2010

Discovering patterns in a survey of secondary injuries due to agricultural assistive technology

Yanjun Shi
Iowa State University

Discovering patterns in a survey of secondary injuries due to agricultural assistive technology

by

Yanjun Shi

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:

Sigurdur Olafsson, Co-major Professor

Nir Keren, Co-major Professor

Gary Mirka

Iowa State University

Ames, Iowa

2010

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my major professor Dr. Sigurdur Olafsson for guiding me into the Data Mining world, for his creative suggestions, guidance, encouragement and patience through this research and the writing of this thesis. I greatly appreciate the assistance of my co-major professor Dr. Nir Keren in providing important knowledge, professional experience and perspective in agriculture field. I appreciate my committee member for his efforts and contributions to this work: Dr. Gary Mirka.

I am most thankful to my husband, parents, grandparents and son for their sacrifice, understanding and support during this study. I would also like to thank my classmates: Yan Wang and Xiang Wu for their help in this research.

ABSTRACT

The research is motivated by the need for hazard assessment in agriculture field. A small and highly-imbalanced dataset, in which negative instances heavily outnumber positive instances, is derived from a survey of secondary injuries induced by implementation of agriculture assistive technology which assists farmers with injuries or disabilities to continue farm-related work. Three data mining approaches are applied to the imbalanced dataset in order to discover patterns contributing to secondary injuries.

All of patterns discovered by the three approaches are compared according to three evaluation measurements: support, confidence and lift, and potentially most interesting patterns are found. Compared to graphical exploratory analysis which figures out causative factors by evaluating the single effects of attributes on the occurrence of secondary injuries, decision tree algorithm and subgroup discovery algorithms are able to find combinational factors by evaluating the interactive effects of attributes on the occurrence of secondary injuries. Graphical exploratory analysis is able to find patterns with highest support and subgroup discovery algorithms are good at finding high lift patterns.

In addition, the experimental analysis of applying subgroup discovery to our secondary injury dataset demonstrates subgroup discovery method's capability of dealing with imbalanced datasets. Therefore, identifying risk factors contributing to secondary injuries, as well as providing a useful alternative method (subgroup discovery) of dealing with small and highly-imbalanced datasets are important outcomes of this thesis.

CHAPTER 1 INTRODUCTION

1.1 Motivation

Agriculture has been estimated to be one of the United States' most hazardous industries [32], ranking second among nation's industries with high risk of occupational injuries [41]. Agricultural workers suffer injuries, disabilities, and fatal injuries at much higher rates than those in other occupations [23]. For example, the possibility for agricultural workers to experience a disabling injury is as twice as the average American workers, and the possibility to experience fatal injury is almost six times [41]. Although permanent or temporary disabilities and other farming injuries, farming workers need to continue farming work [22] as this is the only source for providing to their families.

Assistive technology, which commonly refers to both assistive and adaptive devices and practices, enables farmers, ranchers, and agricultural workers with physical disabilities to continue their farm-related work [43]. In some cases, however, the induction of assistive technology leads to secondary injury. Secondary injury can be defined as follows: "Injury caused by limitations associated with disability conditions and/or resulting from the use of modified practices or AT [Assistive Technology] to compensate for disabling conditions" [43].

Despite the prevalence of assistive technology in agriculture field, little research has explored how the assistive technology employed might lead to secondary injuries [43]. Most of the research in the literature is limited to summaries case studies and surveys to provide examples or list potential causative factors for secondary injuries. Very little research has pointed out interactions between different factors resulting in secondary injuries. This motivates us to employ data mining techniques to discover secondary injury-related patterns which are rules of a conjunction of features. Therefore, it would be interesting and meaningful to research secondary injury dataset with data mining

techniques and address solutions to promote safe-related work practices and prevent secondary injuries while continuing farming work with the help of assistive technology in agriculture field.

Data mining has been applied to very wide fields and has become a research area with increasing importance. In recent years, data mining techniques have been successfully applied to several safety-related areas such as aviation safety [14], road accidents dataset [16], food safety [1], drug safety [4], and traffic safety [21].

Since occurrence of hazards which data mining systems are aimed to detect is rare in a general population, most of datasets in safety-related fields analyzed by data mining techniques have the class imbalance problem [9]. The class imbalance problem can be stated as follows: for datasets labeled with two or more than two classes, the class which users are interested in is named as positive class, and the other class (es) is (are) named as negative class (es). For the imbalanced dataset, the number of positive instances is significantly lower than the number of negative instances [5]. The dataset of secondary injury in this thesis is also of this type.

Current methods for handling the class imbalance problem include re-sampling instances, adjusting misclassification costs and recognition-based learning approaches [42]. All of them have some disadvantages when they handle our imbalanced secondary injury dataset: Re-sampling instances approach requires to manually re-sample the positive or negative instances, which inevitably reduce the credibility of data mining results; both adjusting misclassification costs and recognition-based learning approach have different performance when they are applied to different classification algorithms, which cause the uncertainty of data mining results. This motivates us to employ a new approach: subgroup discovery-to analyze our secondary injury dataset, as a good supplement to existing solutions to deal with imbalanced datasets. To evaluate this approach, the patterns discovered by graphical exploratory analysis, classification algorithm applied to re-sampling dataset, and subgroup discovery algorithms are compared in this thesis.

1.2 Objective

In summary, the objectives are to:

- Find appropriate data mining methods for dealing with imbalanced datasets which are common in safety-related applications.
- Discover interesting and meaningful patterns contributing to secondary injuries induced by implementation of assistive technology.
- Provide objective evaluation measures of resulting patterns.

In all, this thesis not only discovers rules associated with secondary injury induced by the use of AT, but also provide a new direction to handle the class imbalance problem which is very common in safety-related fields: employing subgroup discovery algorithms to analyze imbalanced datasets.

To achieve the research objectives, the following steps are taken by this thesis: the first step is to preprocess data in order to prepare the survey of secondary injuries for the purpose of data mining. After the dataset is preprocessed, we should select data mining methods according to the characteristics of our secondary injuries dataset in order to discover patterns of interest. Then we respectively illustrate the analysis results of three different data mining approaches. Also, comparison of patterns generated by our three approaches testifies subgroup discovery approach's remarkable success in dealing with imbalanced datasets.

1.3 Thesis Organization

Chapter 2 reviews the relevant literature about risk assessment and assistive technologies used in agriculture field, data mining techniques applied to the safety-related fields, and data mining techniques used to deal with imbalanced dataset. Imbalanced dataset of secondary injury induced by implementation of assistive technology and data preprocessing are presented in chapter 3. Chapter 4 provides a detailed description of three approaches that are used to analyze the imbalanced dataset:

graphical exploratory analysis, decision tree algorithm, and subgroup discover algorithms. Visualized results of patterns discovered by these three approaches are presented and compared in chapter 5. Research conclusion and future work are illustrated in chapter 6.

CHAPTER 2 REVIEW OF LITERATURE

2.1 Assistive Technology in Agriculture Field

Agriculture has been estimated to be one of the United States' most hazardous industries, ranking fourth among nation's industries with high risk of occupational fatalities [32]. For example, it has been estimated that 841 deaths and 512,539 non-fatal injuries happened in 1992 for agricultural production [23]. A National Institute for Occupational Safety and Health (NIOSH) reports that approximately 210,000 injured farming workers suffer at least one-half of one-day of work loss every year [24]. Leigh et al. [23] shows that costs for United States' farming occupational injuries are considerable with range from $ 3.14 billion to $ 13.99 billion, and "agriculture injuries contribute to approximately 30% more than the national average to occupational injury costs."

Hancock et al. [22] has estimated that 25% of farmers and farm worker in United State suffer disabilities [22], which make it difficult to continue their farming jobs. Although permanent or temporary disabilities and other farming injuries, farming workers need to continue farming work [11] and often do not have the resources that allow them to recover without continuing farming activities [43]. In agricultural work environment, AT such as modifications of equipment, tools, workplace and so on is very helpful to assist farmers with severe injury or disability in continuing their work activities. Introduction of assistive technology, however, may cause secondary injury.

Assistive Technology has been helping farmers with disability or injury to continue their farming works for more than 25 years, but there is little research on how secondary injuries due to agricultural assistive technology occur. Willkomm et al. [47] have studied how AT equipments such as a wheelchair or a prothetic device might contribute to secondary injuries. Their work focuses on providing ideas and recommendations to improve design and use of AT equipments. In addition, Mathew et al. listed and rated the potential causes of AT-related secondary injury by evaluating the

single effects of causes on the occurrence of secondary injuries. The objective of our study is to analyze and identify the conditions and factors which have single or combinational effects on the occurrence of secondary injuries induced by implementation of assistive technology, and can be used to assess the possibility of a future injury for farmers supported by AT.

2.2 Introduction to Data Mining Techniques

This study employs data mining to achieve the above objective. Data mining functionality is described in [20] as follows: "data mining is the process of discovering interesting knowledge from large amounts of data stored in datasets, data warehouses, or other information repositories". With the explosive growth in volume of data collected and stored, traditional manual data analysis and interpretation became impractical [48]. Moreover, hypothesis driven methods such as most statistical methods and on-line analytical processing will generally fail to uncover hidden and unexpected knowledge especially in many data-rich but hypothesis-poor fields [10]. Inductive data mining methods, however, have the ability to uncover hidden patterns and knowledge [50]. In all, data mining methods are very powerful for indentifying hidden patterns associated with secondary injuries induced by AT.

The two high-level primary tasks of data mining are predictive knowledge discovery and descriptive knowledge discovery [48]. Descriptive knowledge discovery means to "characterize the general properties of the data in the database" [20], while the task of predictive knowledge discovery is to "perform inference on the current data in order to make predictions" [20]. According to what kind of knowledge to be mined [34], Data mining systems can be categorized into the following aspects: data summarization, mining association rules, data classification and prediction, cluster analysis, and so on [20].

Data summarization, which performs a general description for a subset of data, includes data characterization and data discrimination. Data characterization techniques usually use bar charts,

curves, and other multivariate visualization techniques to summarize the general characteristics of a target class of data"[20]. Data discrimination usually provides a general comparative description of the two or more than two target classes of data. In this thesis, the first step of our methodology is to employ one of data characterization techniques: bar charts to analyze our secondary injuries dataset.

Mining association rules is one kind of descriptive knowledge discovery. Its task is to find rules to describe the correlations of frequent attributes which frequently appear together in a dataset. For example: the association rule $buys(X,"laptop\_computer") \Rightarrow buys(X,"HP\_printer")$ found in [20] would indicate that if a customer buys laptop, he or she is likely to also buy HP printer.

Data classification and prediction is one kind of predictive knowledge discovery methods of finding a model to classify data into one of several predefined classes [48]. The output of data classification and prediction can be presented in the following forms: classification (IF-THEN) rules, decision trees, mathematical formulae, and neural networks [20]. Our goal is to classify attributes into the class of experiencing secondary injury, so the data mining task of secondary injury dataset is data classification according to the definition of data classification and prediction. In this thesis, we employ decision-tree algorithm to induce secondary-injuries model according to the characteristics of our database.

Data classification and prediction is used to analyze class-labeled datasets, however, **cluster analysis** aims to analyze dataset where class labels unknown. Cluster analysis is a descriptive knowledge discovery task where data is assigned to different clusters based on similarity, that is, "patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster" [3].

From the classification of data mining system presented above, we know that data classification and prediction algorithms are appropriate for our class-labeled dataset. In addition to decision tree algorithms, we also can use rule learning algorithms to analyze our dataset. Rule

learning, which is a very important forms of knowledge discovery, includes predictive rule learning (for example, the classification (IF-THEN) rules), and descriptive rule learning (for example, association rules learning). There is an intersection of predictive and descriptive induction: subgroup discovery. In this thesis, the final step of our research methodology is to employ subgroup discovery algorithm to deal with the imbalance problem. Subgroup discovery algorithm will be discussed further in chapter 4.

Data mining is an application-dependent issue and different datasets and applications may require different data mining techniques to deal with. We will discuss the data mining techniques applied to our secondary injury dataset in chapter 4.

2.3 Data Mining Methods Dealing with Imbalanced Datasets

Studies show that there are imbalanced datasets needed to deal with in many applications such as fraud detection of telephone calls [46], detection of oil spills in satellite radar images [33], fault monitoring [37], medical decision support [7], and minority class prediction in language field [8].

Imbalanced datasets cause poor performances from standard classification algorithms. Many of standard classification algorithms usually assume that training examples are evenly distributed among different classes. These classification algorithms generate classifiers that maximize the overall classification accuracy. Trivial classifiers that completely ignore the minority class, that is, the concept of minority class with few examples is difficult to be uncovered. This will cause problems when people focus on minority class.

As stated in section 1.1, several methods have previously been proposed to deal with imbalance problem including re-sampling training sets, adjusting misclassification costs and recognition-based learning approaches.

These three methods have their own advantages and disadvantages, and get different results when are applied on the same imbalanced datasets. In all, we may not make a conclusion on which method can get best result.

2.3.1 Re-sampling training dataset

Re-sampling training sets method includes over-sampling minority class examples and under-sampling majority class examples in order to make datasets balanced [5] [2]. Re-sampling methods are usually used for handling imbalanced dataset because such methods are very simple to implement externally. There are two kind of re-sampling methods: over-sampling and under-sampling.

2.3.1.1 Over-sampling examples of minority class

Over-sampling minority class examples method is to keep all positive examples in the training set, but replicate negative examples to form new training sets [2]. It is apparent that adding negative examples does not increase information, but it does increase the misclassification cost.

There are two kind of over-sampling strategies: replicating randomly and guided re-sampling [5]. The first method is a simple method which randomly selects and duplicates negative examples until the two classes are balanced. The second method aims to clear not only between-class imbalances but also within-class imbalances [5]. The guided re-sampling method is much better than random re-sampling method by increasing classification accuracy.

Within-class imbalances refer to the difference between the relative densities of the subcomponents within a single class (either majority class or minority class) [5]. Take a letter recognition dataset for example, which includes the majority class containing letter A and B and the minority class containing letter H and F. In the minority class, examples containing F are much fewer than examples containing H. if a random re-sampling technique is applied, these F examples would

not copied as often as H examples, which increase within-class imbalances and impair performance of re-sampling approach.

When the exact relative densities of components of classes are known in advance, the re-sampling process can be guided by copying more "underrepresented' examples in negative class. It is true for many classification problems that data distribution of each single class are unknown, Nickerson , Japlowicz, and Milios introduced an unsupervised clustering algorithm PDDP to find clusters for each single class and employ the clustering result to resample appropriately [5].

Nickerson et al. compared the performance of guided re-sampling and random re-sampling on two letter recognition domains and found that this approach can get higher classification precision and recall than random re-sampling especially for imbalanced dataset including very-underrepresented examples in minority class [5].

2.3.1.2 Under-sampling examples of majority class

As opposite to over-sampling, the under-sampling method is to remove randomly some examples from majority class in order to balance the two class examples [2]. This method can be applied on imbalanced datasets when people cares more about the minority class because this method may lose information from majority class.

Both over-sampling and under-sampling can be applied on imbalanced datasets and have their own advantages and disadvantages. We cannot tell that oversampling is better than under-sampling or the opposite case. Estabrooks et al. [2] conducted an experimental study on different datasets to show combination of over-sampling and under-sampling is an effective way to deal with imbalance problem and have better results than single over-sampling or under-sampling method.

2.3.2 Adjusting misclassification costs

Adjusting misclassification costs method solves the imbalance problem by setting different misclassification error costs [42]. For example, this method may set high cost to the misclassification of a minority class example when people care about the minority class. Domingos proposed MetaCost which is a new procedure making classifiers cost-sensitive [39]. Sun et al. introduced three cost-sensitive boosting algorithms into the learning framework of AdaBoost algorithm in order to improve the classification accuracy of imbalanced data [52].

2.3.3 Recognition-based learning approach

Recognition-based learning approaches focus on analyzing and identifying rules from minority class but ignore majority class [42].

Zhang et al. [27] designed a novel recognition-based learning algorithm – RLSD (Rule Learning for Skewed Data), which frequent patterns of the minority class. Guo and Viktor [15] present a novel approach named DataBoost-IM approach for learning imbalanced datasets. This approach combining data generation and boosting procedures as an amelioration of DataBoost algorithm also brought forward by them in order to focus on hard examples which are difficult to classify. They tested DataBoost-IM approach on seventeen imbalanced datasets and evaluated its performance in comparison to a couple of traditional decision trees algorithms, and conclude that classification results of this new approach are promising and slightly better for both majority and minority class.

2.3.4 Comparison and connection of the methods above

Adjusting misclassification cost method and recognition-based learning method belong to internal approaches which take the imbalanced problem into consideration by employing new algorithms or improving existing ones. Studies test that internal approaches performs well in handling imbalanced datasets because new algorithms created according to specific dataset effectively take

different characteristic of different dataset into consideration. However, the performance of internal approaches is very dependent on classifiers, that is, for certain imbalanced dataset handled by certain internal approach, different classifier has different results. It is uncertain which classifier can yields the best results.

As opposite to internal approaches, external ones do not modify or create algorithms, but re-sampling the datasets in order to reduce the impact caused by the imbalanced problem. Re-sampling training datasets belongs to external approach, but there are certain connections between internal and external approaches. Maloof [31] states that varying decision threshold or the cost matrix is equivalent to resizing datasets for dealing with imbalanced datasets by analyzing ROC curves. Just as Breiman et al. [28] concluded in their book that there is connection among the error costs, decision threshold changes, distribution of examples in the training set, and the probability distribution of each class. Varying one of these elements has the same effect as varying any other.

CHAPTER 3 DATA PREPROCESSING

3.1 Description of Secondary Injury Dataset

As stated in chapter 1, in agricultural work environment, assistive technology (AT) such as modifications of equipment, tools, workplace and so on, is very helpful to assist farmers with severe injury or disability in continuing their work activities. Introduction of assistive technology, however, may cause secondary injury.

In order to evaluate the risk of secondary injury, the scholars from the Department of Agricultural and Biosystems Engineering at Iowa State University conducted a survey that was distributed to all of the farmers with disabilities that are registered in Iowa's AgrAbility program. The survey was conducted between the end of harvest and beginning of preparation and planting season at late/fall winter of 2007/2008. This survey has 236 respondents out of the sample consisting of approximately 720 individuals.

The survey was organized into four sections as follows: Section1 is demographic information including age, gender, ethnicity and primary source of farm income (cash grain, beef, dairy, swine etc.). Section 2 covers information on initial disability including nature of disability (NOD) and how/where it was acquired (farm work, traffic accident etc). Section 3 covers information on assistive technology including what modifications of equipment/machines were made to accommodate disability (MNA), who designed and made the assistive technology (ATD), and who installed the assistive technology (ATI). Section 4 covers aspects of the secondary injury, if occurred: how many secondary injury events happened, the nature of secondary injury and so on. A complete list of the available independent attributes is shown in Table 3.1.

There is a specific attribute "Experienced injury using AT?" that is of primary interest, and all of the data instances (survey responses) are labeled according to this attribute as either "Yes" or "No". For convenience, we will call the instances in class 0 ("No") negative instances and instances

in class 1 (“Yes”) positive instances in the rest of this thesis. We would like to classify and predict if secondary injury with assistive technology happened based on other independent attributes. This means that the natural learning task is data classification.

3.2 Data Preprocessing

Real-world datasets are generally incomplete, noisy and inconsistent because of the explosive growth in data and dataset [20], so data preprocessing plays a very important role in data mining. Data preprocessing includes data cleaning, data integration, data transformation, data reduction and data discretization [20]. Different datasets need some or all of these data preprocessing operations. Data preprocessing methods applied on our survey dataset are described in the following subsections.

3.2.1 Removal of redundant information

We ignore the attributes related to additional information on secondary injury because they are not useful to predict whether people will experience secondary injury or not.

We also delete all combinational attributes that simply combine other single attributes. For example, under the category of primary source of farm income, "cash grain" and "beef" are listed as two separate attributes and "cash grain and beef" is also listed. Therefore, an instance with "yes" at the first two attributes will has the value of "yes" at the last attribute for sure, which indicates the last one is actually redundant.

In addition, we delete the attribute “Need to install AT?” and all instances with value 0 (no) for attribute “Need to install AT?”, because these instances who do not need to install AT did not use AT at all and are irrelative to AT and the data mining task.

After these modifications, the dataset includes 143 instances and 74 attributes including the class attribute “Experienced injury using AT?” Then we transfer the dataset from XML format to CSV format which is required by data mining software WEKA [17], which was used for the remaining

analysis. WEKA is an open-source data mining software written in Java, developed at the University of Waikato in New Zealand [17]. WEKA "contains a collection of algorithms for data mining tasks, including data preprocessing, association mining, classification, regression, clustering, and visualization"[17].

3.2.2 Discretization

After loading CSV format dataset into WEKA, We have done the following preprocessing steps to prepare dataset for the further classification analysis.

We discretize the attributes "age" and "# of injury years". Nominal attributes with more than three values such as "ethic" and "farmed after injury but now retired" are also transferred through discretization. The attributes that have two values are converted from numeric to binary.

3.3 Imbalanced Dataset of Secondary Injuries

After preprocessing, the dataset has 74 attributes, 130 negative instances and 13 positive instances. From the distribution of class values perspective, our secondary injury dataset is highly-imbalanced dataset; from the aim of data mining perspective, our learning task is classification of two-class dataset.

Demographic	Primary Source of Income (PSOI)	Nature of Disability (NOD)		Initial Disability (HA)	Modification Needed (MNA)	Assistive Technology Designer (ATD)	Assistive Technology Installer (ATI)
Age	Cash Grain	Spinal Cord	Muscular	Farm work	Lift	Home	Self
Gender	Beef	Paraplegia	Dystrophy	Traffic accident	Controls	Family	Family
Ethnicity	Dairy	Spinal Cord	Cancer	Recreational	Machine Mod.	Local	Local
	Forage/hay	Quadriplegia	Diabetes	activity	Mobility	Manufac. Comp.	Manufac. Comp Rep
	Swine	Amputation upper	Respiratory	Health condition	Mod. Tools	Profess. AT	(Not specialized with
	Field Crops	Amputation lower	Cerebral Palsy	At home	Structure Mod.	manufacturer	AT)
	Forestry	Arthritis	Burn	Congenital	Other		Profess. AT manufac.
	Horses	Hearing	Amyotrophic Lateral	Other			Rep
	Truck Crops	Orthopedic Elbow	Cognitive	# years ago			
	Sheep	Orthopedic Hip	Other				
	Tobacco	Orthopedic Knee					
	Poultry	Orthopedic other					
	Nursery/Greenhouse	Visual					
	Fruit tree/Orchard	Polio					
	Horticulture	Head					
	Aquaculture	Heart					
	Cotton	Multiple Sclerosis					
	Other	Hand					

Table 3.1 List of independent attributes of secondary injury dataset

CHAPTER 4 THREE APPORACHES APPLIED TO SECONDARY INJURY DATASET

As stated in chapter 3, our secondary injury dataset is highly imbalanced, in which most of all instances belong to the negative class, but positive class which we are more interested in has only a small portion of all instances. The primary data mining challenge is therefore to determine an appropriate method for handling the class imbalance problem. Our methodology proposes three approaches: graphical exploratory analysis, decision tree algorithm, and subgroup discovery algorithms. The three approaches applied to our imbalanced dataset will be discussed in details in the beginning three sections of this chapter.

4.1 Graphical Exploratory Analysis

Graphical exploratory analysis was described by Tukey as "graphs intended to let us see what may be happening over and above what we have already described" [26]. Graphical exploratory analysis usually employs visualization techniques such as categorized graphs (histograms, scatterplots, pie charts, etc.), brushing, icon plots, etc. to identify the most relevant variables and picture the general natures of models of datasets. As the first stage of data analysis, this approach is often used to filter irrelative variables and make the following data analysis more effective. The analysis results, however, are tentative and need to be confirmed by advanced data mining techniques.

It is all of current research of assessing the potential secondary injuries of assistive technology that use graphical exploratory analysis method (histograms) to simply rate causative factors of secondary injuries and give summary findings from surveys. In chapter 5, we firstly employ exploratory analysis method to analyze our secondary injury dataset.

4.2 Decision Tree Algorithm with Re-sampling Methods

As stated in section 2.2, we know the learning task of secondary injury dataset is data classification. Regarding the data classification methods, there are a couple of options including

classification by decision tree induction, Bayesian classification, rule-based classification, classification by back-propagation, support vector machines, and so on [20]. In this section, we will introduce data classification and decision tree algorithm (J48) which we employ to analyze our secondary injury dataset, while the introduction of other classification methods mentioned above can be found in [20].

As a second approach of our methodology, we will employ decision tree algorithm (J48) to analyze our dataset. The first step is to handle the class imbalance problem. To handle the class imbalance, this thesis choose re-sampling (both over-sampling and under-sampling) the secondary injury dataset before applying decision tree algorithm to the balanced dataset. Over-sampling approach replicates positive instances in order to balance the number of instances in the two classes, while under-sampling approach randomly samples to decrease the number of negative instances. The details of re-sampling of our secondary injury dataset will be introduced in chapter 5. The second step is to import the balanced dataset to WEKA to make classification analysis. The details of results will be addressed in chapter 5.

4.2.1 Introduction to classification

	Age	Income	Student	Credit_rating	**Class:** Buys- comupter
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no

Table 4.1 Class-labeled training dataset from the AllElectronics customer dataset [20]

Classification induction is a data mining technique used to classify instances into different classes from class-labeled dataset [20]. For example, a class-labeled dataset is illustrated in table 4.1, which is made up of five attributes and eight instances. "Buys_computer" is class attribute which has two values: yes or no. Classification, for example, can be employed by sales manger to extract patterns or predict who will buy computers [20].

The form of a classifier can be classification rules, decision trees, or mathematical formulae [20]. In the following section, we will illustrate the decision tree algorithm which is used to analyze our dataset.

4.2.2 Introduction to decision tree algorithms

Figure 4.1 is a typical decision tree structure, where each internal node denoted by a rectangle tests an attribute, each branch corresponds to attribute value, and each leaf node denoted by an oval assigns a classification. The basic decision tree algorithm can be found in [20].

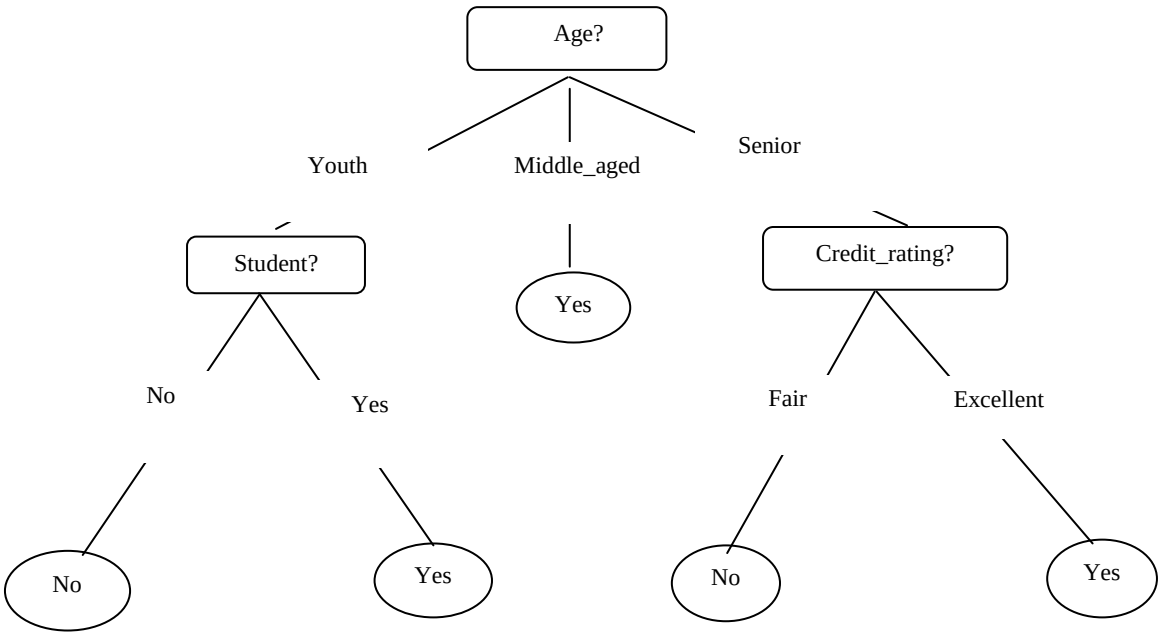

Figure 4.1 Decision tree representation[20]

Decision tree algorithms have the following advantages [20] [30] over other supervised classification algorithms such as classification rules and mathematical formulae. First, decision trees do not require any domain knowledge, parameter setting or any assumptions about distribution of input dataset. Second, decision trees perform very well with high-dimensional dataset. Third, they can handle dataset with missing data. In addition, the learning steps of decision tree induction requires less computer time. Finally, they can be easily interpreted because of the explicit decision tree structure. Therefore, decision tree algorithms are the most popular symbolic learning method.

Decision tree algorithm is the main algorithm for our dataset because of the following reasons:

- Class attribute “Experienced injury using AT?” already exists, so our secondary injury dataset is class-labeled dataset. Moreover, the class attribute is discrete-valued (binary: Yes or No), which is the reason for choosing classification induction to analyze our dataset.

- Our dataset is high-dimension containing 74 attributes.
- Some attribute values of our dataset is missing, because our dataset is converted from a survey data which inevitably contain missing information.
- We need to experiment algorithms many times in order to identify the best patterns associated with secondary injuries, so fast algorithms have to be selected.
- Routes of decision tree leading to confirmation of second injury can be used as rules to discover future injuries.

Decision tree algorithms include C4.5 [25], CART [28] and other new algorithms [51]. We employ J48 decision tree algorithm [18] which is slightly modified C4.5 in WEKA to analyze our dataset.

4.3 Subgroup Discovery Algorithms

As the third approach of our methodology, we will employ subgroup discovery algorithms to analyze our imbalance dataset. The first step is to import the imbalanced dataset into WEKA to discretize the continuous attributes because subgroup discovery algorithms SD and CN2-SD take as their input the training instances described by discrete attributes values [19]. The second step is to import the discretized imbalanced dataset into ORANGE [19] so that we can employ SD and CN2-SD algorithms to analyze our dataset. ORANGE is a data mining software which supports several data mining tasks including data preprocessing, development of classification models, regression methods, association rules, clustering methods, and so on [19]. We use subgroup discovery toolkit [40] for ORANGE including SD, CN2-SD, and Apriori-SD algorithms to analyze our dataset. The details of analysis results will be addressed in chapter 5.

4.3.1 Introduction to subgroup discovery

The definition of subgroup discovery can be stated as follows: “Given a population of individuals and a property of individuals we are interested in, find population subgroups that are statistically ‘most interesting’, e.g., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest”[44] [45].

An induced subgroup discovery rules has the following description form:

$$Cond \rightarrow Class \tag{4.1}$$

Where the rule consequent $Class$ stands for the target class values (target variables), $Cond$ stands for conjunction of attributes-values (independent variables) pairs from training datasets [35]. For example, a subgroup discovery rule is described by “Nature of disability: Amputation upper = true AND Nature of disability: Heart = false $\rightarrow$ Experienced Injury using AT = true”. In the above description form, the values of target class (the target variables), which is predefined a property of interest of subgroup discovery task may be binary, nominal or numeric [29]. Our target variable in this thesis is binary: 0 means the respondents did not experience secondary injuries with AT, while 1 stands for the opposite results.

Although subgroup discovery as descriptive induction language aims at discovering individual rules of interest, it can be used to classify datasets as predictive induction language because of the following reasons [44] [45]. First of all, subgroup rules are induced from labeled (positive instances and negative instances) training datasets; second, the form of induced subgroup rules is $Cond \rightarrow Class$ which belongs to supervised machine learning forms. Therefore, the process of subgroup discovery is limited to finding interesting rule sets for target class and can be used for classification purposes.

4.3.2 Subgroup discovery algorithms

In this section, we will illustrate common subgroup discovery algorithms. Moreover, we will compare subgroup discovery algorithms to classification rule algorithms so that we can testify why subgroup discovery algorithms are able to directly analyze imbalanced dataset without re-sampling datasets.

Earlier subgroup discovery algorithms include EXPLORA developed by Klösgen [49] and MIDOS developed by Wrobel [44] [45]. In addition, there are several advanced subgroup discovery algorithms including CN2-SD [36], APRIORI-SD [6], and SD [12]. In this section, we will illustrate CN2-SD and SD algorithms which we choose to analyze our dataset because APRIORI-SD algorithm takes too much computer time (usually over 2 hours) to analyze our dataset.

4.3.2.1 SD algorithm

SD algorithm, which aimes at searching for rules covering many target class instances and a low number of non-target class instances [13], is developed by Gamberger et al. [12]. SD basic algorithm can be found in [12].

4.3.2.2 CN2-SD algorithm

CN2-SD algorithm is developed by Lavrac et al. [36] by modifying CN2 algorithm, which is a classical classification rule induction algorithm [38]. Lavrac et al. [36] modified parts of CN2 classification rule learner including its covering algorithm and search heuristic to the aim of subgroup discovery. We will illustrate the CN2 and CN2-SD algorithm in details in the rest of this section.

CN2 is a rule-based classifier where induced rules have the form: $Cond \rightarrow Class$. Its algorithm can be found in [38]. After studying at the CN2 algorithm, it is known that CN2 algorithm removes instances covered by the rule at the current iteration from the entire instance subsets, which results in subsequent rules induced from incomplete instance subsets. CN2-SD overcomes this bias by weighting instances in the subsequent iterations instead of deleting covered instances. In the first

iteration all instances are assigned the same weight: $w(e_j, 0) = 1$, which means these instances have not been covered by any rule. In the following iteration, weights of instances covered by one or more rules will decrease according to weighting scheme. It is usual to find that there are two weighting schemes. The first one is multiplicative weights, whose weight equation is defined as follows:

$$w(e_j, i) = \gamma^i \qquad (0 < \gamma < 1) \tag{4.2}$$

where i means the number of rules that cover instance e_j.

The second weighting scheme is additive weights, where the weight of instance decreases according to the following equation:

$$w(e_j, i) = \frac{1}{i+1} \tag{4.3}$$

CN2 employs classification accuracy of the rule as a heuristic function to check if adding such a new rule to rule set will result in an improved rule sets. The accuracy of classification rule is defined as the following [36]:

$$Acc(Cond \rightarrow Class) = p(Class \mid Cond) = \frac{p(Class.Cond)}{P(Cond)} \tag{4.4}$$

The accuracy probability is usually estimated by the following relative frequency [36]:

$$\frac{n(Class.Cond)+1}{n(Cond)+k} \tag{4.5}$$

$n(Cond)$ stands for the number of instances covered by the rule $Cond \rightarrow Class$, $n(Class)$ stands for the number of instances of $Class$, and $n(Class.Cond)$ stands for the number of correctly classified positive instances.

Subgroup discovery algorithm CN2-SD employs weighted relative accuracy as search heuristic function in order to trade off generality of the rule and relative accuracy, since considering

only accuracy of the rule will easily result in highly over-fitted rules with high accuracy. Weighted relative accuracy is defined as the following [36]:

$$WRAcc(Cond \rightarrow Class) = p(Cond) \bullet (p(Class \mid Cond) - p(Class)) \quad (4.6)$$

Lavrač et al. [36] further modified the weighted relative accuracy by incorporating instances weights in order to provide more appropriate rule quality measure for each iteration of the weighted covering algorithm. The modified weighted relative accuracy with instance weights is described as the following:

$$WRAcc(Cond \rightarrow Class) = \frac{n'(Cond)}{N'} \bullet (\frac{n'(Class.Cond)}{n'(Cond)} - \frac{n'(Class)}{N'}) \quad (4.7)$$

In this equation, N' stands for the sum of the weights of all instances $n'(Cond)$ stands for the sum of the weights of all covered instances, and $n'(Class.Cond)$ stands for the sum of the weights of all correctly covered instances by the rule [36].

4.3.3 Applications of subgroup discovery

All papers related to subgroup discovery applications have been reviewed by this thesis. There are just two papers which mentioned that subgroup discovery algorithms have the power to deal with imbalanced dataset by providing empirical evidences.

Kavšek and Lavrač [6] developed subgroup discovery algorithm APRIORI-SD and applied it to UCI datasets and U.K. traffic accident dataset which is highly imbalanced (the distribution of minority class is 6.01%). In **Lavrač's another paper [36]**, another subgroup discovery algorithm CN2-SD has been developed and applied to the same datasets. Their main concern is to testify APRIORI-SD and CN2-SD product better rule-sets compared to other rule learners and are competitive subgroup discovery algorithm compared to other subgroup discovery algorithm. From the

experimental results, however, they found that subgroup discovery algorithms are able to overcome the imbalanced bias and produce interesting rule-sets for minority classes.

4.4 Comparisons of Decision Tree and Subgroup Discovery Algorithms

In this section, we will compare subgroup discovery algorithms to decision tree algorithm when they are applied to imbalanced datasets.

First, subgroup discovery algorithms aim to overcome the problem of inappropriate bias to majority class of the standard classification algorithms when they are used to predict a minority class. First of all, standard classification algorithms use covering algorithm for rule-set construction. Covering algorithm makes the consequent subsets include only positive examples not covered by previously induced rules. This bias causes the loss of positive examples information, which imbalanced datasets cannot afford especially when the positive class is a minority one (our secondary injury dataset only has 13 positive instances). Subgroup discovery algorithms overcome this problem by employing weighted covering algorithm for rule-set construction. In addition, standard classification algorithms (for example, J48 algorithm) use rule-set accuracy for search heuristics, but subgroup discovery algorithms uses weighted relative accuracy WRAcc for search heuristics and objective quality measurement [12].

Second, J48 aims at maximizing classification accuracy of induced rulesets. However, subgroup discovery algorithms focus on finding interesting rulesets for the entire population. This principle allows rule-sets induced from subgroup discovery algorithm tolerate more false positives, which is good for our imbalanced dataset because some of negative examples (majority class) may be become positive examples in future (some of farmers who have not experienced secondary injuries by the time of the survey may experience secondary injuries in future).

Third, another advantage of subgroup discovery algorithms is stated as the following in Nada Lavrač's paper [36]: " subgroup discovery aims at discovering individual rules or 'patterns' of

interest, which must be represented in explicit symbolic form and which must be relatively simple in order to be recognized by potential users." This advantage is important for classification of minority class in imbalanced datasets because users, especially when users are not experts, and cannot afford indirect explanations from standard classification algorithms like J48 decision tree for minority class including less information in imbalanced datasets.

The last but not the least, compared to other traditional methods applied to imbalanced datasets such as sampling down and sampling up, subgroup discovery algorithms are convenient for users because they do not need to replicate minority examples manually or randomly sample down datasets.

4.5 Evaluation Measurements Used to Compare Patterns of Secondary Injuries

In order to compare all of patterns discovered by our three approaches, this thesis introduces three evaluation measurements as follows:

$$Support(X \Rightarrow Y) = P(X \cup Y) \tag{4.8}$$

$X \Rightarrow Y$ stands for a pattern (rule), where X stands for an attribute set (attributes and attributes values) contributing to secondary injuries, and Y stands for the positive class : class 1. $Support$, which "represents the percentage of instances from a database that the given rule satisfies" [20], is "taken to be the probability $P(X \cup Y)$. $X \cup Y$ indicates that an instance contains both X and Y " [20].

$$Confidence(X \Rightarrow Y) = P(Y|X) = \frac{P(X \cup Y)}{P(X)} \tag{4.9}$$

Confidence "is taken to be the conditional probability $P(Y|X)$, that is, the probability that an instance containing X also contains Y" [20].

$$Lift(X \Rightarrow Y) = \frac{Confidence(X \Rightarrow Y)}{P(Y)} = \frac{P(X \cup Y)}{P(X) \bullet P(Y)} \tag{4.10}$$

Lift is "a correlation measure" [20]. If the value of *Lift* is less than 1, the occurrence of X does NOT indicate the occurrence of Y; If the value of *Lift* is equal to 1, the occurrence of X is independent of the occurrence of Y; if the value of *Lift* is greater than 1, X and Y are positively dependent and correlated as events, "which means the occurrence of one implies the occurrence of the other" [20]. The higher the value of *Lift*, the more likely that the existence of X and Y together in an instance is because there is a relationship between them.

We will use the pattern: "If PSOI cash grain=1 then experiencing secondary injuries=1 (support=6.3%, confidence=10.2%, lift=1.125)" to illustrate the meaning of support, confidence and lift. A 6.3% support means that 6.3% of all of the instances under analysis showed that respondents whose primary source of income is cash grain experienced secondary injuries. A confidence of 10.2% means that if a respondent's primary source of income is cash grain, there is a 10.2% chance that she/he will experience secondary injuries. Lift=1.125 >1 means the occurrence of PSOI cash grain has a positive impact on the occurrence of experiencing secondary injuries.

In order to conveniently compare all of patterns according to the three evaluation measures, we will give a sequence number to every pattern according to how they are generated. For example, patterns discovered by graphical exploratory analysis are named as G-1, G-2, and so on; patterns discovered by J48 decision tree algorithm are named as D-1, D-2, and so on; patterns discovered by subgroup discovery algorithms are named as S-1, S-2, and so on.

CHAPTER 5 IMPLEMENTATION AND RESULTS

We will implement the three approaches proposed by this thesis to our secondary injury dataset in this chapter. Section 5.1 will illustrate the experiment results of graphical exploratory analysis, section 5.2 will address the experiment results of J48 decision tree algorithm, and section 5.3 will demonstrate the experiment results of subgroup discovery algorithm (SD algorithm). Comparison of patterns discovered by the three approaches is illustrated and potentially most interesting patterns are found in section 5.4.

5.1 Graphical Exploratory Analysis

For the first approach, exploratory data analysis, we plotted the raw data in terms of each of the main categories of the survey. The categories include Primary Source of Income (PSOI), Nature of Disability (NOD), Initial Disability (HA), Modification Needed (MNA), Assistive Technology Designer (ATD), and Assistive Technology Installer (ATI).

Out of eighteen categories of Primary Source of Farm Income (PSOI) (Table 3.1), there were no respondents that reported their PSOI as truck crops, tobacco, fruit tree/orchard, horticulture, aquaculture, or cotton. There were no recorded secondary injuries when PSOI is dairy, swine, field crops, forestry, poultry, nursery /greenhouse. Among the remaining six categories which relate to secondary injuries, the most common PSOI is cash grain which has nine respondents of secondary injuries (G-1 pattern). PSOI beef with four respondents is the second place (G-2 pattern), followed by PSOI forage/hay (G-3 pattern), other (G-4 pattern), horses (G-5 pattern), and sheep (G-6 pattern) (Figure 5.1). One instance may have more than a single PSOI category, so the total number of positive respondents of PSOI is higher than thirteen.

Nature of Disability (NOD) has twenty nine categories (Table 3.1). There were eleven categories where secondary injuries were reported (Figure 5.2). Seven secondary injuries happened

when NOD is amputation upper (G-7 pattern), followed by NOD spinal cord paraplegia (G-8 pattern), orthopedic other (G-9 pattern), orthopedic hip (G-10 pattern), orthopedic knee (G-11 pattern), spinal cord quadriplegia (G-12 pattern), amputation lower (G-13 pattern), arthritis (G-14 pattern), hearing (G-15 pattern), orthopedic elbow (G-16 pattern), and other (G-17 pattern). There were no recorded secondary injuries for the remaining NOD categories. One instance may have more than a single NOD category, so the total number of positive respondents of NOD is higher than thirteen.

Out of five categories of Initial Disability (HA), the most common category related to secondary injuries is HA farm with eight secondary injury respondents (G-18 pattern). HA traffic has two secondary injury respondents (G-19 pattern), and each of HA recreational (G-20 pattern), health (G-21 pattern), home (G-22 pattern) and other (G-23 pattern) have one secondary injury respondents (Figure 5.3). One instance may have more than a single HA category, so the total number of positive respondents of HA is higher than thirteen.

Out of seven categories of Modification Needed (MNA), there were two categories, namely, MNA lift and MNA mobility which are not related to secondary injuries. Among the remaining five categories, the most common category related to secondary injuries is MNA controls with eight secondary injury respondents (G-24 pattern), followed by MNA modification tools (four seconday injury respondents, G-25 pattern), machine modification (two secondary injury respondents, G-26 pattern), sturcture modification (one secondary injury respondents, G-27 pattern), and MNA other (one secondary injury respondents, G-28 pattern) (Figure 5.4). One instance may have more than a single MNA category, so the total number of positive respondents of MNA is higher than thirteen.

All of five Assistive Technology Designer (ATD) categories have secondary injury respondents. Most secondary injuries happened when ATD is home (G-29 pattern) and professional AT manufacturer (G-30 pattern). Each of them has five secondary injury respondents. ATD local machine shop has two secondary injury respondents (G-31 pattern), followed by ATD family (G-32

pattern) and manufacturing company (G-33 pattern) (Figure 5.5). One instance may have more than a single ATD category, so the total number of positive respondents of ATD is higher than thirteen.

Assistive Technology Installer (ATI) has five categories (Table 3.1). There were four categories where secondary injuries were reported (Figure 5.6). Most secondary injuries (four secondary injuries, G-34 pattern) happened when AT is installed by respondents themselves, followed by local machine shop (3 secondary injuries, G-35 pattern), family (1 secondary injury, G-36 pattern) and professional AT manufacturing representative (1 secondary injury, G-37 pattern). There was no recorded secondary injuries when the AT is installed by a manufacturing company representative not specialized with AT.

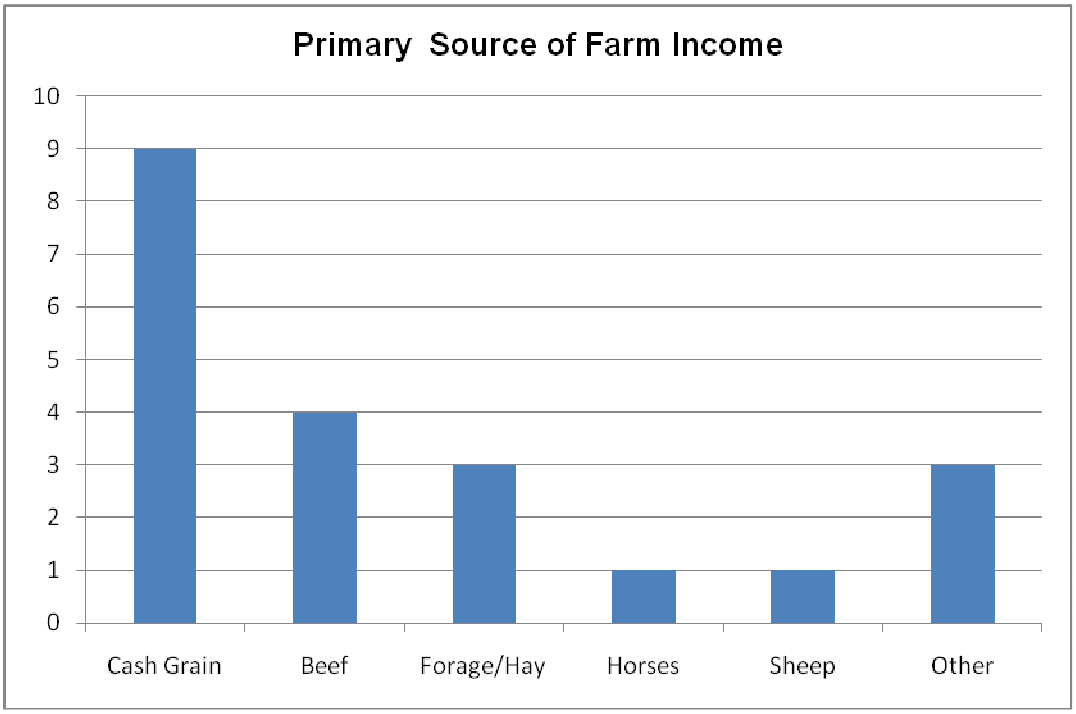

Figure 5.1 Secondary injuries grouped based on Primary Source of Farm Income (PSOI). The most common PSOI is cash grain, followed by beef, forage/hay, other, horses, and sheep . there were no recorded secondery injuries when PSOI is dairy (8), swine(12), field crops(2), forestry(2), poultry(4), nursery/greenhouse(2). There were no respondents that reported their PSOI as truck crops, tobacco, fruit tree/orchard, horticulture, aquaculture, or cotton.

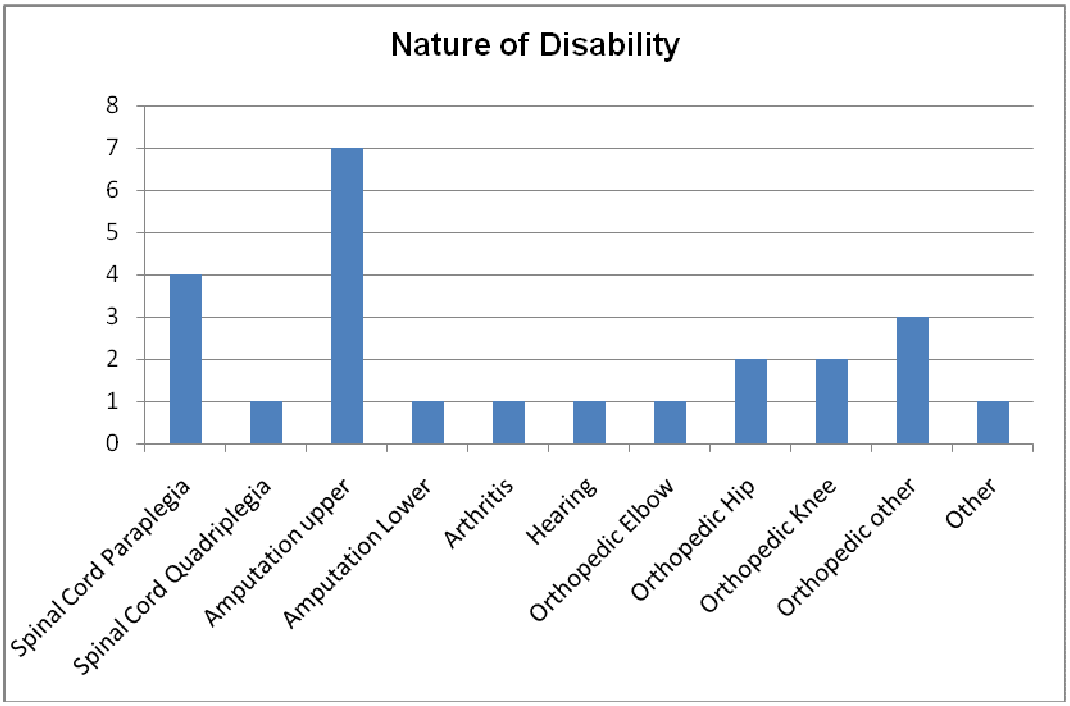

Figure 5.2 Secondary injuries grouped based on Nature of Disability (NOD). The most common NOD related to seconday injuuries is amputation upper, followed by spinal cord paraplegia, orthopedic other and other NOD factors. There were no recorded secondery injuries when NOD is visual(12), polio(7), head(9),heart(14), multiple sclerosis(6), hand(11), muscular dystrophy(3), cancer(3), diabetes(19), respitory(4), cerebral palsy(1), burn(2), amyotropic lateral sclerosis(2), or cognitive (3).

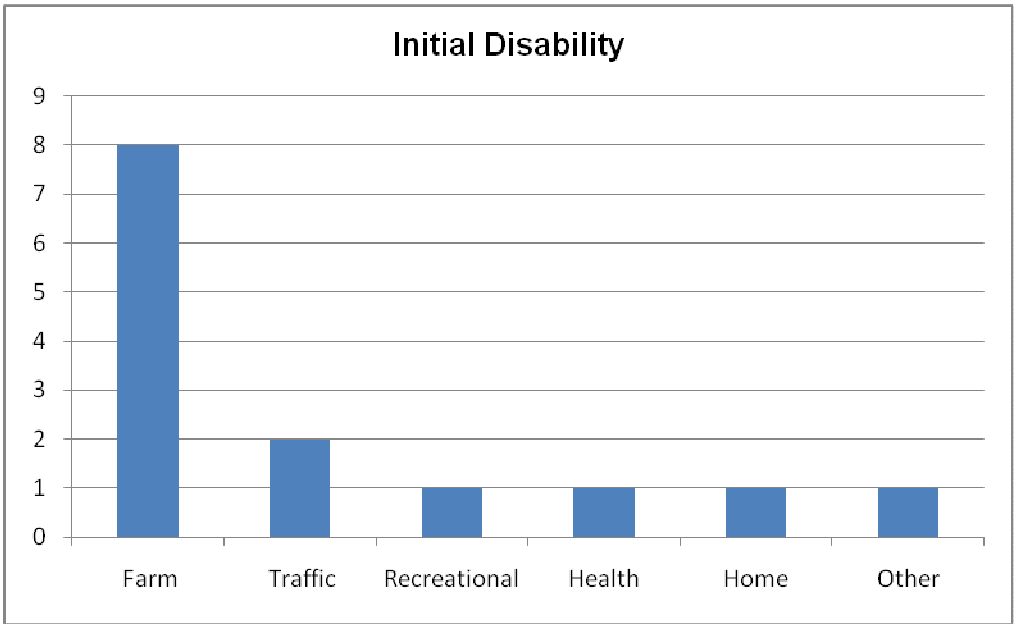

Figure 5.3 Secondary injuries grouped based on Initial Disability (HA). Most secondary injuries occurred when the initial disability is related to farm, followed by initial disabilities related to traffic, recreational, health, home and other.

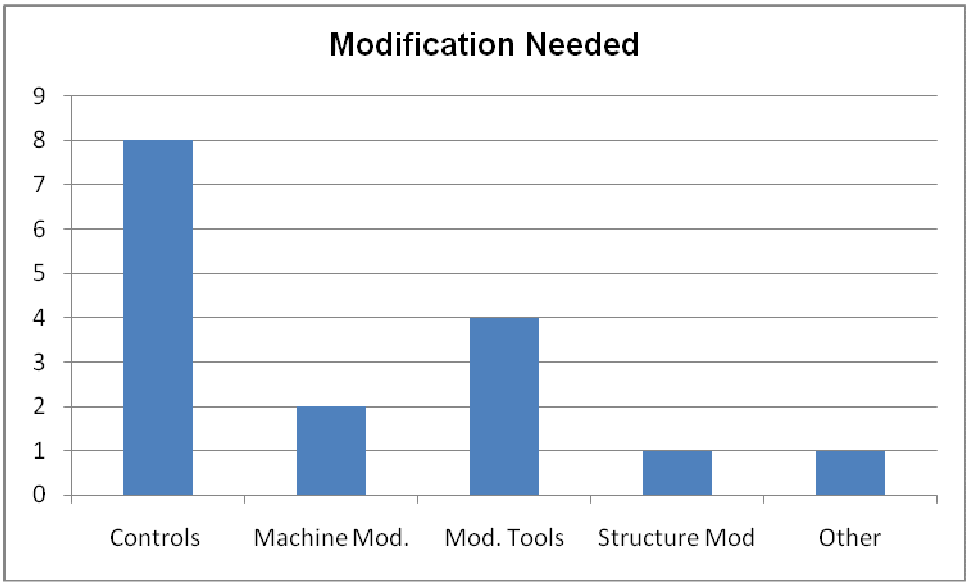

Figure 5.4 Secondary injuries grouped based on Modification Needed (MNA). Most secondary injuries occurred when the MNA is related to controls, followed by MNA related to mod.tools, machine mod. sturcure mod., and other. There were no recorded secondery injuries when MNA is lift(9) and mobility (8).

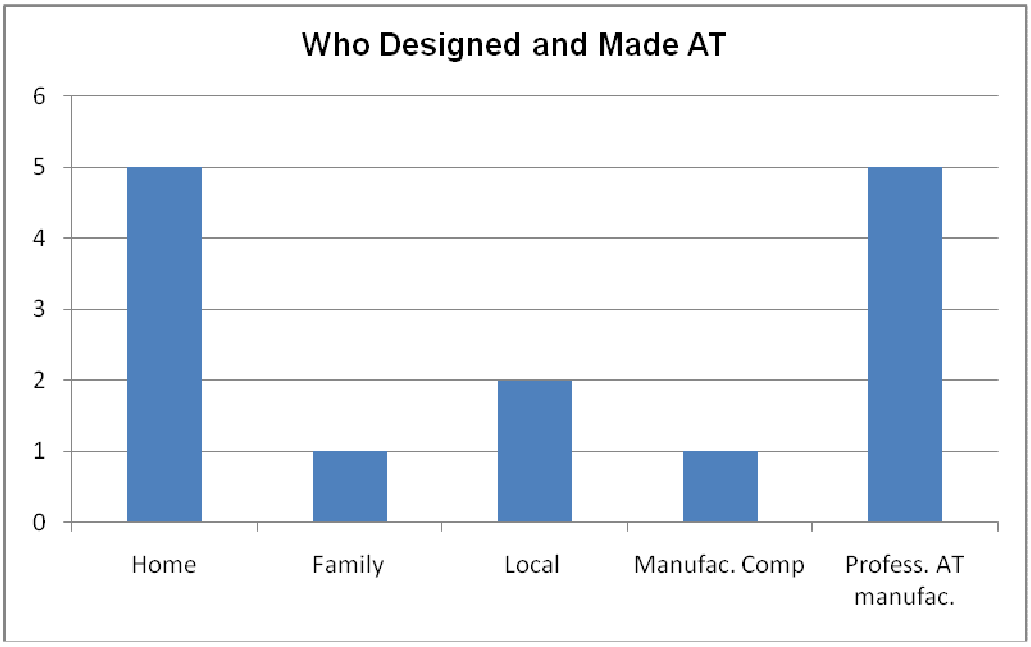

Figure 5.5 Secondary injuries grouped based on Assistive Technology Designer (ATD). Most secondary injuries occurred when the ATD is home and profess.AT manufac. rep., followed by local machine shop, family and manufac. company.

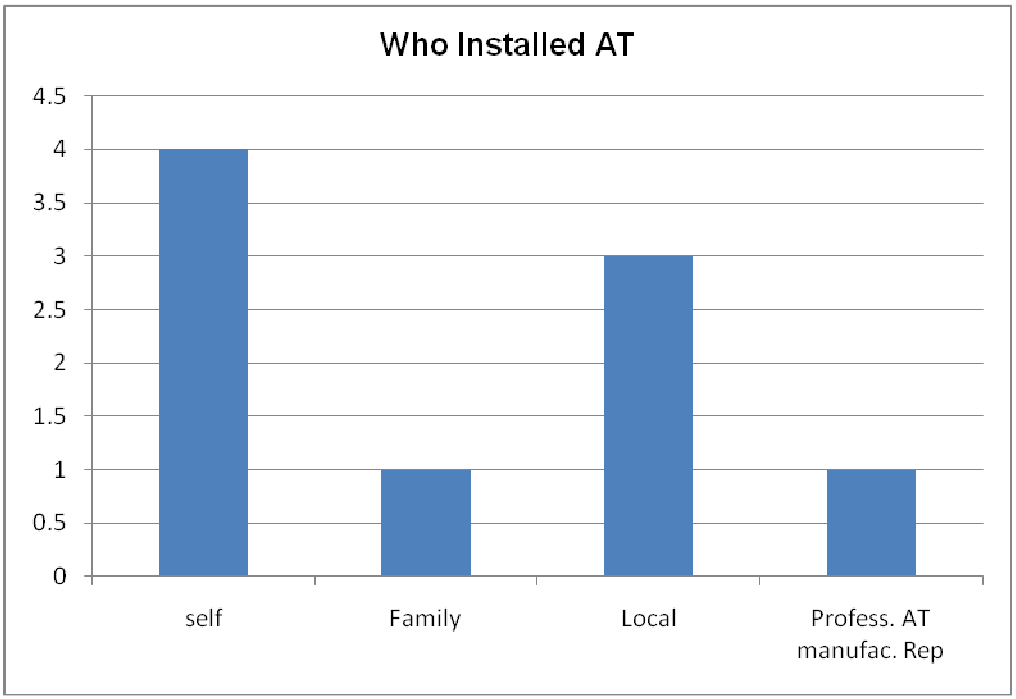

Figure 5.6 Secondary injuries grouped based on Assistive Technology Installer (ATI). Four of nine secondary injuries occurred when assistive techonoly was installed by repondents themselves, followed by local machine shop (3), family (1) and profess.AT manufac. Rep (1). there was no recorded seconday injuries when the AT is installled by a manufacturing company representative not sepcialized with AT, while twenty of the overall respondent reported such installation.

5.2 Decision Tree Algorithm with Re-sampling Methods

After preprocessing stated in chapter 3, the dataset has 143 instances including 130 negative instances and 13 positive instances, and 74 attributes including the class attribute "Experienced injury using AT?" As noted in section 4.2, before classification algorithms can be applied to such imbalanced dataset it is necessary to balance the two classes. One approach to address the class imbalance problem is over-sampling, that is, using replication to increase the number of positive instances. Classification based on this technique and corresponding analysis will be discussed in section 5.2.1. Another approach is under-sampling which uses sampling to decrease the number of negative instances. The corresponding analysis will be addressed in Section 5.2.2.

5.2.1 Decision tree from over-sampling dataset

5.2.1.1 Preparation of over-sampling dataset

Estabrooks et al. did random sampling and get equal numbers of instances from each class [5]. However, due to the limit number of positive instances in our secondary injury, we cannot afford to lose any of the instances experiencing second injuries. Therefore, we manually replicated the positive instances nine times so that there are 130 positive and 130 negative instances. Then we built a J48 decision tree on this enlarged dataset (Figure 5.7).

5.2.1.2 Patterns discovered by J48 algorithm applied to over-sampling dataset

A decision tree was established by applying J48 decision tree algorithm to over-sampling dataset with 74 attributes is illustrated in Figure 5.7. In the decision tree, we highlight the leaf nodes with red color under which secondary injuries are predicted to occur, and we can find there are five routes of trees (decision tree rules or patterns) that lead to the classifications as positive. From the decision tree, interesting patterns are drawn as follows:

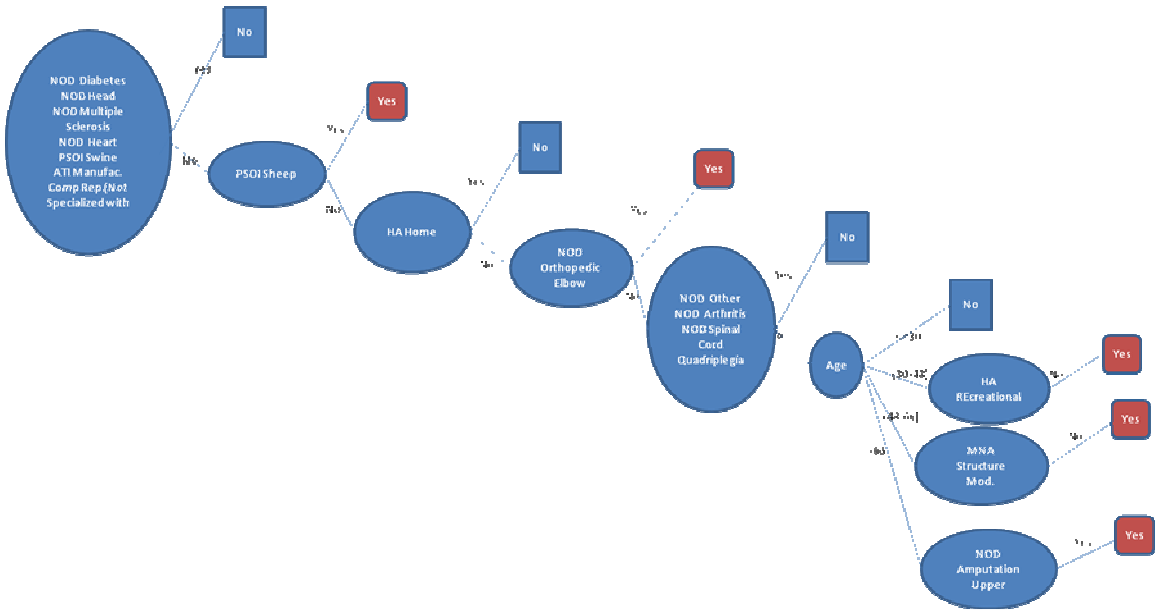

Figure 5.7 A decision tree classifying whether a secondary injury will occur ("Yes") or not ('No")

- If NOD diabetes=0, NOD head=0, NOD multiple sclerosis=0, NOD heart=0, PSOI swine=0, ATI manufacturing company representative=0, and **PSOI sheep=1**, then experiencing secondary injuries=1 (D-1 pattern).
- If NOD diabetes=0, NOD head=0, NOD multiple sclerosis=0, NOD heart=0, PSOI swine=0, ATI manufacturing company representative=0, PSOI sheep=0, HA home=0, **NOD orthopedic elbow=1**, then experiencing secondary injuries=1 (D-2 pattern).
- If NOD diabetes=0, NOD head=0, NOD multiple sclerosis=0, NOD heart=0, NOD other=0, NOD orthopedic elbow=0, NOD arthritis=0, NOD spinal cord quadriplegia=0, PSOI swine=0, PSOI sheep=0, ATI manufacturing company representative=0, HA

home=0, **age is between 30 and 48** (be included), and HA recreational=0, then experiencing secondary injuries=1 (D-3 pattern).

- If NOD diabetes=0, NOD head=0, NOD multiple sclerosis=0, NOD heart=0, NOD other=0, NOD orthopedic elbow=0, NOD arthritis=0, NOD spinal cord quadriplegia=0, PSOI swine=0, PSOI sheep=0, ATI manufacturing company representative=0, HA home=0, **age is between 48 and 66** (be included), and MNA structure Mod.=0, then experiencing secondary injuries=1 (D-4 pattern).
- If NOD diabetes=0, NOD head=0, NOD multiple sclerosis=0, NOD heart=0, NOD other=0, NOD orthopedic elbow=0, NOD arthritis=0, NOD spinal cord quadriplegia=0, PSOI swine=0, PSOI sheep=0, ATI manufacturing company representative=0, HA home=0, **age is greater than 66 and NOD amputation upper=1**, then experiencing secondary injuries=1 (D-5 pattern).

5.2.2 Decision tree from under-sampling dataset

We have been focusing on addressing the problem that the number of instances in the interested class is very limited. In the previous section, the instances where secondary injury was reported are replicated nine times in order to balance the two classes. From the enlarged dataset, the size of the tree is large, which makes it not obvious to identify the key attributes that lead to the secondary injury. There is also a problem with the instances tagged “negative”. Although these instances have not yet experienced a secondary injury associated with AT, some of them may be injured in the future. Therefore, the size of class 0 is actually overestimated.

Considering the problems mentioned above, instead of adding duplicate instances with secondary injuries, we reduced the number of negative instances by sampling, through which the two classes would be balance and moreover, the number of instances that are in potential tagged

"negative" wrong is reduced. We ran classification models on several reduced datasets randomly generated.

5.2.2.1 Preparation of under-sampling dataset

We constructed four reduced datasets by randomly selecting negative instances from the original imbalanced dataset: for the first two datasets, the number of instances in class 0 and class 1 is 2:1 which means that every dataset include 13 positive instances and 26 negative instances. For the other two datasets, the number of instances in class 0 and class 1 is 1:1 which means that every dataset of the rest two datasets include 13 positive instances and 13 negative instances. Then four decision trees were produced from these four datasets, which are presented in Figure 5.8, 5.9, 5.10, and 5.11.

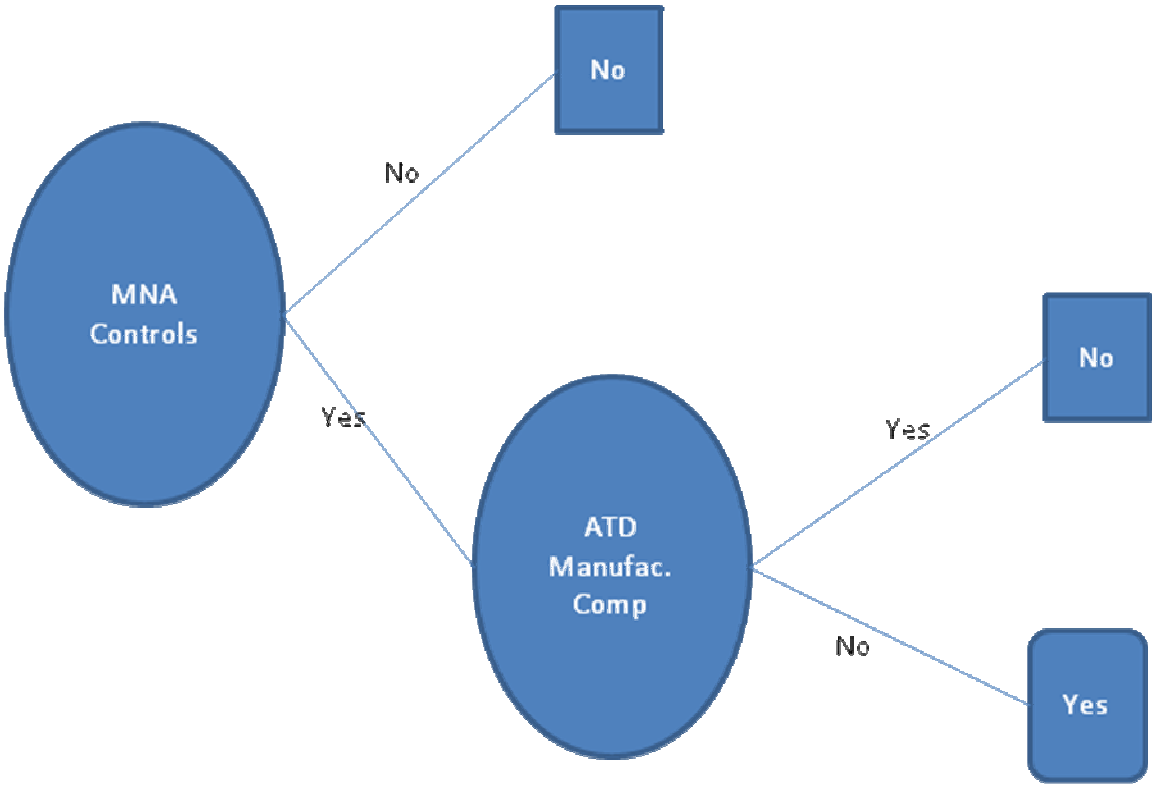

Figure 5.8 Decision tree produced from the dataset with 13 positive and 26 negative instances

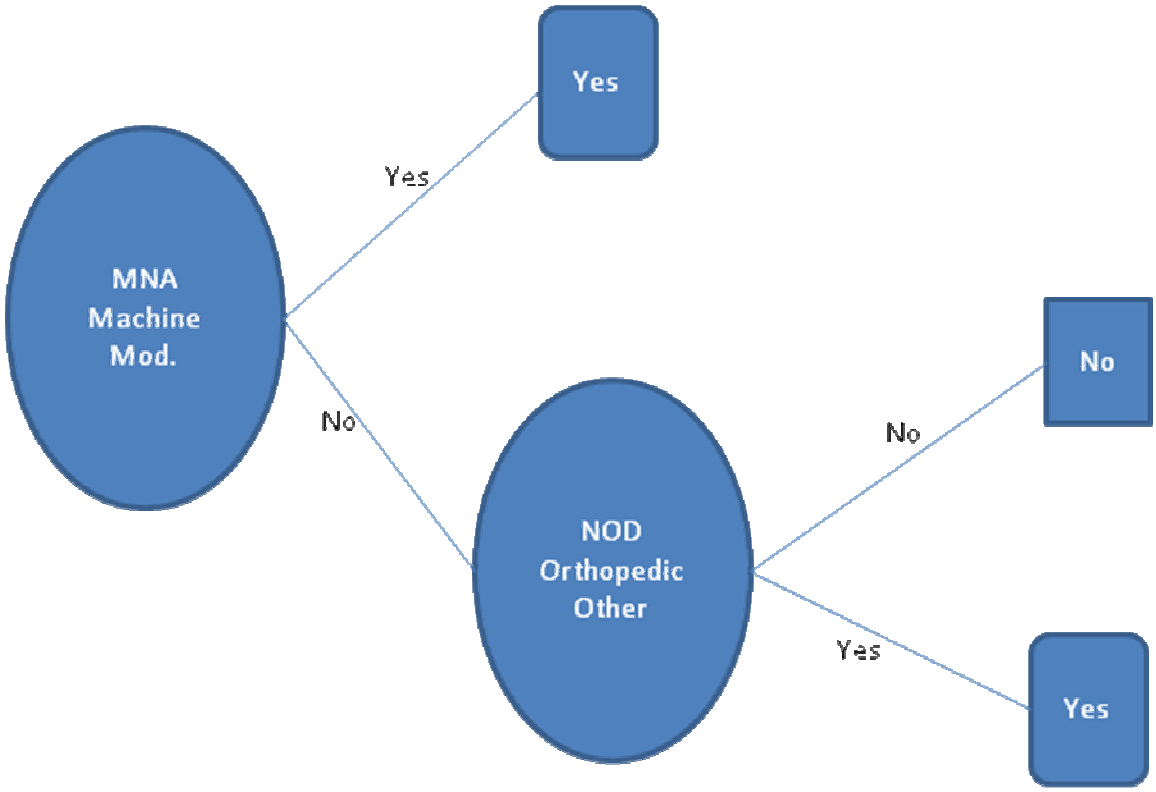

Figure 5.9 Decision tree produced from the dataset with 13 positive and 26 negative instances

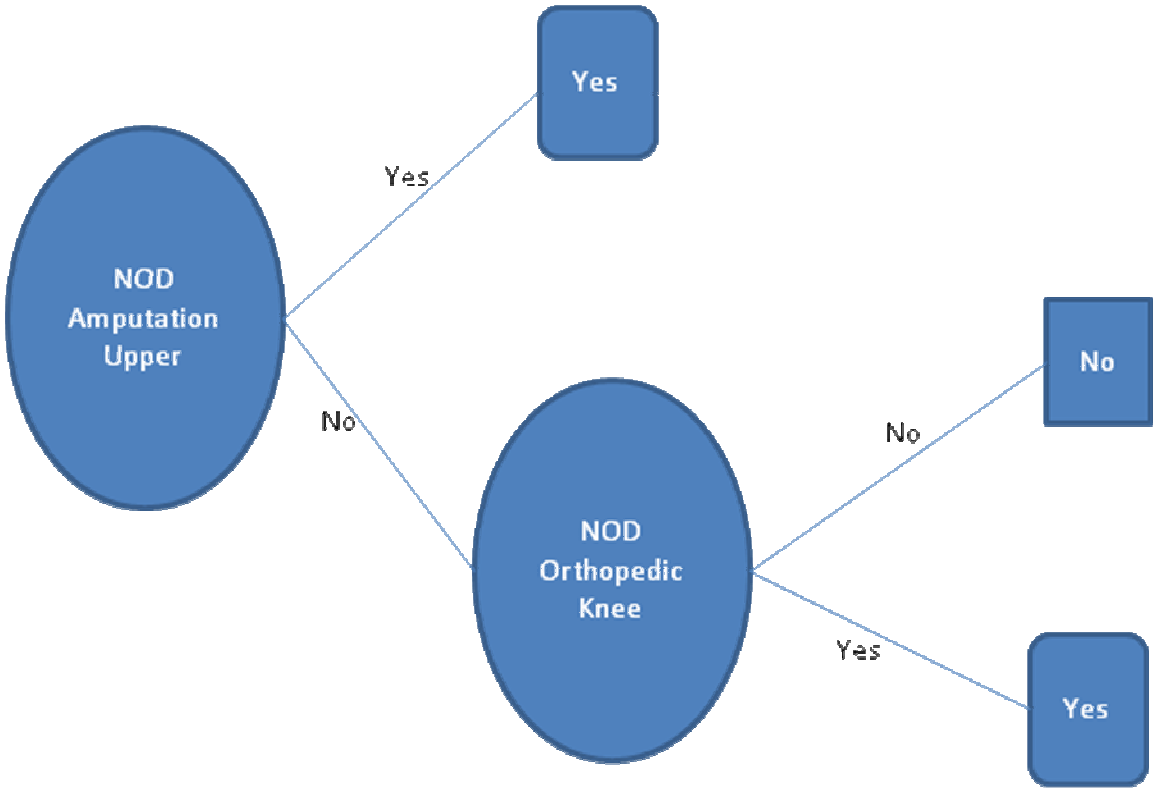

Figure 5.10 Decision tree produced from the dataset with 13 positive and 13 negative instances

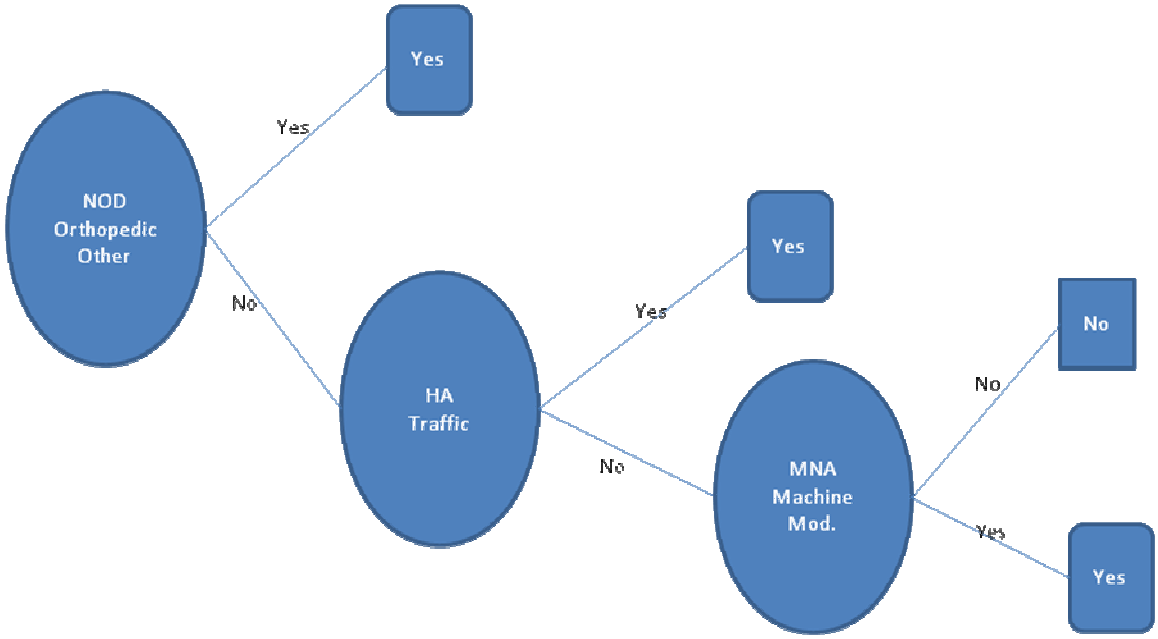

Figure 5.11 Decision tree produced from the dataset with 13 positive and 13 negative instances

5.2.2.2 Patterns discovered by J48 algorithm applied to under-sampling dataset

Compared to the enlarged dataset, the reduced dataset has the following advantages and disadvantages: First, reduced dataset offset some of the impacts introduced by incorrect labeling, which is not addressed in the enlarged datasets. Second, the reduced dataset produces elegant trees which can be explained by very simple decision rules. The major limitation of using sampling down method is that the resulted decision tree is not very stable because it is dependent on the datasets that are randomly generated. Therefore, multiple runs are necessary in order to get credible decision rules.

In summary, reduced datasets with balanced classes can be produced by sampling down the less interesting class. With fewer instances from class 0, the reduced sets focus more on explaining class 1. Some of the interesting rules generated from these datasets include:

- If MNA controls =1 and ATD manufacturing company =0, then secondary injuries will happen (D-6 pattern).

- If MNA machine modification =1, then there is a high possibility to experiencing secondary injuries (D-7 pattern).
- If MNA machine modification =0 and NOD orthopedic other =1, then the possibility of experiencing secondary injuries is high (D-8 pattern).
- If NOD amputation upper =1 then experiencing secondary injuries =1 (D-9 pattern).
- If NOD amputation upper =0 and NOD orthopedic knee =1 then experiencing secondary injuries = 1 (D-10 pattern).
- If NOD orthopedic other =1 then experiencing secondary injuries =1 (D-11 pattern).
- If NOD orthopedic other =0 and HA traffic =1, then experiencing secondary injuries =1 (D-12 pattern).
- If NOD orthopedic other =0 and HA traffic =0 and MNA Machine Modification =1 then experiencing secondary injuries =1 (D-13 pattern).

5.3 Experimental Results of the Application of Subgroup Discovery Algorithms

We applied different subgroup discovery algorithm (SD algorithm and CN2-SD algorithm) to our imbalanced secondary injury dataset in order to discover interesting secondary injuries patterns with which we can analyze what conditions and factors that are related to occurrence of secondary injuries, and evaluate the results obtained. Our final objective is to present the patterns to potential users of agricultural safety field in the form of rules in order to allow them to use this knowledge in the decision making concerning the safety of assistive technology.

5.3.1 Data preprocessing and preparation

First, we need to import the imbalanced dataset into WEKA in order to discretize the continuous attributes, since subgroup discovery algorithm SD take as their input, the training instances described by discrete attributes values. We have discretized the attributes with more than

two values: “age”, “# of injury years”, “ethnicity”, and “farmed after injury but now retired”. The attributes that have two values are converted from numeric to binary. Also, we have discretized the attribute “Experienced injuries using AT?” into two classes (yes, no) in order to codify them as the values of the subgroup discovery rule consequent.

Then we imported the prepared dataset into the ORANGE data mining plat form, which is available via the web, and implemented the subgroup discovery algorithm on the platform. In this section, we will describe subgroup discovery rules obtained and how these can be useful for potential users of safety analysis of assistive technology in agriculture field. Since all of the induced rules got from SD algorithm and CN2-SD algorithm are very similar, we select a small number of distinct rules from SD algorithm to illustrate the experiment results.

5.3.2 Patterns discovered by SD algorithm

The subgroup discovery algorithm SD is applied to the 74-attribute imbalanced dataset, the top ten rules are described below:

- **NOD Amputation upper=1** PSOI Swine=0
  - NOD Arthritis=0 PSOI Poultry=0
    - NOD Hearing=0 NOD Heart=0
      - _ (S-1 pattern)
      - ATD Home=0 (S-2 pattern)
      - ATI Self=0 (S-3 pattern)
      - NOD Muscular Dystrophy=0 Age □(49, 67) (S-4 pattern)
      - ATI Profess. AT manufac. Rep=0 (S-5 pattern)
    - NOD Hearing=0
      - PSOI Other=0 (S-6 pattern)
      - NOD Muscular Dystrophy=0 (S-7 pattern)
    - PSOI Other =0 (S-8 pattern)
    - NOD Heart=0 NOD Muscular Dystrophy=0 (S-9 pattern)
  - PSOI Other=0 NOD Hearing=0 NOD Muscular Dystrophy=0 (S-10 pattern)

5.4 Comparison of All of Patterns Discovered by the Three Approaches

We totally found sixty patterns by applying the three approaches to our secondary injury dataset. In order to filter out patterns with low support, we set a minimum support threshold (support=0.021) for all of patterns discovered by three approaches, that is, we eliminate thirty patterns with support lower than 0.021 out of sixty patterns. In this section, we will illustrate patterns discovered by graphical exploratory analysis (Table 5.1), patterns discovered by decision tree

algorithm (Table 5.2), and patterns discovered by subgroup discovery algorithm (Table 5.3) according to the three evaluation measurements: support, confidence and lift.

# of pattern	Support	Confidence	Lift
G-1	0.063	0.102	1.125
G-2	0.028	0.105	1.158
G-3	0.021	0.2	2.2
G-4	0.021	0.111	1.222
G-7	0.049	0.259	2.852
G-8	0.028	0.143	1.571
G-9	0.021	0.3	3.3
G-18	0.056	0.133	1.467
G-24	0.056	0.17	1.872
G-25	0.028	0.222	2.444
G-29	0.035	0.143	1.571
G-30	0.035	0.156	1.719
G-34	0.028	0.118	1.294
G-35	0.021	0.103	1.138

Table 5.1 Patterns discovered by graphical exploratory analysis

# of pattern	Support	Confidence	Lift
D-3	0.021	0.75	8.25
D-4	0.049	0.318	3.5
D-6	0.056	0.205	2.256
D-8	0.021	0.333	3.667
D-9	0.049	0.259	2.852
D-11	0.021	0.3	3.3

Table 5.2 Patterns discovered by decision tree algorithm: J48

# of pattern	Support	Confidence	Lift
S-1	0.049	0.467	5.133
S-2	0.035	0.5	5.5
S-3	0.042	0.545	6
S-4	0.035	0.556	6.111
S-5	0.042	0.545	6
S-6	0.042	0.5	5.5
S-7	0.049	0.467	5.133
S-8	0.042	0.462	5.077
S-9	0.049	0.467	5.133
S-10	0.042	0.462	5.077

Table 5.3 Patterns discovered by subgroup discovery algorithm: SD

In order to compare the strengths of three approaches, we illustrate five patterns with highest lift (Table 5.4) and five patterns with highest support (Table 5.5).

# of pattern	Support	Confidence	**Lift**
D-3	0.021	0.75	**8.25**
S-4	0.035	0.556	**6.111**
S-3	0.042	0.545	**6**
S-5	0.042	0.545	**6**
S-6	0.042	0.5	**5.5**

Table 5.4 Five patterns with the highest lift

# of pattern	**Support**	Confidence	Lift
G-1	**0.063**	0.102	1.125
G-7	**0.049**	0.259	2.852
G-18	**0.056**	0.133	1.467
G-24	**0.056**	0.17	1.872
G-29	**0.035**	0.143	1.571

Table 5.5 Five patterns with the highest support

From table 5.4 and table 5.5, it is can be concluded that graphical exploratory analysis always generates patterns with highest support and subgroup discovery algorithms are able to generate high-lift patterns.

Then, we use the thirty patterns with support higher than or equal to 0.021 to compare the performance of three approaches (Figure 5.12). Figure 5.12 helps identify which patterns are potentially most interesting (with higher support or higher lift).

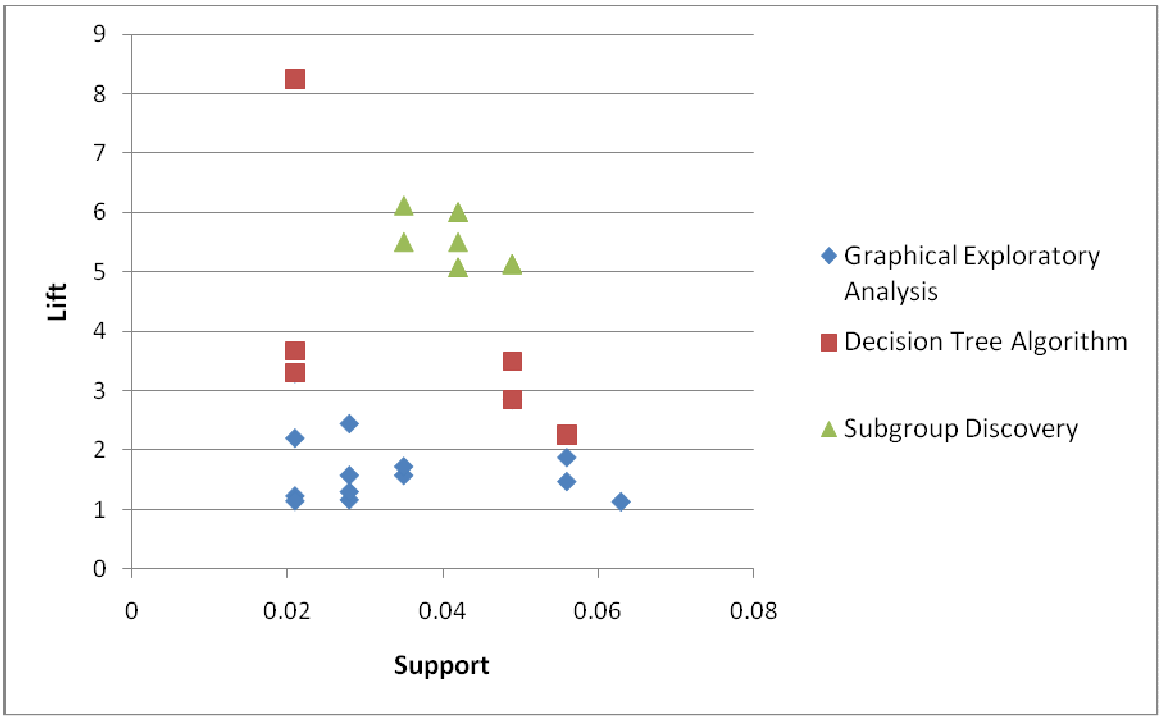

Figure 5.12 The support and lift of all of patterns (supports are greater than or equal to minimum support) discovered by three approaches

Note: a couple of patterns have the same support and same lift

We can conclude the advantages and disadvantage of three approaches as follows:

- The first approach, graphical exploratory analysis, is good at identifying patterns with highest support, but the disadvantage is that the lift of patterns is low. For example, the highest lift of patterns generated by graphical exploratory analysis is 3.3, but most of patterns generated by the other two approaches have greater lift than this value. Some interesting patterns discovered by the first approach are G-1, G-7, G-18, G-24 that are all high-support patterns with acceptable lift.

- The second approach, decision tree algorithm (J48), is able to generate high-support patterns like graphical exploratory analysis, but with consistently higher lift. For example, G-24 graphical exploratory analysis pattern: if MNA controls=1(yes), then experiencing secondary injuries=1(yes), and D-6 decision tree pattern: If MNA controls =1 and ATD manufacturing company =0, then secondary injuries will happen, have the same value of support, but D-6 decision tree pattern has a higher value of lift (2.256) compared to G-24 pattern discovered by graphical exploratory analysis. This comparison illustrates that decision tree algorithm is able to find interesting interactions, not apparent by any of the graphs in the graphical exploratory analysis approach. Some potentially most interesting patterns generated by decision tree algorithm are D-3 with highest lift (8.25) and D-6 with higher support (0.056).

- The third approach, subgroup discovery algorithms, is clearly best at generating patterns with higher lift. For example, the highest value of lift of patterns discovered by graphical exploratory is 3.667 and most of the patterns generated by subgroup discovery algorithms are higher than this value. On the other hand, both of graphical exploratory analysis and decision tree algorithm are able to generate patterns that have higher support than any of

the subgroup discovery patterns. For example, the highest value of support of subgroup discovery patterns is 0.049 and graphical exploratory analysis pattern, G-1, G-18, and G-24 and decision tree algorithm patterns, D-6 have higher support than this value.

Several of the subgroup discovery patterns (S-1, S-7, and S-9) have both high support (0.049) and high lift (great than 4). For example, S-1 subgroup discovery pattern: If NOD amputation upper=1, PSOI swine=0, NOD arthritis=0, PSOI poultry=0, NOD hearing =0, and NOD heart=0, then experiencing secondary injury=1, and D-9 decision tree pattern (G-7 graphical exploratory analysis pattern): If NOD amputation upper=1, then experiencing secondary injury=1, have the same support, but S-1 subgroup discovery pattern has much higher lift. This comparison illustrates that subgroup discovery algorithms are able to find patterns where factors have much closer correlations with target class. Some potentially most interesting patterns generated by subgroup discovery algorithms are S-3, S-4, and S-5 patterns with higher lift.

CHAPTER 6 CONCLUSION AND FUTURE WORK

6.1 Conclusion

Agriculture is one of the United State's most hazardous industries. Most of farm workers suffering disabilities or other injuries need to continue their farming jobs usually without appropriate recovery time. Assistive Technology (AT) has been assisting these farm workers to continue farming work. The use of AT, however, can result in secondary injuries. Scholars from the Department of Agricultural and Biosystems Engineering at Iowa State University conducted a survey related the use of AT and secondary injuries. From a data mining perspective the resulting dataset is challenging because it is small and highly imbalanced. This led us to apply three approaches in order to discover patterns of secondary injuries, namely, graphical exploratory analysis, classification analysis with re-sampled methods, and subgroup discovery.

The specific contributions of this thesis can be summarized as follows:

First, we applied three approaches to our imbalanced secondary injury dataset so that we successfully found not only causative factors which have single effects on the occurrence of secondary injuries but also combinations of factors.

Second, all of patterns discovered by the three approaches are evaluated according to three objective evaluation measurements including support, confidence and lift, and potentially most interesting patterns are found. From the comparison, we can conclude that graphical exploratory analysis is good at finding patterns with highest support and subgroup discovery algorithms are able to find patterns with higher lift.

Third, this thesis provides an alternative method, subgroup discovery algorithms, to deal with small and imbalanced dataset. As safety-related incident data is frequently imbalanced in this manner, our results and comparison of these three approaches is relevant to the analysis of other safety datasets.

Fourth, this thesis identified work hazards of a significant higher-risk subpopulation (workers with disabilities) from a population of agricultural workers that is already at higher working risk. Once the contributing factors to secondary injury are discovered, developing appropriate interventions will become possible.

6.2 Future Work

6.2.1 The ability of subgroup discovery algorithms to handle imbalanced datasets

We have successfully applied the subgroup discovery algorithms (SD and CN2-SD algorithms) to our imbalanced dataset, but if we can apply subgroup discovery algorithms to more imbalanced datasets from different fields, we can make more confident conclusion about the ability of subgroup discovery algorithms to handle imbalanced datasets.

6.2.2 Application fields

This thesis use two data mining approaches include classification analysis with re-sampling methods and applying subgroup discovery to imbalanced dataset. It is obvious that both of the two approaches achieve good results, but future work needs to address what kind of imbalanced dataset is appropriate to classification algorithms and subgroup discovery algorithms works better on what kind of imbalanced dataset.

BIBLIOGRAPHY

[1]A. Dubrawski, K. Elenberg, A. Moore, and M. Sabhnani. Monitoring food safety by detecting patterns in consumer complaints. In *Proceedings of the National Conference on Artificial Intelligence* AAAI/IAAI 2006, 2006.

[2]A. Estabrooks, T. Jo, and N. Japkowicz. A multiple resampling method for learning from imbalanced datasets. *Computational Intelligence*, 20(1):18–36, 2004.

[3]A. K. Jain, M. N. Murty, P. J. Flynn. Data clustering: a review. *ACM Computing Surveys* 31, 264-323, 1999.

[4]A. M. Hochberg, S. J. Reisinger, R. K. Pearson, D. J. o'Hara, K. Hall. Using data mining to predict safety actions from FDA adverse event reporting system data. *Drug Information Journal*, 41: 633-643, 2007.

[5]A. Nickerson, N. Japkowicz, and E. Millos, Using Unsupervised learning to guide resampling in imbalanced datasets", *In Proceedings of the 8th International Workshop on AI and Statistics*, pages 261-265, 2001.

[6]B. Kavšek and N. Lavrač. Apriori-SD: Adapting Association Rule Learning to Subgroup Discovery. *Applied Artificial Intelligence*, 20(7): 543-583, 2006.

[7]B. Mac Namee, P. Cunningham, S. Byrne, O.I. Corrigan. The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, 24(1): 51-70, 2002.

[8]C. Cardie and N. Howe. Improving minority class prediction using casespecific feature weights. *In Proceedings of the Fourteenth International Conference on Machine Learning*, pages 57-65, Morgan Kaufmann, 1997.

[9]C. I. Lee, C. J. Tsai, T. Q. Wu, and W. P. Yang. An approach to mining the multi-relational imbalanced database. *Expert Systems with Applications*, 34(4):3021–3032, 2008.

[10]D. B. Kell and S. G. Oliver. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic data. *BioEssays* 26: 99-105, 2004.

[11]D. B. Reed, and D. T. Claunch. Returning to farming after upper-extremity loss: What farmers say. *Journal of Agricultural Safety and Health Special Issue* (I): 129-137, 1998.

[12]D. Gamberger and N. Lavrač. Expert guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research 17,* 501-527, 2002.

[13]D. Gamberger, N. Lavrač, F. Zelezny, and J. Tolar. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *J. Biomed. Inform.*, 37: 269–284, 2004.

[14]E. Harris, E. Bloedorn, N. Rothleder. Recent experiences with data mining in aviation safety. *SIGMOD-DMKD98 Workshop,* Seattle, WA, June 5, 1998.

[15]H. Guo and H. L.Viktor. Learning from imbalanced datasets with boosting and datageneration: The DataBoost-IM approach. *SIGKDD Explorations*, 6(1):30-39, 2004.

[16]I. Abugessaisa. Knowledge discovery in road accidents database. *International Journal of Public Information Systems*, 1: 59-85, 2008.

[17]I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Morgan Kaufmann, 2005.

[18]I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham. Weka: Practical Machine Learning Tools and Techniques with Java Implementations. In *Proceedings of ANNES' 99 International Workshop on emerging Engineering and Connectionnist-based Information Systems*, pages 192-196, 1999.

[19]J. Demšar and B. Zupan. Orange: From experimental machine learning to interactive data mining, White Paper (http://www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.

[20]J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[21]J. M. Hardin, M. Conerly, and W. Watkins. Traffic Safety Analysis: A Data Mining Approach. *UTCA Project # 02115*, 2003.

[22]J. N. Hancock. Kentucky AgrAbility: Helping Disabled Farmers Return to the Land. *Journal of Agromedicine*, 5(1): 35-41, 1998.

[23]J. P. Leigh, S. A. McCurdy, and M. B. Schenker. Costs of Occupational Injuries in Agriculture. *Public Health Reports* 116(May-June): 235-248, 2001.

[24]J. R. Myers. Injuries Among Farm Workers in the United States, 1993. Cincinnati, OH: U.S. Department of Health and Human Services. *DHHS (NIOSH) Publication Number* 97-115,1997.

[25]J. R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[26]J. W. Tukey. Some graphic and semigraphic displays. In T. A. Bancroft (Ed.) *Statistical Papers in Honor of George W. Snedecor*. (pp.). Ames: Iowa State University Press, 1972.

[27]J. Zhang, E. Bloedorn, L. Rosen, and D. Venese. Learning Rules from Highly Unbalanced Datasets. *Proc. Fourth IEEE Int'l Conf. Data Mining (ICDM '04)*, 571-574, 2004.

[28]L. Breiman, J. H. Friedman, R. Olshen, C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Monterey, 1984.

[29]M. Atzmueller. Subgroup discovery. *Künstliche Intelligenz*, (4):52–53, 2005.

[30]M. A. Friedl and C. E. Brodley. Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61, pages 399–409, 1997.

[31]M. A. Maloof . Learning when datasets are Imbalanced and when costs are unequal and unknown. *ICML-2003 Workshop on Learning from Imbalanced Datasets II*, 2003.

[32]M. Brown, D. Parker, E. Seeland, D. Boyle, G. Wahl. Five years of work-related injuries and fatalities in Minnesota - Agriculture: a high-risk industry, *Minnesota Medicine*, 80(8): 29-32,1997.

[33]M. Kubat, R. C. Holte, S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30 (2–3): 195–215, 1998.

[34]M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):866-883, 1996.

[35]N. Lavrač, B. Cestnik, D. Gamberger, P. Flach. Decision support through subgroup discovery: three case studies and the lessons learned. *Machine Learning*, 57(1-2):115–143, 2004.

[36]N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research, 5*: 153–188, 2004.

[37]N. Japkowicz, C. Myers, and M. Gluck. A novelty detection approach to classification. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, pages* 518–523, 1995.

[38]P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning, 3*, 261-284, 1989.

[39]P. Domingos, (1999). MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155–164, 1999.

[40]P. K. Novak and N. Lavrač. http://kt.ijs.si/petra_kralj/SubgroupDiscovery/

[41]R. L. Miller, J. K. Webster, and S. C. Mariger. Nonfatal injury rates of utah agricultural producers. *Agricultural Safety and Health*, 10(4): 287-295, 2004.

[42]S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, Handling imbalanced datasets: a review. *GESTS International Transactions on Computer Science and Engineering*, 30 (1):25–36, 2006.

[43]S. N. Mathew, W. E. Field, and B. F. French. An assessment Process to estimate the secondary injury potential of assistive technology adopted by farmers with disabilities. *American Society of Agricultural and Biological Engineers,* Annual International Meeting, 2009.

[44]S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proceedings of the First European Conference on Principles of Data Mining and Knowledge Discovery*, pages: 78–87, Springer, 1997.

[45]S. Wrobel. Inductive logic programming for knowledge discovery in databases. In S. Dzeroski & N. Lavrač (Eds.), *Relational data mining*. Springer-Verlag, 2001.

[46]T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3): 291-316, 1997.

[47]T. M. Willkomm and M. Novak. Secondary injuries experienced by farmers using a wheelchair or a prothetic device. *American Society of Agricultural and Biological Engineers*, 2008.

[48]U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in datasets. *American Association for Artificial Intelligence*, pages 37-54, 1996.

[49]W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In U. M. Fayyad, G. Piatetsky- Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*. The MIT Press, 1996.

[50]X. L and S. Olafsson. Discovering dispatching rules using data mining. *Journal of Scheduling*, 8(6), 515-527, 2004.

[51]Y. Freund and L. Mason. The alternating decision tree learning algorithm. In *Machine Learning: Proceedings of the Sixteenth International Conference*, pages 124–133, 1999.

[52]Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition 40*, pages 3358–3378, 2007.