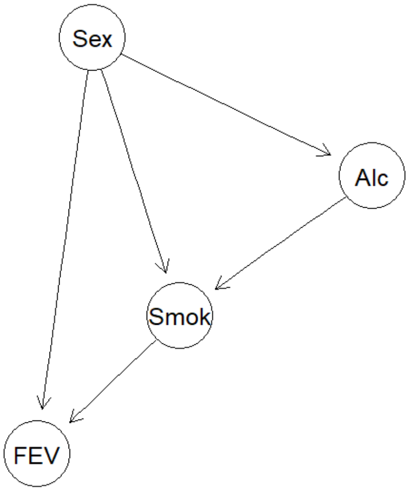


Figure 5-6: Hybrid Bayesian Network on Markov blanket subspace of Smok.

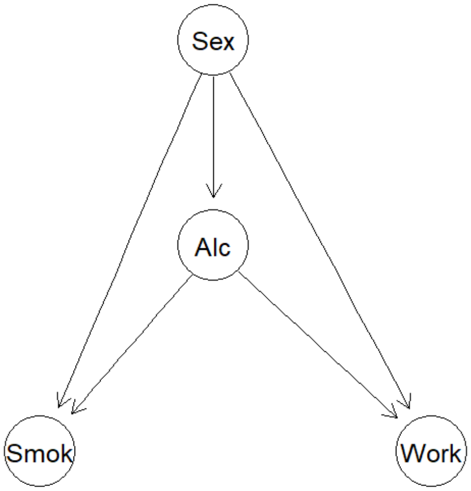


Figure 5-7: Hybrid Bayesian Network on Markov blanket subspace of Alc.

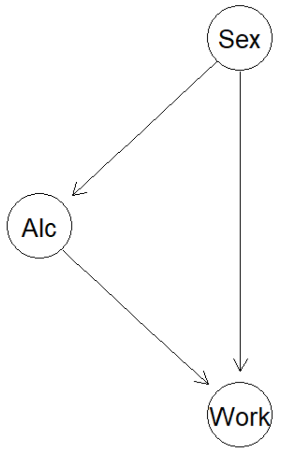


Figure 5-8: Hybrid Bayesian Network on Markov blanket subspace of Work.

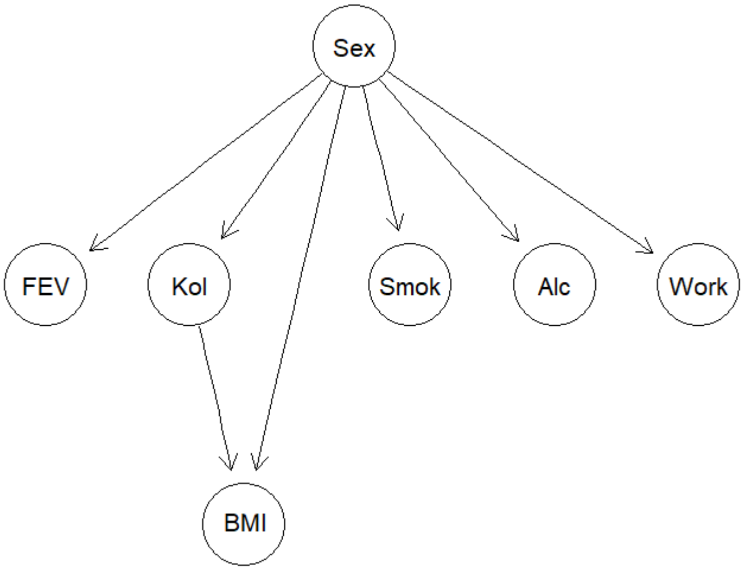


Figure 5-9: Hybrid Bayesian Network on Markov blanket subspace of Sex.

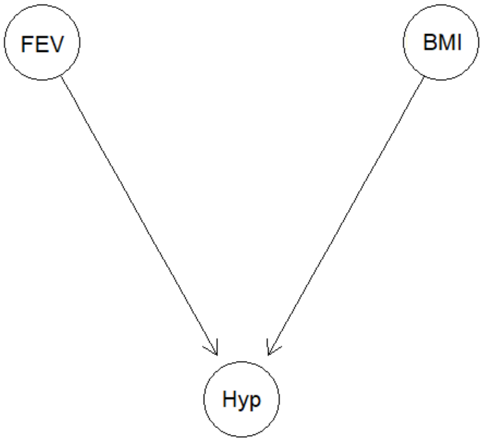


Figure 5-10: Hybrid Bayesian Network on Markov blanket subspace of Hyp.

5.2.4 Analysis of the Results

In this section, we investigate the results gained for the KSL data set. Considering the full attribute space, only top-10 data points with lowest outlier scores are identified as possible outliers using the proposed method. However, we discover 44 outlier points including the ones discovered in full attribute space after searching through different Markov blanket subspaces. The **Table 5-1** shows some of the discovered outliers in full attribute space and Markov blanket subspaces.

Table 5-1: Outliers discovered in the KSL dataset.

Data Point	Is Outlier in Full Attribute Space	Outlier in Markov Blanket Subspace
7	Yes	FEV, BMI, Smok, Sex
11	No	Alc, Work, Sex
20	Yes	Alc, Work, Sex
39	No	FEV
96	Yes	FEV, Sex
97	Yes	Kol, BMI, Sex
242	Yes	FEV, Hyp, BMI, Smok

Furthermore, we observe interesting patterns such as data point 11 is reported as outlier in the Markov blanket subspaces of Alc, Work and Sex. Interestingly, the subspaces FEV, Kol, BMI and Smok do not report the data point 11 as an outlier. This shows a uniqueness in these subspaces because the same data point is reported as outliers in only some subspaces. Note also that this data point is not identified as an outlier in the full attribute space. It only shows outlying behavior within specific subspaces.

Moreover, the following **Table 5-2** represents the relevance and quality of the results by discussing an outlier instance discovered by this approach. For discovered outlier 7, the person has high FEV, high cholesterol levels, normal BMI and is a non-smoker, non-alcoholic, not-working, but has hypertension. For discovered outlier 11, the person has high FEV, high cholesterol levels, low BMI and is a smoker, alcoholic, working, but has no hypertension. Similarly, for discovered outlier 20, the person has high FEV, normal BMI and is a non-smoker, non-alcoholic, but has hypertension and high cholesterol levels.

Table 5-2: Outlier analysis on the KSL dataset using proposed method.

Data Point	Outlier Characterization	Outlier Score
7	FEV = 314, Kol = 755, BMI = 21.91, Smok = no, Alc = no, Work = no, Hyp = yes	2.49
11	FEV = 311, Kol = 719, BMI = 17.9, Smok = yes, Alc = yes, Work = yes, Hyp = no	2.64
20	FEV = 213, Kol = 797, BMI = 18.27, Smok = no, Alc = no, Work = yes, Hyp = yes	2.91
39	FEV = 295, Kol = 891, BMI = 23.6, Smok = no, Alc = yes, Work = no, Hyp = yes	3.12
96	FEV = 304, Kol = 810, BMI = 30.12, Smok = yes, Alc = no, Work = no, Hyp = no	3.16
97	FEV = 227, Kol = 799, BMI = 46.87, Smok = no, Alc = no, Work = no, Hyp = yes	3.31
242	FEV = 38, Kol = 385, BMI = 30, Smok = no, Alc = no, Work = no, Hyp = yes	3.55

We compared the results of our method with results of Local Outlier Factor and Subspace Outlier Detection methods. The following **Table 5-3** shows the analysis of the results for LOF and **Table 5-4** shows the results for SOD. We can observe that the detected points are not interesting as they represent the already known knowledge. This is due to the fact that LOF uses densities to compute outliers in the nearest neighbors, whereas SOD uses the shared nearest neighbor approach.

Table 5-3: Outlier analysis on the KSL dataset using LOF.

Data Point	Outlier Characterization
17	FEV = 155, Kol = 656, BMI = 21.83, Smok = no, Alc = no, Work = no, Hyp = no
124	FEV = 201, Kol = 539, BMI = 22.72, Smok = no, Alc = no, Work = no, Hyp = no
141	FEV = 133, Kol = 759, BMI = 33.67, Smok = no, Alc = yes, Work = yes, Hyp = yes
142	FEV = 163, Kol = 717, BMI = 22.15, Smok = yes, Alc = yes, Work = yes, Hyp = yes
162	FEV = 252, Kol = 675, BMI = 23.66, Smok = yes, Alc = yes, Work = no, Hyp = no
227	FEV = 176, Kol = 643, BMI = 26.86, Smok = yes, Alc = yes, Work = no, Hyp = yes
291	FEV = 136, Kol = 850, BMI = 24.24, Smok = yes, Alc = yes, Work = yes, Hyp = yes

Table 5-4: Outlier analysis on the KSL dataset using SOD.

Data Point	Outlier Characterization
30	FEV = 220, Kol = 348, BMI = 23.71, Smok = no, Alc = no, Work = no, Hyp = no
27	FEV = 250, Kol = 141, BMI = 23.71, Smok = yes, Alc = no, Work = no, Hyp = no
49	FEV = 109, Kol = 896, BMI = 27.48, Smok = yes, Alc = no, Work = no, Hyp = yes
82	FEV = 287, Kol = 347, BMI = 24.38, Smok = yes, Alc = no, Work = no, Hyp = no
171	FEV = 75, Kol = 840, BMI = 29.78, Smok = yes, Alc = no, Work = yes, Hyp = yes
235	FEV = 140, Kol = 697, BMI = 27.89, Smok = yes, Alc = yes, Work = no, Hyp = yes
287	FEV = 225, Kol = 413, BMI = 19.63, Smok = yes, Alc = no, Work = yes, Hyp = no

5.2.5 Evaluation of the Proposed Model

Furthermore, we remove the outliers which were discovered using our model, LOF and SOD, and evaluate the classification accuracy. We use Logistic Regression as the classifier. The data set is divided into train and test sets with 70% for training and 30% for testing the model.

Figure 5-11 and demonstrates the results obtained for the KSL dataset. The definition of outlier aspect chosen to search for outliers in the Local Outlier Factor (LOF) [8]. The number of nearest neighbors chosen for LOF and SOD is 5. To get even results, we performed LOF with the above chosen Markov blanket subspaces. As is evident from the figure, the ROC curve is the higher our model, with AUC value 0.731, whereas the AUC for LOF method is 0.612. The precision and recall obtained for our model is 0.727 and 0.711, respectively, compared to 0.576 and 0.483 obtained for LOF. Furthermore, the proposed approach performs well compared to Subspace Outlier Detection method as is evident from **Figure 5-12**.

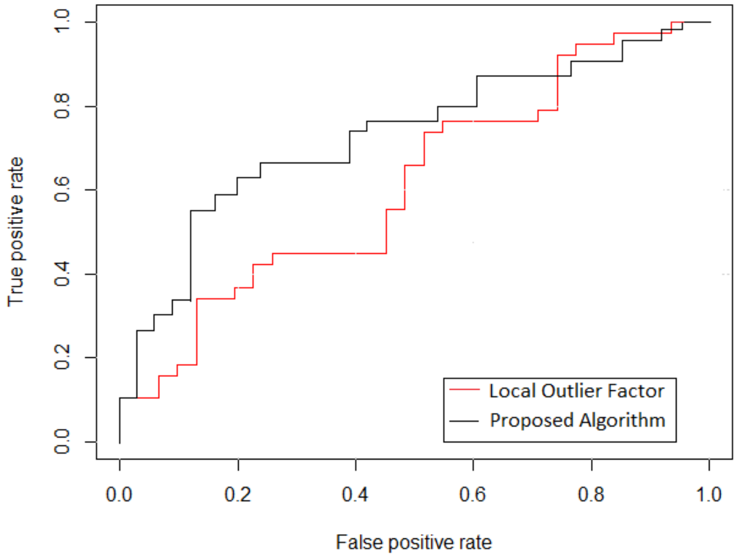


Figure 5-11: ROC curve of proposed algorithm against LOF.

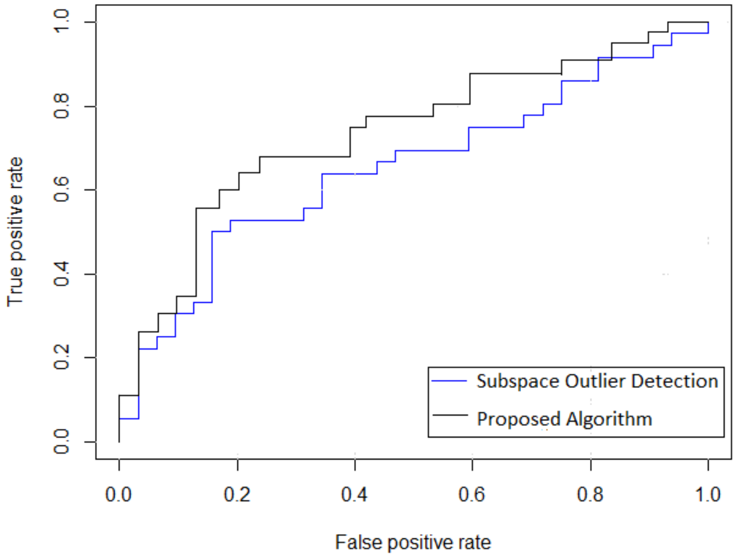


Figure 5-12: ROC curve of proposed algorithm against SOD.

Table 5-5: Summary of the results obtained for the KSL dataset.

Evaluation Metric	**Proposed Approach**	**LOF**	**SOD**
Precision	0.727	0.576	0.586
Recall	0.711	0.483	0.531
F-Measure	0.718	0.526	0.557
AUC	0.731	0.612	0.652

5.2.6 Visualization of the Reported Outliers

In **Figure 5-13** and **Figure 5-14**, we present two-dimensional visualization of data points in the causal subspace of FEV, Kol and BMI for the KSL data set. For the subspace of FEV and Kol, the already established contextual belief is, when the values of FEV are high the cholesterol levels must be low and vice-versa. For the subspace of Kol and BMI, the already established contextual belief is, when the values of Kol are high the BMI levels must be high and vice-versa. Therefore, from the scatter plots, we can see that, our method identifies true outliers which violate the above contextual beliefs and are also sparse; i.e. they are far away from their neighbors. Contrary to our method, the SOD technique and LOF approaches failed to discover these data points even though most of them are away from their nearest neighbors.

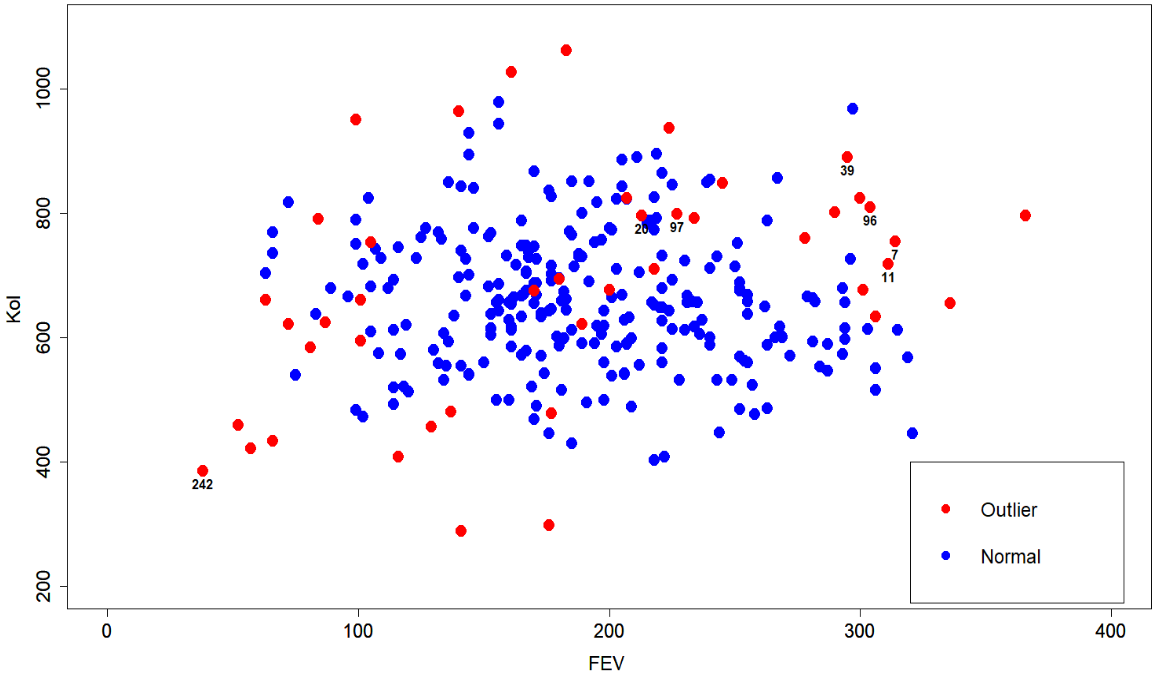


Figure 5-13: 2D visualization of causal subspace of FEV and Kol in the KSL data set.

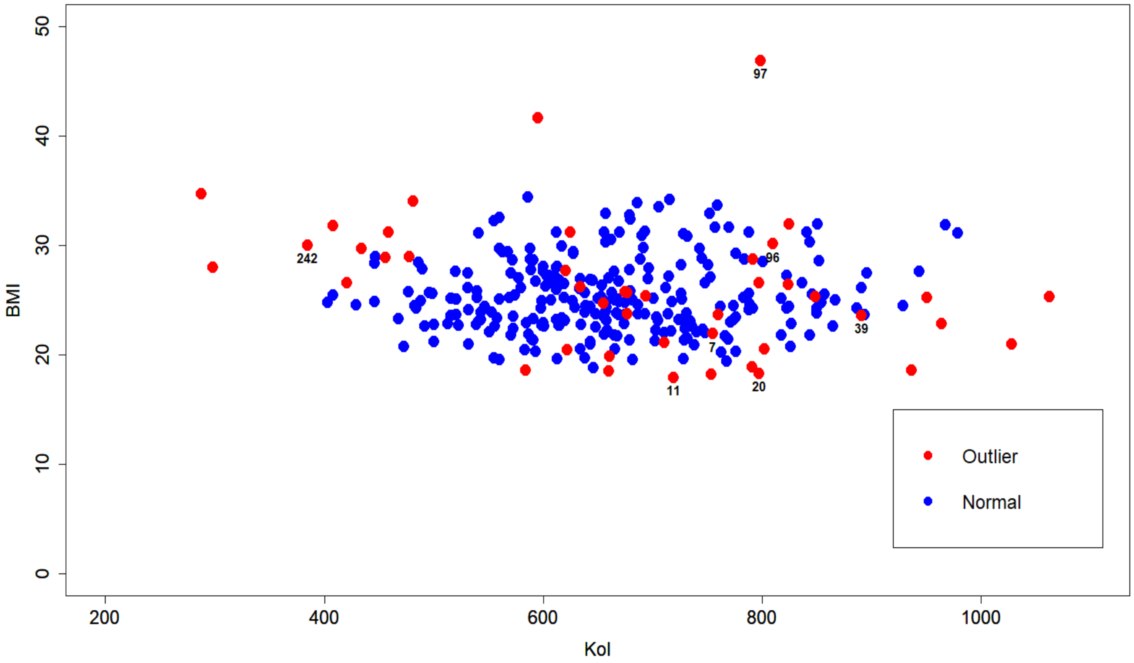


Figure 5-14: 2D visualization of causal subspace of Kol and BMI in the KSL data set.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

In this thesis, we hypothesize that true and meaningful outliers are likely to be identified in highly correlated feature subspaces by considering both contextual beliefs and similarity of the data points. In this regard, we propose a comprehensive approach to exploit the underlying contextual beliefs for detecting outliers, particularly dealing with the existing issue caused by the sparsity of contexts in high dimensional data. Specifically, we introduce Hybrid Bayesian Networks to capture the contextual beliefs and angle-based similarity measure to tackle sparse contexts and describe an algorithm to fuse them. Experimental results show that our approach detects outliers more accurately and efficiently than previous methods.

6.2 Future Work

The approaches used in this thesis are designed for static mixed attribute data sets. However, we can extend this methodology to detect outliers in streaming data. Due to transient nature of the streaming data, the complexity of both inference and representation grow multi-fold. This aspect of handling time series data for multiple data types has not been addressed in this research work and left as an area of further research.

APPENDIX A

A.1 KSL data set description

Table A-1: Description of attribtes in the KSL data set.

Variable Name	Description	Type
FEV	Forced ejection volume of person’s lung	Continuous
Kol	Cholesterol level	Continuous
BMI	Body Mass Index	Continuous
Smok	Smoking 1 = no, 2 = yes	Categorical
Alc	Alcohol consumption 1 = no, 2 = yes	Categorical
Work	Working 1 = yes, 2 = no	Categorical
Sex	Gender 1 = male, 2 = female	Categorical
Year	Survey Year 1 = 1967, 2 = 1984	Categorical
Hyp	Hypertension 0 = no, 1 = yes	Categorical

A.2 Summary of notations

Table A-2: Notations.

Notation	Description
n	A data point
C	Child node
Pa(C)	Parent of the child node
PC(X)	Parent-Child of node X
CS	Causal Subspaces in Hybrid Bayesian Network
|CS|	Total number of causal subspaces in Hybrid Bayesian Network
$CS_{\blacklozenge}$	Subspaces involving only categorical attributes
CS_{τ}	Subspaces involving only continuous attributes
CS_{λ}	Subspaces involving mixed attributes
MB	Markov Blanket
HBN	Hybrid Bayesian Network

BIBLIOGRAPHY

[1] A.K. Patnaik, B. Nayak and S. Prasad, “Data mining and its current research directions.”

[2] D. Hawkins, Identification of outliers, Chapman and Hall: London, 1980.

[3] J. R. Dorronsoro, F. Ginel, C. Sanchez, and C. S. Cruz, "Neural fraud detection in credit card operations," *IEEE Transactions on Neural Networks*, vol. 8, no. 4, pp. 827–834, 1997.

[4] U. S. Kameswari, and I. R. Babu, “Sensor data analysis and anomaly detection using predictive analytics for process industries,” *IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions*, 2015.

[5] W. Wong, A. Moore, G. Cooper, and M. Wagner, "Bayesian network anomaly pattern detection for disease outbreaks," in *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[6] D. Barbara, N. Wu, and S. Jajodia, "Detecting novel network intrusions using bayes estimators," in *Proceedings of SIAM International Conference on Data Mining*, 2001.

[7] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The VLDB Journal*, vol. 8, pp. 237–253, 2000.

[8] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2000.

[9] S. Babbar, and S. Chawla, "On bayesian network and outlier detection," in *Proceedings of the 16th International Conference on Management of Data*, 2010.

[10] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 5, pp. 631–645, 2007.

[11] C. Aggarwal, Outlier analysis, Springer, 2013.

[12] A. C. Sanchez, J. Aracil, and J. R. Santiago, "Proposal of a new information-theory based technique and analysis of traffic anomaly detection," *International Conference on Smart Communication in Network Technologies, 2014*.

[13] W. Lee, and D. Xiang, "Information theoretic measures for anomaly detection," *IEEE Symposium on Security and Privacy, 2001*.

[14] M. Ester, H. P. Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[15] A. L. M. Chiu, and A. W. C, Fu, "Enhancements on local outlier detection," in *Proceedings of the Seventh International Database Engineering and Applications Symposium, 2003*.

[16] J. Tang, Z. Chen, A. W. C. Fu, and D. W. L. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2002.

[17] C. Varun, B. Arindam, and K. Vipin, "Anomaly detection: A survey," *ACM Computing Survey*, vol. 41, no. 3, pp. 1-58, 2009.

[18] P. N. Tan, M. Steinbach, and V. Kumar, Introduction to data mining, 1st ed., Boston, MA: Addison-Wesley Longman Publishing Co. Inc., 2005.

[19] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," in *Proceedings of the 8th SIAM International Conference on Data Mining*, 2008.

[20] E. M. Knorr, and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proceedings of the 24rd International Conference on Very Large Data Bases*, 1998.

[21] T. T. Dang, H. Y. T. Ngan, and W. Liu, "Distance-based k-nearest neighbors outlier detection method in large-scale traffic data," *IEEE International Conference on Digital Signal Processing*, 2015.

[22] H. M. Shirazi, "Anomaly intrusion detection system using information theory, K-NN and KMC algorithms," *Australian Journal of Basic and Applied Sciences*, vol. 3, no. 3, pp. 2581-2597, 2009.

[23] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 5, pp. 631–645, 2007.

[24] A. M. Hayes, A. M. M. Capretz, "Contextual anomaly detection in big sensor data," *IEEE International Congress on Big* Data, 2014.

[25] Y. Kou, C. Lu, and D. Chen, "Spatial weighted outlier detection," *in Proceedings of SIAM Conference on Data Mining*, 2006.

[26] G. Sadhana, A. Muralidhar, and M. Swamynathan, "Context based anomaly detection in images," *International Journal of Control Theory and Applications*, vol. 9, no. 51, pp. 165-172, 2016.

[27] L. Wei, W. Qian, A. Zhou, W. Jin, and J. X. Yu, "HOT: Hypergraph based outlier test for categorical data," *in Proceedings of the 7th Pacific-Asia Conference on Advances in knowledge Discovery and Data Mining*, 2003.

[28] M. Valko, H. Valizadegan, B. Kveton, G. Cooper, and M. Hauskrecht, "Conditional anomaly detection with soft harmonic functions," *IEEE 11th International Conference on Data Mining*, 2011.

[29] X. Wang, and I. Davidson, "Discovering contexts and contextual outliers using random walks in graphs," *IEEE 9th International Conference on Data Mining*, 2009.

[30] S. Babbar, D. Surian, and S. Chawla, "A Causal Approach for Mining Interesting Anomalies," *in Proceedings of the 26th Canadian Conference on Artificial Intelligence*, 2013.

[31] A. H. Nicholas, D. J. Weston, K. Platanioti and D. J. Hand, "Bayesian anomaly detection methods for social networks," *The Annals of Applied Statistics*, vol. 4, no. 2, pp. 645-662, 2010.

[32] L. Rashidi, S. Hashemi, and A. Hamzeh, "Anomaly detection in categorical datasets using bayesian networks," *International Conference on Artificial Intelligence and Computational Intelligence*, 2011.

[33] A. Masood, and W. Li, " Interestingness filtering engine: Mining Bayesian networks for interesting patterns," *SoutheastCon*, 2015.

[34] R. Malhas, and Z. A. Aghbari, "Finding interesting outliers - a belief network-based approach," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5137-5145, 2009.

[35] C. C. Aggarwal, and P. S. Yu, "Outlier detection for high dimensional data," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2001.

[36] E. M. Knorr, and T. Ng, "Finding intentional knowledge of distance-based outliers," *Proceedings of 25th International Conference on Very Large Data Bases*, 1999.

[37] F. Keller, E. Muller, and K. Bohm, "HICS: High contrast subspaces for density-based outlier ranking," in *IEEE 28th International Conference on Data Engineering*, 2012.

[38] A. Lazarevic, and V. Kumar, "Feature bagging for outlier detection," in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge discovery in data mining*, 2005.

[39] H. P. Kriegel, P. Kroger, E. Schubert, and A. Zimek, "Outlier detection in arbitrarily oriented subspaces," in *Proceedings of the 12th IEEE International Conference on Data Mining*, 2012.

[40] S. Chebrolu, A. Abraham, and P. Thomas, "Feature deduction and ensemble design of intrusion detection systems," *Computers & Security*, vol. 24, no. 4, pp. 295-307, 2005.

[41] X. Wang, T. Lin, and J. Wong, "Feature selection in intrusion detection system over mobile ad-hoc network," *Computer Science Technical Reports*, 2005.

[42] L. Duan, G. Tang, J. Pei, J. Bailey, G. Dong, X. V. Nguyen, A. Campbell, C. Tang, "Efficient discovery of contrast subspaces for object explanation and characterization," *Knowledge on Information Systems*, vol. 47, no. 1, pp. 99-129, 2016.

[43] V. Joshi, and R. Bhatnagar, "Outlier analysis using lattice of contiguous subspaces," *Knowledge on Information Systems*, vol. 47, no. 1, pp. 99-129, 2016.

[44] D. Koller and N. Friedman, Probabilistic graphical models: Principles and techniques, The MIT Press, 2009.

[45] J. Pearl, Probabilistic reasoning in intelligent systems: Networks of plausible inference, Morgan Kaufmann Publishers Inc., 1988.

[46] T. K. Laga, "The Markov blanket concept in Bayesian networks and dynamic Bayesian networks and convergence assessment in graphical model selection problems," 2008.

[47] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Machine Learning*, vol. 65, pp. 31–78, 2006.

[48] Y. Zhang, Z. Zhang, K. Liu, and G. Qian, "An improved IAMB algorithm for markov blanket discovery," *Journal of Computers*, vol. 5, no. 11, pp. 2011.

[49] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian Networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, pp. 197-243, 1995.

[50] G. F. Cooper, and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, pp. 309-347, 1992.

[51] S. Lauritzen, “Propagation of probabilities, means, and variances in mixed graphical association models,” *Journal of the American Statistical Association,* vol. 87, no. 420, pp. 1098-1108, 1992.

[52] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, “Detecting novel associations in large datasets,” *Science*, vol. 334, no. 6062, pp. 1518-1524, 2011.

[53] H. P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," *in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data* Mining, 2008.

[54] S. G. Bottcher, and C. Dethlefsen, "Deal: A package for learning Bayesian networks," URL http://cran.r-project.org/.

[55] H. P. Kriegel, P. Kroger, E. Schubert, and A. Zimek, “Outlier detection in axis-parallel subspaces of high dimensional data,” *13th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, 2009.

[56] A. M. Šimundić, “Measures of diagnostic accuracy: Basic definitions.,” *The Electronic Journal of the International Federation of Clinical Chemistry and Laboratory Medicine*, vol. 19, no. 4, pp. 203–11, 2009.