CHAPTER 3

PROBLEM STATEMENT AND
PROPOSED METHODOLOGY

3.1 GENERAL

This chapter deals with the problem statement and technological background required for the present study. The objectives and methodology are briefly described in the first part. As mentioned in Chapter 2, association rule mining which comes under the category of data-driven modeling has been proposed to be used as the tool for the generation of reservoir operation rules. Hence the details of data mining and classification using association rule mining which are needed for the present study are highlighted in the second part.

3.2 PROBLEM STATEMENT

Water management in irrigation consists of determining when to irrigate, the amount to be applied during each stage of plant growth, and the operation and maintenance of the system. The shortage of reservoir water is the main constraint in establishing stable water management programs.

In developing countries like India irrigation is the primary purpose for which most of the reservoirs are operated. It is essential to develop procedures to optimally allocate the available water for various needs among the different periods of a year keeping the long term benefit of the system in
view. Irrigation planners need to analyse complex climate-soil-plant relationships and apply mathematical optimisation technique to determine the optimally beneficial crop pattern and water allocation. A computer based model could facilitate irrigation planners in reaching sound decision prior to each crop season.

Reservoir operating rules guide release decisions prior to each crop season. Good reservoir management therefore requires the creation of “a set of operation procedures, rules, schedules or plans that best meet a set of objectives”. Rapid advances in computer technologies and learning algorithms have the potential to revolutionize the derivation of the reservoir operation rule using implicit stochastic optimisation.

A search in the literature reveals that specific system level operating policy improves the system performance which can be implemented easily using latest methodologies. Thus the present study focuses on the improvement of the system performance of an existing reservoir for single purpose namely, irrigation, and development of a Decision Support Model (DSM) for optimal reservoir operation using the association rule mining which come under the category of data-driven modeling.

3.3 OBJECTIVES

The broad objective of the present study is to develop a Decision Support Model (DSM) for optimal reservoir operation for irrigation planning.

The specific objectives of the present study are

i) To formulate a mixed integer linear programming model to obtain the optimal reservoir operation and to determine whether significant improvements might be realized from
optimisation of operation of the existing reservoir system with
different management strategies.

ii) To apply predictive data mining approach (association rule
mining) for the generation of reservoir operation rules by
implicit stochastic optimisation with streamflow sequences
generated from semi-parametric approach.

iii) To develop a decision support model for reservoir operators
by bridging the gap between predictive data mining model and
decision support model.

iv) To demonstrate the expected performance benefit of the
proposed model in practical applications by a case study.

3.4 PROPOSED METHODOLOGY

The proposed methodology is based on the synergistic action of an
optimisation approach via simulation to search for performance and a learning
approach to search for knowledge. First of all the optimisation generates high
performance solutions. The synthetic generation model generates the flow
sequences. These generated sequences are used by the optimisation to
generate diversified set of high performance solution. Now by applying a
learning approach to this solution set, knowledge is extracted in the form of
reservoir operation rules. The knowledge discovery method is based on
association rule mining which discover relationship between influential
parameters. Finally the knowledge extracted is used to develop the decision
support model for the reservoir operation. The methodology adopted in this
research work is shown in Figure 3.1.
3.4.1 Optimisation

Using a mixed integer linear programming based optimisation model, scope of the improvement of the performances of an existing reservoir system has been investigated. Alternate management strategies are considered for selecting the best viable irrigation release policy. A strategy for decision making under deficit supply is also considered, since full crop water demands cannot be met in all season due to scarcity of water resources.

The results obtained from the MILP model for each strategy are evaluated using the following performance indicators as proposed by Hashimoto et al (1982), Mujumdar and Vedula (1992) and Mannocchi and Todisco (2006).

3.4.1.1 Reliability

The reliability of a system under a given operating policy is defined as the probability that the system actual demand is satisfactory (Hashimoto et al 1982). It can be expressed as a ratio of the number of non-failure time
intervals to the total number of time intervals in the period under consideration, i.e.:

\[ \alpha = \frac{1}{N} \sum_{t=1}^{N} Z_t \]  

(3.1)

with

\[ Z_t = 1 \quad \forall x_t = S \]  

(3.2)

\[ Z_t = 0 \quad \forall x_t = F \]  

(3.3)

where \( \alpha \) is the reliability,

\( Z_t \) is the state of the irrigation release system in the time interval \( t \),

\( S \) is the satisfactory state,

\( F \) is the failure state,

\( N \) is the total number of periods.

Failure states in this study are considered to be time intervals during which the total irrigation release is less than 75% of total demand.

The performance indicator reliability simply reflects the likelihood of non-failure without specifying the extent of failure when one occurs. Even with high reliability, crop yields may still be low if the failure occurs in the critical periods of the growing season of a crop. It still provides a good measure of the ability of the system in providing the required irrigation.
3.4.1.2 Maximum Seasonal Deficit in %

This measures the severity of failure. In this study, it is simply defined as the maximum difference between the demand and release by considering all the years. It is calculated as:

\[ \beta = \text{Max} \left( \frac{D_t - R_t}{D_t} \right) \times 100 \]  

(3.4)

where \( \beta \) is the maximum seasonal deficit,

\( D_t \) is the total demand during the period,

\( R_t \) the release from the reservoir during that period.

3.4.1.3 Mean Value of Deficit in Time and Space

It is given by the formula

\[ \mu = \frac{\sum_{i=1}^{N} (D_t - R_t)}{N \times A}, \text{deficit(Mm}^3/\text{ha}) \]  

(3.5)

where \( \mu \) is mean value of deficit in time and space,

\( D_t \) is the total demand during the period,

\( R_t \) the release from the reservoir during that period,

\( N \) is the total number of periods,

\( A \) is the total command area in ha.

3.4.2 Synthetic Generation of Stream Flow

Generally, the available historic al stream flow record is not long enough to contain extreme conditions of high flow and low flows, hence it is...
not possible to assess completely the performances or reliability of a reservoir system situated across such a river. Stochastic methods provide a powerful tool for the water resources planners to effectively and efficiently formulate the development proposals by testing different operating policies with generated sequences of hydrologic inputs. The choice of a streamflow generation model for the purpose of planning and operation has very important practical implications in terms of system reliability and investment decisions (Pereira et al 1984). Efficient model selection procedures are essential for the successful application of stochastic models in planning studies.

The most effective stochastic flow generation model should preserve the skewness and extended tail behavior seen in the natural flow data. Preserving the tails is of particular interest because the tails exhibit the probability of extreme low or high flow. Here, models which preserve skewness also, such as PAR(1) with logarithmic transformation and the semi-parametric model are selected and performances of these models are assessed based on the various statistical tests. A semi-parametric approach in PAR(1) uses, PAR(1) as a parametric constituent of the model, and the residual resampling scheme based on the Moving Block Bootstrap (MBB) as a nonparametric constituent. The bootstrap (Peter Hall and Abhinanda Sarkar 2000) is a computational procedure that uses intensive resampling, with replacement, to reduce uncertainties. The aim of re-sampling is to mimic the random component of a process and to reduce variance through averaging over numerous different partitions of the data. The advantage of bootstrapping is that it does not rely on the assumption of normality and parameter estimation is simple. A moving blocks bootstrap differs from a regular bootstrap in that the data are resampled in contiguous blocks, rather than by individual values. This technique helps to preserve the autoregressive structure within the data.
3.4.3 Derivation of Reservoir Operation Rules

In implicit stochastic approach for deriving operation rules, data driven modeling technique, which is defined as a model connecting the system state variables (input, internal and output variables) with only a limited knowledge of the details about the “physical behavior of the system”, are generally employed. Earlier, finding dependencies from data is by using statistics, as represented by multivariate regression and classification. In the 60s and 70s, new techniques which were often not based on the assumption of “well behaved” statistical distribution of random process started to emerge, and these were used in many successful applications. In literature, several such approaches such as fuzzy logic, neural networks, data mining approaches etc have been used for deriving the reservoir operation rules. One of the new methods in this category is the association rule-based classification based on association rule mining, which has been proposed to be used here.

3.4.4 Development of Decision Support Model

Decision Support Model (DSM) is a specific class of computerized information system that supports decision-making activities. DSM are interactive computer-based systems and subsystems intended to help decision makers use data, documents, knowledge and/or models to identify and solve problems and make decisions. In DSM it is possible to provide a graphical user interface for the presentation of data and results in a form that is easily recognize and adopt by the stakeholders. Thus to aid the real time reservoir operator, an easy to use decision support model by bridging the gap between the predictive data mining model and decision support is proposed here.

The “If-Then” rules in the knowledge base can be utilized for the prediction. A hybrid method for rule ordering is used here. For easy use they can be encoded in Excel – VBA platform. It acts as a data base and a
prediction model, thus a Decision Support Model (DSM) for the real time reservoir operation.

Thus the step by step procedure adopted for mining association rules and building Decision Support Model (DSM) is shown in Figure 3.2.

**Figure 3.2** Step by step procedure adopted for building Decision Support Model (DSM)
3.4.4.1 Validation of DSM

For validation purpose, using the historical data, the Decision Support Model is run and the results were compared with optimisation model and the historical releases on the basis of the following performance indices.

i) Reliability: The probability that the system actual demand is satisfactory

ii) Maximum seasonal deficit in %

iii) Mean value of deficit in time and space.

3.5 DATA MINING

This section provides a short introduction to data mining and gives a precise definition of association rules. Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner (Hand et al 2001). It is often set in the broader context of Knowledge Discovery in Databases (KDD). Knowledge Discovery in Databases (KDD) is a data exploration methodology that is defined to be the non-trivial extraction of implicit, previously unknown, potentially useful, ‘relatively simple’, and not predefined information from large databases. The KDD process consists of several stages. Hand et al (2001) mention the following ones: selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to extract patterns and relationships, and interpreting and assessing the discovered structures as shown in Figure 3.3.
Figure 3.3 Schematic views of stages in KDD process (Hand et al. 2001)

KDD starts with the goal definition, which must be formalised and made executable so that it can be related to relevant data, which are hopefully present in the database. The data preprocessing step prepares and reshapes the data for subsequent processing. It involves data and attributes focusing, data cleaning, data projection, and data augmentation. The data mining step induces the model. It consists of a model specification, model fitting, model evaluation, and model refinement.

The boundaries of the data mining step in the KDD process are not clearly defined. Data mining is the knowledge discovery approach in KDD, which tries to recognise patterns in data. Thus, Data mining is the process of discovering valuable information from large amounts of data stored in databases, data warehouses, or other information repositories. This valuable information can have the form of associations, patterns, changes, anomalies and significant structures (Frawley et al. 1992). Data mining is a multi-disciplinary research field and many research fields have their contributions, such as database, machine learning, statistics, and artificial intelligence. It
includes several sub fields such as rule generation, classification and clustering, probabilistic modeling and visualisation. In brief, data mining tasks can be classified in two categories (1) descriptive data mining and (2) predictive data mining. The first describes the dataset in a concise and brief manner and presents general properties of the data; whereas the second constructs one or a set of, models, performs inference on the available dataset, and attempts to predict the behavior of new datasets (Chen et al 1996). The application fields of data mining are class description, association analysis, classification, prediction, clustering, time series analysis and outlier analysis (Han and Kamber 2001).

Rule mining is one of the central tasks in data mining since the rule is one of the most expressive and human readable representation for knowledge. There are fundamentally two categories of rule generation techniques: association rule mining and classification rule mining.

3.5.1 Association Rule Mining

Association rule induction introduced by Agrawal et al (1993) is a powerful method, which aims at finding regularities in the trends of the data and exhaustively looks for hidden patterns, making them suitable for discovering predictive rules. They are often expressed in a rule form showing feature-value conditions occurring frequently in a given dataset. The formal description of this technique is as follows.

Let \( I = \{i_1,i_2,i_3,\ldots,i_n\} \) be a set of items. Let \( D \) be a set of transactions (the dataset), where each transaction \( d \) (a data record) is a set of items such that \( d \not\in I \). An association rule is an implication of the form, \( X \rightarrow Y \), where \( X \in I, Y \in I, \) and \( X \cap Y = \emptyset \). The rule \( X \rightarrow Y \) holds in the transaction set \( D \) with
confidence c if c% of transaction in D that contain X also contain Y. The rule
X → Y has support s in the transaction set D if s% of transactions in D
contains both X and Y.

Given a set of transactions D, the problem of mining association
rules is to generate all association rules that have support and confidence
greater than a user-specified minimum support and minimum confidence,
where the thresholds for support and confidence are user specified values. The
task of discovering association rules can be done in three steps.

i) Find all sets items that have transaction support above
   minimum support

ii) For each item obtained, find all nonempty subsets

iii) For each such subset X of item I, if the confidence is bigger
   than the given threshold, produce the rule X → (I-X).

3.5.2 Classification Rule Mining

Classification is one of the key issues in the field of decision
sciences, a field which plays an important role in supporting business and
scientific decision-making. In recent years, it has also been one of the focal
points in data mining and knowledge discovery. Classification tries to predict
the class labels of unseen cases. In order to construct a classifier, a training
dataset is usually used to extract features to build the classification model.
The classification performance is then measured on another dataset that
is reserved for algorithm testing.

Classification has been studied extensively in literatures and
techniques developed for classification include a variety of learning
algorithms, such as k-nearest neighbors, decision tree induction, Bayesian
classification, neural networks, hidden Markov models, support vector
machines, etc. Neural-based approaches have resulted in high performance but they do not provide explanations for the classification decisions. Classification rule mining algorithm in a rule-based approach exhibit the highly desirable feature of interpreting the decisions. Decision trees classifiers, such as Quinlan’s C4.5/5.0 classifier and its extensions (Quinlan 1996), have received considerable attention due to its speed and understandability. Most classification rule mining algorithm have some problem such as, they may miss some global optimal rules and they may suffer from over-fitting problem. Another type of classification technique that has attracted an increasing number of attempts in recent years is finding classification rules based on association rule mining techniques (Ali et al 1997, Liu et al 1998, Li et al 2001).

Association rules will search globally for all rules that satisfy minimum support and minimum confidence norms. They will therefore contain the full set of rules, which may incorporate important information. The richness of the rules gives this technique the potential of reflecting the true classification structure in the data (Wang et al 2000). Associative classification is therefore gaining increasing popularity. However, association rules as such is not possible to use for classification problems since it contain a very large set of rules. Also, for association rule mining the target of mining is not predetermined while for classification rule mining there is one and only one pre-determined target, i.e., the class. While no single technique is proven to be the best in all situations, association rule-based classifiers have been found to be very useful in many situations (Ali et al 1997, Liu et al 1998 and Wang et al 2000). Both classification rule mining and association rule mining are indispensable to practical applications. Thus, great savings and conveniences to the user could result if the two mining techniques can somehow be integrated (Liu. et al 1998).
3.5.3 Classification based upon Association rule mining

Recent studies in classification have proposed ways to exploit the paradigm of association rule mining for the problem of classification. Thus Liu et al (1998) proposed a new algorithm which integrates two major techniques in Data mining which are Data Classifications and Association Rule.

3.5.3.1 Class association rules

The integration of association rule mining and classification rule mining is done by focusing on a special subset of association rules whose right hand side is restricted to the classification class attribute. These subsets of rules are known as the class association rules (CARs) and they are in the form $X \rightarrow C$, where $C = \{c_1, c_2, \ldots, c_m\}$ be a set of class labels, thus a CAR describes an implicative co-occurring relationship between a set of binary-valued data attributes (items in a transaction database) and a pre-defined class, expressed in the form of an “antecedent $\rightarrow$ consequent-class” rule. With such a rule, a transaction or data record $t$ in a given database could be classified into class $C$ if $t$ contains $X$. Apparently, a classification rule could be regarded as an association rule of a special kind.

The formal description of which is as follows:

Let $D = \{d_1, d_2, \ldots, d_i\}$ is a database that consist of set of data that have $n$ attributes and class label, where $d = \{x_1, x_2, \ldots, x_n, c_k\}; k = 1, \ldots, m$,

$I = \{x_1, x_2, \ldots, x_j\}$ is set of all items that appears in $D$

$C = \{c_1, c_2, \ldots, c_m\}$ is set of class labels.
Class Association that derive from Association Rule has a format, which is $X \rightarrow c$ with support value = $s\%$ and confidence value = $c\%$ where $X \in I$ and $c \in C$.

### 3.6 INTEGRATION OF CLASSIFICATION AND ASSOCIATION RULE MINING

The following step by step procedure is adopted for the integration of classification and association rule mining.

**Step 1:** Data selection - Relevant data is to be selected for the training set and the class label attribute also to be selected from the data set according to the goal.

**Step 2:** Data preprocessing - Many factors affect the success of class association rule mining. If there is much irrelevant and redundant information present then knowledge discovery during the training phase is more difficult. Hence data preprocessing step can often have a significant impact on generalisation performance.

For real world applications, the number of attributes of a data record could be very large. It is not uncommon that the class label of a data record (tuple) depends only on the value of a few attributes. In such cases, directly feeding the original data records with the entire attribute into a classifier may “confuse” the classifier to generate unnecessary complex classification rules. There are a few advantages if a small set of attributes can be determined before the actual data records are fed into the classifier. First the time required to extract rule is reduced because of smaller input data set compared to the original one. Second, as the attributes that do not contribute to the classification functions are removed, the rules generated by the classifier are expected more concise than the original data set used.
Now coming to the case of continuous attributes, the number of values that it can take is infinite. Even with a large sample size, it is still possible that almost all the data records have different values for the attribute. If we build contingency table using the distinct attributes values, it is very likely that there are no conflict entries in the contingency table. Especially if there are more than one continuous attributes in the data tuple, the combination of those continuous attribute can often uniquely determine the class label. Obviously, it may not be correct. Furthermore, even this is the case; the result of feature selection may not be interesting as we will end up with large classification rules. The solution to this problem is data discretisation- mapping the values of continuous attributes to integer intervals. The sole purpose of discretisation for feature selection is to make the number of distinct values of a continuous attribute small enough compared to the sample space. The process of discretisation involves the transformation of a quantitative variable into a qualitative one (Catlett 1991) and it is described below.

Considering A to be a numerical attribute of a set of objects, the set of values of the components of these objects that correspond to the attribute a is the active domain of A and is denoted $\text{ActDom}(A)$. To discretize A we select a sequence of numbers $n_1 < n_2 < \ldots < n_m$ in $\text{ActDom}(A)$. Then, the attribute A is replaced by the nominal attribute $\bar{A}$ having $m+1$ distinct values in its active domain, which is denoted as $\{K_0, K_1, \ldots, K_m\}$. Each component $a$ of attribute A for an object $o$ is replaced by the discretised component $K$ of $\bar{A}$ defined as

$$
K = \begin{cases} 
K_0 & \text{if} \quad a \leq n_1 \\
K_i & \text{if} \quad n_i \leq a \leq n_{i+1}, \text{ for } 1 \leq i \leq m-1 \\
K_m & \text{if} \quad n_m \leq a
\end{cases}
$$

(3.6)
The numbers \( n_1, n_2, \ldots, n_m \) define the discretisation process and are known as class separators. There are two types of discretisation: unsupervised discretisation, where the discretisation takes place without any knowledge of the classes to which objects belong and supervised discretisation which takes into account the classes of the objects.

In this study, supervised discretisation which is based on Decision tree modeling is used for the Preprocessing of the data. The program See5 (Quinlan 1997) is run over the training set and the features that appear in the pruned decision tree are selected.

Step 3: Data transformation - The preprocessed data set is converted into appropriate forms for the next operation of class association rule mining.

Step 4: Data mining – The goal of data mining is to automate the process of finding interesting patterns and trends. Association rule mining is used here for this purpose. Many algorithms are present for association rule mining. APRIORI algorithm proposed by Agarwal et al (1993) is the mostly used one. Here an improved version of APRIORI algorithm known as OPUS_ AR rule-discovery algorithm. Webb (2000) is adapted to mine all the CARs that satisfy the minimum support and minimum confidence constraints.

3.7 BUILDING CLASSIFIER USING CARS

CARs have two important drawbacks when used for building classifier: they do not capture the sequentiality of both antecedent and consequent and they contains over fitting rules. Thus for building classifier using CARs, pruning to remove over fitting rules and ordering the rules to capture sequentiality is required. Thus the algorithmic approach for classification using association rules can be divided into three fundamental parts: association rule mining, pruning and classification.
As explained above, the mining of association rules is a typical data mining task that works in an unsupervised manner. A major advantage of association rules is that they are theoretically capable of revealing all interesting relationships in a database. But for practical applications the number of mined rules is usually too large to be exploited entirely. This is why the pruning phase is stringent in order to build accurate and compact classifiers.

The association rule-based classification usually contains following steps:

i) Compute frequent itemsets that occur together in the training dataset at least as frequently as a pre-determined minimum support percentage. The itemsets mined must also contain the class labels.

ii) Generate association rules from the frequent itemsets, where the right hand side of the rules only contains class labels. In addition to the minimum support threshold, these rules must also satisfy a minimum confidence.

iii) As association rules generated from step (ii) usually are of huge amount, and many of them are overfitting and do not generalize to unseen testing cases, pruning needs to be done in order to select appropriate association rules to be used for classification. The basic principle is to prune rules that are not helpful in reducing the error rate of the final classifier.

iv) Finally classification is done by using remaining classification rules to predict the class labels of the test data. This can be
done by either using a collection of matching rules and taking the majority vote or using the highest ranked matching rule.

So for making the class association rules for prediction purpose they have to be represented in a new form, where a prediction model makes use of the rules in the form of a decision list.

There are two types of rule-based classification models:

i) Ordered rule-based classifiers: Rules are organized as a sequence, e.g., in the descending accuracy order. When classifying a test record, the first rule covering the record in the sequence makes the prediction. This sequence is usually tailed by a default class (prediction). When there are no rules in the sequence covering the test record, the record is predicted to belong to the default class.

ii) Unordered rule-based classifiers: Rules are not organized in a sequence and all (or some) rules covering a test record participate in the determination of the class of the record. A straightforward way is to adopt the majority vote of rules. A more complex way is to build a model to compute the combined accuracy of multiple rules.

Here the first method was used to build the rule based model since the first model is simple and effective. It makes a prediction based on the maximum likelihood. This is because rules with higher accuracy usually precede rules with lower accuracy and the accuracy approximates the conditional probability when the data set is large. Also according to Li (2006), there is no uniform form for the second model. Methods of voting vary in different proposals. An important condition for using the second model, independence of rules, is normally not satisfied. For example, in a complete
class association rule set, the conditions of most rules are correlated. Further, voting may be bias against small distributed classes.

3.7.1 Rule Ordering

Once the classifier has been established (usually presented in the form of a list of rules which are the generated CARs), regardless of the methodology used to generate it, a classifier is usually presented as an ordered CAR list, based on an applied rule ordering mechanism. According to Coenen and Leng (2004) five existing rule ordering mechanisms are there. They are, Confidence-Support size of Antecedent (CSA), size of Antecedent-Confidence-Support (ACS), Weighted Relative Accuracy (WRA), Laplace Accuracy, and $\chi^2$ Testing.

3.7.1.1 CSA Ordering

CSA (Confidence, Support, size of Antecedent) ordering is defined as follows:

In a set of CARs

i) A rule $r_1$ has priority over a rule $r_2$, if $Conf(r_1) > Conf(r_2)$.

ii) When, $Conf(r_1) = Conf(r_2)$, then a rule $r_1$ has priority over a rule $r_2$, if $Sup(r_1) > Sup(r_2)$.

iii) When, $Conf(r_1) = Conf(r_2)$ and $Sup(r_1) = Sup(r_2)$, then a rule $r_1$ has priority over a rule $r_2$, if $|r_1 . A| < |r_2 . A|$.

where The notation $r. A$ and $r. C$ are used to indicate respectively the antecedent (attribute) and consequent (class) of a rule $r$.

The support ($Sup(r)$) is the proportion of occurrences of the set $r.A \cup r.C$ in the input data compared to the number of records in the input.
The confidence of \( r \) (\( \text{Conf}(r) \)) is given as \( \frac{\text{Sup}(r)}{\text{Sup}(r.A)} \) and
\(|r. A|\) is the size of the antecedent that is the number of attributes in the antecedent.

### 3.7.1.2 ACS Ordering

ACS (size of Antecedent, Confidence and Support) ordering, which is also known as Specificity ordering, is an alternative ordering to CSA, which is defined as follows

i) A rule \( r_1 \) has priority over a rule \( r_2 \), if \(|r_1.A| < |r_2.A|\)

ii) When, \(|r_1.A| = |r_2.A|\) then a rule \( r_1 \) has priority over a rule \( r_2 \), if \( \text{Conf}(r_1) > \text{Conf}(r_2) \).

iii) When, \(|r_1.A| = |r_2.A|\) and \( \text{Conf}(r_1) = \text{Conf}(r_2) \), then a rule \( r_1 \) has priority over a rule \( r_2 \), if \( \text{Sup}(r_1) > \text{Sup}(r_2) \)

### 3.7.1.3 Weighted Relative Accuracy (WRA)

WRA was specifically designed as a rule ordering mechanism that reflects a number of rule measures. In its rule weighing stage, WRA assigns an additive weighting score to each CAR.

The WRA for a rule \( r \) is calculated as follows:

\[
\text{wra}(r) = \text{Sup}(r.A) \times (\text{Conf}(r) - \text{Sup}(r.C)) \tag{3.7}
\]

In the rule re-ordering stage the original CAR list is simply sorted in a descending order based on the assigned \( \text{wra} \) value of each CAR.
3.7.1.4 Laplace Accuracy

The principle of applying this rule ordering mechanism is similar to WRA. The calculation of the Laplace value of a CAR $r$ is

$$Laplace\ Accuracy\ (r) = \frac{Sup(r.A \cup r.C) + 1}{Sup(r.A) + k}$$

(3.8)

where $k$ is the number of predefined classes.

3.7.1.5 $\chi^2$ Testing

$\chi^2$ Testing is a well known technique in statistics, which can be used to determine whether two variables are independent of one another by comparing a set of observed values (O) against a set of expected values (E) — values that would be expected if there were no association between the variables. $\chi^2$ value is calculated using the following identity:

$$\chi^2 = \sum_{i \in S \cap A} (O_i - E_i)^2 / E_i$$

(3.9)

where $n$ is the number of observed/expected values. After assigning an additive $\chi^2$ value to each CAR, it can be used to re-order the CAR list in a descending basis.

According to Wang et al (2007) the above rule ordering mechanism can be subdivided into two major categories such as (i) pure “support-confidence” framework like and (ii) additive score assigning like. The first two methods (CSA and ACS) come under the first category and the other three in the second category. They found that the hybrid rule ordering strategy by combining one rule ordering mechanism taken from the pure “support-confidence” framework like, and another rule ordering mechanism taken from
the additive score assigning like gave better accuracy than their parent rule ordering approach. Hence here a hybrid rule ordering strategy is utilized.

3.8 SUMMARY

The details regarding the methodology adopted for the development of Decision Support Model (DSM) are discussed. A brief description of the classification using association rule mining method is also presented.