2. LITERATURE REVIEW

2.1 INTRODUCTION

This chapter deals with the review of papers in the area of constraints based data mining. Association rule mining is an important technique of data mining which discovers association among the set of items [22]. Association rule mining was introduced in 1993 by Agarwal et al. [23]. A typical example of association rule mining is market basket analysis. It refers discovering association between various items stored in a transaction database, such as buying butter with bread. The discovery of such associations can help retailers in developing market strategies on the various permutations and combinations of items bought by customer. In 1994 researchers proposed apriori a level wise algorithm, which generate frequent item sets. It used downward closure property according to which any non empty sub set of a frequent item set is also frequent. At each level, the frequent item sets are generated based on frequent item sets found in the previous level [24]. About the same time, in 1994 another study was carried out in which a variation of apriori, the Offline Candidate Determination (OCD) algorithm was proposed by H. Mannila et al. [25]. Since then several algorithms have been proposed such as Frequent Pattern (FP) Growth algorithm [26]. Researchers addressed several problems from various application domains such as temporal [27] and inter transaction rules [28].

In recent data mining researches, mining of single dimensional association rules from transaction databases have been a focused issue [23, 24, 29]. Sale, purchase and other information can be stored also in relational database known as multi dimension database [8]. Relational database records data in various attributes associated with the entity. Each attribute of the database is considered as
a predicate. Association rules that involve two or more predicates or dimensions can be referred to as multi dimensional association rules [8, 30]. In the present study single dimensional and multi dimensional association rules are discovered from relational database consisting of various attributes.

The association rule mining algorithm produce large sets of association rules which consume more time, occupy considerable storage and create confusion for the decision maker to reach the final outcome. The solution of this problem is to enforce logical constraints on association rule mining. Such strategy is called constraints based data mining.

### 2.2 CONSTRAINTS BASED DATA MINING

Constraints based data mining has been said to be the “best division of work” where the computer does the complex and lengthy computations and user provide the line of focus by supplying search constraints [8]. Constraints work as a guide to address the search combining with knowledge and inductive logic. Knowledge directs the search and limits the result. Constraints perform following tasks:

- They limit the space where the algorithm can look: User or experts impose constraints in the form of parameters embedded in query that limit the search area of the algorithm to discover the desired patterns.
- They give indication about what to look: Constraints guide the algorithm to search as per user input information.
- Help user to formulate effective query: Constraints can be given in the form of query through the input output interface which produce interactive environment to help the user to formulate a query.

If every possible pattern is selected and the constraint tested afterwards, it make search space very large, memory needed to store the results become huge and the time required to search them become excessive. The principle behind the constraints based data mining is to guide the focused search, which to focus and
what to ignore. The application of user-input knowledge in the form of constraints, reduces the search space, storage area, processing time and focuses the results to improve the decision making process. Constraints can be given in different forms such as simple threshold value, rule and complicated logic to formulate various conditions and queries. Different types of constraints are as follows:

### 2.2.1 Knowledge Constraints

Knowledge constraints specify the type of knowledge to be desired, such as characterization, discrimination, link analysis, association and correlation analysis etc. These constraints unlike other constraints are specified in the beginning of the query because different types of knowledge can require different constraints at later stage [31].

In the present study algorithms are designed to establish mutual relationships between Inter Disciplined Independent Variables (IDIV) and Dependent Variable (DV). To discover the mutual relationships constraints are provided in two phases as shown in Fig. 2.1.

![Fig. 2.1 Knowledge Constraints used to Establish Mutual Relationship among IDIV/DV](image)

In the first phase user inputs the type of knowledge such as association rules and in the second phase inputs the IDIV/DV. Hence, the identification of mutual relationship among IDIV/DV is considered in the category of knowledge constraints. In this field a very few papers are available which are as follows:

Discovering mutual relationships among items is useful to diagnose the faults in various applications. One of such application area is computer networks. In
computer networks sometimes fault occurs in one component due to the fault in other component. A study was carried out to identify the mutual dependency in faults occurring in computer networks. Association rule mining was used to discover mutual dependency in faults and concludes that the interrelated components are impacted by the same failure and strong mutual dependencies are common in computer networks [32, 33].

Association rule mining generally used to discover the co-occurrence of items in transaction database. A study was carried out by researchers in which they proposed a probability based evaluation metric which produces the probability of one item when another is present and also presented a mining algorithm as a extension of apriori algorithm to mine mutually exclusive items from transaction database and conclude that two items are mutually exclusive means two items cannot occurs at the same time in the database. This study provides an important indication when looking for previously unknown taxonomical relationships among items [34].

Another study was carried out to describe the future trends of data mining applications and data mining patterns. In this study Kriegel et al. emphasized the future store of different complex types of data which need more preprocessing of data before discovering the complex patterns and also specify that future patterns will be more complex with mutual relationships and more valuable [35].

Various objective and subjective measures are used to measure the attributes inter dependencies. A study undertaken by Nazareth et al. proposed that Pearson correlation coefficient can be used to identify attributes interdependencies to discuss mutual relationship among item sets [36].

Researchers in all these studies mentioned the need of a mechanism that allows the user to determine the complex, understandable and advanced patterns, which can reveal the hidden valuable patterns, interpret them and discover the mutual relationships among various items to help the domain expert in decision making,
but didn’t suggest any model or algorithm to find mutual relationships in items to the satisfaction of user. In the present work association rule based classification technique is used to mine the mutual relationships among item sets. The concept of threshold value which is unique and beneficial to make the algorithm general has been proposed. In addition to this IDIV are also introduced and assigned threshold values.

2.2.2 Dimension Constraints

Association rules which contain only one dimension are called single dimensional or intra dimensional rules. These rules contain one predicate. In some applications database is multi dimension which allows data mining using multi dimensional rules. Association rules containing more than two dimensions or attributes are known as multi dimensional rules. These rules contain more than one predicates. Dimension constraints specify desired dimension (or attribute) of the data to be used in mining.

Multi dimensional association rule mining was used to diagnose the diseases from medical database. A study was carried out in which an algorithm is proposed to detect breast cancer [37].

Sug [38] proposed multidimensional association rules and used as preprocessing method to reduce the mining time and generate only interesting rules and avoiding useless uninteresting rules. Instead of enforcing the constraints researcher extract the rules which have common attributes and assign these sets of rules to a common class name and avoid the association rules having uncommon attributes.

Swami et al. proposed multidimensional rules to prevent the smoking habits in youth. The study concludes that multi dimensional rules facilitate in identifying various relations in various factors responsible for certain conditions or habit formation [39].
Lu et al. described n-dimensional inter-transaction association rules to predict the stock movement [40].

2.2.3 Level Constraints

Concepts hierarchy is a tree like hierarchical structure of concepts and items reside in the database. Items at the lower levels are descendents and items at the higher levels contain many lower level items. It is a popular form to represent the items of transaction database into a hierarchical tree structure which allows data to be mined at multiple levels and cross levels of abstraction. In some applications transaction database is used at multiple levels of abstractions which require mining at multiple levels and cross levels such as sales and purchase transaction database.

Liu et al. discovered high utility items from transaction database which is organized in multiple levels of abstractions. They used two phase algorithm. In the application of market basket analysis mostly frequent patterns are discovered to make the strategy for increasing the sales. Sometimes infrequent patterns are also very important to improve the profit as infrequent patterns specify the sale from the rich class only. Considering the same concept researchers constructed the table for high utility items sets and discovered high utility patterns to improve the profit [41].

In another study cross level frequent patterns mining technique was used to give suggestions to e-learners and helping them in learning more effectively from web based materials [42].

Shaw et al. and Prakash et al. proposed diversity and peculiarity to measure the rule interestingness. Diversity measures the distance between two items within a rule, based on their position in the hierarchy. Peculiarity determines the distance between the two rules. The further away the rule is the more peculiar the rule. These peculiar rules are generated from the outlying data and hence few in
number. Study concludes that these two measures are important for discovering useful patterns [43, 44].

Gouider et al. discovered frequent multilevel patterns under the user input constraints. Researchers proposed two algorithms. One algorithm discovers the rules and tests the interestingness of the rules by the user given constraints and another algorithm removed the uninteresting rules [45].

Shaharanee et al. combined data mining and statistical techniques such as chi square, regression analysis and hypothesis testing and provide some control to minimize the risk of discovering uninteresting rules and patterns from transaction database [46].

Han et al. developed a group of algorithms to mine interesting rules from transaction database and conclude that different algorithms may have the best performance for different distribution of data [31].

Bhasgi et al. used apriory algorithm to find multi level association rules from transaction database to increase the sale and profit of items [57].

### 2.2.4 Meta-Rule Constraints

Meta-rule guided mining is an interactive approach to data mining, where user can probe the data under analysis by specifying the hypothesis in the form of meta-rule or pattern template. Data mining system attempts to confirm the hypothesis by searching for patterns that matches the given meta-rule. Meta-rule guided mining increases the possibility of discovering rules that are of interest to the user and can make the discovery process more effective by using the meta-rule to constraint the search space [48]. Meta-rule may be based on the user’s knowledge, understanding, experience or expectations. Generally, meta-rule forms a hypothesis regarding the relationships that the user is interested in confirming. The data mining system then searches for rules that match the given meta-rule.
Gouider et al. proposed algorithms for meta-rule guided mining in which $p$ predicates are input to guide the algorithm where $p < n$, and $n$ was total number of parameters. Researchers conclude that it is more efficient to construct and mine multiple $p$ cube chunks when number of dimensions and attributes per dimension are large as comparison to that when number of dimensions and attributes per dimension are smaller [48].

2.2.5 Interestingness Constraints

These constraints specify what ranges of measures associated with discovered patterns are useful statistically. For example the percentage of threshold value of support and confidence. The percentage of support and confidence threshold values below user-specified thresholds is considered uninteresting rules. Support and confidence reflect the usefulness and certainty of the rule. Association rules are considered interesting if they satisfy both a minimum support threshold and minimum confidence threshold [8]. Such threshold can be set by the user or domain experts.

Cooper carried out a study in which a simple constraints based method was used for discovering causal relationships between any two variables existing in observational database. Researcher used seven assumptions which were required to identify causal relationships between the variables [49].

Constraints based data mining allows users to describe the rule that they would like to mine, thereby making the mining process more focused, relevant and effective. Constraints can be implemented using high-level declarative data mining query language, user interface or query optimizer. Dong et al. studied issues and methods for efficient mining of multidimensional constrained gradient in data cubes [50].
Leung and Brajczuk developed algorithm for mining constrained frequent patterns from uncertain correlated data and conclude that constraints based data mining discovers only desired patterns [51].

### 2.2.6 Data Constraints

Data constraints specify the type of task related data to be used in mining in the form of structured query language. For example range of values, dates, time etc. A study was carried out in which researchers discussed about domain driven data mining. Domain driven data mining provide complementary support to data driven data mining by utilizing domain intelligence including domain expertise and knowledge constraints to discover hidden and valuable patterns [52].

Leung et al. developed an algorithm to mine uncertain data for frequent patterns that satisfy user specified succinct constraints [53].

Cao suggested a domain driven data mining framework which include (1) mining with constraints based context, (2) mining in-depth patterns, (3) supporting human-machine co-operated interactive knowledge discovery, (4) enhancing knowledge actionability and (5) supporting loop closed iterative refinement. The framework was explained using the real world stock market example and concludes that the framework can be used for the dynamic business process and requirements [54].

In the present work a mechanism is designed and algorithms are developed to apply the constraints based data mining on a relational database related to Socio-Economic factors of farmers to make the strategies to minimize the Socio-Economic Risk Factors (SERF) and improve the agriculture production. Among all the above mentioned constraints, algorithms for following constraints: Knowledge Constraints, Single Dimensional Constraints, Multi Dimensional Constraints and Interestingness Constraints are developed and assessed through real life database as shown in Fig. 2.2.
2.3 DATA MINING TECHNIQUES AND ALGORITHMS TO OPTIMIZE SOCIO-ECONOMIC CONDITIONS IN INDIAN AGRICULTURE SYSTEM

Agriculture is a combined system of independent component such as soil, water, crops, livestock, labor, environmental resources etc. The complete agriculture system can be divided into two main elements: technology and human. Technology includes the use of pest, seed, livestock potentials, water soil etc. The application of these factors can be improved by the improvement in technology through teaching and training. Human is exogenous factor that is farmer. Farmer has to decide adoption of technology and related resources. The decision depends upon the understanding and awareness of the technology. Farmer perception and understanding are influenced from a variety of Socio-Economic factors such as age, education, dependents, experience, livestock, innovativeness, extension-services etc. Review is undertaken to explore the role of data mining in the field of agriculture with a view to explore the efficient data mining model and developing mining algorithms to improve the Socio-Economy of farmers with a view to improve the county economy which was a neglected area because the emphasis was always given to production: quantity and quality. This is first time when
efforts are made to use data mining techniques in minimizing SERF of farmers. This section with a brief review related to the applications of data mining in the field of technology used in the agriculture and Socio-Economic factors related to farmers.

### 2.3.1 Data Mining Applications for Agriculture

Agriculture technology deals with the techniques used to improve quantity and quality of agricultural products focusing the new seeds, harvesting techniques and the management of water, soil, pest etc.

In the past various studies were carried using mining techniques applied to manage the pest. Combined technique of hierarchical clustering and association rules were used to develop the model for pest management [55]. Unsupervised clustering recursive noise removal technique was used to find the circumstances that lead the farmers to an excessive pesticide usage [56].

Weather conditions play an important role in agriculture production. A study was carried out in which unsupervised data mining technique K-means clustering was used to analyze the agricultural meteorology for the enhancement of crop yields and the reduction of crop loses [11]. In a study apple images acquired by CCD(Charge Coupled Device) camera were analyzed and classified by K-means clustering according to their quality (good or bad apples) [57].

Vegetable and fruits grading are the fields that also benefited from the data mining. A study was carried out in which classification algorithm based on decision tree named j4.8 classifier was used to grade the mushroom according to the mushroom quality [58]. Ekasingh et al. used decision tree algorithm C4.5 classifier to know the Socio-Economic factors responsible for farmers’ decisions for land use decisions [59].

Soil is the special entity with fuzzy boundaries and need the development of data mining techniques to extract the knowledge for soil classification. A study
was carried out in which physical and chemical soil properties of soil such as soil acidity, hydrogen, aluminum, calcium etc. (17 attributes) were analyzed using decision rules to discover the ways for improving productivity identifying causes of low yields. It has been observed that data mining as an effective technique to analyze data from precision farming [60]. Feng used hybrid data mining methods using Fuzzy Set theory and Probability theory for soil classification to improve the production [61]. Ramesh and Ramar assessed different data mining classification techniques such as J48, Naives Bayes, Random Forest etc. and applied to soil science database consisting 1500 soil samples and found that the Naïve Base is a useful classifier [62].

A combination of GPS-based technology and Support Vector Machine was used to classify crops through hyper spectral data [63]. Support Vector Machine was also used to detect weed and nitrogen stress in corn [64].

Literature survey shows that researchers also used neural network based technique to classify apples as good and bad apples [65], eggs as fertile and non fertile eggs [66], wine fermentation process and used taste sensors [67].

2.3.2 Data Mining Applications for Analyzing Socio-Economic Factors of Farmers

In the present study new algorithms have been developed to discover complex hidden patterns from huge database and algorithms have been assessed on the real life data related to Socio-Economic conditions of farmers collected through the survey. The reason behind the selection of this application area is that this area has been remained a neglected area. In this section a brief review is given related to Socio-Economic condition of farmers.

Chi and Yamada studied factors affecting the rice farmers’ adoption of new technology such as Integrated Pest Management (IPM), row seeding, certified seeds and rice varieties etc. They used group discussion method, and found that
low education of farmer, weak perception of new technology, weak teaching capacity, lack of capital, small land, poor infrastructures and limited capacity of extension staff led to low technology adoption [21].

Malathesh et al. assessed the contribution of socio-economic factors towards the extension use efficiency of farmers in selected farming systems in eastern dry zone of Karnataka. Sample data was collected from 120 respondents by using pretested structured interview schedule. The 4 farming system were identified: Crop, Crop+Dairy, crop+Dairy+Sericulture and Crop+Sericulture. Socio-Economic factors: age, education, land holding, family size, social participation, management orientation, scientific orientation, risk willingness and innovativeness, livestock possession, Cosmo politeness, cropping intensity were considered. Land holding, cropping intensity and scientific orientation were found to be significant at 5% level, Cosmo politeness at 1% level and age education, family size, social participation livestock possession, management orientation, risk willingness and innovativeness were found to be non-significant under crop farming system [17].

Matata et al. identified the factors that affect the adoption of new technology among smallholder farmers in western Tanzania and resulted that the lack of technology knowledge and inability of farmers to wait for two years before getting direct benefits from the technology were the main factors in the planting improvement. The study also identified that the farmers’ training through workshops and seminars, enforcement of laws for animal grazing and facilitation to farmers in getting more credits, are the major approaches to enhance the adoption of the new technology [20].

Idrisa et al. ascertained the Socio-Economic and technology specific factors that influenced the farmers’ decision to adopt the improved soybean seed as a production technology and concludes that farm size and expenditure on hired labor were significant at 5% level [68].

Some times farmers periodically improve their land which is allowing land to lie idle for one or more seasons to restore its fertility. Franzel assessed the
feasibility, profitability and accessibility of improved tree fallows. Labor constraints and institutional support were found to greatly influence the feasibility of improved tree fallows. Low returns to cropping, low base yields and high opportunity cost of labors increases the returns to improve fallow. Important factor associated with acceptability include past perception of soil fertility problems, past use of measures for improving soil fertility, current fallowing, economic importance of annual cropping and wealth level [69].

Boardman et al. studied the influence of Socio-Economic factors on erosion processes and conservation measures, mainly in a Western European farming context and found that agricultural subsidies, quotas and guaranteed prices are important for farmers in developed world [18].

Stocking and Murnaghan studied Socio-Economic Risk factors: land tenure, poverty, pressure on the land, labor availability, economic incentives, technical knowledge, appropriateness of used technology, economic and financial returns, off-site vs. on-site costs and social status affecting land users and land degradation in the context of developing countries, and found that farmers were strongly influenced by economic incentives [70].

Aslam et al., studied personal, physical and Socio-Economic factors affecting farmers’ adoption of land consolidation to reduce the conflicts related to irrigation, inter-farmer conflicts and shaping parcels into a form proper for mechanized agriculture. SERF related to an individual’s management skills. SERF include education, farmer experience, age and vocational training etc. These factors are important to represent a farmer’s ability to understand farm technologies. Age and education of farmer are found to be significant in land adoption technology [71].

Omobolanle examined Socio-Economic conditions of peasant farmers and the consequences on agricultural technology in South west Nigeria. In depth study and structured interview schedule were used to collect the sample. Sample was analyzed using statistics and found that sustained farmer and abandoned farmers had most features in common. Both cultivated similar crops, used family land or
inherited land and they heavily depended on family labor or hired labor. The study also revealed significant positive correlation between age and soybean adoption level, age and cassava adoption level, organizational membership and extension contact, factor affecting sustained use of maize and cassava technology while a negative correlation was reported between factors affecting sustained use of maize technology and extension contacts [19].

Sikorska analyzed the situation of agricultural land market in Poland. Researcher identified the factors affecting the turnover and the supply and demand relationships in trade and characteristics of parties to contracts of sale of agricultural land. Changes in supply and demand are main factors affecting the modernization of the agricultural sector and strongly influence the Socio-Economic development of rural areas [15].

Raju et al. studied SERF and livestock aspects of different production systems. Age, occupation, education, family size, farm holding, income from different sources, number and type of different animals, milk, production, marketing, breeding, feeding aspects etc. were studied with the help of pre-tested schedule, and found that type of livestock play an important role in various production systems. Mixed animal system proved best in the development of farming community and overall development of farming community based on local livestock aspects and Socio-Economic characteristics leads to sustainability [16].

Food security is the term used to describe whether people have received sufficient quality of food. Oni et al. investigated the Socio-Economic issues related to the state of food security of small holder farmers in the Thulamela local municipality of Vhembe district of South Africa and identified the factors that can be adopted by smallholder farmers in addressing the problems of household food insecurity[72].

Ngamsomsuke and Ekasingh studied the impact of Socio-Economic issues on farmers’ crop choice decisions using decision tree based algorithm. Socio-Economic factors including land size, off-farm income, land-labor ratio and
estimated profit are found important for farmers taking the crop choice decision [73].

2.4 TECHNICAL GAPS

After the extensive literature reviews following technical gaps are identified:

(1) Existing association rule mining system mine relevant and irreverent both the rules simultaneously which create difficulty for the user in decision making.

(2) Existing association rule mining system takes more memory to store the results as compare to constraints based mining as it extract less number of patterns and more valuable patterns.

(3) Existing data mining system mines only frequent patterns, however infrequent patterns are also important for right decision making.

(4) Existing mining system is not able to mine mutual relationships in various items while some items are mutually related and interesting for experts and users.

(5) Existing mining system is not able to discover multidimensional relationships in various items. While it takes important role in expert decision making.

(6) Existing mining system ignores the multi levels mining while sometimes data sets are stored in hierarchical form, where multi level rule mining is important for ensuring the correct decision making.

(7) Cross levels mining also take important part in discovering hidden complex patterns which are neglected by existing mining system.
Expert or user is responsible to take the decisions from discovered results, hence his/her participation in discovering process can provide more accurate results, which is ignored by existing mining system.

There are so many applications which remain neglected by the use of information technology, one such application is Socio-Economic factors affecting farmers’ income and farming practices, therefore in this study all the algorithms developed to discover complex patterns are applied on live data related to Socio-Economic conditions of farmers with a view to increase the production and the quality.