Ontology-based Constrained Anonymization for Domain-Driven Data Mining Outsourcing

Brian Loh Chung Shiong

School of Engineering, Computing and Science
Swinburne University of Technology
Kuching, Sarawak, Malaysia

Submitted for the degree of Master of Science

2012

Dedicated to my family and friends

Abstract

Introduction. This thesis focuses on the data mining outsourcing scenario whereby a data owner publishes data to an application service provider who returns mining results. To ensure data privacy against an un-trusted party, protection techniques are required. Anonymization, a widely used method provides the benefit of preserving true attribute values as well as the capability of supporting various data mining algorithms. Although this is so, several issues emerge when anonymization is applied in a real world outsourcing scenario. Most methods have focused on the traditional data mining paradigm, therefore they do not implement domain knowledge nor optimize data for domain-driven purposes. Furthermore, existing techniques limit users' control while assuming their natural capability of producing Domain Generalization Hierarchies (DGH). Moreover, previous utility metrics have not considered attribute correlations during generalization.

Objective. The research objective is to create an ontology-based constrained anonymization framework which aims to preserve meaningful and actionable models for domain-driven data mining while protecting privacy.

Framework. In contrast with existing works, this framework integrates the Unified Medical Language Systems (UMLS) as a form of domain ontology knowledge during DGH creation to preserve value meanings. Furthermore, it allows for user constraints based on attribute semantic types and relations to suit physician mining tasks. Also, attribute correlations are determined with external domain knowledge in the form of MEDLINE literatures to improve attribute selection during anonymization.

Results. Experiments show that ontology-based DGHs manage to preserve semantic meaning after attribute generalization. Additionally, by setting constraints, important attributes for specific mining tasks can be preserved. Finally, utilizing a correlation-based measure can improve attribute selection during anonymization for domain-driven purposes.

Conclusion. There is an urgent need for privacy preserving methods capable of anonymizing data for domain-driven usage. The proposed framework proves

the benefit of integrating domain ontology knowledge and external literatures in improving utility for domain-driven purposes. Therefore, it is expected that by utilizing such a framework, data owners can protect data while maintaining utility for real world requirements.

Acknowledgements

I would like to thank my supervisor, Dr. Patrick Then Hang Hui, for his support and friendship throughout the years. Thanks goes to my co-supervisor, Associate Professor Enn Ong as well.

I am grateful for my fellow friends, Yakub Sebastian, Sam Seo Wei Jye, Lai Chee Ping, Lesley Lu, Vong Wan Tze and Wendy Japutra Jap who have provided me helpful inputs and given me happy times.

Declaration

I declare that this thesis contains no material that has been accepted for the award of any other degree or diploma and to the best of my knowledge contains no material previously published or written by another person except where due reference is made in the text of this thesis.

Brian Loh Chung Shiong

Publications Arising from this Thesis

Parts of this thesis has been published in the following publication:

Conference paper

1. B. C. S. Loh and P. H. H. Then. Ontology-enhanced interactive anonymization in domain-driven data mining outsourcing. In *2010 Second International Symposium on Data, Privacy and E-Commerce (ISDPE)*, pages 9-14, 2010.

Contents

List of Figures

List of Tables

Commonly Used Acronyms

ASP	Application Service Provider
CSV	Comma-Separated Values
CUI	Concept Unique Identifier
DDDM	Domain-Driven Data Mining
DGH	Domain Generalization Hierarchy
eUtils	Entrez Programming Utilities
HIPAA	Health Insurance Portability and Accountability Act
ICD	International Classification of Disease
InfoGain	Information Gain
JNI	Java Native Interface
KDD	Knowledge Discovery in Databases
MeSH	Medical Subject Headings
MIM	Mutual Information Measure
NIH	National Institutes of Health
PHI	Protected Health Information
PPDM	Privacy Preserving Data Mining
PPDP	Privacy Preserving Data Publishing
QID	Quasi-Identifier
TDS	Top-Down Specialization
UMLS	Unified Medical Language System
UMLSKS	UMLS Knowledge Source Server

CHAPTER 1

Introduction

Knowledge discovery (KDD) also known as data mining allows for the extraction of knowledge from various domains including medical, financial, marketing, etc. The data mining process involves seeking of relationships and global patterns that are present within large databases but may be hidden within vast quantities of data [1]. Most applications of KDD focus on discovering interesting data patterns to solve problems related to a specific field. In the medical field, data mining plays an important role by enhancing the quality and efficacy of healthcare. For instance, a hospital intends to study a group of patients to predict their probability of having heart disease. Through data mining, this task can be solved by utilizing classification algorithms. Once results are obtained, appropriate actions can be taken based on a patient's condition, thus enhancing treatment as well as saving time and costs.

1.1 Data Mining

In the real world, data mining is highly constraint-based as opposed to traditional data mining which is a data-driven trial-and-error process [2], [3], [4], [5].

1.1.1 Traditional

In traditional data mining, knowledge discovery is usually performed without the aid of domain intelligence thus affecting model or rule interestingness for real business needs. Moreover, traditional mining techniques aim to satisfy technical significance such as accuracy or confidence when discovering patterns. Although accurate results

can be gained with these methods, the correlation between predictive accuracy and usefulness of discovered models remain unclear [6]. Because of this, patterns generated may be of no interest to users in realistic business scenarios due to the lack of appropriate domain knowledge.

1.1.2 Domain-driven

On the other hand, in domain-driven data mining (DDDM), domain experts and knowledge are involved to obtain results applicable to real world business requirements. Instead of satisfying technical interestingness measures, the ability to make decisions from rules, and the actionability of patterns is focused upon. As a result, DDDM aims to bridge the gap between academic objectives and outputs as well as business goals and expectations.

1.2 Data Mining Options

When implementing data analysis and mining technology for decision-making purposes, an organization can choose to perform analysis in-house or pursue outsourcing options [7]. In-house solutions involve development of custom data mining applications by the organization or purchasing existing software suites [8]. By developing in-house mining applications, organizations achieve the best security guarantee and specific customization options. To perform development, particular skills and knowledge are required, which an organization may or may not have. In the latter case, existing data mining software suites can be purchased instead. These off-the-shelf applications offer a more convenient approach as they demand usage understanding only. Although in-house mining techniques offer high security or adaptability, they generally possess characteristics such as large initial costs, specific hardware or software requirements, and considerable costs for system maintenance [9], [10]. Furthermore, some organizations lack in-house expertise required for performing data mining although they do hold domain knowledge [11], [12], [13], [14], [15].

Alternatively, outsourcing solutions involve outsider employment of mining experts or interaction with application service providers (ASP). Organizations with insufficient proficiencies can choose to hire experts from third parties for a one-time

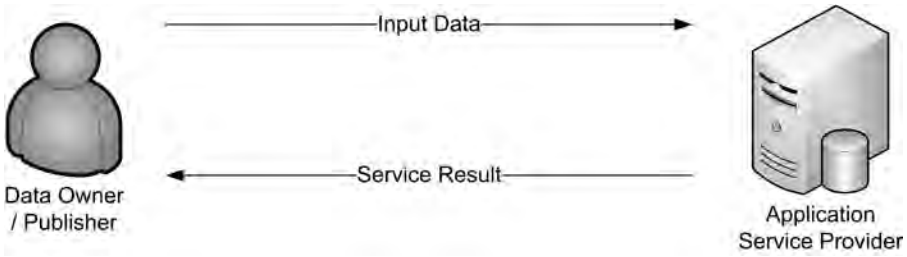


Figure 1.1: 2-party outsourcing scenario

cost. This reduces maintenance costs incurred for employing specialized staff. Additionally, mining tasks can be outsourced by engaging an application service provider. Figure 1.1 describes a simple outsourcing scenario involving two parties, the data owner or publisher (hospital, clinic, etc) who provides input data and the ASP who returns service results (patterns, models, rules, etc) [16], [17]. Several benefits emerge from this option including reduced mining cost, decrease in resource demand and effective centralized mining for multiple distributed owners [13], [18], [19]. Through outsourcing, an organization utilizes minimal computational resources since mining shall be performed by the service provider. Moreover, assume that an organization owns several hospitals in multiple locations. All patient records can be sent to the service provider who would compute patterns local to individual hospitals, or global for the whole organization.

In-house data mining can be an expensive and potentially complicated process which poses a problem for small to medium scale companies. Consequently, service providers are usually seen as a faster, more cost-effective solution compared to in-house application development or software purchases [7], [13]. Furthermore, several data mining characteristics including diverse requirements, need for immediate benefits and specialized tasks fit intuitively into the ASP model [9]. To acquire business intelligence, organizations require diverse tools supporting a variety of algorithms and techniques. Also, organizations desire immediate benefits without long term implementation or learning. Service providers address these needs by hosting multiple data mining systems with infrastructure and expertise in place. Lastly, certain specialized mining tasks are performed once-off and would incur higher costs if organizations created custom applications or purchased software. As a result, service providers fit well into this role by offering a onetime cost effective solution.

1.3 Data Mining Outsourcing Issues

Although outsourcing offers these advantages compared to in-house mining, there are far more privacy and security concerns. There exist three main issues in an outsourcing scenario: data owner's willingness to share sensitive data, un-trusted service provider and laws forbidding the sharing of individually identifiable data [20], [17], [21], [18], [19]. Consider the previous example regarding classification of heart disease, in an outsourcing scenario. Due to privacy concerns or fear of leakage, the hospital may be unwilling to share patient data though it still wishes to investigate the occurrences of heart disease. To promote sharing, some form of data protection would be needed to reduce privacy risks while outsourcing. Additionally, the service provider would most likely be an un-trusted entity who should be denied access to certain private or sensitive information. The assumption here is that trust between data owner and service provider is unattainable, thus the objective remains to protect personal data rather than to create trust. Finally, in certain countries, laws prevent the sharing of identifiable data and removal is required through the use of various protection techniques. The issues discussed relate to data privacy or to be more precise, privacy preserving data publishing (PPDP) and mining (PPDM).

1.4 Anonymization Techniques

Anonymization methods including k-anonymity, l-diversity and t-closeness have been created to preserve privacy through generalization [22], [23], [24]. These techniques were designed to work with static datasets meaning that whenever new data was published, previous releases of the dataset were not considered during anonymization. In a real world data mining outsourcing scenario, data is constantly changing within a dynamic environment. For instance, a hospital outsources their patient data yearly with the intention of generating interesting models. Since modifications have occurred over time, each year, a new dataset is published to the service provider. Therefore, recent dynamic anonymization techniques including m-invariance, ϵ-inclusion and m-distinct have been created to deal with insertions as well as deletions of new records or values [25], [26], [27].

Both static and dynamic techniques mentioned work as a "1 size fits all" solution,

meaning that it supports any mining operation though data quality may not be optimal for each task. Several researchers have developed techniques to anonymize data based on specific applications or workloads to ensure the best data utility for a particular task [28], [20], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39]. As discussed earlier, in an outsourcing scenario a data owner would already have a mining task in mind when publishing data to a service provider. Thus, the precise anonymization techniques that should be applied fall into this category.

1.5 Research Motives and Contributions

The majority of anonymization methods have focused on the traditional data mining paradigm, protecting data for general purposes or specific mining tasks such as classification. Although models obtained from these datasets may possess similar accuracy with their original counterparts, they do not necessarily contain actionable rules usable in real world settings. Furthermore, existing techniques limit user control, only allowing the selection of privacy parameters [40], [41]. Users are also assumed to be fully capable of creating Domain Generalization Hierarchies (DGH) based upon their own knowledge [42], [43]. Lastly, previous studies on utility metrics have not considered attribute correlations during anonymization thus leading to possible over generalization of important attributes [44].

1.5.1 Research Problem

The main research problem is to discover a practicable anonymization technique capable of preserving individual privacy while maintaining utility for domain driven data mining in an outsourcing scenario.

1.5.2 Research Objective

The objective of this research is to create an anonymization framework capable of protecting data while maintaining rule meaningfulness and actionability after mining. The term meaningfulness refers to semantic meaning while actionability represents a user's ability in utilizing mined rules for a particular task. The framework shall incorporate domain ontology knowledge during the anonymization

process as a means to improve data mining results. Findings obtained from the framework shall provide a basis for the development of anonymization techniques specifically for domain driven data mining.

1.5.3 Significance and Contributions

Several factors support the significance of this work. Traditional data mining, being a data centric process produces numerous patterns but most are of no interest to the user and do not meet real world business requirements [45]. Therefore, a main challenge of KDD in realistic scenarios is to obtain actionable knowledge capable of benefiting an organization. Actionable in this case strongly relies on domain knowledge such as domain ontologies, expert knowledge and business logic. In the field of PPDP and PPDM, anonymization techniques once focused on general purpose usage have shifted to data mining applications. Numerous methods have been created that provide strong guarantees against privacy risks while maintaining reliable utility for traditional mining purposes. However, none of these techniques have focused upon maintaining utility for domain driven mining tasks. Furthermore, most of these methods are still in the research paper or prototype stage with little impact in real world commercial terms [46]. As a result, it is essential that anonymization methods integrate domain knowledge as a means to preserve actionable patterns while protecting individual privacy.

This research is significant as it aims to produce an anonymization framework capable of utilizing domain ontology knowledge for attribute DGH creation to maintain and improve value meanings. Furthermore, attribute semantic types and relations are evaluated to enable user constraints, thereby allowing for task specific anonymization. Finally, a correlation-based measure integrating literature domain knowledge is utilized to ensure the preservation of important attributes.

1.6 Thesis Structure

The thesis structure is as follows:

Chapter 2 Literature Review presents related works. Data privacy guidelines and protection, the data mining outsourcing scenario, the anonymization process and

issues, as well as interestingness measures are reviewed.

Chapter 3 Framework presents the proposed framework's design and components.

Chapter 4 Experiments and Results describes experiments and discusses their results.

Chapter 5 Conclusion concludes the thesis. Framework limitations and recommendations for future research are also discussed.

Chapter 6 Appendix describes various software components utilized in the proposed framework.

CHAPTER 2

Literature Review

2.1 Privacy Guidelines

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) ensures the removal of patients' protected health information (PHI) from medical records which are to be used in research [47]. 18 specific categories of PHI are removed including names, geographic locations, dates, social security numbers, telephone numbers, etc. In Malaysia, the Personal Data Protection Bill regulates the collection, possession, processing and use of personal data [48]. According to the bill, personal data is defined as any information recorded in a document which can be practically processed by any automatic means or otherwise, which relates directly or indirectly to a living individual who can be identified from that information. The 8th data principle of the bill concerns the security of personal data such that practical steps are required to protect personal data against unauthorized or accidental access, processing or erasure, alteration, disclosure, or destruction. Both the HIPAA and Personal Data Protection Bill aim to protect the identity of individuals thus preserving their personal privacy.

2.2 Data Privacy and Protection

PPDP involves the development of methods or tools for releasing data that remains practically useful, while still preserving individual privacy. Conversely, PPDM focuses on the issue of recovering useful mining results from modified data [14]. Both PPDP and PPDM share the common goal of balancing privacy preservation

and data mining utility. The difference between both areas lies with the data mining task at hand. In PPDP, data mining tasks may be unknown during publishing while in PPDM, solutions are often coupled with a particular data mining algorithm. For instance, when a mining task is known (association/classification rule mining, etc), specialized PPDM techniques can be utilized to ensure maximal utility towards their mining algorithms. On the other hand, when mining tasks are unknown, PPDP methods aim to provide reliable utility for a variety of mining algorithms. The previously discussed outsourcing scenario is a mixture of both areas whereby the hospital publishes data to a service provider with the intention of performing a specific mining task.

Data owned by the hospital may contain three types of attributes which can be divided into the following categories. Explicit identifiers which can clearly identify an individual (e.g. name or social security number). Quasi-identifiers (QID) whose values when taken together can potentially identify individuals (e.g. zip code, gender, or date of birth). Sensitive attributes that represent private information of individuals (e.g. disease or salary). Guidelines such as the HIPAA aim to preserve individual privacy by removing protected health information through de-identification. Although this can prevent direct identification of patients from a medical dataset, Sweeney has shown that 87% of United States individuals can be uniquely identified based on a set of QID attributes which includes zip code, gender, and date of birth [24].

Table 2.1: Sample raw data

	Quasi-identifier			Sensitive Attribute
ID	**Age**	**Gender**	**Zip Code**	**Disease**
1	45	M	93350	Cancer
2	27	M	93300	Fever
3	45	M	93350	Cancer
4	30	F	93300	Flu
5	30	F	93000	Fever
6	35	F	93000	Cancer

Privacy threats occur when an adversary links an individual in published data to their record or sensitive attribute. These attacks are referred to as identity linkage, record linkage, and attribute linkage. In all cases, it is assumed that an adversary knows an individual's QID attributes. Furthermore, it is assumed that an attacker is aware of

the individual's existence in the released data. In an identity linkage attack, if a record is very specific whereby only few patients match it, an adversary with background knowledge can identify that particular individual. Consider the sample raw data in Table 2.1 where each record represents patients of the hospital. Age, gender, and zip code are QID attributes while disease represents a sensitive attribute. Suppose an adversary knows that their victim is a 27 year old male living in an area with zip code 93300. Since only one individual possesses that information, the adversary can easily infer that their target has fever. Record linkage is similar whereby an adversary attempts to identify an individual by linking their QID attributes with externally available information. For example, assume the previous victim appears in a voter's list which contains their explicit identifiers and QID attributes. As shown in Figure 2.1, by matching QID attributes from both sample raw data and voter's list, the adversary can identify the particular person. Attribute linkage happens when sensitive values occur recurrently with a specific set of QID attributes therefore allowing inferences to be made without exact matches. For instance, an adversary knows that their target is a 45 year old male. Although the raw data contains two records, both records show that the patient is suffering from cancer. Consequently, the adversary can safely assume that their victim has cancer.

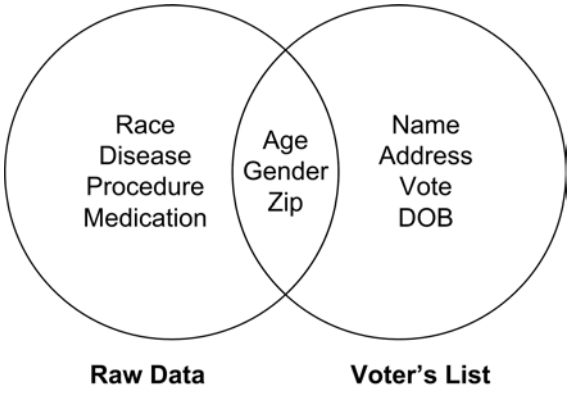


Figure 2.1: Record linkage attack

In order to protect sensitive information, the hospital can apply techniques such as perturbation, encryption, generalization or suppression on their input data before publishing [49]. Perturbation modifies data through value randomization while still maintaining its original distribution. By distorting original values, even if adversarial attacks are performed, the truthfulness of those inferences remains uncertain. Therefore, it is claimed that original data would remain hidden while

any added noise would average out in data mining output. One limitation of the perturbation approach is that perturbed records do not correspond to real world entities represented by original data; therefore, individual records are meaningless to data miners [14]. Furthermore, the perturbation approach lacks a formal framework for proving the amount of privacy guaranteed, with recent evidence showing that certain perturbation approaches offer no privacy at all [50]. Alternatively, encryption transforms data into a new format such that original values are unreadable to all except the data owner. This prevents linkage attacks as long as data remains encrypted. The encryption process generally occurs before outsourcing and results are decrypted once returned. Compared to perturbation, cryptography provides methodologies for proving and quantifying privacy. Moreover, there are numerous encryption techniques for implementing privacy-preserving data mining algorithms. However, it has been shown that encryption only prevents privacy leaks during the process of computation, and not for mining outputs [50]. Lastly, generalization transforms QID record values by replacing them with a corresponding generalized value. For instance, the gender *male* or *female* can be generalized to *person*. Suppression works similarly to generalization, but instead of substituting a generalized value, a special value is used instead. For example, the age *25* can be suppressed to *2**, indicating a range of *20 - 29*. This reduces the adversary's probability of correctly inferring sensitive information. For the remainder of this thesis, suppression shall be grouped under generalization as a single protection technique.

Generalization has been widely utilized in safeguarding individual privacy with the anonymization model. Compared to perturbation, it provides the advantage of preserving true attribute values, hence allowing for higher accuracy of data mining results [38], [49], [39]. This is an important factor for data owners since their main objective for outsourcing is to obtain models or patterns. Moreover, encryption techniques require data processing before and after outsourcing, hence causing inefficiency issues [51], [52]. As mentioned previously, some organizations require immediate results or manage data in multiple locations. Encryption may slow down the outsourcing process thereby preventing the data owner from achieving their needs. Also, both perturbation and encryption cannot be applied for all data mining functionalities [53], [52]. This conflicts with the need for diverse or specialized

mining tasks since a data owner can perform any sort of request. Based on these reasons, generalization appears to be a more suitable choice as it complements both data mining and protection characteristics in an outsourcing scenario.

2.3 Data Mining Outsourcing

The data mining outsourcing scenario with regards to data protection, though not widely discussed has been the focus of several researchers from past to present. The overall scheme involving data owner and service provider has been presented in several papers [11], [15], [54], [18], [55]. These studies relate directly to the scenario mentioned in this thesis and provide supporting reasons for choosing such a setting. Brumen et al. [11] proposed database transformation on both data structure and values before outsourcing for analysis. After mining, the process is reversed hence allowing data owners to decode results for interpretation. On the other hand, Wong et al. [18], and Qiu, Li & Wu [15], focused on mining association rules from transformed data. Wong et al. [18] suggested the use of substitution cipher methods to encrypt data while Qiu et al. [15] used Bloom filters. In both cases, transformed data is sent to a service provider who computes rules and returns them to the owner. Algorithms are then used to convert the rules to actual association rules involving original values. Although similar in circumstances, these papers utilize techniques other than anonymization to handle privacy issues. A major difference between these works and the proposed framework is that results returned do not require conversion since rules obtained from anonymized data remain comprehensible.

2.4 Data Anonymization

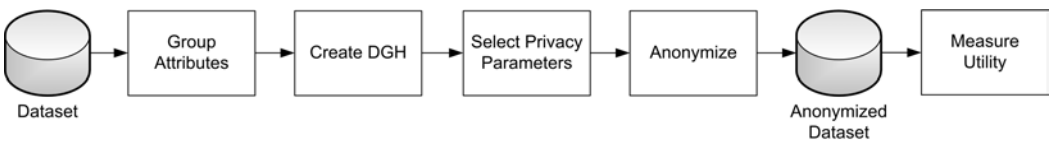


Figure 2.2: Anonymization process

To prevent inference or individual identification when outsourcing, data can be protected via anonymization techniques. Anonymization allows concealment of patient identities or sensitive data, assuming that this information is required for

data analysis [14]. In general, the steps involved in an anonymization process are as shown in Figure 2.2. First, the data owner groups each attribute in their chosen dataset into explicit, QID or sensitive attributes. It is usually assumed that all dataset attributes can be correctly categorized by the data publisher. For instance, explicit attributes such as names or addresses are removed while sensitive attributes including disease or salary are not confused for QIDs. After that, a DGH is created for each attribute. Normally, these hierarchies are created manually but certain methods allow for dynamically created hierarchies. Next, an anonymization algorithm and privacy requirement is chosen based on the data mining purpose or potential linkage attacks. At this stage, privacy parameters can be set which control the level of anonymization on the dataset. The QID attributes would then be generalized according to the chosen algorithm, privacy requirement, and their respective DGH. During the anonymization process, a utility measure is employed to guide the algorithm towards transformations that provide the best utility and privacy trade-offs. An anonymized version of the initial dataset would be the result of this generalization. After anonymization, another utility measure would then be used to determine data quality as compared to the original dataset, or result accuracy for a particular mining task. The objective of this approach is to ensure that utility remains after data protection.

2.4.1 *K*-anonymization

One commonly implemented privacy requirement is k-anonymization which transforms a table so that no party can make high-probability associations between records and their sensitive attributes through QIDs [24], [56]. By enforcing k-anonymity, it is guaranteed that even if an adversary knows that a table contains a particular individual record as well as the quasi-identifier value for the individual, he/she cannot determine which record corresponds to the individual with a probability greater than $1/k$.

Table 2.2 represents a 3-anonymous table of the previous sample raw patient data. Both age and zip code have been generalized to protect individual identities. The first half (1 - 3) and last three (3 - 6) records of the table are indistinguishable from one another as their QID values are the same. These groups of records are referred to as equivalence classes. By anonymizing the dataset, an adversary would be unable to

Table 2.2: Sample 3-anonymous table

	Quasi-identifier			Sensitive Attribute
ID	**Age**	**Gender**	**Zip Code**	**Disease**
1	20-50	M	933**	Cancer
2	20-50	M	933**	Fever
3	20-50	M	933**	Cancer
4	30-35	F	93***	Flu
5	30-35	F	93***	Fever
6	30-35	F	93***	Cancer

perform the previously mentioned linkage attacks therefore preserving privacy.

2.4.2 *L*-diversity

According to Machanavajjhala et al. [23], *k*-anonymity is still susceptible to attribute linkage and background knowledge attacks. Based on the 3-anonymous data in Table 2.2, although records 1 and 3 have been generalized, an adversary still has a 2/3 chance of inferring that their target has cancer. This conflicts with the *k*-anonymization property where the chances of privacy threats occuring cannot be greater than $1/k$. Apart from linkage attacks, an adversary can also use background knowledge to discover sensitive information. For instance, if the adversary is certain that the target individual does not have fever, they can conclude that the victim has cancer. In order to address the limitations of *k*-anonymity, Machanavajjhala et al. [23] introduced *l*-diversity which represents a stronger notion of privacy. An equivalence class possesses *l*-diversity when there are at least *l* "well-represented" values for the sensitive attribute. A table possesses *l*-diversity when every equivalence class of the table has *l*-diversity. Basically, *l*-diversity ensures that each equivalence class contains diverse sensitive attribute values, therefore reducing privacy attacks.

As seen in Table 2.3, an adversary cannot perform either linkage or background knowledge attacks since each equivalence class has diverse sensitive attributes. For example, there are a total of three equivalence classes with two records each. Since the sensitive attributes for both records differ, the adversary would have a lower chance of successfully inferring sensitive information. However, it can be noticed that as privacy requirements increase or as stricter privacy models are implemented, data becomes more generalized therefore leading to some utility loss.

Table 2.3: Sample 2-diverse table

	Quasi-identifier			Sensitive Attribute
ID	**Age**	**Gender**	**Zip Code**	**Disease**
1	20-50	M	933**	Cancer
2	20-50	M	933**	Fever
3	20-50	*	933**	Cancer
4	20-50	*	933**	Flu
5	30-35	F	93***	Fever
6	30-35	F	93***	Cancer

2.5 Anonymization Process

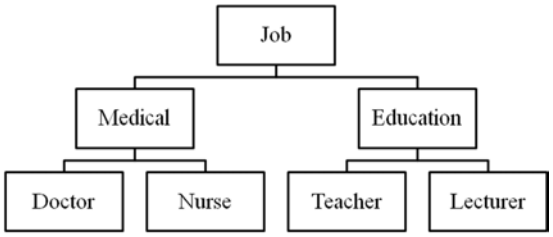
2.5.1 Domain Ontologies

Medical applications often involve various datasets containing complex information with differing structures and semantics. Because of this, knowledge organisation systems such as ontologies, taxonomies and thesauri are required. According to Gruber [57], an ontology can be defined as an explicit, formal specification of a shared conceptualisation. To further explain, an ontology refers to a generally agreed domain model with explicitly defined concepts and relationships that are machine-readable and machine-understandable [58]. Taxonomies establish a classification hierarchy of terms by grouping similar objects under classes while thesauri refine these hierarchies by providing a fixed set of predefined relations between concepts [59]. There exist several biomedical knowledge repositories such as the Unified Medical Language System (UMLS), the International Classification of Diseases (ICD) and Medical Subject Headings (MeSH) which have been used with knowledge-based systems.

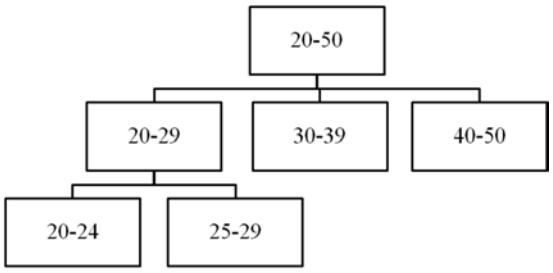
2.5.2 Domain Generalization Hierarchies

The majority of anonymization methods employ DGHs for generalizing attribute values. Initially, DGHs were used for categorical attributes but were extended for numerical attributes [60]. Most techniques featured predefined DGHs which were manually created based on prior user knowledge. Figure 2.3 shows a predefined categorical and numerical DGH. A hierarchy for the job attribute is shown in Figure 2.3a, with *doctor* and *nurse* being under the *medical* field while *teacher* and *lecturer* are under *education*. Figure 2.3b presents the age attribute with evenly distributed ranges

at each hierarchy level.



(a) Categorical DGH for job



(b) Numerical DGH for age

Figure 2.3: Sample domain generalization hierarchies

Other works implemented hierarchy-free models which could dynamically generate DGHs. These automated methods utilize specific metrics to determine the optimal split for numerical values. One disadvantage with predefined DGHs was that experts faced issues among themselves when determining the correct structure of hierarchies [61]. As for dynamically grown DGHs, although efficient, they could only be created for numerical attributes. Because of this, a number of studies focusing on pre-defined DGHs chose to implement domain ontology knowledge as a source for hierarchy ranges and values. Both Lin, Hewett & Altman [62] and Lin [63] utilized the ICD medical data hierarchy as a mapping for their DGHs. Similarly, Bertino, Lin & Jiang [64] based their DGH on the ICD hierarchy as well. The proposed framework intends to follow in the same direction and furthermore, study the effect of ontology created DGHs on data mining results.

2.5.3 Generalization Schemes

Generalization involves replacing attribute values with a parent value found in their DGH. The reverse operation, specialization, involves replacing attribute values with

a child value. There are two general categories of generalization schemes which are global recoding and local recoding. In global recoding schemes, when a value is generalized, all instances of that value are generalized. For local recoding, some value instances can remain ungeneralized while others are generalized. Four generalization schemes shall be discussed.

1. Full-domain generalization (Global recoding) [65], [24]. In this scheme, all attribute values are generalized to the same levels in their DGH. For instance, based on Figure 2.3a, when *doctor* or *nurse* is generalized to *medical*, both *teacher* and *lecturer* would also be generalized to *education*. Using this scheme causes the largest data distortion since generalization of one value affects all others.

2. Subtree generalization (Global recoding) [28], [30], [29], [34], [65], [37]. For this scheme, when at a parent node, either all child values or none are generalized. For example, if *doctor* is generalized to *medical*, *nurse* is also required to be generalized. But for *teacher* and *lecturer*, since both values are from a different parent node, they can remain ungeneralized.

3. Sibling generalization (Global recoding) [65]. This scheme is similar to subtree generalization with the exception that some child values can remain ungeneralized. A parent value is then used to represent all generalized values. For instance, if *doctor* is generalized to *medical* and *nurse* remains ungeneralized, *medical* represents all jobs except for *nurse*. This scheme causes less distortion compared to subtree generalization as it does not need to generalize all child values.

4. Cell generalization (Local recoding) [65], [66], [67]. This scheme, being a local recoding scheme, allows some instances of values to remain ungeneralized after the generalization process. For example, two records containing the value *doctor* are generalized. In the first record, *doctor* is generalized to *medical* while a same *doctor* value in the second record can remain ungeneralized. This scheme produces less data distortion but can affect data utility through the data exploration problem. Data mining algorithms would treat *doctor* and *medical* as independent values, when in fact they are not. Consequently, classification accuracy and rule meaningfulness would be affected.

The proposed framework implements the subtree generalization scheme as it provides appropriate levels of generalization without affecting the validity of data mining results. Although other generalization schemes may allow for a lower data distortion rate, the importance of maintaining meaningful classification rules has to be considered as well.

2.5.4 Anonymization Algorithms

Anonymization algorithms can be classified into two categories which are optimal anonymization and minimal anonymization. Optimal anonymization algorithms aim to discover optimal solutions when anonymizing datasets. They perform anonymization based on a given metric and utilize full-domain generalization schemes. Finding an optimal solution for small datasets is achievable since full-domain generalization schemes perform exhaustive searches. However, optimal algorithms are not scalable to large datasets due to time and efficiency constraints. Minimal anonymization algorithms aim to produce minimal solutions when anonymizing datasets by using greedy search methods. These heuristic algorithms are more scalable than optimal algorithms as they do not need to find an ideal solution. Three minimal anonymization algorithms shall be examined.

1. Genetic [34]. This algorithm was among the first that focused on preserving classification utility. It employed a stochastic search based on a classification metric and also implemented a subtree generalization scheme. The genetic algorithm worked by representing generalization states as chromosomes which evolved to find the fittest solution. Experiments suggested that classifiers built from anonymous data using classification based utility metrics produced models with lower classification error.

2. Bottom-up generalization [38]. The bottom-up generalization algorithm was proposed for finding minimal solutions to address efficiency issues in *k*-anonymization. It starts from a raw dataset which is in violation of the *k*-anonymity requirement. At each step, an attribute is selected for generalization based on a utility metric. Each operation increases the size of equivalence classes and the process is terminated once all groups have reached the minimum *k* parameter.

3. Top-down specialization (TDS) [30], [29], [37]. As opposed to bottom-up generalization, top-down specialization specializes a table from its most general state where all values have been generalized to the highest values in their DGH. Each operation is performed according to a utility metric and is terminated once no specialization can be done without violating any privacy requirements.

The proposed framework utilizes the TDS algorithm for anonymizing datasets due to several benefits it provides. First, since all attribute values start from a generalized state, the specialization process produces an anonymous dataset after each specialization. Furthermore, specialization can be stopped at any moment, and would result in an anonymized solution. Second, TDS can handle multiple QID groups instead of just a single high dimensional QID group, therefore avoiding excessive distortion. Third, TDS is more efficient compared to other algorithms including genetic or bottom-up as it requires less memory for storing data records.

2.5.5 Utility Metrics

Numerous metrics have been proposed for both general and data-specific purposes. General-purpose metrics which include average equivalence class size [23], discernibility metric [28], minimal distortion [24], and information loss [41] aim to measure utility loss caused by generalizations during anonymization. In scenarios where data mining purpose is unknown, general-purpose metrics aim to measure the closeness of values between original and anonymous datasets. The minimal distortion metric operates by charging a penalty for each generalized value. For instance, generalizing ten values of *doctor* to *medical* would incur a penalty of ten distortion units. The information loss metric works similarly by measuring the amount of information loss when generalizing a specific value to a more general value. Both minimal distortion and information loss penalizes the generalization of values independently of other records. On the other hand, the discernibility metric charges a penalty to each record for being indistinguishable from other records. Although useful for capturing similarities between datasets, these measures do not necessarily indicate quality with respect to a particular mining task [35]. For instance, unmodified data containing noise often has worse classification compared to generalized data where noise has been masked [29]. Because of this, general-purpose

metrics may indicate reduced utility after generalization when in fact, data mining utility has risen. Data-specific metrics avoid this error by measuring the ability of an anonymized dataset to build accurate models. For classification tasks, numerous papers have adopted the Information Gain (InfoGain) measure as a score for attribute selection and generalization [30], [35], [29], [37], [68]. The InfoGain measure is based on information theory and is used to assess the purity of an attribute. During data mining, classification algorithms are capable of building decision trees based on attribute InfoGain scores. Therefore the use of InfoGain in the anonymization process may lead to a generalized dataset with higher predictive accuracy. Although existing metrics manage to compare generalization levels or mining accuracy, even if they indicated high utility, rules obtained from anonymized data may not necessarily be actionable.

2.6 Anonymization Issues

2.6.1 Quasi-identifier Selection

According to Dalenius [69], QID represents a set of attribute values in census records which may be used for individual or group re-identification. In the works of Samarati [70] and Sweeney [24], the notion is extended to a set of attributes whose combined values may be used for re-identification with the aid of external information. Many definitions of QID which differ in concepts have since appeared in literatures [71], [72]. QID selection plays an important part in the anonymization process and misclassification of attributes can cause privacy or utility loss [14], [73]. For instance, classifying a QID as a sensitive attribute can allow an attacker to easily perform record linkage. Then again, classifying a sensitive attribute as a QID would cause unnecessary information loss due to generalization.

The categorization of attributes is commonly assumed to be performed manually by the data owner through the help of domain knowledge or publicly available datasets. For example, a hospital outsources patient information containing attributes such as age, gender, ethnicity, height, weight, blood pressure, and disease. By applying simple logic or examining a publicly available source of information, a publisher can straightforwardly determine that age, gender, and ethnicity can identify a patient, hence should be QID. But the attributes height, weight, and blood pressure pose a

more difficult categorization task. Although they can possibly identify a patient, their existence in public datasets is uncertain. In principle, QID should contain attributes which an adversary can potentially obtain from external sources. The problem here is that there exists no definite solution as to how a data owner can determine if information on a particular attribute can be found by an attacker. Another method of attribute categorization is through utilization of algorithms that aim to find a minimal set of QID based on certain measures or requirements [72], [40], [73]. These automated techniques can aid publishers during attribute grouping by indicating previously unconsidered attributes as QID. However, they do not take into account adversarial background knowledge or correlations between attributes.

2.6.2 Parameter Selection

Numerous anonymization models have been created for use with static, dynamic and workload oriented applications. These algorithms aim to prevent individual identification and unwanted disclosure of sensitive information by generalizing QID attributes within a dataset. Though privacy is protected, a direct consequence of anonymization is the loss of data quality due to simplification of values. When selecting a privacy requirement, a data publisher is required to specify certain parameters which affect both data quality and privacy. Take for example, the k-anonymity model. According to Sweeney [24], a dataset is k-anonymous if for each record in a table, there exist k-1 other records with identical QID attributes. This means that an adversary would have a $1/k$ chance of successfully performing identity linkage attacks. For this model, a publisher would specify the value of k, ranging from one to the total amount of records, with either larger values meaning more privacy or smaller implying more utility. This causes a privacy utility trade-off dilemma since the choice of k is not clear. If the data owner desires higher utility, they may be willing to accept more privacy loss in order to gain better results. If their concern is privacy, then the value of k should be high enough to prevent attacks while still maintaining data usefulness. This issue extends to other privacy requirements as well. For instance, in the l-diversity model, the value of l can range from one to the total number of sensitive values. Not only does this affect the user's ability to select an appropriate parameter, but it prevents comparison of privacy requirements which possess differing parameter meanings [74]. Therefore, to better select a privacy

parameter, a publisher needs to understand the trade-off from both privacy and utility aspects.

2.6.3 User Constraints

Several works have studied the need for user-specified constraints or preferences to control the generalization process [39], [60], [75], [76], [77], [20]. The general idea was that most anonymization techniques did not allow users to express their specific protection or usage needs. For instance, data owners may require prioritization of certain attributes which are important in accomplishing their data mining task. The majority of methods restrict the anonymization process based on the utility metrics chosen. Because of this, users would have no control over the generalization of attributes and may encounter a loss in utility.

Both Miller, Campan & Truta [60] and Loukides, Tziatzios & Shao [75] suggest the use of attribute constraints which impose a limit on the level of generalization allowed for a particular attribute. Moreover, Loukides et al. [75] introduces the use of value constraints which specifies allowable generalizations for a chosen attribute value. Attribute constraints enabled users to control generalization ranges for selected attributes. For example, based on Figure 2.3b, a ten year constraint set on the age attribute would prevent any values from being generalized to *20 - 50*. Value constraints allowed selected attribute values to be generalized to a certain preset level. For instance, referring to Figure 2.3a, an attribute constraint set on *medical* would prevent *doctor* or *nurse* from being generalized to *job*. Xiong & Rangachari [39] on the other hand, enabled users to create preferences based on their mining purpose. Depending on the task, each attribute would be associated with priority weights that determined which are to be preserved. Dewri et al. [76] proposed a similar approach by enforcing weights on attributes based on their significance. According to Sun, Wang & Li [77], one issue with such a method is that users are required to specify attribute priorities beforehand. Because of this, they proposed an anonymization scheme that derived priorities automatically by calculating attribute dependencies based on their entropy levels. Bhumiratana & Bishop [20] contributed a different constraint method by allowing negotiations between data owner and miner through the use of policies. It enabled users to create privacy or analysis policies which controlled the generalization of specific attributes. The proposed framework

differs from previous solutions by allowing prioritization of attributes based on their semantic relation or type and also on a particular mining purpose. By doing so, it avoids the problem of requiring user specified weights since attributes are chosen based on their importance towards a task.

2.7 Interestingness Measures

Interestingness measures allow for the selection and ranking of mined models or rules based on potential user interest. According to Geng & Hamilton [78], there are nine specific criteria for determining pattern interestingness. They include conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility and actionability. These criteria can be further classified into objective, subjective and semantics-based. An objective measure is usually based on statistics or information theory and focuses only on the raw data. This means that any knowledge regarding the user or application is not required. A subjective measure considers both data and user domain knowledge which is obtained through user interaction in the mining process. A semantic measure evaluates the semantics and explanations of patterns as well as user goals.

1. Conciseness (Objective). A concise pattern is easily understandable and remembered as it contains less rules. Therefore, it can undoubtedly be added to user beliefs or knowledge.

2. Coverage (Objective). Coverage measures how much of all records are covered by a particular pattern. If a pattern characterizes a majority of information in a dataset, it can be considered as more interesting.

3. Reliability (Objective). A reliable pattern is one that occurs frequently in the majority of applicable cases. For instance, accurate classification rules capable of making predictions are considered reliable.

4. Peculiarity (Objective). Peculiar patterns which are generated from outliers may be unknown to users since they are significantly different from the rest of the data. Because of this, they can be considered as interesting.

5. Diversity (Objective). A diverse pattern contains elements which differ

significantly from each other and can be considered interesting in the absence of relevant knowledge.

6. Novelty (Subjective). A pattern is novel if a person did not know it before and is unable to infer it from known patterns. Novelty cannot be measured explicitly based on a user's knowledge or ignorance since no known system represents everything a user knows or does not. Because of this, novelty can be detected by having a user confirm it or noticing that a pattern does not come from or contradict a previously discovered pattern.

7. Surprisingness (Subjective). A surprising pattern contradicts a person's existing knowledge or expectations and can be considered interesting since it identifies failings in previous knowledge which suggest the need for further study.

8. Utility (Semantics-based). A pattern is of utility if it enables a person to achieve a certain goal. Since goals depend on a user, this interestingness criteria is based on user-defined utility functions in addition to the raw data.

9. Actionability (Semantics-based). An actionable pattern allows for decision making concerning future actions in a particular domain. Currently, there are no general methods for measuring actionability, with existing methods relying upon application goals.

The framework selects the semantics-based actionability criteria as the interestingness measure of choice due to several reasons. First, objective measures only focus on the raw data without considering user needs. Previous anonymization techniques have mostly aimed to preserve model reliability by ensuring that anonymized patterns possess similar accuracies to the original output. Since domain knowledge is not considered in these methods, although accurate rules are maintained, it does not necessarily mean that mined patterns can be applied to real world situations or user tasks. Second, though subjective measures implement domain knowledge, they strongly depend on the actual input data. If the raw data does not contain any novel or surprising rules in the first place, it can be expected that after anonymization, the situation would still remain the same. Therefore, by choosing to satisfy actionability, the framework integrates domain knowledge and user interaction in the form of constraints to preserve actionable rules for user goals.

2.8 Summary

The data mining outsourcing scenario has been the interest of several researchers. Although their techniques differ from the proposed framework, they aim to protect privacy and maintain data utility as well. The framework intends to utilize predefined DGHs with domain ontology knowledge for generalizing attribute values. This avoids the issue of having to correctly determine an appropriate structure for hierarchies since ontologies can provide suitable ranges. Additionally, the subtree generalization scheme is chosen for its ability in generalizing attribute values without negatively affecting data mining results. Furthermore, the TDS algorithm is selected as it outperforms previous algorithms in terms of efficiency and scalability. For measuring utility, the framework adopts the existing data-specific metric, InfoGain, together with a proposed correlation measure. Although three anonymization issues are presented, the main focus of the framework is to enable user constraints through attribute semantic types and relations towards a specific mining task. Lastly, the framework measures interesting patterns by their actionability towards a particular user goal.

CHAPTER 3

Framework

Figure 3.1: Proposed framework anonymization process

The main motivation of this framework is to preserve privacy with the *LKC*-privacy model and maintain utility for domain-driven data mining. Figure 3.1 illustrates a typical anonymization process with highlighted squares indicating the proposed framework's contributions (ontology-based DGH, user-specified constraints, and correlation-based anonymization algorithm), previously proposed in [79]. As mentioned earlier, a typical anonymization process first involves the categorization of attributes for a particular dataset. In this scenario, it is assumed that attributes are properly grouped with no misclassification of QIDs or sensitive attributes. In the DGH creation phase, ontology-based hierarchies are manually created with domain knowledge to retain attribute meanings. Next, constraints are specified based on attribute semantic relations and types, for a chosen data mining task. After that, appropriate privacy parameters are selected to ensure sufficient data protection and utility. Lastly, anonymization occurs and attributes are generalized according to specified constraints as well as the literature aided correlation-based measure.

3.1 Ontology-based Domain Generalization Hierarchy

In data mining, domain knowledge includes comprehension of a dataset, variable relationships, variable ranges, known causal relations, etc [80]. Domain ontologies

represent a promising source of knowledge as they express domain concepts and relationships in a comprehensible way to a particular professional community [81], [82]. One of the world's most comprehensive medical domain ontologies, the UMLS is a suitable choice for semantically mapping attributes to ranges or concepts while preserving meaning. There are three major components in the UMLS which are Metathesaurus, Semantic Network and Specialist Lexicon [83]. The Metathesaurus contains inter-related biomedical concepts and is utilized for DGH creation. Additionally, the Semantic Network assigns Metathesaurus concepts high-level categories while the Specialist Lexicon generates lexical variants of biomedical terms. Table 3.1 describes the UMLS semantic mapping for age with eight concepts and their year ranges.

Table 3.1: UMLS semantic mapping

Attribute	Definition	Concept
Age	Age [1 - 23 months]	Infant
	Age [2 - 5 years]	Child, Preschool
	Age [6 - 12 years]	Child
	Age [13 - 18 years]	Adolescent
	Age [19 - 44 years]	Adult
	Age [45 - 64 years]	Middle Aged
	Age [65 - 79 years]	Elderly
	Age [80 over]	Ages, 80 and over

By employing this semantic mapping during hierarchy construction, the DGH for age can be improved. Figure 3.2 illustrates a basic DGH which discretizes age into ranges of five years and more while Figure 3.3 displays an ontology-based DGH which discretizes values into appropriate ranges following the UMLS concepts. One difference of the ontology-based DGH is that there are both numerical ranges and categorical concepts which values can be generalized to. The benefit of such an approach can be seen through a simple example where the attribute age, with a value of *42* needs to be generalized. Based on the basic DGH, it can be generalized to either range *40 - 45* or *1 - 50*. On the other hand, if the ontology-based DGH is applied, *42* can be generalized to either *38 - 44* or *adult*. According to the UMLS semantic mapping for age, an adult is between 19 to 44 years old while a middle aged person is between 45 to 64. The basic DGH fails to capture this semantic information since the range *40 - 45* can mean either adult or middle aged. In this case, although the basic DGH provides

more specific ranges, five years as opposed to seven, the ontology-based DGH retains more meaning.

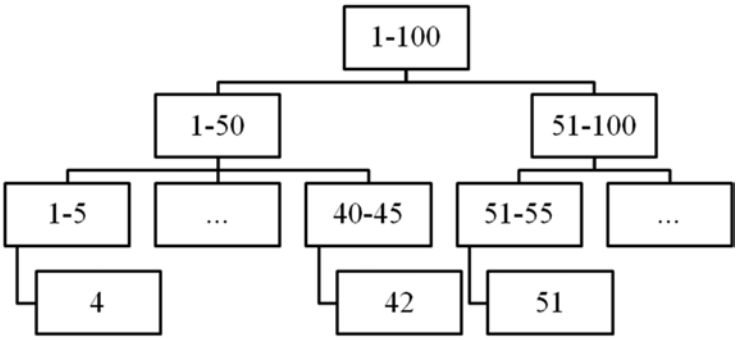


Figure 3.2: Basic domain generalization hierarchy

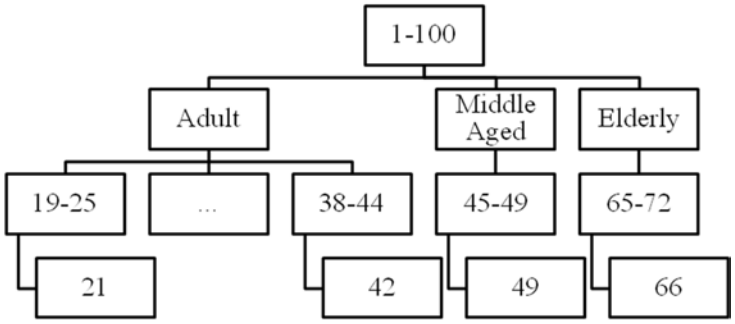


Figure 3.3: Ontology-based domain generalization hierarchy

3.2 User-specified Constraints

Previous anonymization techniques have mostly restricted users to specifying privacy requirement parameters which would affect both data protection and utility factors. Their objective was to obtain anonymous data and conserve as much information as possible with a chosen algorithm. The produced anonymized dataset while sufficiently protected and satisfying utility or information loss metrics, may still be unusable for intended mining tasks if important attributes have been over generalized. Before outsourcing, a publisher would already have a mining purpose in mind for their dataset. Therefore, it is essential that control be given to the user for determining attribute priorities during generalization. For instance, picture a scenario

where a hospital intends to perform research on their patient data. Physicians usually engage in four main clinical tasks which are therapy, diagnosis, etiology, and prognosis [84]. Therapy involves selecting treatments for patients while diagnosis predicts the likelihood of diseases on an observed patient. Furthermore, etiology identifies factors causing a particular disease or condition in a patient whereas prognosis anticipates potential complications of a patient over the course of time.

A typical patient dataset may include the attributes age, gender, race, cholesterol, blood sugar, and heart disease. The study aims to solve an etiology task by examining the relationships of cholesterol and blood sugar with the occurrence of heart disease. As a result, cholesterol and blood sugar are considered important to the task and should be preserved. One issue with previous classification-based anonymization algorithms is that attributes are automatically selected based on metric scores, thus leading to possible over generalization of important attributes. If cholesterol and blood sugar were considered unimportant due to low scores, dataset utility would be reduced for the chosen mining task. There is a need for user-specified constraints which allow for the prioritization of attributes to suit a mining task. By providing such an option, instead of relying solely on privacy utility tradeoff metrics, a user can effectively constrain the attribute generalization process. In this case, physician clinical tasks including therapy, diagnosis, etiology, and prognosis are considered as the data mining purpose. The framework implements the UMLS Metathesaurus and Semantic Network to obtain attribute semantic types as well as relations, for specification of user constraints. Table 3.2 presents attribute semantic types and relations of the discussed sample patient dataset.

The semantic type corresponds to a high level concept of an attribute, for instance age or gender being an organism attribute. Additionally, the semantic relation represents the relationship between two concepts such as age being *associated with* heart disease or blood sugar *affecting* heart disease. A user has the choice of constraining attributes based on their semantic type, semantic relation or both. To constrain the anonymization process for an etiology task, the semantic relation *causes* can be chosen so that both blood sugar and cholesterol attributes are given a higher priority. For example, Table 3.3 shows a possible ranking of attributes with sample utility scores when *causes* is set as the semantic relation constraint. In this scenario, although cholesterol and blood sugar have lower scores as compared to age or gender,

Table 3.2: Attribute semantic types and relations

Attribute	**Semantic Type**	**Semantic Relation**
Age	Organism Attribute	Organism Attribute *result_of* Disease or Syndrome Organism Attribute *associated_with* Disease or Syndrome
Gender	Organism Attribute	Organism Attribute *result_of* Disease or Syndrome Organism Attribute *associated_with* Disease or Syndrome
Race	Population Group	Population Group *associated_with* Disease or Syndrome
Cholesterol	Biologically Active Substance	Biologically Active Substance *affects* Disease or Syndrome Biologically Active Substance *causes* Disease or Syndrome Biologically Active Substance *complicates* Disease or Syndrome
Blood Sugar	Carbohydrate	Carbohydrate *causes* Disease or Syndrome Carbohydrate *affects* Disease or Syndrome
Heart Disease (Target)	Disease or Syndrome	

they are prioritized first due to the constraints set.

Table 3.3: Semantic relation attribute constraint

Rank	4	3	5	1	2
Score	0.75	0.8	0.2	0.5	0.4
Attribute	Age	Gender	Race	Cholesterol	Blood Sugar
Type	Organism Attribute	Organism Attribute	Population Group	Biologically Active Substance	Carbohydrate
Relation	Result of	Result of	Associated with	Affects	*Causes*
	Associated with	Associated with		*Causes*	Affects
				Complicates	

If patient related attributes are of more importance towards a task, the semantic type of *organism attribute* can be selected instead. Table 3.4 presents attribute rankings based on the *organism attribute* semantic type constraint. In this case, age and gender have been ranked according to their scores as well as semantic constraints, while

remaining attributes are prioritized solely based on their scores.

Table 3.4: Semantic type attribute constraint

Rank	2	1	5	3	4
Score	0.75	0.8	0.2	0.5	0.4
Attribute	Age	Gender	Race	Cholesterol	Blood Sugar
Type	*Organism Attribute*	*Organism Attribute*	Population Group	Biologically Active Substance	Carbohydrate
Relation	Result of	Result of	Associated with	Affects	Causes
	Associated with	Associated with		Causes	Affects
				Complicates	

In a situation where the majority of attributes possess similar relations, both semantic type and relation can be used as constraints. For example, the attributes age, gender and race, though having different semantic types are all *associated with* heart disease. Assume that a user intends to solve a task involving all attributes *associated with* heart disease, and also requires a patient's race. In this case, *population group* and *associated with* can be chosen as the semantic type and relation constraints. By doing so, the race attribute would be given first priority while age and gender are prioritized next. The resulting rankings are shown in Table 3.5.

Table 3.5: Semantic type and relation attribute constraint

Rank	3	2	1	4	5
Score	0.75	0.8	0.2	0.5	0.4
Attribute	Age	Gender	Race	Cholesterol	Blood Sugar
Type	Organism Attribute	Organism Attribute	*Population Group*	Biologically Active Substance	Carbohydrate
Relation	Result of	Result of	*Associated with*	Affects	Causes
	Associated with	*Associated with*		Causes	Affects
				Complicates	

These examples indicate the possible benefits of constraining the anonymization process based on attribute semantic types or relations. By enforcing constraints, important attributes related to specific tasks can be preserved according to user needs,

therefore enhancing actionability of mining results.

3.3 Correlation-based Metric

Even with user constraints, generalization algorithms still require privacy or utility measures for optimal anonymization. Past works for classification tasks have implemented metrics involving InfoGain which are reminiscent of decision tree construction. Basically, InfoGain shows how much information would be gained when selecting an attribute for splitting. When creating a decision tree, the purest attribute which provides the best classification would be chosen. The InfoGain score is calculated as shown in Equation 3.3.1.

$$InfoGain(v) = E(T[v]) - \sum_{c} \frac{|T[c]|}{|T[v]|} E(T[c]) \tag{3.3.1}$$

T[v] denotes the set of records in a Table T generalized to the value v while *T[c]* denotes a set of records specialized to a child value c. $|T[v]| = \sum_c |T[c]|$, where $c \in$ child(v). The entropy of *E(T[x])* is shown in Equation 3.3.2.

$$E(T[x]) = -\sum_{cls} \frac{freq(T[x], cls)}{|T[x]|} \times log_2 \frac{freq(T[x], cls)}{|T[x]|} \tag{3.3.2}$$

freq(T[x], cls) denotes the number of records in *T[x]* having the class *cls*. *E(T[x])* measures the entropy or impurity of classes in *T[x]* while *InfoGain(v)* measures the reduction of entropy after refinement.

Referring to the previous scenario, imagine a situation where no constraints are specified and users rely purely on the anonymization algorithm. Metrics incorporating InfoGain focus on attribute purity instead of correlation, and as a result, important attributes related to the task may be over generalized if deemed impure. Furthermore, prior metrics have only evaluated attributes based on their characteristics within a dataset. What this means, is that determining attribute associations solely on their distributions in a dataset may fail to reflect real world relationships. Therefore, in DDDM, the inclusion of external domain knowledge such as ontologies or literature sources can aid in obtaining more actionable results. Although both DGH creation and user constraint components integrate the UMLS,

interesting rules may not necessarily be derived from domain ontologies alone [80]. Consequently, it is important to implement a combination of both domain-specific literatures and ontologies into the anonymization process.

The Mutual Information Measure (MIM) is often used to estimate dependencies or strength of associations between co-occurring variables within literature. It is a straightforward and well-established means of measuring information content between two terms [85]. The MIM formula is shown in Equation 3.3.3.

$$MIM(A,B) = log_2 \frac{P_{AB}}{P_A \cdot P_B} \tag{3.3.3}$$

P_A and P_B indicate the probability of terms A or B occurring in a given literature while P_{AB} signifies the probability of terms A and B co-occurring together. To avoid a negative score, the log function is removed as performed by Wren [85] and Hu, Zhang & Zhou [86]. The modified MIM equation is shown in Equation 3.3.4.

$$MIM(A,B) = \frac{P_{AB}}{P_A \cdot P_B} \tag{3.3.4}$$

By using MIM, users are given the ability to determine attribute relationships with real world domain knowledge. Furthermore, it has been shown that MIM outperforms other methods in ranking rare terms [86]. A collection of literature in a particular field can be seen as an embodiment of the entire domain knowledge, just as biomedical literatures are a representation of knowledge in the biomedical field [87]. The Pubmed search engine is capable of retrieving over 19 million biomedical article citations from the U.S. National Library of Medicine (MEDLINE database) as well as other journals, and is directly accessed by the framework to evaluate attribute correlations.

For example, based on Figure 3.4, assume that term A corresponds to cholesterol while term B represents heart disease. Both terms are queried on the MEDLINE database in order to retrieve their literatures for processing. The association strength between cholesterol and heart disease would then be determined by counting the number of literatures containing both words, with a high count implying the existence of a potentially strong relation. One disadvantage with such an approach is that no specifics are given as to how these attributes are related. Moreover,

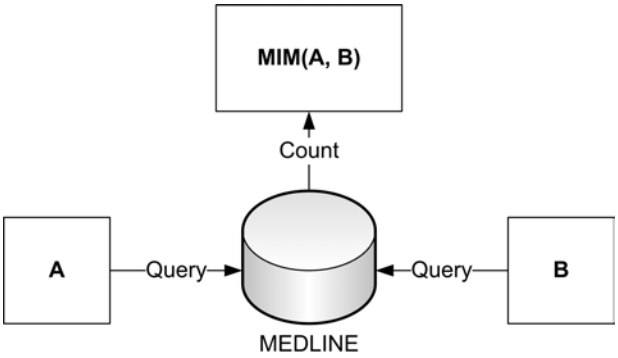


Figure 3.4: Calculating MIM score

false positives may arise since co-occurring terms with no apparent connections may be included as relationships [85]. To alleviate this issue, the framework implements search filters to improve retrieval of relevant articles from MEDLINE. [88], [89] developed search filters capable of constraining literatures to four types of articles (therapy, diagnosis, etiology and prognosis), thereby reducing the number of irrelevant publications as well as improving search results for clinical tasks. Table 3.6 describes a potential ranking of attributes based on their InfoGain scores.

Table 3.6: InfoGain score attribute ranking

Rank	2	1	5	3	4
InfoGain Score	0.75	0.8	0.2	0.5	0.4
Attribute	Age	Gender	Race	Cholesterol	Blood Sugar
Type	Organism Attribute	Organism Attribute	Population Group	Biologically Active Substance	Carbohydrate
Relation	Result of	Result of	Associated with	Affects	Causes
	Associated with	Associated with		Causes	Affects
				Complicates	

As can be seen, attributes are prioritized in descending order from the highest InfoGain score (gender) to the lowest (race). Since InfoGain measures the purity of attributes for classification purposes, it may not provide the best selection for domain-driven tasks.

Table 3.7 shows attribute rankings based on MIM scores which have been calculated from attribute correlations in literature.

Table 3.7: MIM score attribute ranking

Rank	3	4	5	1	2
MIM Score	12	7	2	23	15
Attribute	Age	Gender	Race	Cholesterol	Blood Sugar
Type	Organism Attribute	Organism Attribute	Population Group	Biologically Active Substance	Carbohydrate
Relation	Result of	Result of	Associated with	Affects	Causes
	Associated with	Associated with		Causes	Affects
				Complicates	

Several attributes have been prioritized differently compared to InfoGain rankings. For example, cholesterol and blood sugar are now the top ranking attributes due to stronger correlations in external domain knowledge. Since a different ranking does not prove that MIM is a better score than InfoGain, the framework combines both to improve attribute selection.

3.4 Privacy Model and Anonymization Algorithm

Numerous privacy models have been proposed to thwart privacy threats caused by linkage attacks. These threats as mentioned previously, allow an attacker to link or infer individual information from a published dataset. The most commonly used model, *k*-anonymity, though capable of preventing such attacks, falls victim to the curse of dimensionality [90]. Experiments have shown that applying *k*-anonymity to a large number of QID attributes (high dimensional data) can significantly degrade data quality, thus leading to inferior data mining results. Furthermore, the *k*-anonymity model is still susceptible to threats such as attribute linkage or background knowledge attacks. Machanavajjhala et al. [23] has shown that *k*-anonymous datasets, though satisfying privacy requirements, may still allow attackers to infer certain information. Moreover, previous anonymization models require that every QID attribute be generalized to satisfy their privacy requirements. To overcome these issues, the framework implements the *LKC*-privacy model and TDS algorithm as used in [37]. In real world privacy attacks, it may be difficult

for an adversary to acquire all knowledge regarding their target since it requires a non-trivial effort to gather information. Because of this, it can be assumed that an adversary's prior knowledge is bounded by at most L values of QID attributes. The *LKC*-privacy model reflects this situation by allowing users to specify the maximum number of prior knowledge an adversary possesses. Additionally, *LKC*-privacy guarantees the probability of successful attacks to be $\leq 1/K$ for identity linkage and $\leq C$ for attribute linkage, thereby providing protection similar to traditional models such as k-anonymity and l-diversity. The computational complexity of *LKC*-privacy is NP-hard as it represents an instance of the (α, k)-anonymity problem, with $\alpha = C$ and $k = K$ [37].

To achieve *LKC*-privacy the TDS algorithm is incorporated for dataset anonymization. Basically, the idea is to generalize all attributes to their topmost values based on their respective DGHs, and then, sequentially specializing them by their scores. A specialization is performed by replacing a parent value with a child value. For example, the topmost value for gender may be *person* or *any*. Once a specialization has been performed, the value would be replaced with either *male* or *female*. A specialization is valid only if it results in a dataset satisfying the specified privacy requirements. Each specialization decreases privacy but increases data utility as values become more specific. Both InfoGain and MIM are used to score the goodness of a specialization on a particular attribute. The reason for this is to ensure that utility gained reflects external knowledge shown by literature co-occurrences, and predictive information contained within the dataset. Furthermore, although able to determine attribute relationships, MIM does not take into account attribute values when searching for occurrences. Consequently, InfoGain is used to measure value changes after specialization while MIM ranks correlated attributes.

Algorithm 1 provides an overview of the constrained correlation-based anonymization algorithm. Initially, all QID attribute values in the dataset are generalized to their topmost values. Next, both MIM and InfoGain scores are calculated for each attribute and their values. MIM scores are evaluated based on the relationship between a certain attribute with the selected target attribute. As for InfoGain, attribute values are evaluated based on their purity. During the first iteration, attribute semantic types and relations are compared with user specified constraints. Attributes with matching constraints would be considered depending

Algorithm 1 Constrained correlation-based anonymization algorithm

Input:
Raw dataset
Domain generalization hierarchy

Output:
Anonymized dataset

Procedures:
Initialize every value in dataset to topmost value;
Calculate attribute MIM score towards target attribute;
Calculate attribute value InfoGain score;
while attributes can be specialized **do**
    Check attributes semantic type and relation;
    if match user constraints **then**
        Find best attribute for specialization based on MIM score;
        Find best attribute value for specialization based on InfoGain score;
        Perform specialization on attribute;
        Update score and validity of attribute;
    end if
end while
while remaining attributes can be specialized **do**
    Find best attribute for specialization based on MIM score;
    Find best attribute value for specialization based on InfoGain score;
    Perform specialization on attribute;
    Update score and validity of attribute;
end while
Output anonymized dataset;

on their scores. Attributes with high MIM and InfoGain scores would be given first priority during specialization. Once completed, affected attribute scores are updated along with their validity. During the second iteration, remaining attributes with differing semantic types and relations would be specialized if possible. The algorithm terminates when any further specialization leads to a violation of the *LKC*-privacy model. An important characteristic of this algorithm is the anti-monotonic property of the *LKC*-privacy model. If a generalized table violates *LKC*-privacy before any specialization, it would remain violated since specialization does not increase the size of equivalence classes. Therefore, this guarantees a sub-optimal solution once specialization has ended.

There are several differences between the proposed algorithm and existing TDS algorithm. First, a separate module is implemented due to the addition of the MIM score. This component enables the querying of MEDLINE literatures for

evaluation of attribute associations. Second, instead of relying purely on InfoGain, attribute constraints and MIM score are added to allow for prioritization of important attributes. Because of this, the anonymization process is divided into two specialization phases, with the first iteration focusing on prioritized attributes while the second on remaining unconstrained attributes.

3.5 Framework Design

A detailed framework design explaining user actions in the anonymization process is presented in Figure 3.5.

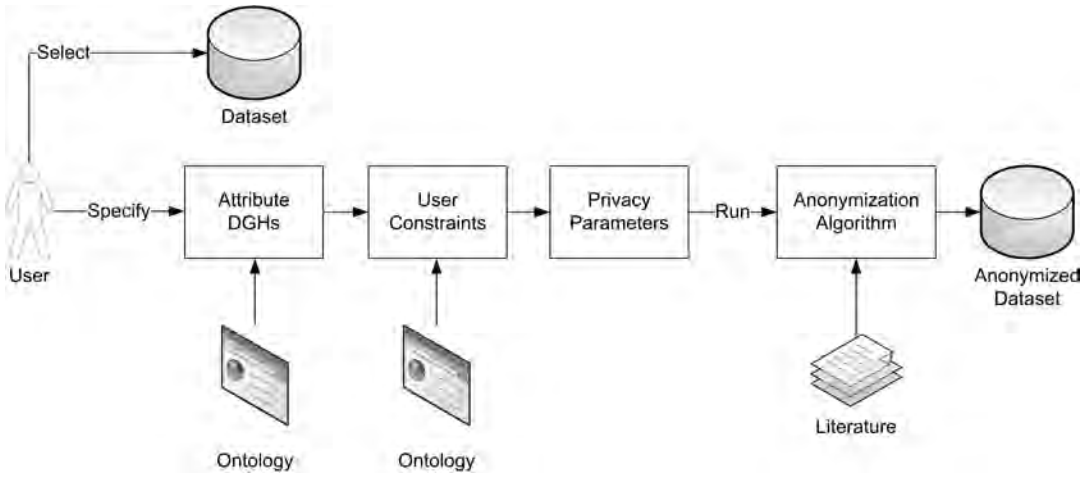


Figure 3.5: Proposed framework design

The user starts by selecting a dataset to anonymize for DDDM outsourcing purposes. Next, attributes are categorized or removed depending on the intended task. For selected QID attributes, DGHs are manually created with help from the UMLS ontology. To achieve this, attribute semantic mappings such as the one shown in Table 3.1 are obtained through the UMLS Knowledge Source Server (UMLSKS). By entering attribute names into the online metathesaurus search engine, the user can acquire important information such as attribute semantic types, relations and ranges. Once DGHs are made, the user can choose to specify constraints based on attribute semantic types or relations which were previously acquired from the UMLSKS. It is also at this stage where the choice between a pure InfoGain or combined MIM score is given. A final step before anonymization is the specification of privacy parameters. Here, the user can control both protection and utility requirements through manipulation of L, K and C parameters. After previous phases have

been completed, the anonymization algorithm is run. If the combined MIM score was selected, during the specialization process, the MEDLINE database would be automatically queried to calculate attribute correlations. Once privacy and utility requirements are met, the algorithm terminates and an anonymized dataset is produced.

3.6 Technological Architecture

The proposed framework was designed as a semi-automated anonymization tool capable of protecting data while maintaining utility. It was implemented with both C++ and Java programming languages to allow for external domain knowledge integration. Figure 3.6 describes the technological architecture.

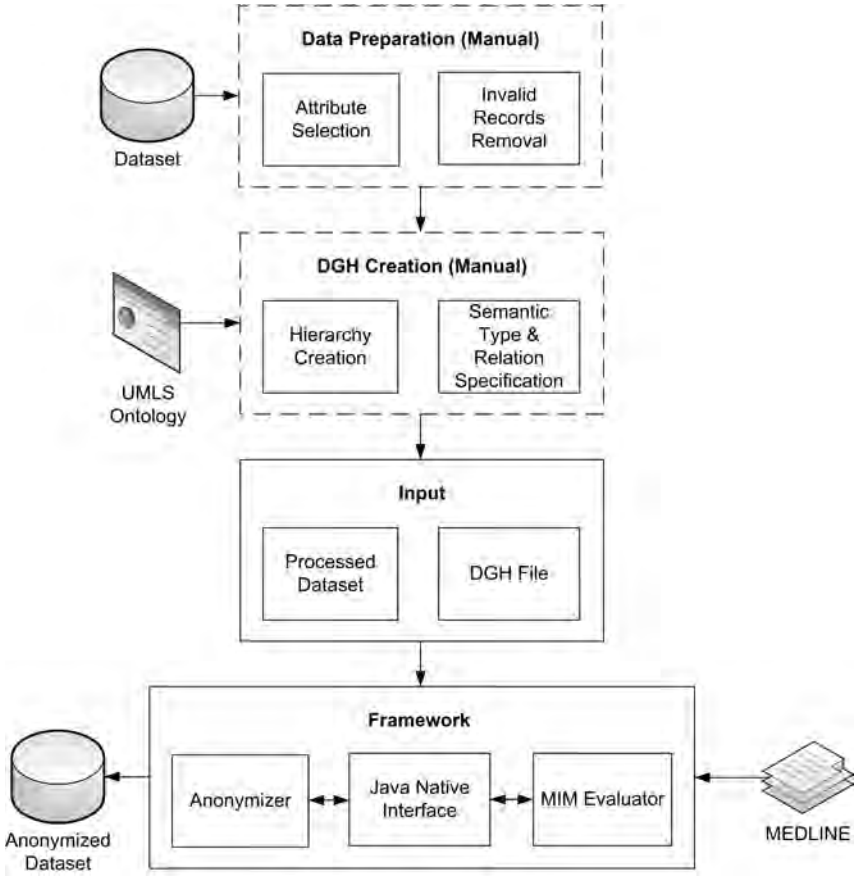


Figure 3.6: Proposed framework technological architecture

The data preparation phase is to be performed manually by the user. In this stage a subset of attributes are selected from a dataset and invalid records are removed. Datasets in comma-separated values (CSV) format were filtered using Microsoft Excel

and records containing blank or invalid values were removed.

During the DGH creation stage, the user is also required to manually created hierarchies for each attribute. Furthermore, attribute semantic types and relations are needed as well. The UMLS Metathesaurus and Semantic Network can be accessed through an online interface on the official UMLSKS website. Apart from this option, a downloadable release of UMLS archives is available together with the MetamorphoSys application which allows browsing of the Metathesaurus.

Figure 3.7 shows a sample DGH file containing a target attribute as well as age, sex, chest pain, and blood pressure attributes. There are three sections that must be specified for each attribute. First, the attribute name and its value type which can be either discrete or continuous. Second, the attribute semantic type and semantic relation towards the target attribute. Third, the hierarchy for value generalization and specialization. This file would be used together with the processed dataset as input to the anonymization framework.

```
classes:discrete
Disease-or-Syndrome:NONE
{Any Class {yes} {no}}

age:continuous
Organism-Attribute:Result-of, Associated-with
{25-80 {25-50 {25-35 {25-30} {30-35}} {35-50 {35-40} {40-45} {45-50}} {50-80 {50-65 {50-55} {55-60} {60-65}} {65-80 {65-70} {70-75} {75-80}}}}

sex:discrete:generalization
Organism-Attribute:Result-of, Associated-with
{any {male} {female}}

chest-pain:discrete:generalization
Sign-Or-Symptom:Diagnoses, Evaluation-of, Manifestation-of, Associated-with
{any {angina {typical-angina} {atypical-angina}} {non-angina {non-anginal-pain} {asymptomatic}}}

blood-pressure:continuous
Organism-Function:Result-of, Affects
{90-205 {90-150 {90-100} {100-110} {110-120} {120-130} {130-140} {140-150}} {150-205 {150-160} {160-170} {170-180} {180-190} {190-205}}}

```

Figure 3.7: Sample DGH file

The framework consists of two major components which are the anonymization application (anonymizer) and MIM score evaluator. The anonymizer was modified from the existing TDS algorithm with added functionalities such as semantic constraints and MIM utility measure. The MIM evaluator was utilized to access MEDLINE literatures and calculate attribute scores. C++ programming language was used to create the anonymizer while the MIM evaluator was coded in Java. Since both modules were written in different programming languages, the Java Native Interface (JNI) was utilized to enable function calls from either application. To access literatures, the MIM evaluator integrated Java classes from Entrez Programming Utilities (eUtils) which provide entry to the MEDLINE database outside of the regular Pubmed web query interface. The three packages implemented were EInfo for

retrieving article counts, ESearch for obtaining citation identifiers, and EFetch to extract article details such as title, authors, abstract and headings.

3.7 Summary

The proposed framework aims to protect privacy through the *LKC*-privacy model and maintain utility for DDDM purposes. It contributes three major components, ontology-based DGH, user-specified constraints and correlation-based anonymization algorithm. These components integrate domain knowledge in the form of the UMLS ontology and external literatures from MEDLINE database. During DGH creation, attribute values are mapped to UMLS concepts to retain value meanings. As for user constraints, attribute semantic types and relations are used to prioritize important attributes for a particular mining task. To enhance attribute selection, MIM is combined with InfoGain as a score for measuring attribute correlations in literature.

CHAPTER 4

Experiments and Results

Experiments were performed to evaluate the framework in terms of data quality from the traditional and domain driven data mining perspective. First, from a traditional standpoint, classification accuracy was compared between the original TDS algorithm and proposed framework. The objective of this experiment was to discover whether models generated from anonymized datasets retained similar accuracies with their original counterparts. Also, both InfoGain and MIM scores were compared to determine their ability in maintaining technical significance. From a DDDM viewpoint, the capability of ontology-based DGHs in preserving semantic meanings was tested. The aim here was to compare value meaningfulness of decision tree rules generated from ontology-based and dynamically created DGHs. Next, the effectiveness of user constraints in maintaining task important attributes was examined. The point of such an experiment was to prove the need for constraints capable of prioritizing attributes for a particular purpose based on their relationship towards the target. Finally, the use of a correlation-based measure for improving attribute selection was analyzed. The intention was to show differences in attribute prioritization when utilizing InfoGain and MIM as scores. Efficiency and scalability experiments were not run as Mohammed et al. [37] had already shown the capabilities of the *LKC*-privacy model in anonymizing large datasets containing at least one million records, with total processing times running below 120 seconds. Since all datasets used for the experiments contained less than a million records, processing times were at a minimal.

4.1 Datasets

We utilized three datasets, Cath, Cleveland, and CHD_DB for testing. All records with missing or invalid values were removed from the datasets.

4.1.1 Cath

Cath, a real-life patient dataset was extracted from a cardiovascular database of a local health institution. The dataset contained 1,589 records of patients who have undergone cardiac catheterization, and has a total of 11 QID attributes with coronary artery obstruction being the target attribute. The Cath dataset attributes along with their semantic types and relations are described in Tables 6.1 and 6.4, in the Appendix section.

4.1.2 Cleveland

The Cleveland Heart Disease dataset which has been used in numerous studies was obtained from the UC Irvine Machine Learning Repository. In total, there are 297 patient records with 13 attributes being QIDs and heart disease being the target attribute. The Cleveland dataset attributes along with their semantic types and relations are described in Tables 6.2 and 6.5, in the Appendix section.

4.1.3 CHD_DB

The Coronary Heart Disease Database (CHD_DB) is a synthetic cardiovascular dataset which has been replicated from the original Framingham Heart Study data. 26,000 records were extracted for training as well as testing purposes with 8 attributes chosen as QID and heart disease as the target attribute. The CHD_DB dataset attributes along with their semantic types and relations are described in Tables 6.3 and 6.6, in the Appendix section.

4.2 Experiment Design

In general these steps were taken for dataset preparation and analysis:

1. Removal of records containing invalid or missing values. This was done to prevent any errors from occurring during anonymization and also to avoid bias in data mining results.

2. Renaming attribute names and values to improve comprehensibility. Certain datasets had short formed attribute names as well as numerically represented attribute values. Because of this, they were renamed to allow for easier interpretation of mined rules.

3. Creation of DGH file for dataset attributes. A sample is shown in Figure 3.7.

4. Setting of *LKC*-privacy model parameters, attribute semantic constraints and utility score. L, K and C parameters as well as the utility score must be defined before anonymization while constraints are optional.

5. Dataset anonymization. The dataset is anonymized based on the chosen parameters, constraints and score.

6. Classification model generation. RapidMiner is used to generate a decision tree model from both original and anonymized datasets.

7. Assessing classification accuracy. Each model is processed through RapidMiner to determine their average classification accuracy.

8. Analyzing classification model rules. The rules extracted from decision trees are compared manually to determine their utility and actionability.

4.3 Classification Accuracy

To assess classification accuracy, each dataset was generalized based on the *LKC*-privacy model. For Cath and Cleveland datasets, parameters were set at L = 2, 4, 6 for adversary knowledge and K = 2, 4, 6 for anonymity thresholds. On the other hand, for CHD_DB, L parameter remains the same while K = 20, 40, 60, 80, 100. Both Cath and Cleveland possessed a significantly lower number of records compared to CHD_DB, therefore a decrease in K values were required to prevent total dataset generalization. The C parameter which controls diversification of values was not set because it is claimed to contradict the homogeneous values requirement thus lowering classification accuracy [74]. Although the specified parameters may

seem as a relaxation of the *K*-anonymity model, in real world situations, an adversary would not necessarily possess knowledge of all QID attributes. Therefore, these settings aim to minimize attacker inferences for realistic scenarios while maintaining rule actionability. We utilized RapidMiner, an open-source data mining system for classifying each dataset. The Decision Tree learner which works similarly to the well-known C4.5 classifier was chosen to create classification models. Ten fold cross-validations were performed on Cath and Cleveland datasets to obtain an average accuracy value while CHD_DB was split in half for training and testing. The average accuracy results for each dataset are shown in the figures below.

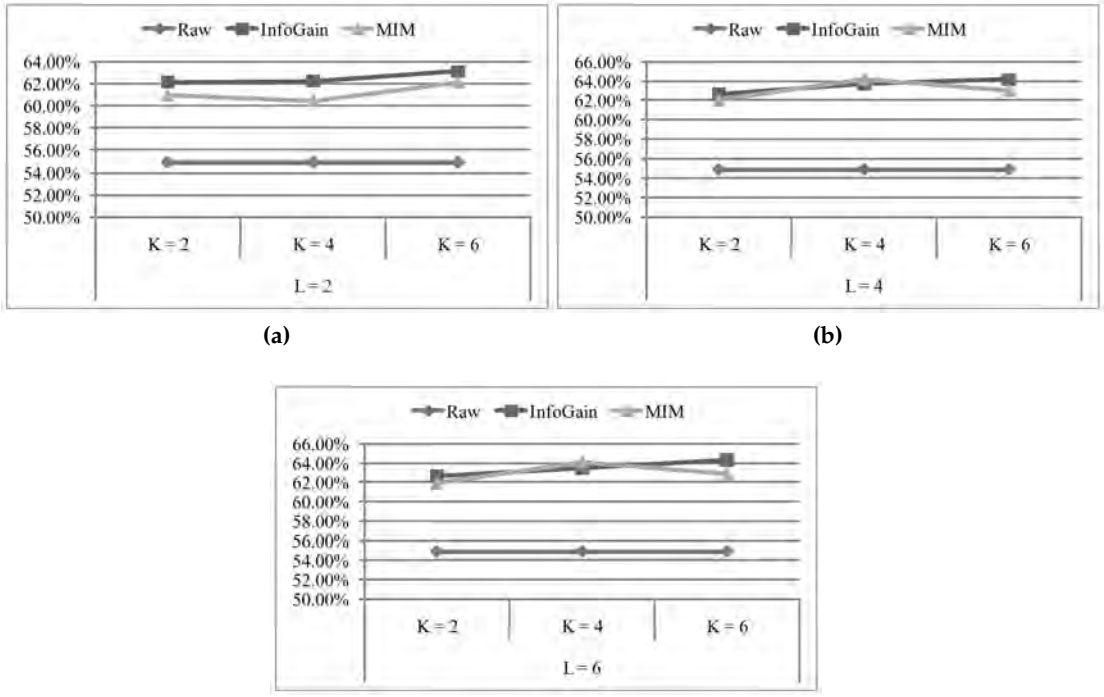


Figure 4.1: Cath dataset classification accuracy

Figures 4.1, 4.2 and 4.3 show the classification accuracy of each dataset with increasing *L* and *K* parameters. Each line in the graph represents the baseline, InfoGain and MIM accuracy levels. In theory, classification accuracy should decrease as privacy is increased, but as seen in Figures 4.1, 4.2a, 4.3a and 4.3b, InfoGain models show a gain in utility when *K* rises. Furthermore, both InfoGain and MIM models surpass the baseline accuracy in Figures 4.1 and 4.2. This can be due to noise reduction in data caused by generalization, thereby forming a better classification structure. Additionally, in most cases, models created with MIM score would possess

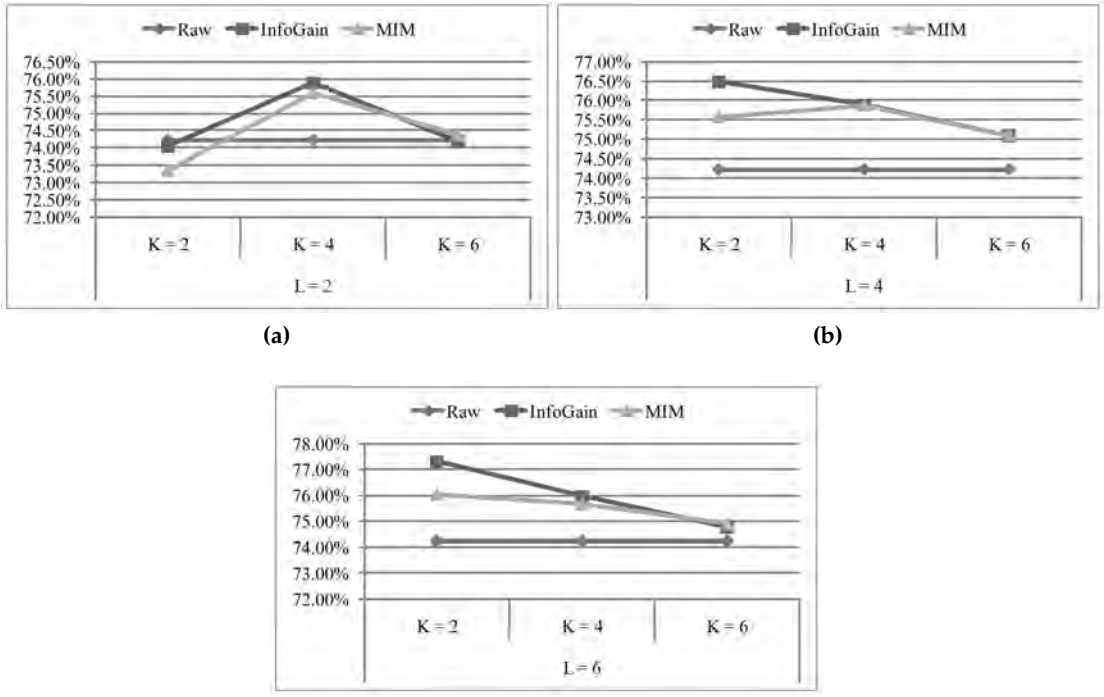


Figure 4.2: Cleveland dataset classification accuracy

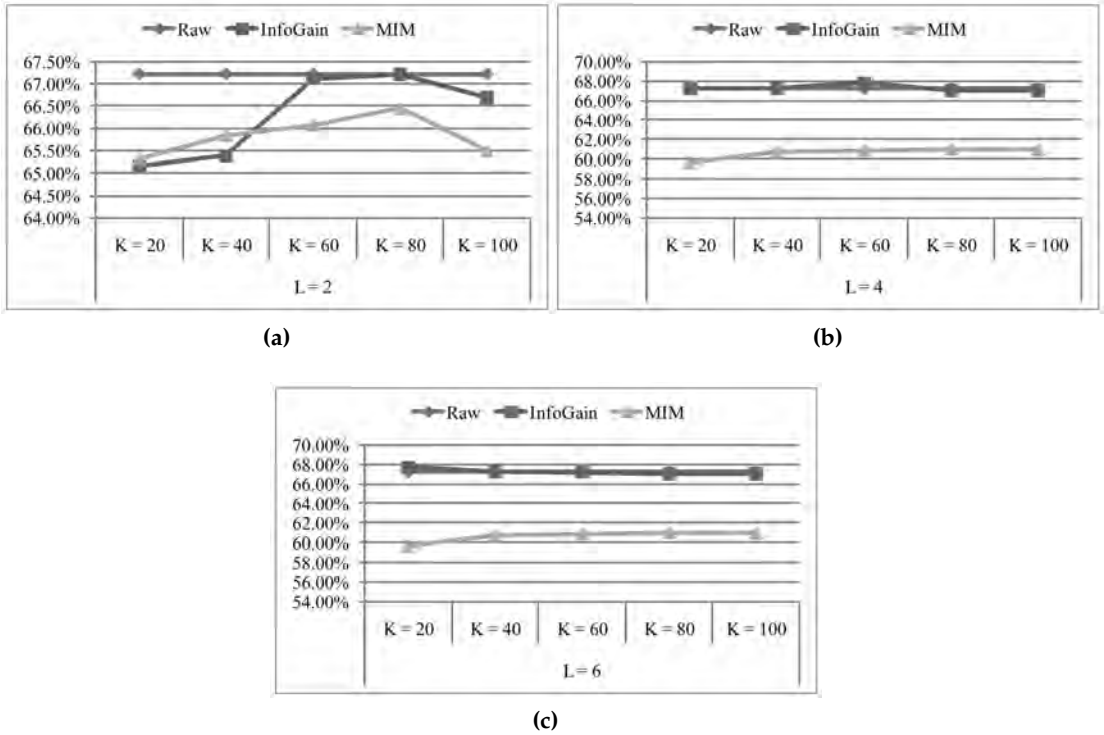


Figure 4.3: CHD_DB dataset classification accuracy

lower accuracy since the RapidMiner Decision Tree learner splits attributes based on InfoGain instead of MIM. From the results, it can be assumed that each generalized

dataset would contain sufficient data mining utility since they boast higher or similar accuracies with the baseline. But, as mentioned before, technical significance such as accuracy may not necessarily indicate usefulness of discovered models.

4.4 Basic VS. Ontology-based Domain Generalization Hierarchy

To determine the effectiveness of ontology-based DGHs, we first selected an attribute from each dataset to be compared. For both Cath and Cleveland datasets, the age attribute was chosen and an ontology-based DGH was created similar to the one shown in Figure 3.3. As for CHD_DB, cholesterol was picked for comparison. To perform the test, each dataset was anonymized twice, with and without an ontology-based DGH. Since the TDS algorithm implemented a hierarchy-free model, the baseline DGH was dynamically created without the need for user specification. Both versions of datasets were then mined with RapidMiner to generate a decision tree. The resulting rules from each dataset are shown in Table 4.1.

Table 4.1 shows part of the decision trees generated from the anonymized datasets. In all cases, it can be seen that both types of DGHs manage to retain value meanings for the selected attributes. But the key difference here is that using a dynamically or manually created DGH, without domain knowledge, can lead to less precise rules. For example, in Table 4.1(a) patients are grouped under the age range *66 - 95* which is a cross between *elderly* and *ages, 80 and over* concepts. Furthermore, Table 4.1(b) shows a range of *25 - 55* which is a combination of *adult* and *middle age* ranges. Table 4.1(c) suffers from the same problem since the cholesterol range of *233 - 252* falls under the concepts of *normal* and *high*. For rules created from the ontology-based DGH, attribute ranges remain within their preset UMLS semantic mappings no matter the level of generalization. Therefore, if decisions were to be made, ontology-based DGH created rules would provide greater precision and meaning for real world usage.

4.5 User Constraints Evaluation

To evaluate the potential of user constraints for maintaining important attributes, we analyzed rules obtained from decision tree results. The actionability of a model can

Table 4.1: Decision tree rules

Dynamic DGH	Ontology-based DGH
(a) Cath (L = 2, K = 2)	
Age = 66.00-95.00 | Post-Diastolic Blood Pressure = 5.00-68.00 | | Pre-Diastolic Blood Pressure = 30.00-67.00 | | | Body Surface Area = 0.80-1.49: yes. | | | Body Surface Area = 1.49-1.63 | | | | Post-Heart Rate = 30.00-78.00: yes. | | | | Post-Heart Rate = 78.00-165.00: no.	*Age = adolescent: no.* *Age = ages-80-and-over: yes.*
(b) Cleveland (L = 2, K = 2)	
Heart Rate = 137.00-148.00 | *Age = 25.00-55.00* | | ST Depression = 0.00-1.80 | | | Angina = no: no. | | | Angina = yes: yes. | | ST Depression = 1.80-7.00: yes. | *Age = 55.00-64.00: yes.* | *Age = 64.00-80.00: no.*	Angina = no | *Age = 45-49* | | Sex = female: no. | | Sex = male | | | Heart Rate = 148.00-210.00: no. | | | Heart Rate = 70.00-148.00: yes. | *Age = 50-54: no.* | *Age = 55-59: no.*
(c) Framingham (L = 2, K = 20)	
systolic-blood-pressure = 112.00-121.00 | *cholesterol = 233.00-252.00* | | smoking-habit = cigar-or-pipe: no. | | smoking-habit = never-smoked: yes. | | smoking-habit = stopped: no. | | smoking-habit = tobacco(<20/day): no. | | smoking-habit = tobacco(>=20/day): no.	systolic-blood-pressure = 112.00-121.00 | *cholesterol = 230-239: no.* | *cholesterol = 240-249: no.* | *cholesterol = 250-259: no.*

be determined by assessing the attributes contained in each rule. For instance, when a decision tree is created, certain attributes may be missing due to generalization. Because of this, if the model is applied for a particular task requiring the missing attributes, it may provide low actionability. We first began by creating decision trees from each dataset with previously set parameters and using InfoGain as score. Two sets of models were created for each parameter, one with attribute constraints and the other without. Next, remaining attributes in the models were analyzed and compared.

4.5.1 Cath

As shown in Table 4.2, all attributes are preserved in the raw Cath dataset model. At each L and K increment, less attributes remain in the resulting decision tree due to

generalization. For example, at $L = 4$ and $L = 6$, only three attributes, gender (1), age(3) and body mass index(5) remain in the decision tree. As a result, if this model were to be used for a task requiring other attributes, it would be deemed useless. Moreover, as presented previously in Figure 4.1, the accuracy for these models are higher than the baseline, and yet differs significantly in actionability. This supports the idea that accuracy, though indicative of a model's predictive capabilities, may not necessarily prove that actionable rules exist.

Table 4.2: Cath decision tree attributes

Attributes		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Raw		X	X	X	X	X	X	X	X	X	X	X
L=2	K=2	X	X	X	X	X		X	X	X		
	K=4	X		X		X	X	X	X	X		
	K=6	X		X		X	X	X	X			
L=4	K=2	X		X		X						
	K=4	X		X		X						
	K=6	X		X		X						
L=6	K=2	X		X		X						
	K=4	X		X		X						
	K=6	X		X		X						

Semantic Type

Table 4.3 presents a decision tree constrained by the attribute semantic type *population group*. When comparing both decision trees, it can be seen that the race(2) attribute is missing from parameters $L = 2$ and $K = 4$ onwards in the unconstrained tree. On the other hand, in the constrained tree, the race attribute still remains at parameters $K = 4$.

Table 4.4 shows another decision tree with *clinical attribute* chosen as the semantic type constraint. In this scenario, although five attributes have been constrained, only body mass index(5) remains in the resulting model.

4.5.2 Cleveland

Table 4.5 shows a similar trend whereby the majority of attributes are preserved in the raw model but reduces as privacy parameters increase. Also, accuracy levels are

Table 4.3: Cath InfoGain constrained (population group) decision tree attributes

Attributes		1	2	3	4	5	6	7	8	9	10	11
L=2	K=2	X	X	X	X		X	X	X			
	K=4	X	X		X							
	K=6											
L=4	K=2			X								
	K=4	X	X		X							
	K=6											
L=6	K=2			X								
	K=4	X	X		X							
	K=6											

Table 4.4: Cath InfoGain constrained (clinical attribute) decision tree attributes

Attributes		1	2	3	4	5	6	7	8	9	10	11
L=2	K=2					X						
	K=4					X						
	K=6					X						
L=4	K=2					X						
	K=4					X						
	K=6					X						
L=6	K=2					X						
	K=4					X						
	K=6					X						

similar or higher than the baseline. In this case, when $K = 6$, only two attributes, angina(9) and thallium test(13) remain.

Table 4.5: Cleveland decision tree attributes

Attributes		1	2	3	4	5	6	7	8	9	10	11	12	13
Raw		X	X	X	X	X			X	X	X	X	X	X
L=2	K=2	X				X	X	X	X	X	X		X	X
	K=4	X		X			X	X	X	X	X			X
	K=6									X				X
L=4	K=2							X					X	X
	K=4								X					X
	K=6									X				X
L=6	K=2							X					X	X
	K=4								X					X
	K=6									X				X

Semantic Type

Table 4.6 shows the remaining attributes of a constrained decision tree with *sign or symptom* semantic type constraint. The attributes chest pain(3) and angina(9) fall under this semantic type. One major difference in the constrained decision tree can be noticed in parameters $L = 4$ and $L = 6$. The chest pain attribute is non-existent in unconstrained models at these levels. But, by specifying constraints, it can be preserved at all levels thus allowing users to apply the rules for tasks requiring that particular attribute.

Table 4.6: Cleveland InfoGain constrained (sign or symptom) decision tree attributes

Attributes		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
L=2	K=2	X		X		X		X	X	X		X		X
	K=4	X	X	X		X		X	X					
	K=6	X	X	X		X		X	X	X	X			
L=4	K=2			X										
	K=4			X										
	K=6			X										
L=6	K=2			X										
	K=4			X										
	K=6			X										

Table 4.7 presents a constrained tree with *diagnostic procedure* as the constraint. In comparison, both the constrained and unconstrained trees are alike for the prioritized attributes.

Table 4.7: Cleveland InfoGain constrained (diagnostic procedure) decision tree attributes

Attributes		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
L=2	K=2	X				X	X	X	X	X	X		X	X
	K=4	X		X			X	X	X	X	X			X
	K=6									X				X
L=4	K=2	X						X					X	X
	K=4								X					X
	K=6									X				X
L=6	K=2	X						X					X	X
	K=4								X					X
	K=6									X				X

Semantic Relation

In Table 4.8, all attributes sharing a *result of* relation to the target attribute (heart disease), would be prioritized during anonymization. Consequently, the attributes age(1), sex(2) and heart rate(8), are maintained at most parameter levels in the constrained decision tree.

Table 4.8: Cleveland InfoGain constrained (result of) decision tree attributes

Attributes		1	2	3	4	5	6	7	8	9	10	11	12	13
L=2	K=2		X				X	X	X	X	X			
	K=4	X	X						X		X			
	K=6								X					
L=4	K=2	X	X						X					
	K=4								X					
	K=6								X					
L=6	K=2	X	X						X					
	K=4								X					
	K=6								X					

Table 4.9 presents a similar case to Table 4.7 where both constrained and unconstrained models are the same.

Table 4.9: Cleveland InfoGain constrained (diagnoses) decision tree attributes

Attributes		1	2	3	4	5	6	7	8	9	10	11	12	13
L=2	K=2	X				X	X	X	X	X	X		X	X
	K=4	X		X			X	X	X	X	X			X
	K=6									X				X
L=4	K=2							X					X	X
	K=4								X					X
	K=6									X				X
L=6	K=2							X					X	X
	K=4								X					X
	K=6									X				X

4.5.3 CHD_DB

The CHD_DB models as shown in Table 4.10 differs from both Cath and Cleveland trees since most attributes are preserved when privacy requirements increase. Although this is so, three attributes, national origin(5), education(6) and drinking

habit(8) are still missing at certain parameter levels.

Table 4.10: CHD_DB decision tree attributes

Attributes		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Raw		X	X	X	X		X	X	X
L=2	K=20	X	X	X	X	X	X	X	
	K=40	X	X	X	X	X	X	X	
	K=60	X	X	X	X	X	X	X	
	K=80	X	X	X	X	X	X	X	
	K=100	X	X	X	X	X	X	X	
L=4	K=20	X	X	X	X	X	X	X	
	K=40	X	X	X	X			X	
	K=60	X	X	X	X		X	X	
	K=80	X	X	X	X			X	
	K=100	X	X	X	X			X	
L=6	K=20	X	X	X	X			X	
	K=40	X	X	X	X			X	
	K=60	X	X	X	X			X	
	K=80	X	X	X	X			X	
	K=100	X	X	X	X			X	

Semantic Type

The decision tree in Table 4.11 shows remaining attributes for a *finding* semantic type constraint. The affected attributes include education(6) and smoking habit(7). Compared to the unconstrained model, this tree preserves the education attribute at parameters $K = 40$ and above.

Table 4.12 presents a *clinical attribute* semantic type constrained tree. In this scenario, remaining constrained attributes are matching those in the unconstrained model except for national origin(5) which is missing from parameters $L = 4$ and $K = 20$.

Semantic Relation

Tables 4.13 and 4.14 show semantic relation constrained trees for *associated with* as well as *result of* relations. In this case, both models possess similar results to the unconstrained model with the exception of national origin(5) and education(6) attributes. The constrained trees are missing those attributes at higher privacy levels.

Table 4.11: CHD_DB InfoGain constrained (finding) decision tree attributes

Attributes		1	2	3	4	5	6	7	8
L=2	K=20	X	X	X	X	X	X	X	
	K=40	X	X	X	X	X	X	X	
	K=60	X	X	X	X	X	X	X	
	K=80	X	X	X	X	X	X	X	
	K=100	X	X	X	X	X	X	X	
L=4	K=20		X				X	X	
	K=40			X			X	X	X
	K=60						X	X	X
	K=80						X	X	X
	K=100	X	X	X			X	X	
L=6	K=20		X				X	X	
	K=40			X			X	X	X
	K=60						X	X	X
	K=80						X	X	X
	K=100	X	X	X			X	X	

Table 4.12: CHD_DB InfoGain constrained (clinical attribute) decision tree attributes

Attributes		1	2	3	4	5	6	7	8
L=2	K=20	X	X	X	X	X	X	X	
	K=40	X	X	X	X	X	X	X	
	K=60	X	X	X		X	X	X	
	K=80	X	X	X		X	X	X	
	K=100	X	X	X	X	X	X	X	
L=4	K=20		X	X				X	
	K=40		X	X				X	
	K=60		X	X				X	
	K=80		X	X				X	
	K=100		X	X				X	
L=6	K=20		X	X				X	
	K=40		X	X				X	
	K=60		X	X				X	
	K=80		X	X				X	
	K=100		X	X				X	

4.6 MIM Score Evaluation

Our main objective when evaluating the MIM score was not to prove its performance against InfoGain as a utility measure, but to demonstrate its abilities in selecting correlated attributes. We repeated the same experiments as done during user constraints evaluation with the only difference being the combination of MIM with

Table 4.13: CHD_DB InfoGain Constrained (Associated With) Decision Tree Attributes

Attributes		1	2	3	4	5	6	7	8
L=2	K=20	X	X	X	X	X	X	X	
	K=40	X	X	X	X	X	X	X	
	K=60	X	X	X	X	X	X	X	
	K=80	X	X	X	X	X	X	X	
	K=100	X	X	X	X	X	X	X	
L=4	K=20		X	X	X			X	
	K=40		X	X	X			X	
	K=60		X	X	X			X	
	K=80		X	X	X			X	
	K=100		X	X	X			X	
L=6	K=20		X	X	X			X	
	K=40		X	X	X			X	
	K=60		X	X	X			X	
	K=80		X	X	X			X	
	K=100		X	X	X			X	

Table 4.14: CHD_DB InfoGain Constrained (Result Of) Decision Tree Attributes

Attributes		1	2	3	4	5	6	7	8
L=2	K=20	X	X	X	X	X	X	X	
	K=40	X	X	X	X	X	X	X	
	K=60	X	X	X	X	X	X	X	
	K=80	X	X	X	X	X	X	X	
	K=100	X	X	X	X	X	X	X	
L=4	K=20		X	X	X			X	
	K=40		X	X	X			X	
	K=60		X	X	X			X	
	K=80		X	X	X			X	
	K=100		X	X	X			X	
L=6	K=20		X	X	X			X	
	K=40		X	X	X			X	
	K=60		X	X	X			X	
	K=80		X	X	X			X	
	K=100		X	X	X			X	

InfoGain. MIM was utilized for attribute selection while InfoGain measured the purity of attribute values during specialization.

4.6.1 Cath

Table 4.15 shows a MIM created unconstrained decision tree with several differences compared to the pure InfoGain model in Table 4.2. The InfoGain tree prioritizes body mass index(5), pre-heart rate(6) and pre-diastolic blood pressure(8) attributes while the MIM model focuses on body surface area(4), post-heart rate(7) and post-diastolic blood pressure(9) attributes. It can be assumed that remaining attributes in the MIM model possess a stronger relationship towards heart disease compared to the InfoGain tree. To prove this point, take for example the attributes body mass index and body surface area which are used for dividing measurements among individuals. According to [91], the American Society of Echocardiography recommends the use of body surface area as an indexing method when predicting heart failure. Because of this, the MIM score prioritizes body surface area since most literatures have used that attribute when performing prediction tasks.

Table 4.15: Cath MIM decision tree attributes

Attributes		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
L=2	K=2	X		X	X			X		X		
	K=4	X		X	X			X		X		
	K=6	X		X	X			X		X		
L=4	K=2	X		X	X					X		
	K=4	X		X	X			X				
	K=6	X		X	X			X				
L=6	K=2	X		X	X			X		X		
	K=4	X		X	X			X				
	K=6	X		X	X			X				

Table 4.16 details attribute rankings for both InfoGain and MIM scores. It should be noted that the InfoGain scores shown are based on fully generalized attribute values and are updated as values are specialized. The InfoGain tree ranks both body mass index and body surface area as the top three attributes but gives a slightly higher score to body mass index. On the other hand, the MIM tree provides a much higher score to body surface area compared to body mass index.

Table 4.16: Cath attribute scores

Rank	Attribute	InfoGain	Attribute	MIM
1	Age	0.049434	Post-Heart Rate	17.3243
2	*Body Mass Index*	*0.0157186*	Post-Diastolic Blood Pressure	6.37481
3	*Body Surface Area*	*0.0139282*	Post-Systolic Blood Pressure	6.37249
4	Race	0.00970125	*Body Surface Area*	*5.41926*
5	Post-Diastolic Blood Pressure	0.00705695	Age	1.88044
6	Post-Heart Rate	0.00633723	Gender	1.19236
7	Post-Systolic Blood Pressure	0.00542349	*Body Mass Index*	*0.87215*
8	Pre-Diastolic Blood Pressure	0.00533694	Race	0.535098
9	Pre-Systolic Blood Pressure	0.00432754	Pre-Heart Rate	0.0
10	Pre-Heart Rate	0.00283128	Pre-Diastolic Blood Pressure	0.0
11	Gender	0.0	Pre-Systolic Blood Pressure	0.0

4.6.2 Cleveland

The Cleveland MIM tree with a *result of* semantic relation constraint is shown in Table 4.17. When compared to Table 4.8, it can be seen that different attributes have been preserved. Table 4.8 which uses a pure InfoGain score indicates that age(1), sex(2) and heart rate(8) remain while Table 4.17 which uses the combined MIM and InfoGain score shows only age and heart rate. While it may seem that MIM created rules are less actionable since only two attributes remain, logically, age should have a higher correlation with heart disease compared to sex. It is known that for both genders, the risk of heart disease increases with age [92]. Therefore, prioritizing age over gender can benefit mining results.

Table 4.18 shows Cleveland attribute rankings based on InfoGain and MIM scores. For trees utilizing InfoGain, the prioritization of attributes can be easily seen through their scores. Heart rate has the highest score among the constrained attributes, therefore is specialized first. Next, age and sex share similar scores which cause them to be specialized after heart rate. Finally, since blood pressure possesses the lowest score, it has not been chosen for specialization due to a possible privacy breach

Table 4.17: Cleveland MIM constrained (result of) decision tree attributes

Attributes		1	2	3	4	5	6	7	8	9	10	11	12	13
L=2	K=2	X							X					
	K=4	X				X		X	X	X	X			
	K=6	X							X					
L=4	K=2	X							X					
	K=4	X							X					
	K=6	X							X					
L=6	K=2	X							X					
	K=4	X							X					
	K=6	X							X					

or minimal utility gains. MIM trees share a similar ranking with the exception of blood pressure being ranked second. Although it scores higher than age or sex in MIM, the low InfoGain score prevents any specialization. Lastly, age has a higher correlation to heart disease as compared to sex, and consequently is specialized. The main difference between both scores, as mentioned previously, is that InfoGain scores change whenever updates occur while MIM does not. This allows strongly correlated attributes to be focused on as long as they maintain adequate utility after specialization.

Table 4.18: Cleveland attribute scores

Rank	Attribute	InfoGain	Attribute	MIM
1	Thallium Test	0.210233	ST Depression	22.1838
2	Major Vessels Colored	0.184635	Angina	21.2427
3	Angina	0.132295	Thallium Test	20.9199
4	*Heart Rate*	*0.131957*	Resting ECG	20.2425
5	ST Depression	0.123189	Chest Pain	18.4808
6	Slope	0.108775	*Heart Rate*	*4.65253*
7	*Age*	*0.0631621*	Cholesterol	3.61077
8	*Sex*	*0.057874*	*Blood Pressure*	*3.20631*
9	Chest Pain	0.0559548	Major Vessels Colored	1.98734
10	Cholesterol	0.0170486	Slope	1.82277
11	*Blood Pressure*	*0.0163683*	*Age*	*1.71381*
12	Blood Sugar	0.0000072122	*Sex*	*1.36481*
13	Resting ECG	0.0	Blood Sugar	0.626439

4.6.3 CHD_DB

Table 4.19 presents a MIM decision tree with a *finding* semantic type constraint. Compared to Table 4.11, both trees preserve the constrained attributes education(6) and smoking habit(7). Although this is so, certain attributes such as cholesterol(1), systolic blood pressure(2), diastolic blood pressure(3) and left ventricular hypertrophy(4) are prioritized differently in both models.

Table 4.19: CHD_DB MIM constrained (finding) decision tree attributes

Attributes		1	2	3	4	5	6	7	8
L=2	K=20	X	X	X	X	X	X	X	
	K=40	X	X	X	X	X	X	X	
	K=60	X	X	X	X	X	X	X	
	K=80	X	X	X		X	X	X	
	K=100	X	X	X	X	X	X	X	
L=4	K=20	X		X			X	X	X
	K=40	X					X	X	X
	K=60						X	X	X
	K=80						X	X	X
	K=100	X			X		X	X	
L=6	K=20	X		X			X	X	X
	K=40	X					X	X	X
	K=60						X	X	X
	K=80						X	X	X
	K=100	X			X		X	X	

As shown in Table 4.20, the InfoGain tree ranks systolic and diastolic blood pressure as top attributes while the MIM tree chooses left ventricular hypertrophy and cholesterol. This causes the differences in attribute selection for both trees.

4.7 Discussions

4.7.1 Ontology-based Domain Generalization Hierarchy

The ontology-based DGH, though capable of maintaining rule meaningfulness, relies strongly on the chosen ontology. In certain cases, some attributes may not be found in a particular domain ontology knowledge. Furthermore, although they may be available, it does not mean that appropriate ranges have been specified for these attributes. Also, some attributes are only able to have two level DGHs. Take for

Table 4.20: CHD_DB attribute scores

Rank	Attribute	InfoGain	Attribute	MIM
1	Systolic Blood Pressure	0.0459124	Left Ventricular Hypertrophy	15.14
2	Diastolic Blood Pressure	0.0333328	Cholesterol	3.61133
3	Cholesterol	0.0231234	Smoking Habit	3.17224
4	Left Ventricular Hypertrophy	0.0221661	Systolic Blood Pressure	2.39787
5	Education	0.00948066	Diastolic Blood Pressure	2.28178
6	National Origin	0.000910938	Drinking Habit	1.47404
7	Drinking Habit	0.000174105	National Origin	0.726797
8	Smoking Habit	0.0	Education	0.488273

example the attribute gender. The values *male* and *female* can only be generalized once to a value such as *person* or *any*. In situations such as these, the user would need to refer to other sources of knowledge such as literature for constructing DGHs.

4.7.2 User Constraints

Although user constraints are effective at prioritizing task important attributes, it does not mean that every constrained attribute would be preserved. For instance, Tables 4.4, 4.8 and 4.13 are missing certain constrained attributes. This is due to the anonymization process whereby each constrained attribute is specialized till it violates an anonymity threshold or provides no further utility gain. As a result, certain prioritized attributes may remain generalized since any further specialization would lead to a breach of anonymity. Another occurrence is when constrained models remain similar to unconstrained trees such as in Tables 4.7, 4.9, 4.13 and 4.14. This is caused by two factors. Firstly, if selected constraints cover a majority of attributes, the constrained tree would most likely share similar attributes with the unconstrained model. Secondly, if high ranking attributes are part of the constrained group, they would cause specialization to occur similarly to an unconstrained tree. For example, Tables 4.7 and 4.9 both share thallium test as one of their constrained attributes. In Table 4.18, it can be seen that thallium test is the top ranking attribute when scored with InfoGain. Because of this, it would be specialized first for both unconstrained and constrained datasets.

In certain scenarios, semantic relation constraints may be too general. For instance, the age attribute shares the relations *result_of* and *associated_with* towards heart disease. This is because age is categorized under the *organism attribute* semantic type which in general shares those relations with *disease or syndrome*. In this case, the *result_of* relation between age and heart disease should be regarded as invalid. To overcome this issue, users would need to filter the semantic relations of each attribute specifically for their mining task to ensure proper prioritization.

4.7.3 MIM Score

As shown from results, implementing MIM as a utility score can lead to better prioritization of important attributes. One issue when using MIM is that a realistic selection of highly correlated attributes may not always occur since relationships are determined through occurrences in literature. Furthermore, these occurrences do not mean that there exists a direct association between both attributes. For example, two attributes co-occurring frequently within article abstracts may not necessarily be related and could even have a negative association. Because of this, the use of InfoGain to measure value changes is required together with MIM attribute rankings.

4.8 Summary

Experimental results prove our framework's ability in anonymizing data for domain-driven purposes. First, classification accuracy experiments indicate that although model accuracies remain similar to or are higher than the baseline, actionability reduces as anonymization increases. This supports the idea that accuracy though indicative of a model's predictive capabilities, may not necessarily prove that actionable rules exist. Second, by utilizing ontology-based DGHs, attribute semantic meanings and precision can be improved. Third, applying user constraints to preserve important attributes can increase actionability of mined models for specific mining tasks. Lastly, integrating MIM with InfoGain as a correlation-based score can improve attribute selection therefore leading to improved mining results.

CHAPTER 5

Conclusion

In conclusion, there is an urgent need for privacy preserving methods capable of anonymizing data for domain-driven usage. We have discussed the importance of attribute semantic meanings in mining results, user constraints for maintaining task important attributes, and attribute correlations based on external knowledge. Furthermore, experimental results have shown that integrating domain knowledge in the form of ontologies and external literatures can improve the anonymization process for domain-driven purposes. These factors justify the benefit of creating an anonymization framework for DDDM. By utilizing such a framework, publishers are given the ability to protect data while maintaining utility for real world requirements.

5.1 Recapitulation

This thesis proposes an ontology-based constrained anonymization framework for domain-driven data mining outsourcing. In the real world, data mining requires the use of domain knowledge to obtain results applicable to business goals and expectations. Furthermore, issues in data mining outsourcing such as data owner's willingness to share sensitive data, un-trusted service provider, and laws forbidding the sharing of individually identifiable data, show the need for data protection. Various anonymization techniques have been created that provide protection against privacy risks while maintaining reliable utility for traditional mining purposes. However, none of these methods have focused on the DDDM paradigm. The aim of this framework is to integrate domain knowledge into the anonymization process as a means to protect data privacy and preserve actionable models for DDDM tasks.

The framework consists of three main components: ontology-based DGH, user-specified constraints, and correlation-based anonymization algorithm. Domain knowledge was integrated in the form of domain ontologies and external literatures. Both the UMLS ontology and MEDLINE database were implemented in each component. The UMLS Metathesaurus which contained biomedical concepts was used to map attribute values during DGH creation. This allowed for more precise hierarchies based on predefined ranges from the UMLS ontology. To control attribute prioritization, users can set constraints on attribute semantic types or relations obtained from the UMLS Semantic Network. By doing so, important attributes related to or required for a certain task can be preserved. Even with user constraints, anonymization algorithms require utility measures when generalizing attributes. The MIM score was used to determine attribute relationships through co-occurrences in literature. This way, a more precise selection of attributes can be achieved through the evaluation of real world associations.

Four types of experiments were conducted on three datasets. First, classification accuracy was measured for each dataset. Results show that after anonymization, classification models retained similar or even higher accuracies compared to the baseline accuracy. Also, although classification accuracy may indicate a model's predictive capabilities, it may not necessarily mean that actionable rules exists. Second, ontology-based DGHs were evaluated by comparing value meanings in decision tree rules. It was shown that attribute semantic meanings and precision can be improved through the use of ontologies. Third, user constraints were set, and constrained decision trees were compared with normally anonymized trees. The experiment proved that at increasing levels of anonymization, rule actionability would decrease due to generalization. Furthermore, applying user constraints on task related attributes can improve actionability of mined rules for a particular mining purpose. Fourth, the MIM score was assessed to determine its abilities in selecting correlated attributes. Results indicate that utilizing MIM can improve attribute selection due to external literature knowledge, consequently leading to improved mining results.

5.2 Significance

The significance of the proposed framework can be summarized in the following points:

1. Most anonymization techniques have focused on satisfying traditional data mining utility. The framework differs as it aims to improve domain driven data mining utility by increasing rule actionability.
2. The framework is enhanced through the use of domain knowledge in the form of domain ontologies (UMLS) as well as domain relevant literatures (MEDLINE database).
3. In combination with domain knowledge, user constraints allow control over the anonymization process. Additionally, a correlation-based measure (MIM) aids in the attribute selection process thereby leading to more actionable results.

5.3 Implications

The outcome obtained from this thesis implies several important points:

1. An anonymization framework integrating domain knowledge in the form of ontology and literature can improve actionability for domain-driven data mining purposes.
2. The actionability of decision tree rules may decrease due to anonymization even if model accuracies are similar to or higher than baseline accuracies.
3. Ontology-based domain generalization hierarchies can maintain more meaning and precision compared to basic or dynamically created hierarchies.
4. Constraints provide users the ability to control the anonymization process and suit it towards their mining task.
5. Both technical significance and real world interestingness should be taken into account when measuring utility during anonymization to enhance attribute selection.

5.4 Limitations

The study is limited by some factors:

1. The main objective of the framework is to protect data through the *LKC*-privacy model and to maintain utility for domain-driven data mining purposes. It focuses on integrating domain knowledge in the form of ontology and literature to preserve meaningful and actionable mining results. Because of this, the study does not introduce new privacy attacks that may occur in a data mining outsourcing scenario. Furthermore, it does not intend to create a new privacy model as existing methods are capable of handling current threats.

2. The quality of domain generalization hierarchies strongly depend on attribute value distributions and on the ontology used to create them. This work aims to show the benefits of using an ontology-based DGH to preserve meaningful rules. Therefore, only attributes possessing appropriate ranges in the UMLS are created with the ontology while others are manually created.

3. User constraints based on attribute semantic types or relations are limited to concepts available within the UMLS Metathesaurus and Semantic Network.

4. The MIM score which affects attribute prioritization is dependent on literatures retrieved from MEDLINE. The study does not intend to evaluate the validity of literatures obtained and only focuses on calculating attribute correlations.

5.5 Future Works

The following areas should be considered in future works:

1. The integration of multiple ontologies in the DGH creation phase is important to ensure all attributes are mapped to appropriate ranges and contain semantically valid values. Also, automation of the hierarchy creation process can be enhanced by dynamically creating hierarchies through ontology knowledge.

2. The use of other ontologies during the user constraints phase to allow for more varied attribute prioritization. Instead of relying solely on attribute semantic

types or relations, alternative ontology components such as restrictions or axioms can be applied to better suit other mining tasks.

3. Natural language processing or text mining techniques can be employed to filter literature sources and obtain important or interesting attribute associations. This may improve attribute prioritization as it takes into account real attribute relationships.

4. Other generalization methods and anonymization algorithms should be experimented with to determine the effectiveness of each technique in maintaining DDDM utility.

CHAPTER 6

Appendix

6.1 Datasets

6.1.1 Attribute Descriptions

Table 6.1: Cath attribute descriptions

No	Attribute	Type	Values
1	Gender	Categorical	male, female
2	Race	Categorical	bidayuh, chinese, iban, indian, malay, melanau, others
3	Age	Numerical	13 - 90
4	Body Surface Area	Numerical	0.83 - 2.59
5	Body Mass Index	Numerical	11.03704986 - 43.37107378
6	Pre-Heart Rate	Numerical	36 - 150
7	Post-Heart Rate	Numerical	33 - 161
8	Pre-Diastolic Blood Pressure	Numerical	32 - 120
9	Post-Diastolic Blood Pressure	Numerical	7 - 129
10	Pre-Systolic Blood Pressure	Numerical	60 - 221
11	Post-Systolic Blood Pressure	Numerical	49 - 233
12	Coronary Artery Obstruction (Target)	Categorical	yes, no

Table 6.2: Cleveland attribute descriptions

No	Attribute	Type	Values
1	Age	Numerical	29 - 77
2	Sex	Categorical	male, female
3	Chest Pain	Categorical	typical angina, atypical angina, non-anginal pain, asymptomatic
4	Blood Pressure	Numerical	94 - 200
5	Cholesterol	Numerical	126 - 564
6	Blood Sugar	Categorical	>120, <120
7	Resting ECG	Categorical	normal, abnormal, hypertrophy
8	Heart Rate	Numerical	71 - 202
9	Angina	Categorical	yes, no
10	ST Depression	Numerical	0 - 6.2
11	Slope	Categorical	upsloping, flat, downsloping
12	Major Vessels Colored	Numerical	0 - 3
13	Thallium Test	Categorical	normal, fixed defect, reversable defect
14	Heart Disease (Target)	Categorical	yes, no

Table 6.3: Framingham attribute descriptions

No	Attribute	Type	Values
1	Cholesterol	Numerical	83 - 430
2	Systolic Blood Pressure	Numerical	48 - 261
3	Diastolic Blood Pressure	Numerical	32 - 152
4	Left Ventricular Hypertrophy	Categorical	negative, positive
5	National Origin	Categorical	native born, foreign born
6	Education	Categorical	grade school or less, high school not graduate, high school graduate, college or more
7	Smoking Habit	Categorical	never smoked, stopped, cigar or pipe, tobacco(<20/day), tobacco(>=20/day)
8	Drinking Habit	Numerical	0 - 60.7
9	Heart Disease (Target)	Categorical	yes,no

6.1.2 Attribute Semantic Types and Relations

Table 6.4: Cath attribute semantic types and relations

No	Attribute	Semantic Type	Semantic Relation
1	Gender	Organism Attribute	Organism Attribute *result_of* Disease or Syndrome Organism Attribute *associated_with* Disease or Syndrome
2	Race	Population Group	Population Group *associated_with* Disease or Syndrome
3	Age	Organism Attribute	Organism Attribute *result_of* Disease or Syndrome Organism Attribute *associated_with* Disease or Syndrome
4	Body Surface Area	Organism Attribute	Organism Attribute *result_of* Disease or Syndrome Organism Attribute *associated_with* Disease or Syndrome
5	Body Mass Index	Clinical Attribute	Clinical Attribute *associated_with* Disease or Syndrome Clinical Attribute *result_of* Disease or Syndrome
6	Pre-Heart Rate	Organism Attribute	Organism Attribute *result_of* Disease or Syndrome Organism Attribute *associated_with* Disease or Syndrome
7	Post-Heart Rate	Organism Attribute	Organism Attribute *result_of* Disease or Syndrome Organism Attribute *associated_with* Disease or Syndrome
8	Pre-Diastolic Blood Pressure	Clinical Attribute	Clinical Attribute *associated_with* Disease or Syndrome Clinical Attribute *result_of* Disease or Syndrome
9	Post-Diastolic Blood Pressure	Clinical Attribute	Clinical Attribute *associated_with* Disease or Syndrome Clinical Attribute *result_of* Disease or Syndrome

Continued on Next Page...

Table 6.4 – Continued

No	**Attribute**	**Semantic Type**	**Semantic Relation**
10	Pre-Systolic Blood Pressure	Clinical Attribute	Clinical Attribute *associated_with* Disease or Syndrome Clinical Attribute *result_of* Disease or Syndrome
11	Post-Systolic Blood Pressure	Clinical Attribute	Clinical Attribute *associated_with* Disease or Syndrome Clinical Attribute *result_of* Disease or Syndrome
12	Coronary Artery Obstruction (Target)	Disease or Syndrome	

Table 6.5: Cleveland attribute semantic types and relations

No	Attribute	Semantic Type	Semantic Relation
1	Age	Organism Attribute	Organism Attribute *result_of* Disease or Syndrome Organism Attribute *associated_with* Disease or Syndrome
2	Sex	Organism Attribute	Organism Attribute *result_of* Disease or Syndrome Organism Attribute *associated_with* Disease or Syndrome
3	Chest Pain	Sign or Symptom	Sign or Symptom *diagnoses* Disease or Syndrome Sign or Symptom *evaluation_of* Disease or Syndrome Sign or Symptom *manifestation_of* Disease or Syndrome Sign or Symptom *associated_with* Disease or Syndrome
4	Blood Pressure	Organism Function	Organism Function *result_of* Disease or Syndrome Organism Function *affects* Disease or Syndrome
5	Cholesterol	Biologically Active Substance	Biologically Active Substance *affects* Disease or Syndrome Biologically Active Substance *causes* Disease or Syndrome Biologically Active Substance *complicates* Disease or Syndrome
6	Blood Sugar	Carbohydrate	Carbohydrate *causes* Disease or Syndrome Carbohydrate *affects* Disease or Syndrome
7	Resting ECG	Diagnostic Procedure	Diagnostic Procedure *affects* Disease or Syndrome Diagnostic Procedure *associated_with* Disease or Syndrome

Continued on Next Page...

Table 6.5 – Continued

No	Attribute	Semantic Type	Semantic Relation
			Diagnostic Procedure *diagnoses* Disease or Syndrome Diagnostic Procedure *measures* Disease or Syndrome
8	Heart Rate	Organism Attribute	Organism Attribute *result_of* Disease or Syndrome Organism Attribute *associated_with* Disease or Syndrome
9	Angina	Sign or Symptom	Sign or Symptom *diagnoses* Disease or Syndrome Sign or Symptom *evaluation_of* Disease or Syndrome Sign or Symptom *manifestation_of* Disease or Syndrome Sign or Symptom *associated_with* Disease or Syndrome
10	ST Depression	Finding	Finding associated_with Disease or Syndrome Finding *manifestation_of* Disease or Syndrome Finding *evaluation_of* Disease or Syndrome
11	Slope	Quantitative Concept	None
12	Major Vessels Colored	Diagnostic Procedure	Diagnostic Procedure *affects* Disease or Syndrome Diagnostic Procedure *associated_with* Disease or Syndrome Diagnostic Procedure *diagnoses* Disease or Syndrome Diagnostic Procedure *measures* Disease or Syndrome

Continued on Next Page...

Table 6.5 – Continued

No	Attribute	Semantic Type	Semantic Relation
13	Thallium Test	Diagnostic Procedure	Diagnostic Procedure *affects* Disease or Syndrome Diagnostic Procedure *associated_with* Disease or Syndrome Diagnostic Procedure *associated_with* Disease or Syndrome Diagnostic Procedure *diagnoses* Disease or Syndrome Diagnostic Procedure *measures* Disease or Syndrome
14	Heart Disease (Target)	Disease or Syndrome	

Table 6.6: CHD_DB attribute semantic types and relations

No	Attribute	Semantic Type	Semantic Relation
1	Cholesterol	Biologically Active Substance	Biologically Active Substance *affects* Disease or Syndrome Biologically Active Substance *causes* Disease or Syndrome Biologically Active Substance *complicates* Disease or Syndrome
2	Systolic Blood Pressure	Clinical Attribute	Clinical Attribute *associated_with* Disease or Syndrome Clinical Attribute *result_of* Disease or Syndrome
3	Diastolic Blood Pressure	Clinical Attribute	Clinical Attribute *associated_with* Disease or Syndrome Clinical Attribute *result_of* Disease or Syndrome
4	Left Ventricular Hypertrophy	Disease or Syndrome	Disease or Syndrome *co-occurs_with* Disease or Syndrome Disease or Syndrome *precedes* Disease or Syndrome Disease or Syndrome *process_of* Disease or Syndrome Disease or Syndrome *associated_with* Disease or Syndrome Disease or Syndrome *complicates* Disease or Syndrome Disease or Syndrome *result_of* Disease or Syndrome Disease or Syndrome *degree_of* Disease or Syndrome

Continued on Next Page...

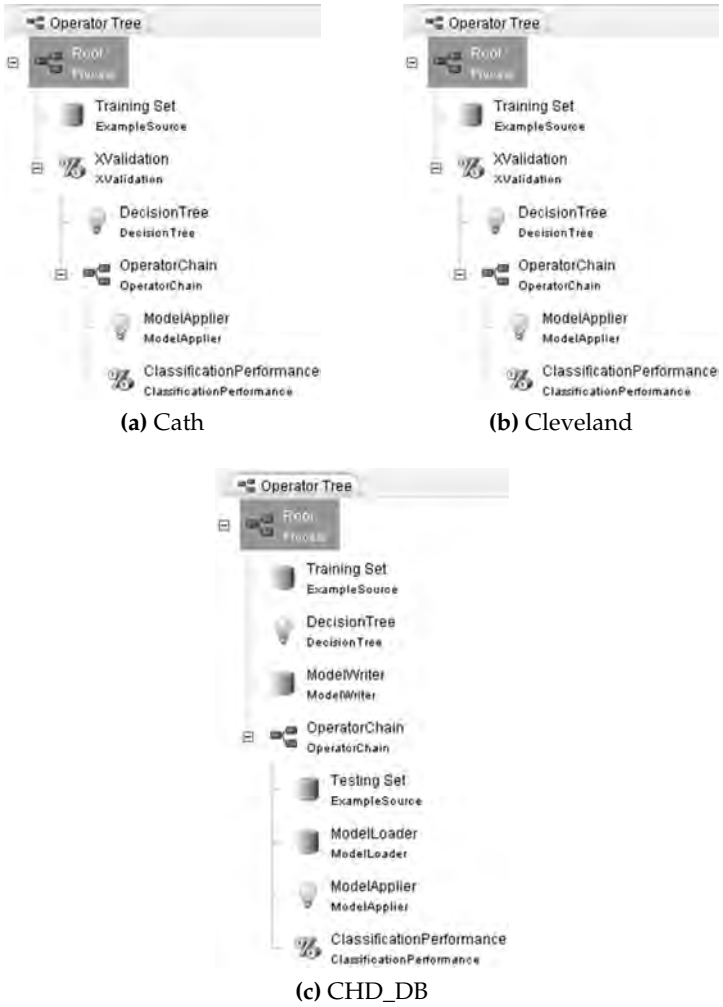
Table 6.6 – Continued

No	Attribute	Semantic Type	Semantic Relation
			Disease or Syndrome *occurs_in* Disease or Syndrome Disease or Syndrome *affects* Disease or Syndrome Disease or Syndrome *manifestation_of* Disease or Syndrome
5	National Origin	Clinical Attribute	Clinical Attribute *associated_with* Disease or Syndrome Clinical Attribute *result_of* Disease or Syndrome
6	Education	Finding	Finding associated_with Disease or Syndrome Finding *manifestation_of* Disease or Syndrome Finding *evaluation_of* Disease or Syndrome
7	Smoking Habit	Finding	Finding associated_with Disease or Syndrome Finding *manifestation_of* Disease or Syndrome Finding *evaluation_of* Disease or Syndrome
8	Drinking Habit	Finding	Finding associated_with Disease or Syndrome Finding *manifestation_of* Disease or Syndrome Finding *evaluation_of* Disease or Syndrome
9	Heart Disease (Target)	Disease or Syndrome	

6.2 RapidMiner

RapidMiner is an open-source system for data mining with applications capable of analyzing and processing data. A free community edition is available for download from the official RapidMiner website. For this thesis, RapidMiner has been utilized for creating classification decision trees as well as calculating model accuracies. Decision trees were constructed through the simple process of specifying a dataset and adjusting certain parameters including minimal gain and maximal depth. These

thresholds controlled which attributes were chosen based on their InfoGain scores and also how large a decision tree would be. Default values were used for the Cleveland dataset while Cath and CHD_DB used lower thresholds to reduce decision tree size.



(a) Cath **(b)** Cleveland

(c) CHD_DB

Figure 6.1: RapidMiner operator tree

Figure 6.1 shows operator trees which represent the knowledge discovery process performed by RapidMiner on Cath, Cleveland and CHD_DB datasets. According to Figures 6.1a and 6.1b, first, a training set is selected for mining. A decision tree is generated from the data and ten fold cross-validation is performed to ensure a proper average is obtained. Next, the decision tree model is assessed to determine classification performance. Figure 6.1c presents a similar process with the only exception being that a testing set is specified beforehand. The reason for this, is because CHD_DB already has a predefined testing dataset which can be used for assessing classification performance.

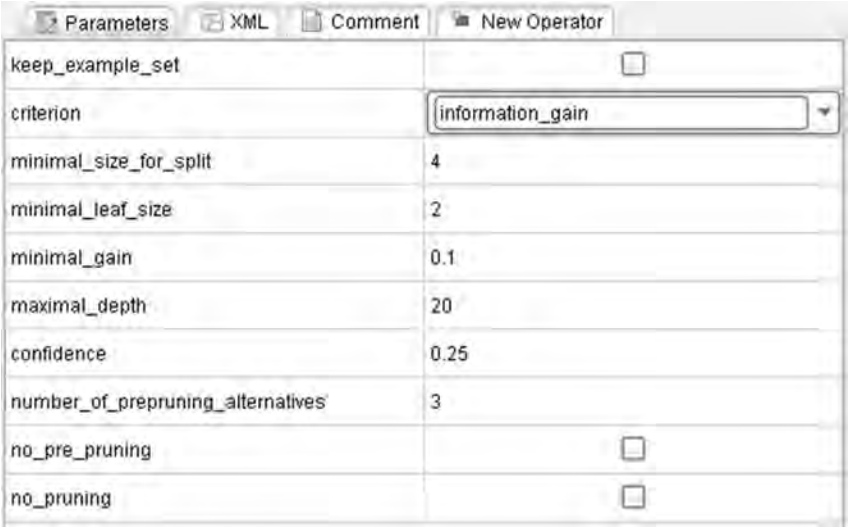


Figure 6.2: Decision tree operator

The decision tree operator details are shown in Figure 6.2. Here, the criterion for splitting decision trees can be set as well as other parameters for controlling decision tree size. InfoGain is chosen as the default splitting criterion while other parameters are left at default.

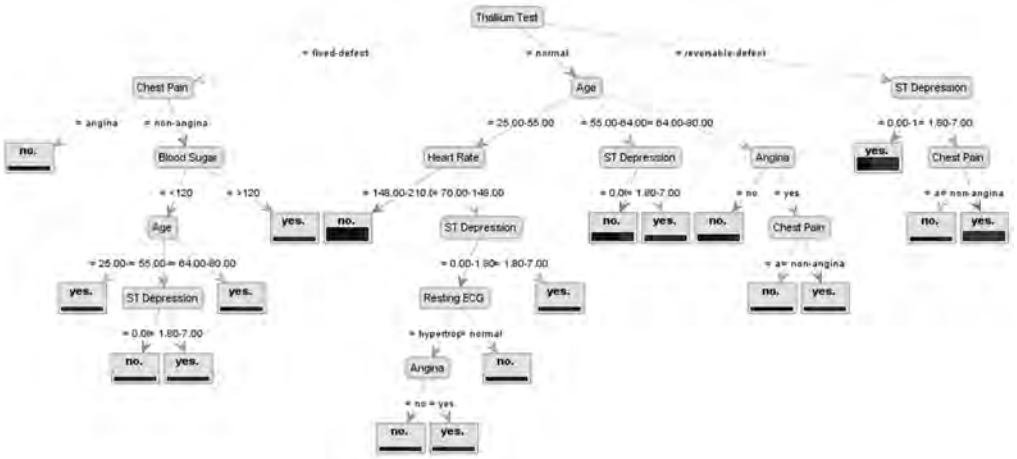


Figure 6.3: Decision tree graphic

After the data mining process is completed, a decision tree such as the one in Figure 6.3 would be generated. This example, taken from an anonymized Cleveland dataset is shown in a graphical format. Figure 6.4 presents the same tree in text form which allows for easier comparison.

Tree

```
Thallium Test = fixed-defect
|   Chest Pain = angina: no. (no.=3, yes.=1)
|   Chest Pain = non-angina
|   |   Blood Sugar = <120
|   |   |   Age = 25.00-55.00: yes. (no.=0, yes.=2)
|   |   |   Age = 55.00-64.00
|   |   |   |   ST Depression = 0.00-1.80: no. (no.=2, yes.=1)
|   |   |   |   ST Depression = 1.80-7.00: yes. (no.=0, yes.=2)
|   |   |   Age = 64.00-80.00: yes. (no.=1, yes.=2)
|   |   Blood Sugar = >120: yes. (no.=0, yes.=4)
Thallium Test = normal
|   Age = 25.00-55.00
|   |   Heart Rate = 148.00-210.00: no. (no.=72, yes.=4)
|   |   Heart Rate = 70.00-148.00
|   |   |   ST Depression = 0.00-1.80
|   |   |   |   Resting ECG = hypertrophy
|   |   |   |   |   Angina = no: no. (no.=3, yes.=1)
|   |   |   |   |   Angina = yes: yes. (no.=1, yes.=2)
|   |   |   |   Resting ECG = normal: no. (no.=5, yes.=0)
|   |   |   ST Depression = 1.80-7.00: yes. (no.=0, yes.=2)
|   Age = 55.00-64.00
|   |   ST Depression = 0.00-1.80: no. (no.=26, yes.=11)
|   |   ST Depression = 1.80-7.00: yes. (no.=0, yes.=8)
|   Age = 64.00-80.00
|   |   Angina = no: no. (no.=17, yes.=4)
|   |   Angina = yes
|   |   |   Chest Pain = angina: no. (no.=2, yes.=0)
|   |   |   Chest Pain = non-angina: yes. (no.=1, yes.=5)
Thallium Test = reversable-defect
|   ST Depression = 0.00-1.80: yes. (no.=24, yes.=47)
|   ST Depression = 1.80-7.00
|   |   Chest Pain = angina: no. (no.=2, yes.=1)
|   |   Chest Pain = non-angina: yes. (no.=1, yes.=40)
```

Figure 6.4: Decision tree text

6.3 Unified Medical Language System

The Unified Medical Language System (UMLS) is one of the world's most comprehensive medical domain ontologies designed by the US National Library of Medicine. The framework employs two major components of the UMLS which are the Metathesaurus and Semantic Network. The Metathesaurus comprises of over a million inter-related biomedical concepts. Each concept is categorized by its semantic type and is identified through its Concept Unique Identifier (CUI). Figure 6.5 presents a sample Metathesaurus output for the heart disease concept. The concept is identified by the code *C0018799* with other details including date added and revision date. Furthermore, the concept of heart disease is defined by several thesauruses and its semantic type is shown.

The Semantic Network assigns Metathesaurus concepts high-level categories and stores the links between them in a hierarchical format. There are a total of 135 semantic types for each concept and 54 relations that exist among these types. Figure

Concept Information for Heart Diseases - UMLS Release 2010AA

[C0018799] Heart Diseases
DA Date Added 19900930 [AT00008091/Metathesaurus Names]
MR Major Revision Date 20100225 [AT121415363/Metathesaurus Names]
ST Status R [AT01958799/Metathesaurus Names]

Definition:
MeSH/A0066404
Pathological conditions involving the HEART including its structural and functional abnormalities.
CRISP Thesaurus/A0319133
Impairment of health or a condition of abnormal functioning of the heart.
NCI Thesaurus/A10798721
Any deviation from the normal structure or function of the cardiac system that is manifested by a characteristic set of symptoms and signs. (NCI)
NCI Thesaurus/A7599175
Any deviation from the normal structure or function of the cardiac system that is manifested by a characteristic set of symptoms and signs.
Semantic Types:
Disease or Syndrome

Figure 6.5: Metathesaurus output

6.6 shows a sample Semantic Network output for the relation *disease or syndrome*. The output provides information regarding relationship definition and also relations to other semantic types. For instance, disease or syndrome *associated_with* organism attribute or virus *causes* disease or syndrome. Moreover, children of the concept are shown with details of their definition and relations.

Semantic Network Types
Semantic Types
Disease or Syndrome (Event-->Phenomenon or Process-->Natural Phenomenon or Process-->Biologic Function-->Pathologic Function)
Definition
A condition which alters or interferes with a normal process, state, or activity of an organism. It is usually characterized by the abnormal functioning of one or more of the host`s systems, parts, or organs. Included here is a complex of symptoms descriptive of a disorder.
Relations
Selected semantic type as left-hand side of the relationship triplet
Selected semantic type as right-hand side of the relationship triplet
Children
Neoplastic Process
Definition
Relations
Mental or Behavioral Dysfunction
Definition
Relations

Figure 6.6: Semantic Network output

The framework utilizes the Metathesaurus and Semantic Network during DGH creation as well as user constraints specification. Attribute semantic maps are used to generate attribute hierarchies so that values are categorized into particular ranges.

Additionally, attributes are constrained by their semantic types and relations which are obtained from the Semantic Network.

6.4 Literature

The Pubmed search engine maintained by the National Institutes of Health (NIH) is capable of retrieving over 19 million biomedical article citations from the U.S. National Library of Medicine (MEDLINE database) as well as other journals. The MEDLINE database includes articles from academic journals covering numerous topics including medicine, nursing, pharmacy and health care. There are approximately 5,000 biomedical journals indexed in MEDLINE with articles from 1950 to the present. Figure 6.7 presents the results of the query phrase: (cholesterol[MeSH]+OR+"cholesterol") AND (heart-disease[MeSH]+OR+"heart-disease"). A total of 2,443 articles containing the words cholesterol and heart disease are found.

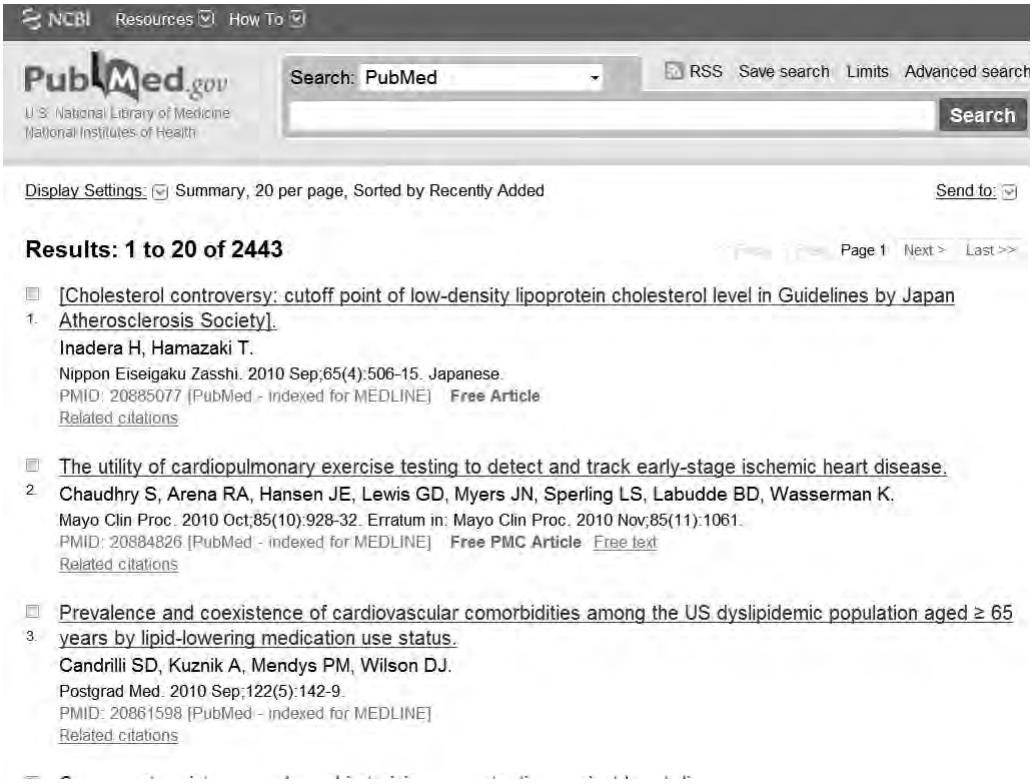


Figure 6.7: Pubmed query result

To reduce the number of articles and focus the query on a particular task, search filters can be implemented [88]. Figure 6.8 shows the results of the previous

query with an added diagnosis filter: (sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnos*[Title/Abstract] OR diagnosis[MeSH:noexp] OR diagnostic * [MeSH:noexp] OR diagnosis,differential[MeSH:noexp] OR diagnosis[Subheading:noexp]). As seen from the results, the total number of literatures have been reduced to 449 while article topics are concentrated on diagnosis tasks.

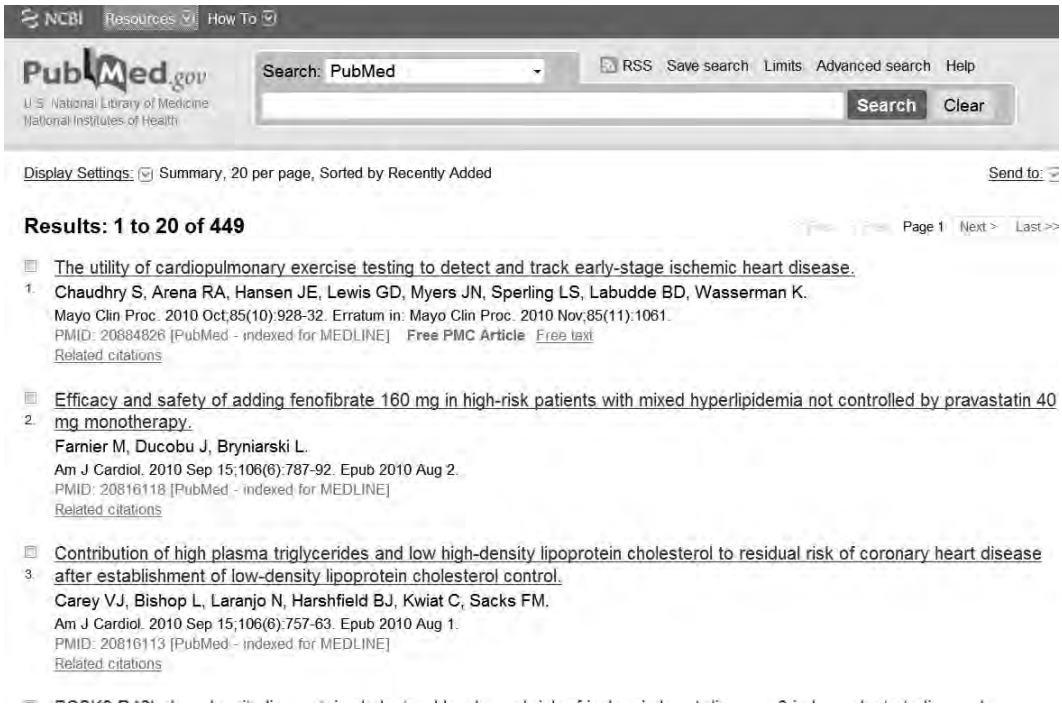


Figure 6.8: Pubmed filtered query result

Instead of manually searching for articles, the framework integrates Entrez Programming Utilities (eUtils) to automatically retrieve query results. Three packages were used from eUtils which are EInfo for retrieving article counts, ESearch for obtaining citation identifiers and EFetch to extract article details such as title, authors, abstract and headings. Article titles and abstracts were processed to determine the relationship of queried attributes, thereby allowing the calculation of MIM scores.

6.5 C++

C++ is a widely used programming language in the software industry which has influenced other popular languages, most notably Java. It was primarily used to

create the anonymization tools available in the proposed framework.

6.6 Java

Java is a programming language developed by Sun Microsystems and derives much of its syntax from C++. A main advantage of Java applications comes from their ability to be run on any Java Virtual Machine regardless of computer architecture or operating system. It was used to develop the MIM evaluator which implemented Java classes from Entrez Programming Utilities (eUtils).

6.6.1 Java Native Interface

The Java Native Interface (JNI) is a programming framework that enables Java codes to be called by applications written in other languages such as C++. Since both the anonymizer and MIM evaluator are written in different programming languages, the JNI is required. The anonymizer makes use of the JNI by calling eUtils classes which pass calculated MIM scores, therefore enabling the prioritization of attributes.

References

[1] J C Prather, D F Lobach, L K Goodwin, J W Hales, M L Hage, and W E Hammond. Medical data mining: knowledge discovery in a clinical data warehouse. In *Proc AMIA Annu Fall Symp*, pages 101–105, 1997.

[2] L Cao, L Lin, and C Zhang. Domain-driven in-depth pattern discovery: A practical methodology. In *Proceedings of AusDM*, pages 101–114, 2005.

[3] L Cao and C Zhang. Domain-driven data mining: A practical methodology. *International Journal of Data Warehousing & Mining*, 2(4):49–65, 2006.

[4] L Cao and C Zhang. The evolution of kdd: Towards domain-driven data mining. *International Journal of Pattern Recognition*, 21(4):677–692, 2007.

[5] L Cao. Domain driven data mining: Challenges and prospects. *IEEE Transactions on Knowledge and Data Engineering*, 22(6):755–769, 2010.

[6] A A Freitas. Are we really discovering interesting knowledge from data? *Expert Update Special Issue on the 2nd UK KDD Workshop*, 9(1):41–47, 2006.

[7] P S Bradley. Data mining as an automated service. In *Proceedings of the 7th Pacific-Asia conference on Advances in knowledge discovery and data mining*, pages 1–13, 2003.

[8] I Bhattacharya, S Godbole, A Gupta, A Verma, J Achtermann, and K English. Enabling analysts in managed services for crm analytics. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1077–1086, 2009.

[9] S Krishnaswamy, S W Loke, and A Zaslavsky. *Knowledge Elicitation through Web-Based Data Mining Services*, pages 120–134. Springer, 2001.

[10] S E Pin. Negotiation for data mining e-services. Technical report, 2002.

[11] B Brumen, T Welzer, M Druzovec, I Golob, and H Jaakkola. Protecting medical data for analyses. In *Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems (CBMS'02)*, pages 102–107, 2002.

[12] B Brumen, T Welzer, M Druzovec, I Golob, H Jaakkola, I Rozman, and J Kubalik. Protecting medical data for decision-making analyses. *Journal of Medical Systems*, 29(1):65–80, 2005.

[13] T Falkowski. Application service providing for data mining applications. In *Proceedings 7. Gottinger Symposium Soft-Computing*, pages 23–40, 2003.

[14] B C M Fung, K Wang, R Chen, and P S Yu. Privacy-preserving data publishing: A survey on recent developments. *ACM Computing Surveys*, 42(4):1–53, 2010.

[15] L Qiu, Y J Li, and X Wu. Preserving privacy in association rule mining with bloom filters. *Journal of Intelligent Information Systems*, 29(3):253–278, 2007.

[16] C Boyens. On privacy trade-offs in web-based services, 2004.

[17] S Evdokimov. *Secure Outsourcing of IT Services in a Non-Trusted Environment*. PhD thesis, Humboldt-University, 2008.

[18] W K Wong, D W Cheung, E Hung, B Kao, and N Mamoulis. Security in outsourcing of association rule mining. In *Proceedings of the 33rd international conference on Very large data bases*, pages 111–122, 2007.

[19] W K Wong and D W Cheung. Security and integrity of association rule mining, 2009.

[20] B Bhumiratana and M Bishop. Privacy aware data sharing: balancing the usability and privacy of datasets. In *Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments*, pages 1–8, 2009.

[21] A A Hintoglu, Y Saygin, S Benbernou, and M S Hacid. *Privacy Preserving Data Mining Services on the Web*, pages 246–255. Springer, 2005.

[22] N Li and T Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE 23rd International Conference on Data Engineering*, pages 106–115, 2007.

[23] A Machanavajjhala, D Kifer, J Gehrke, and M Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 2007.

[24] L Sweeney. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[25] F Li and S Zhou. Challenging more updates: Towards anonymous re-publication of fully dynamic datasets, 2008.

[26] Q Wei, Y Lu, and L Zou. E-inclusion: privacy preserving re-publication of dynamic datasets. *Journal of Zhejiang University - Science A*, 9(8):1124–1133, 2008.

[27] X Xiao and Y Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 689–700, 2007.

[28] R J Bayardo and R Agrawal. Data privacy through optimal k-anonymization. In *Proceedings of the 21st International Conference on Data Engineering*, pages 217–228, 2005.

[29] B C M Fung, K Wang, and P S Yu. Anonymizing classification data for privacy preservation. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):711–725, 2007.

[30] B C M Fung, K Wang, and P S Yu. Top-down specialization for information and privacy preservation. In *Proceedings of the 21st International Conference on Data Engineering*, pages 205–216, 2005.

[31] N Harnsamut, J Natwichai, X Sun, and X Li. *Data Quality in Privacy Preservation for Associative Classification*, pages 111–122. Springer, 2008.

[32] N Harnsamut and J Natwichai. *A Novel Heuristic Algorithm for Privacy Preserving of Associative Classification*, pages 273–283. Springer, 2008.

[33] A Inan, M Kantarcioglu, and E Bertino. Using anonymized data for classification. In *Proceedings of the 2009 IEEE International Conference on Data Engineering*, pages 429–440, 2009.

[34] V S Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 278–288, 2002.

[35] K LeFevre, D J DeWitt, and R Ramakrishnan. Workload-aware anonymization. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 277–286, 2006.

[36] K LeFevre, D J DeWitt, and R Ramakrishnan. Workload-aware anonymization techniques for large-scale datasets. *ACM Transactions on Database Systems (TODS)*, 33(3):1–47, 2008.

[37] N Mohammed, B C M Fung, P C K Hung, and C K Lee. Anonymizing healthcare data: a case study on the blood transfusion service. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1285–1294, 2009.

[38] K Wang, P S Yu, and S Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, pages 249–256, 2004.

[39] L Xiong and K Rangachari. Towards application-oriented data anonymization. In *Proceedings of the 4th SIAM Workshop on Practical Privacy-Preserving Data Mining*, pages 1–47, 2008.

[40] M R Z Mirakabad, A Jantan, and S Bressan. K-anonymity diagnosis centre. *International Journal of Security and Its Applications (IJSIA)*, 3(1):47–63, 2009.

[41] X Xiao, G Wang, and J Gehrke. Interactive anonymization of sensitive data. In *Proceedings of the 35th SIGMOD international conference on Management of data*, pages 1051–1054, 2009.

[42] M E Nergiz and C Clifton. Thoughts on k-anonymization. *Data & Knowledge Engineering*, 63(3):622–645, 2006.

[43] S Kisilevich, Y Elovici, B Shapira, and L Rokach. *kACTUS 2: Privacy Preserving in Classification Tasks Using k-Anonymity*, pages 63–81. Springer, 2009.

[44] Y Liu, T Wang, and J Feng. *A Semantic Information Loss Metric for Privacy Preserving Publication*, pages 138–152. Springer, 2010.

[45] Z Zhu, J Gu, L Zhang, J Li, and Y Zeng. *Research on Domain-Driven Actionable Knowledge Discovery*, pages 176–183. Springer, 2009.

[46] C Clifton, W Jiang, M Muruguesan, and M E Nergiz. Is privacy still an issue for data mining? In *National science foundation symposium on next generation of data mining and cyber enabled discovery for innovation*, 2007.

[47] I Neamatullah, M M Douglass, L H Lehman, A Reisner, M Villarroel, W J Long, P Szolovits, G B Moody, R G Mark, and G D Clifford. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32, 2008.

[48] J Sarabdeen and M M M Ishak. E-health data privacy: How far is protected?, 2008.

[49] X Wu, C Chu, Y Wang, F Liu, and D Yue. *Privacy Preserving Data Mining Research: Current Status and Key Issues*, pages 762–772. Springer, 2007.

[50] A Friedman, R Wolff, and A Schuster. Providing k-anonymity in data mining. *VLDB*, 17(4):789–804, 2008.

[51] D E Bakken, R Parameswaran, D M Blough, A A Franz, and T J Palmer. Data obfuscation: Anonymity and desensitization of usable data sets. *IEEE Security and Privacy*, 2(6):34–41, 2004.

[52] L Qiu, Y Li, and X Wu. Protecting business intelligence and customer privacy while outsourcing data mining tasks. *Knowledge and Information Systems*, 17(1):99–120, 2007.

[53] J Li, H Wang, H Jin, and J Yong. Current developments of k-anonymous data releasing, 2006.

[54] W K Wong, D W Cheung, E Hung, B Kao, and N Mamoulis. An audit environment for outsourcing of frequent itemset mining. *Proceedings of the VLDB Endowment*, 2(1):1162–1173, 2009.

[55] W K Wong and D W Cheung. Security and integrity of association rule mining. In *ACM-HK Student Research and Career Day*, 2009.

[56] J-W Byun, Y Sohn, E Bertino, and N Li. *Secure Anonymization for Incremental Datasets*, pages 48–63. Springer, 2006.

[57] T R Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, 1995.

[58] I Yoo and M Song. Biomedical ontologies and text mining for biomedicine and healthcare: A survey. *JCSE*, 2(2):109–136, 2008.

[59] A Tsymbal, S Zillner, and M Huber. Ontology - supported machine learning and decision support in biomedicine. In *Proceedings of the 4th international conference on Data integration in the life sciences*, pages 156–171, 2007.

[60] J Miller, A Campan, and T M Truta. Constrained k-anonymity: Privacy with generalization boundaries. In *Proceedings of the Practical Preserving Data Mining Workshop*, 2008.

[61] N Matatov, L Rokach, and O Maimon. Privacy-preserving data mining: A feature set partitioning approach. *Information Sciences*, 180(14):2696–2720, 2010.

[62] Z Lin, M Hewett, and R B Altman. Using binning to maintain confidentiality of medical data. In *American Medical Informatics Association Annual Symposium*, pages 454–459, 2002.

[63] Z Lin. *Balancing utility and anonymity in public biomedical databases*. PhD thesis, Standford University, 2005.

[64] E Bertino, D Lin, and W Jiang. A survey of quantification of privacy preserving data mining algorithms. *Privacy Preserving Data Mining*, 34:183–205, 2008.

[65] K LeFevre, D J Dewitt, and R Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of ACM SIGMOD*, pages 49–60, 2005.

[66] R C W Wong, J Li, A W C Fu, and K Wang. (a,k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD*, pages 754–759, 2006.

[67] J Xu, W Wang, J Pei, X Wang, B Shi, and A W C Fu. Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD*, pages 785–790, 2006.

[68] R Vidyabanu, D S Thomas, and N Nagaveni. Enhancing privacy of confidential data using k anonymization. *International Journal of Recent Trends in Engineering*, 2(1):130–133, 2009.

[69] T Dalenius. Finding a needle in a haystack - or identifying anonymous census record. *Journal of Official Statistics*, 2(3):329–336, 1986.

[70] P Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[71] C Bettini, X S Wang, and S Jajodia. How anonymous is k-anonymous? look at your quasi-id. *LNCS*, 5159:1–15, 2008.

[72] S Lodha and D Thomas. Probabilistic anonymity. In *Proceedings of the First International Workshop on Privacy, Security, and Trust in KDD*, pages 56–79, 2007.

[73] R Motwani and Y Xu. Efficient algorithms for masking and finding quasi-identiiňĄers. In *SIAM International Workshop on Practical Privacy-Preserving Data Mining*, 2008.

[74] T Li and N Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–526, 2009.

[75] G Loukides, A Tziatzios, and J Shao. Towards preference-constrained k-anonymisation. *LNCS*, 5667:231–245, 2009.

[76] R Dewri, D Whitley, I Ray, and I Ray. A multi-objective approach to data sharing with privacy constraints and preference based objectives. In *Genetic and Evolutionary Computation Conference (GECCO 2009)*, pages 1499–1506, 2009.

[77] X Sun, H Wang, and J Li. Injecting purposes and trust into data anonymization. In *CIKM 2009*, pages 1541–1544, 2009.

[78] L Geng and H J Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys*, 38(3), 2006.

[79] B C S Loh and P H H Then. Ontology-enhanced interactive anonymization in domain-driven data mining outsourcing. In *2010 Second International Symposium on Data, Privacy and E-Commerce (ISDPE)*, pages 9–14, 2010.

[80] Y-T Kuo, A Lonie, L Sonenberg, and K Paizis. Domain ontology driven data mining: a medical case study. In *Proc. the ACM SIGKDD International Workshop on Domain-Driven Data Mining (DDDM '07)*, pages 11–17, 2007.

[81] H Cespivova, J Rauch, V Svatek, M Kejkula, and M Tomeckova. Roles of medical ontology in association mining crisp-dm cycle. In *Proceedings of the ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies*, 2004.

[82] V Svatek, J Rauch, and M Ralbovsky. Ontology-enhanced association mining, 2006.

[83] O Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, 2004.

[84] D Demner-Fushman and J Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Liguistics*, 33(1):63–103, 2007.

[85] J D Wren. Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics*, 5(1):145, 2004.

[86] X Hu, X Zhang, and X Zhou. *Comparison of seven methods for mining hidden links*, pages 27–44. Wiley Interscience, 2007.

[87] Y Sebastian, B C S Loh, and P H H Then. A paradigm shift: Combined literature and ontology-driven data mining for discovering novel relations in biomedical domain. In *The 3rd IEEE ICDM International Workshop on Domain Driven Data Mining (DDDM 09)*, pages 51–57, 2009.

[88] R B Haynes and N L Wilczynski. Optimal search strategies for retrieving scientifically strong studies of diagnosis from medline: analytical survey, 2004.

[89] R B Haynes, K A McKibbon, N L Wilczynski, S D Walter, and S R Werre. Optimal search strategies for retrieving scientifically strong studies of treatment from medline: analytical survey, 2005.

[90] C C Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st VLDB Conference*, pages 901–909, 2005.

[91] B Ristow, M P Turakhia S Ali, B Na, M A Whooley, and N B Schiller. Predicting heart failure hospitalization and mortality by quantitative echocardiography: Is body surface area the indexing method of choice? the heart and soul study. *J Am Soc Echocardiogr*, 23(4):406–419, 2010.

[92] P Jousilahti, E Vartiainen, J Tuomilehto, and P Puska. Sex, age, cardiovascular risk factors, and coronary heart disease: a prospective follow-up study of 14 786 middle-aged men and women in finland. *Circulation*, 99(9):1165–1172, 1999.