Macro Association Rule Discovery: Impact of Environmental indicators Changes on Life Assurance Business

Suzanne Shafic
Misr Insurance Comp.
Computers sector
44A El Dokki St.
Giza – Egypt
Misrins@tedata.com.eg

Dr .Khaled Shaalan
Computer Science Dept.,
Faculty of Computers and Information
Cairo Univ., 5 Tarwat St., Orman,
Giza, Egypt
Shaalan@mail.claes.sci.eg

Prof. Dr.Ahmed Rafea
Computer Science Dept.,
American University in Cairo 113,
Kasr El-Aini St,
P.O. Box 2511, 11511,
Cairo, Egypt
Rafea@aucegypt.edu

Abstract: Knowledge discovery in financial organization have been built and operated mainly to support decision making using knowledge as strategic factor.
In this paper, we investigate the use of association rule mining as an underlying technology for knowledge discovery in insurance business. Existing association rule algorithms and its extensions are inefficient in mining association rules in such data characteristics. We introduce algorithms for discovering knowledge in the form of association rules, suitable for data characteristics. Proposed data mining techniques is a hybrid of clustering partitioning and multi level rule induction. The proposed tool is managed by a repository meta model instantiated by meta-data libraries specific to insurance domain. It is implemented on a PC running on Ms Windows 2000. Samples of life data are extracted from different geographical locations of an Egyptian insurance company covering ten years. By using the induced rules, the decision- maker can define the horizontal expansion of marketing activities on new geographical area, or vertically empower the marketing forces in existing geographical area.
Keywords: insurance data characteristics, macro association rules, clustering partitioning, preprocessing &transformation, OLAP aggregation, ontology, data warehouse .

اكتشاف قواعد المعرفة على مستوى الماكرو لدراسة اثر التغيرات البيئية على صناعة التأمين

يهدف البحث إلى دراسة إمكانية استخدام الأساليب الفنية فى اكتشاف المعرفة باستخدام البيانات التاريخية لحركة محفظة تأمينات الحياة المتاحة والمؤشرات الأقتصادية والديموغرافية وذلك لمساعدة المستويات الإدارية المختلفة باعتبار المعلومات عامل استراتيجى لنجاح المؤسسة، من حيث اثر العوامل الخارجية على صناعة التأمين. وقد تم تحديد احتياجات متخذى القرارات فى اكتشاف المعرفة عن طريق تحليل منظومة العلاقات بين مختلف مكونات النظام وذلك بعمل Ontology Analysis متخصصة لمجال التأمين و تناول البحث وصف مرحلة إعداد البيانات وبناء مستودعات البيانات.وقد تم استعراض استخدام أسلوبين متكاملين لاكتشاف المعرفة من بيانات الحياة أسلوب تجزئة البيانات إلى مجموعات متجانسة clustering partitioning وأسلوب استنباط القواعد rule induction استعراض استنباط القواعد على مستوى الماكرو Macro level وتنفيذ أداة متخصصة لإدارة الأداة المقترحة واستنباط قواعد المعرفة واستعراضها بحيث تراعى المستويات المختلفة للمستخدمين.

1- Introduction

Many problems face the insurance companies due to the particularities of insurance industry. These problems are three types: Problems related to environmental factors, such as economic developments, social demographic changes, and competition. Problems related to companies' internal factors such as low performance of distribution channels and underwriting concerns. Problems related to technical issues such as availability of large volume of data and prior domain knowledge, overcome the gap between different management levels in the insurance business and usage of available data mining tool-box in the market, which require high skilled users to fulfill tools input restrictions. These problems

emphasis the importance of using the knowledge embedded in the historical data to support the decision makers to gain competitive edges, to increase efficiency of market channels, and provide more valuable services to acquire and keep customers.

The main contribution of this work is the presentation of a new technique in the mining phase. This phase is based on integrating clustering partition and discovery of association rules among different attributes in a data warehouse. Nowadays, there are many techniques to obtain association rules. These techniques can be shown on the well-known Apriori algorithm[27], [1], [3], [26], [16], [18] and many extensions shown in[5],[6],[33],[34],[10]. These techniques are modified to suite dynamic insurance data characteristics. The proposed technique is to perform a multi level rule-induction based on prior domain knowledge. All valid historical transactions of each insurance policy are appended into one composite record holding two parts. The first part represents the assurance policy fixed-length dimensions and the second part represents the variable-length episode of events related to the policy. These composite records are aggregated on different levels to be used in different rule induction levels micro-aggregated, and macro- aggregated level. Macro-level handles summarized high level view of transaction data describes the impact of environmental indicators changes such as economic indicators, competitors saving channels, demographic indicators of the whole country or different geographical areas, on the development of insurance business. Micro-level defines patterns of customer loyalty to predict unloyable customers to take preventive actions, patterns of different marketing channels and insurance product development. Demographic and economic indicators stored as percentage or continuous values, and aggregated insurance data are transformed into Boolean attributes representing the calculated direction of variance ratio of different attributes over times. A comparison of rules antecedent and consequent generated by selecting different aggregation levels are displayed. Each level has two mining tasks level, fixed mining task and variable mining task. For each fixed mining task, the target item may be different. We focus on mining rules for only one target item at a time, such rules could be mined more efficiently than rules with arbitrary heads [31]. In this paper, we will focus on inducing macro level association rules.

Another contribution is to provide a tool for human guidance . The design of the tool allows for different levels of users to derive the tool functions specific to insurance domain. The beginner level can use the predefined mining task, while the expert user can impose his prior domain knowledge in the following functions provided by the tool user interface: (1) Maintenance of meta-data libraries stored in the repository to extend and to customize the tool to meet new or changing requirements of business. (2) Direct the tool during mining phase by changing parameters. (3) Adding new mining task to the mining task library stored in the tool repository.

The rest of the paper is organized as follows. In section 2, we discuss the limitations of some of the existing techniques used in rule association. Section 3 presents the problem definition and the applied methodology. Section 4 describes the preprocessing steps needed to convert operational data to be loaded in data warehouse in order to be used in macro level rule induction. In section 5, we present the data mining algorithms. This is followed by a case study in section 6, while section 7 concludes with suggestion for future directions.

2-Related work

Knowledge discovery in database (KDD) process is dynamic, interactive and iterative. It consists of preprocessing, building a data warehouse, data mining and post processing [27],[13],[22]. Most previous work on KDD has focused on the data mining phase e.g. [2],[32],[4],[7],[29],[5], [30],[6], [33],[18],[10],[21],[19], [20]. However, in practice the other steps are as important for the successful operation of KDD. Preprocessing is necessary to find useful features to represent the operational data suitable to the goal of mining task.

Preprocessing operations are cleaning, dimensionality selection, reduction and transformation methods [13], [8]. Many algorithms have been introduced in handling corrections and completeness of raw data operations, scaling of some attributes, discretization of continuous attributes, feature extraction, and reduction. In our work, preprocessing is an essential factor in the success of mining phase. It's design is based on insurance data characteristics to reconcile the syntax and semantic of operational data to be loaded into a central data warehouse [11], [12], [24] that contains data from many operational systems. Preprocessing and building a data warehouse are followed by a data-mining step. Data mining step takes a lot of interest of researchers. It consists of data analysis and discovery algorithms, which under acceptable computational efficiency limitations, produce a particular enumeration of patterns. Many data mining techniques are reviewed, decision trees [1], [13], [23], neural network [1], [13], [32], [26], clustering [35], [1], [13], [4], rule induction [16], [5], [57], [19], [20]. By matching the goal of knowledge discovery tool in life assurance, we select a hybrid technique consisting of clustering partitioning and rule induction techniques based on Apriori algorithm and its extension. The domain expert can impose his interesting measures, by selecting appropriate attributes for each mining task to guide it.

This hybrid technique is proposed in order to solve rule induction problems. These problems are speed and clarity. Rule induction systems are often used in high dimensional spaces with high cardinality, they construct rules independently by extracting all possible patterns from the database. More dimensions slow down the process, the increase in time is usually linear with the number of predictors. The user is overwhelmed with obscure rules, which affects the clarity of the system, he is often required to review a large number of rules that are proposed as interesting in order to actually determine whether it is important. Speed and clarity problems are solved by clustering partitioning systems, where data are divided mutually exclusive and are grouped before rule induction process to eliminate redundant rules, and improve performance in mining phase. Dimensions reduction is used to speed the mining process, by the selection of relevant attributes of some dimensions based on the expert domain knowledge

3- Problem definition

Due to environmental, companies internal problems and technical considerations stated above, a tool is designed to use the huge amounts of historical transactions data in the life assurance to support the decision makers and the underwriters.

Each transaction has a transaction type associated with time stamp, which represents an event to the policy. Sequence of events that occurs during policy life are logically consistent. These serial events are a subset of the following transactions list: emission of new business, premature withdrawal, amendments of policy statements and conditions, acquire a loan at below market rate, diminution, liquidation of the policy, reemission, claim by early death, claim by death, and maturity at the end policy date. The final status of the enforce policies is effected dynamically by the episode of sequential order of events [28] occurred during the policy life since its emission.

The insurance companies, in order to keep its profitability, try to maximize the events causing company portfolio addition and minimize the events causing company portfolio withdrawal. Companies focus on keeping the customer loyalty; enhance its market activity, and review its underwriting basis. These activities are supported by using the available transactions to induce the pattern of insurance type, customer, market channel which act positive or negative on the company portfolio.

In order to build a knowledge discovery tool dedicated to insurance business we address some problems due to data characteristics, these problems are:

- Mining association rules in large multidimensional database tables containing continuos numeric and categorical attributes. Mining association rules from data representing a sequence of events, where each event has an associated time of occurrence. These events compose an episode of events that have a certain pattern. It is required to define the different patterns to support the decision making. Handling of dynamic nature of data, where appending a new event to an existing episode of events may change the pattern and affect the support of mined rules.
- Reconciliation of different types of data used in association rule mining. Two types of data are used in mining rules, global data holding time series of economic and demographic indicators, and insurance historical detailed data. It is required to define different macro patterns in insurance historical data and global indicators.

Applied methodology for knowledge discovery in life assurance is as follows: 1) Devise domain ontology analysis to identify decisions needed by insurance decision-makers to keep companies profitability. 2) Conduct a search to find out what algorithms are needed to discover such knowledge in different phases. Preprocessing and transformation of operational and external data phase, building a data warehouse phase, data mining, and post processing. (3) Manage Meta-data libraries used for knowledge discovery that assists the decision-makers as consistent documentation and as control information for tools. 4) Implement an integrated tool to support this methodology.

In this paper, we will focus on algorithms related to macro level association rules induction.

4-Preprocessing and build a data warehouse.

4-1- Devise insurance ontology analysis

The first step in devising an effective knowledge representation system and vocabulary is to perform an effective ontological analysis of the domain [15], [17]. Ontology representation is a structured guideline to fully documented conceptual model upon which to formalize the ontology [14]. It consists of requirement specification phase and conceptualization phase. Specifications of ontology describe goal of ontology analysis and scope. Conceptualization phase organizes and structures the acquired knowledge by using external representations that are independent of the implementation. Ontology definition in domain knowledge representation has two dimensions. The first dimension is static domain factual knowledge: which provides knowledge about the objective realities of the domain. As an example of life assurance, domain concepts are: (1) Life policy concept holding technical and financial transaction data of assurance policy, (2) type of production concept holding ways of acquiring the policy new business, and (3) producer's concept holding the main data of company producers. The second dimension is dynamic domain factual knowledge, it defines the properties and relations that can change over time. There are two types of relations: relations between concepts and relations between properties expressions. Is-a relation is a taxonomy relation, relation between concepts such as def, isa, has_a, and belong to.

4-2- Preprocessing and transformation

Transformation model describes extraction, transformation and loading functions. Transformation procedures of basic operational deals with transformation done to reconcile the syntactic and semantic differences between operation sources and data warehouse in order to load the data warehouse. Aggregation and selection components function to prepare data to fed into analysis applications.

Life assurance transformation operations

Figure 1 depicts transformation groups used in the preprocessing data life assurance. These groups are: basic operational transformation**,** feature construction, feature selection, transformation of condense rows, and Olap transformation.

In the following, we state two examples of transformation procedures, feature Construction , and condense rows used in macro level rule induction.

4-2-1 Transformation for feature construction.

Example of feature construction used as preprocessing in macro rule induction is the calculation of the variance ratio conversion operation. Steps of the calculation are as follows:

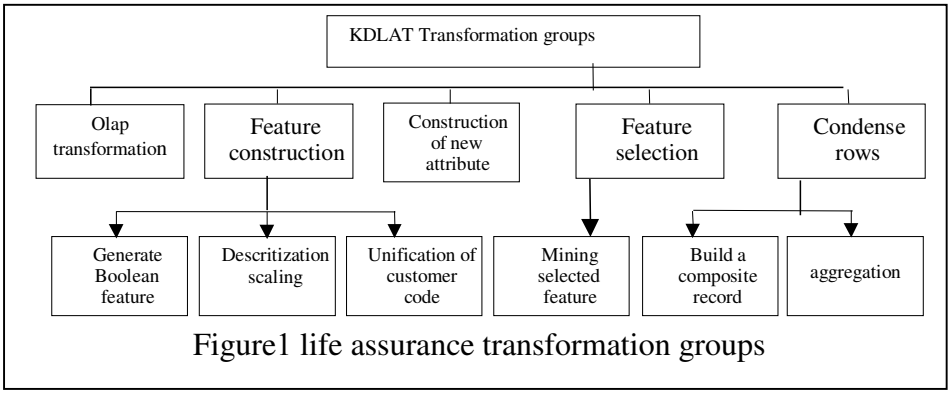


Figure1 life assurance transformation groups

Input parameters: macro economic indicators, demographic indicators, and aggregated insurance data.

Output parameters: ratio of variance of macro economic demographic time series.

Procedures: This procedures is used to define the direction of variance of the value of attributes over time series.

Calculation of the ratio of variance is performed by using the values of each two adjacent periods in the time series in the same table, by using the following formula:

$$\text{Ratio of attribute variance a} = \frac{\text{Attribute value of the period} - \text{attribute value of previous period}}{\text{Attribute value of previous period}}$$

Transform the direction of ratio of attribute variance a into boolean values as follows:

if a is negative then a_1 true , a2 and a3 are false.

if a is zero then a_2 true , a1 and a3 are false.

if a is positive then a_3 true , a1 and a2 are false.

Figure 2 depicts the ratio variance calculation algorithm.

```
Algorithm name    ratio-variance (input, count, output)
input             input-file, count
output            output-file
Procedure
        Read input-file in ratio-of-previous- period-rec
        Repeat until-E.O.F,
            n=0
            Read input-file in ratio-of-period rec
            Repeat until n = count
                Variance-ratio (n) = (ratio-of period(n) _ ratio-of-previous- period(n))
                / ratio-of-previous- period(n)
                if variance- ratio (n) < 0 then variance-ratio-desig (n) = 1
                if variance- ratio (n) = 0 then variance-ratio desig (n+1) = 1
                if variance- ratio (n) > 0 then variance-ratio desig (n+2) = 1
                n = n + 1
```

end repeat,
write output-file,
move ratio-of –period-rec to ratio-of-previous- period-rec
end repeat,
End Procedure

Figure 2 The ratio variance calculation algorithm

4-2-2 OLAP transformation operation.

On line analytical processing (OLAP) function is to enable users to examine data within a multi dimensional model to retrieve and summarize data. We distinguish three types of cube operations [25].

- Specialization operation, where the set of descriptors of the target cubes is a superset of the set of descriptors of the source table.
- Generalizations operation, where the set of descriptors of the target cubes is a subset of the set of descriptors of the source table.
- Mutation operations where the target cube and source cube have the same set of attributes but differ on the descriptor value.

Aggregation functions present the general characteristics or a summarized high level view over a set of user specified data in database.

Aggregation in life assurance data has many levels:

- **Aggregation based on policy characteristics:**

Input : policy composite table,

Output : aggregated table on different policy measures.

Procedures**:** This aggregation is performed based on the measures of different dimensions. These dimensions are insurance type, geographical area, customer characteristics, market dimension, data policy financial dimension, and episode of events dimension. The applied technique of descritization of continuous values and categories grouping similar transactions in generalized customer profile, salesmen profile, sequence of event.

- **Aggregation based on transaction type during a period.**

Input : historical transaction,

Output: aggregated table based on transaction type,

Procedures**:** Aggregate policy sum assured, total premiums for each transaction type. Types are emission of new business, withdrawal from the company portfolio by annulation, diminution, liquidation, maturity, death or disability. Aggregation has many levels, aggregation at the level of governorate, company, sector, and whole market. This aggregation has different times intervals.

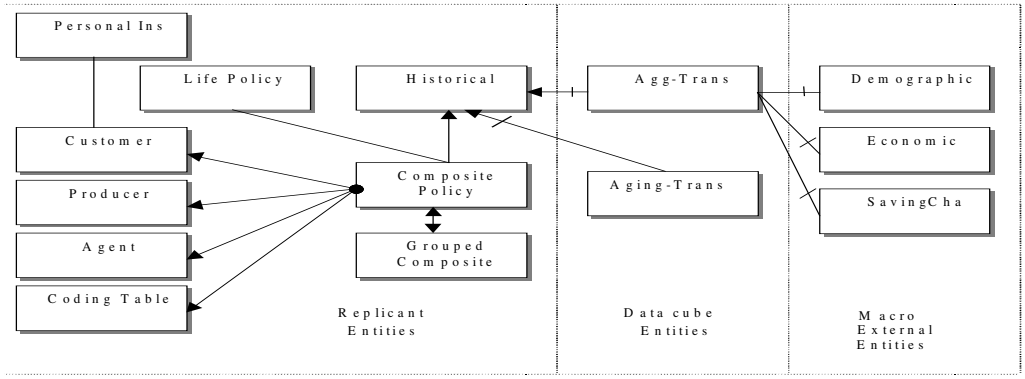


Figure 3 Life assurance Data Warehouse schema

4-3- Data warehouse

Database for data warehouse has many characteristics [9]: 1) It is primarily read only, insertion and append are limited to periodic load or refresh operations. 2) The integrity of different attributes is checked when the database is loaded or refresh. 3) Most of data are timestamps. and 4) Data are large and heavily indexed, and involves controlled redundancy by maintaining copies of the base data, derived data, summary tables.

Life assurance data warehouse

Based on the semantic of life assurance ontology model, an entity relationship model of life assurance data warehouse is designed. It consists of three types of entities: replicates of operational base entities, a derived data cubes, and external entities. Figure 3 depicts life assurance data warehouse schema

- Replicate of operational database: that consists of operational base entities, it consists of life policy multi-dimensional entity, personal insurance, customer, producer, agent, coding table, composite life policy multi-dimensional entity, and grouped composite life policy multi-dimensional entity.

- Data cube Data cubes entities consists of aggregated life transactions, and aging life transaction entities.

Macro external entities Macro external entities consist of demographic indicators, economic indicators and saving channel indicators.

5 -Life assurance Data mining phase

Life assurance data mining phase is supported by inductive methods, that aims to extract patterns from the data warehouse. Efficient preprocessing phase is a mandatory step toward the success of data mining module.

Life assurance mining phase is based on the integration of clustering partitioning and multi-association rules induction levels. This integration implies the generation of rules describing a cluster of identical users patterns .The used technique eliminates the combinatorial explosion of generated rules, improves understandability and interesting, and reduce the validation effort. Life assurance mining phase takes the dynamic nature of insurance business into consideration and deals with different levels of data in a multi layers mining module. Macro level handles summarized high level view of transaction data. The data are aggregated on transaction year, transaction major types, geographical area for both insurance market and economic indicators, and other factors, which can be taken into consideration. The mining rules of this layer describe some factors affect the insurance business, in global trends at macro level. Micro level handles detailed historical transactions stored in different partitions to mine association rules of different mining tasks depending on user selection. For each level many mining tasks are performed. Depending on the function of mining task, a view is created by selecting the relevant attributes stored in different data warehouse tables. Usage of views impact the performance of rule induction, the time and space required depends on the size and number of attributes in the view, not the size and number of attributes in the database.

Mining module is composed of three layers. For each layer, we state the following definitions:

Item-set I = (a, b, c…y), be a set of k elements.
Transaction X is a subset of item-set, where X belonging to I.

Support ($x \rightarrow y$) = P (x, y) or the percentage of transactions in the database that contains both x and y.

Confidence ($x \rightarrow y$) = P (x, y) / p (x) or the percentage of transaction containing y in transaction those containing x.

5-1 First layer: Partitioning composite transactions database

Problem definition

Item set I1= (P, I, C, S, F, T)

Where P = (portfolio code, policy identification code)

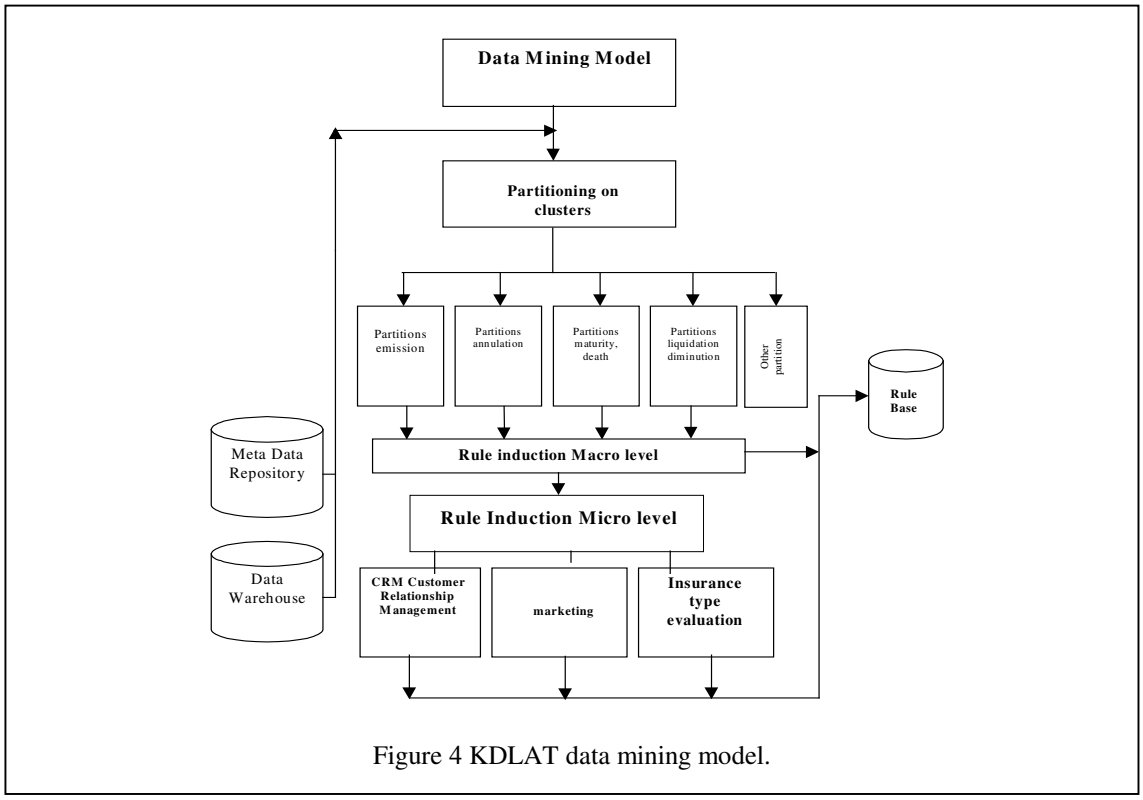


Figure 4 KDLAT data mining model.

I = (Insurance-code, tariff-code)

C = (Customer-ID, Customer-sort-name, Customer-age at transaction time, Customer- gender, Customer-career category)

S (Geog-area, agent-code, salesmen-ID, salesmen-inf.)

F (Currency-code, Premium, occurrence of (Transaction-extra-premium-types, extra-premium-value)

Item set I2= (Transaction-type, transaction date, affected attribute, cardinality, positive or negative designator)

Transaction T= I1 U I2 Grouped by zone code, agent code, insurance type code, customer age-cat, career cat, sum-assured cat, extra-premium-code, production type, salesman type and set of transaction code.

Weight of the record = the number of composite transactions classified and grouped in one record.

Support of partition = Sum of weights of groups of each partition divided by total no of transactions in database. Support of each partition is compared with minimum support defined by the user to select frequent episodes.

In the maintenance phase, the weight of groups and the support of each partition is modified by addition and subtraction of number of transactions as reflected by the movement of different composite transact. A clustering partitioning is performed on preprocessed descritized aggregated group composite table. Each partition represents a cluster of similar record effects on company portfolio episode of events.

Steps performed to clustering partitioning are:

a- Accept minimum partition support.

b- Read policy group composite table, select the defined partition depending on the value of episode of events of the grouped row and the content of valid episode of event table.

c- If the episode of events do not exist into valid episode of events table ,then

c-1 check the validity of logical sequence of historical transactions of group composite record,

c-2 if it is a valid sequence , add the new episode of events to valid episode of event table, check the last transaction code to define the partition number,

c-3 else write the erroneous sequence of event into invalid episode of event partition to be examined by the expert user.

d- Accumulate weights of each partition to calculate partition support.

e- Repeat the steps a to d until the completion of policy group composite table.

f- Select relevant partition by comparing the calculated partition count and accepted minimum support. If minimum support > each partition support, move reject flag into partition header else move accept flag into partition header. We note that irrelevant partition is not pruned, it is used later in the incremental updating of maintenance phase.

In the maintenance phase, the incremental updating method take into consideration the occurrence of one of the following states:

- A composite grouped record is moved from one group to another in the same partition, which affects weight of both groups and the support and confidence of rules generated from these partitions.
- A group is moved from one partition to another, which affects support (partition) of partitions, the support and confidence of rules generated from this partition.
- A new group is added to a partition, which affects partition support and the support, confidence of rules generated from these partitions.
- An existing group is deleted from a partition, which affects partition support between groups and partitions.

5-2 Second layer : discover association rules at macro level

This layer uses some dimensions in data cube of aggregated data and the time series of economic indicator data concepts. The aim of this layer is to induce association rules in the defined dataset by using algorithm based on A priori-like. The modification to A priori is based on the definition of the item set used in the algorithm. The item-set in the algorithm is a set of Boolean values defining the direction of variance ratio for each item in the data cube and economic indicators. Value of variance ratio is calculated as shown in the preprocessing phase.

Definition of macro level mining problem:

Let item-set I = year, company code, (i_1, i_2, i_3, i_4…) \i is a boolean variable representing the direction of ratio variance of aggregated life assurance data calculated during preprocessing phase.

Let item-set E = year (e_1, e_2, e_3, e_4…)

Where e is a boolean variable representing the direction of economic indicator ratio variance calculated during preprocessing phase.

Transaction T = I U E

- _Rule format x ---T--→ y =where x subset of E and y subset of I, if x occurs then y occurs within t times.

- Support of the rule is the percentage of transactions in the database that contains both x and y as true variables.
- Confidence (x --→ y) = P (x, y) / P (x) or percentage of transactions containing y in transactions those contain x.
- Interest (x --→ y) = P (x, y)/ P(x) * P(y).

Steps performed in macro level rule induction are as follows:

a- Accept the user type to define the selected path of macro level induction task. User type may be ordinary user or expert user.

b- If user type is ordinary user , the user may select a tool predefined task

c- if user type is expert the proposed tool accept a selected features as antecedent and consequent to be used in macro rule induction algorithms, and these selected features are added as a new task to the business meta data of the repository.

d- Accept minimum support and confidence.

e- Define the 1-largest-item set, by using the Boolean attributes value of the direction of variance ratio generated in preprocessing phase in the calculation of 1-largest-item set support instead of original attribute.

f- Select the relevant attributes where the calculated attribute support is greater than accepted support.

g- Apply Apriori-like algorithm to induce macro level rules on different aggregation levels for insurance market, insurance sector (private or public), and individual insurance company and governorate geographical areas.

As post processing: Compare rules induced from each aggregated level to define the impact of different scope of aggregation on the induced rules as selected by the user in post processing phase. Figure 5 depicts algorithm pseudo code of macro level rule association mining algorithm

```
Algorithm name      macro–level–association–rule mining
                    /* discover the association rule at macro economic level  */
Input               economic–indicator–table, -aggregate-data-cube
input output        economic–variance–ratio,  -aggregate-variance-ratio, economic–agg–series
Output              Macro-rule base
Procedure           Accept min-support
                Call calculate-ratio-variance (economic-indicator-table, count, ec-ratio-
variance),
                Call calculate ratio -variance (aggregate-data cube, count, ag-ratio variance),
                Call concatinate-ration(ec-ratio-variance,eg-ratio-variance,concatinated
                      table),
                   Repeat until concatinated table EOF
                      Read concatinated table, n=1, count-rec=count-rec+1
                      \* calculate frequency of true attribute*\
                      Repeat until n = no-of-attributes,
                         If attribute (n)  true then      add 1 to count(n),
                      End repeat,
                      \* calculate support of 1-item-set*\
                      n=0
                      Repeat until n = no-of-attributes
                           Add 1 to n
                           support (n) = count (n) / count-rec
                           if support (n) ≥ min-support  move attribute (n)to accepted att (n1),
                                 add 1 to n1,
```

```
                    end-repeat
                    /*define n-large-item set*/
                    Repeat until concatinated table EOF
                        Read concatinated table,
                        Repeat until  n = no-of-attributes
                            If attribute (n) belong to accepted-att-list append attribute (n)
                              to accepted-att-value
                            accepeted-flag(n)  = 1
                        end repeat,
                        If accepted-attr-value  exist in accepted-array index
                        then add 1 to count-array  corresponding to accepted-array-
                        index, else add new entry to accepted-array-index
                        add 1 to count-array corresponding to new-entry
                    end repeat
                    /*Calculate support of each entry of count array*/
                    Check support  > min-support
                    generates rules, write rules in macro-rule base sorted  by  support
                end repeat
End procedure
```

Figure 5 Macro level rule association mining algorithm

Figure 6 depicts an example of induced macro rules

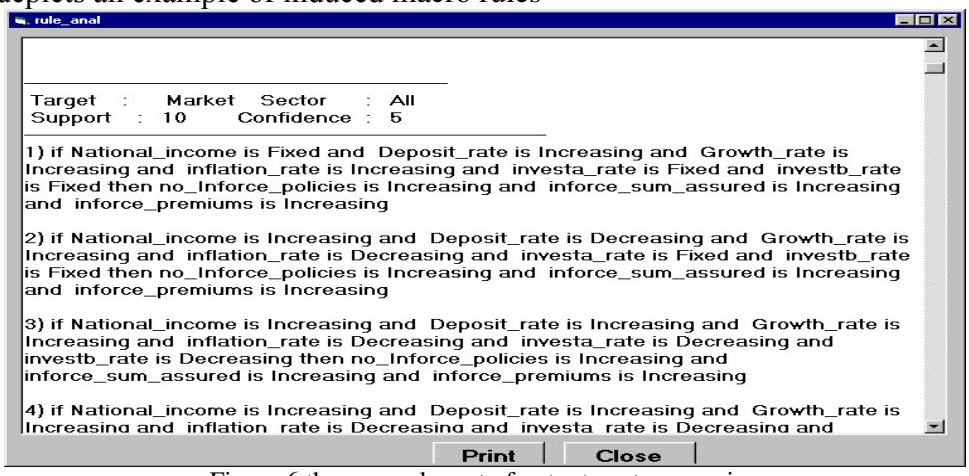


Figure 6 the screen layout of output post processing

macro level the user select to mine data relevant to whole insurance market, a
At macro level, the user select mining scope of data relevant to whole insurance market, a given sector or a given company. A comparison of induced rule can be stated to analyze the impact of different scope of aggregated data, the support and confidence values . Figure 7 depicts the comparison algorithm.

```
Algorithm name   Rule comparison
Input            user-input (session no, date)
                  KDALT-rule-base table
Output           difference- table
Procedure
 Compare(X,Y)  /*(source —> target)*/
      Read rule base table
       /*Calculate x , y difference: compare x*
          Group all rules have same Y
```

```
    /*Check existence of different fields in x and its direction into source rules & target
      rules*/.
     m = 1
   Repeat until no-of-x-source  = 0
       Repeat until no-of-x-target = n
          If x (m) in source not exist in x (n) in target move 0 to difference design
          If x (m) in source exist in x(n) in target and directions is different move 1 to
            different design,  n = n + 1
          end repeat
           write  difference table
            m = m + 1
          No- of- x-source = no- of- x –source _ 1
   End Repeat
End compare
     Display x difference table
/*(  compare target  —>source) Calculate y, x difference, Group all rules have same   x
          check existence of different fields in y and its direction into target  rules   &
          source rules  */
move y to work-area, move x to y , move  work-area to x
             call  compare,
             Display y difference table,
End
```

Figure 7 Comparison algorithm pseudo code

6 - Case study

6-1 Implementation of Knowledge discovery in life assurance data tool

KDLAT Knowledge discovery in life assurance data tool KDLAT is implemented in an experimental PC environment. The data sources are: -

- Operational life assurance stored in a IDMS network database representing a sample of operational data extracted from different geographical locations from a given company.
- Selected life assurance historical data corresponding to the extracted sample of operational data stored in a back up flat files dated 1990. Egyptian insurance market statistics for life assurance business, published by Egyptian Insurance Supervising Authority (EISA), available statistics starting from 1990. Egyptian macro economic and demographic statistics published by CAMPAS, and National Planing Institute..

Figure 8, depicts KDLAT architecture, divided into preprocessing, data mining, and post processing.

The proposed tool is implemented on a sample of operational life assurance and selected life

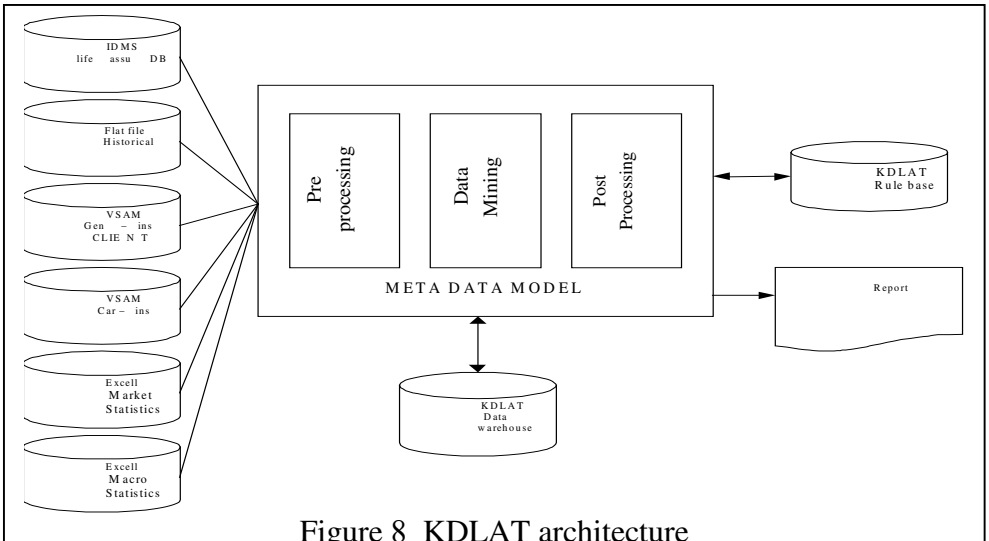


Figure 8 KDLAT architecture

assurance historical data. This sample of used operational records are selected to represent the different geographical areas, and randomly selected customers, financial, marketing channels characteristics. Number of operational records are 177446, historical file are 352681, composite file are 283365, grouped composite records are 131353. The presented KDLAT mining tasks are predefined macro mining task, a new task added by the expert user to mining task library. Two types of rules have been induced, macro rules which describe the impact of economic indicators on the development of insurance business to support strategic decision making, a comparison of the rules antecedent and consequent generated by selecting different attributes aggregation scope are displayed, and micro rules defining pattern of customer loyalty to predict unloyal customers to take preventive actions ,and company' market expansion to support both strategic and tactical decision making.

The user defines the scope of aggregation of used data, it can be all market data, public or private sector companies, or a given company data. In post processing, a comparison of induced rules is stated to analyze the impact of the scope of data, the support and confidence values in the induced rules.

Figure 9 depicts the screen layout of input new emission for task no-1. It represents the x attribute and y attribute for market.

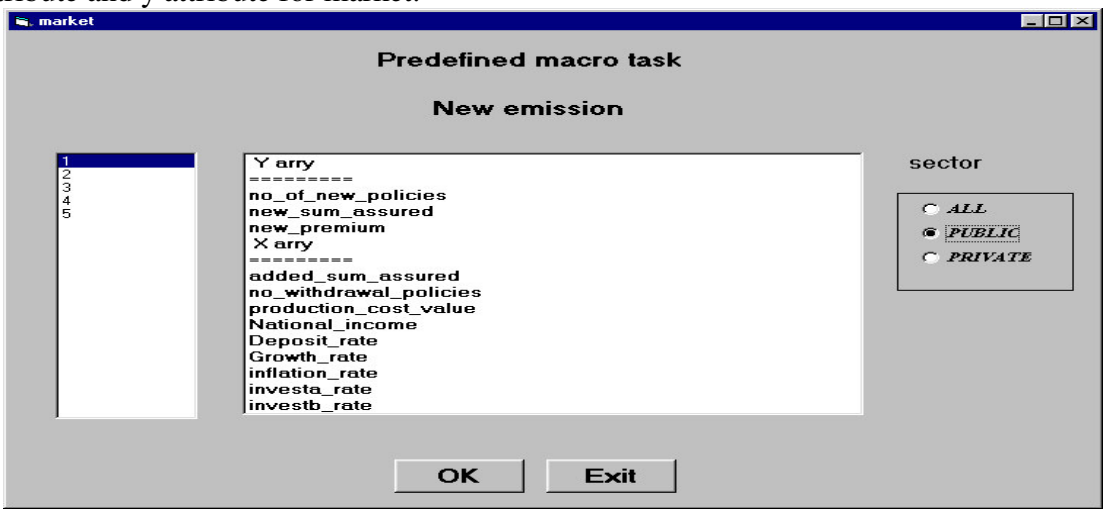


Figure 9 Predefined macro task new emission

Figure 10 depicts the screen layout of output rules for new emission for task no. 1 for market data

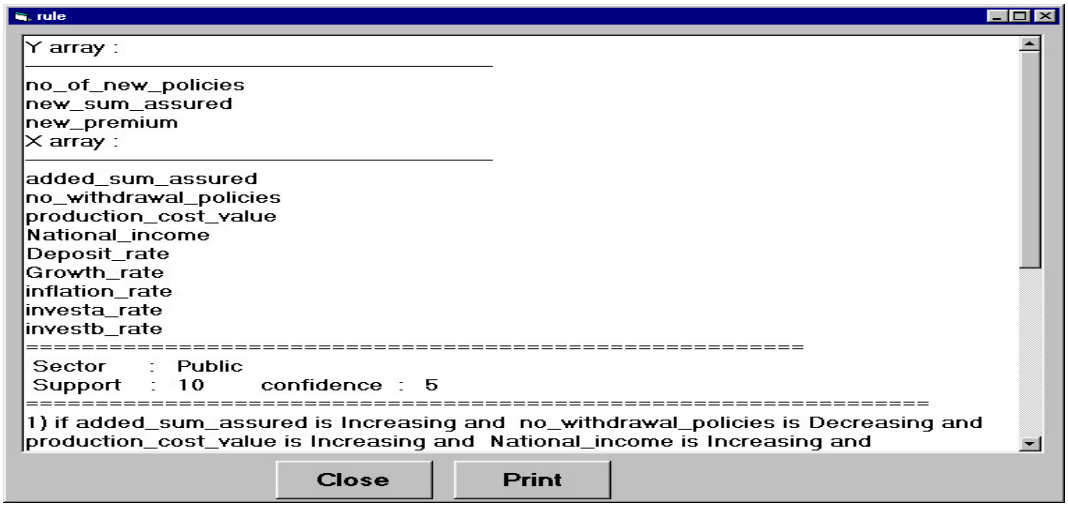


Figure 10 output rules for new emission for task no. 1 for market

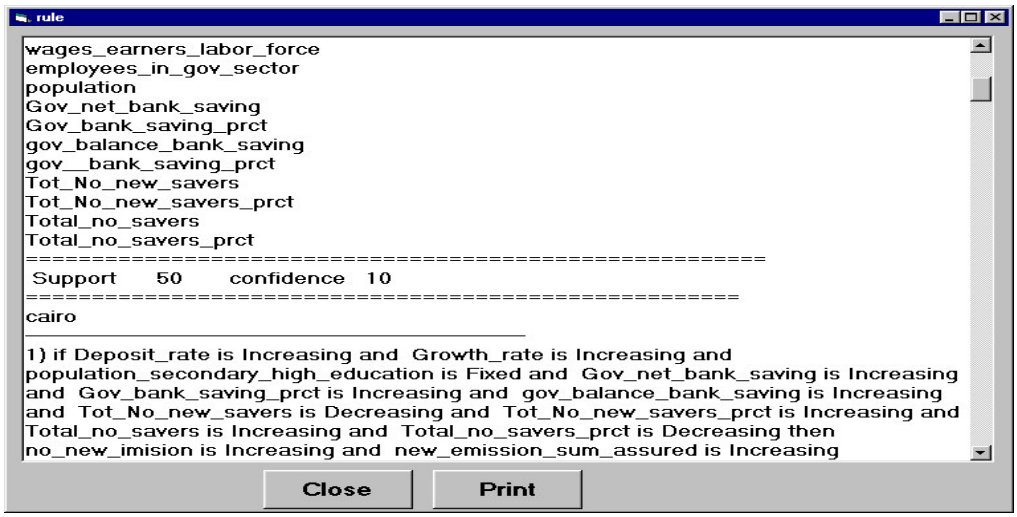


Figure 11 the screen layout of output rules for new emission

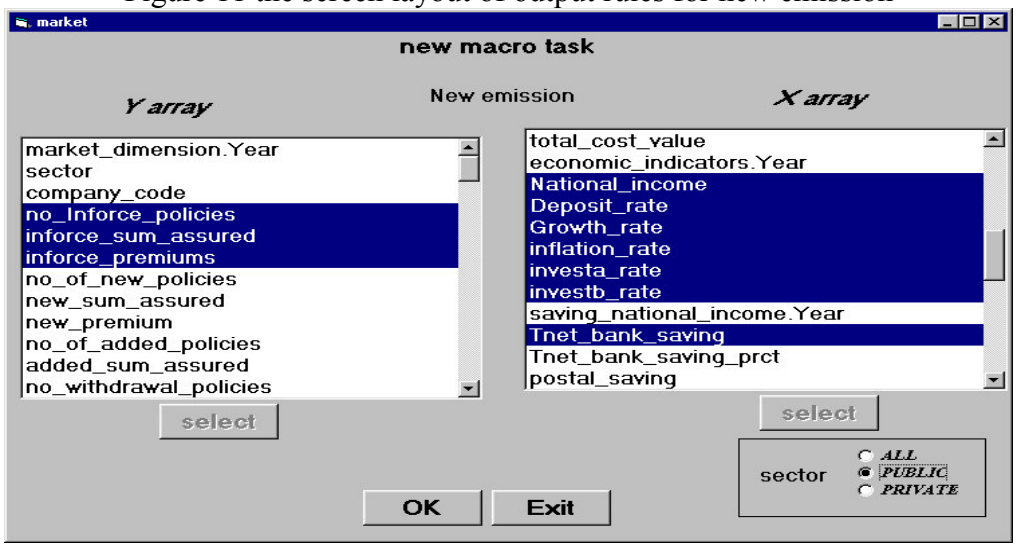


Figure 12 the screen layout of input new emission for expert user represent

Figure 12 depicts the screen layout of input new emission for expert user represent selected X attributes and Y attributes for market.Figure 13 depicts the screen layout of output rules for new emission for expert user for market

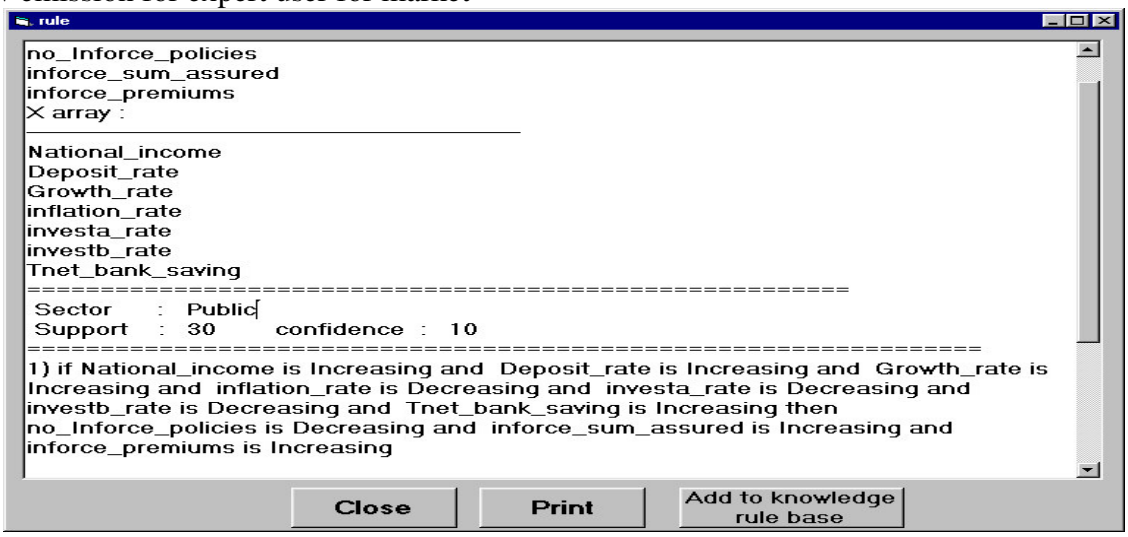


Figure 13 the screen layout of output rules for new emission

Figure 14 depicts the screen layout of selection of rule scope analysis . Figure 15 depicts the screen layout of output rule scope analysis

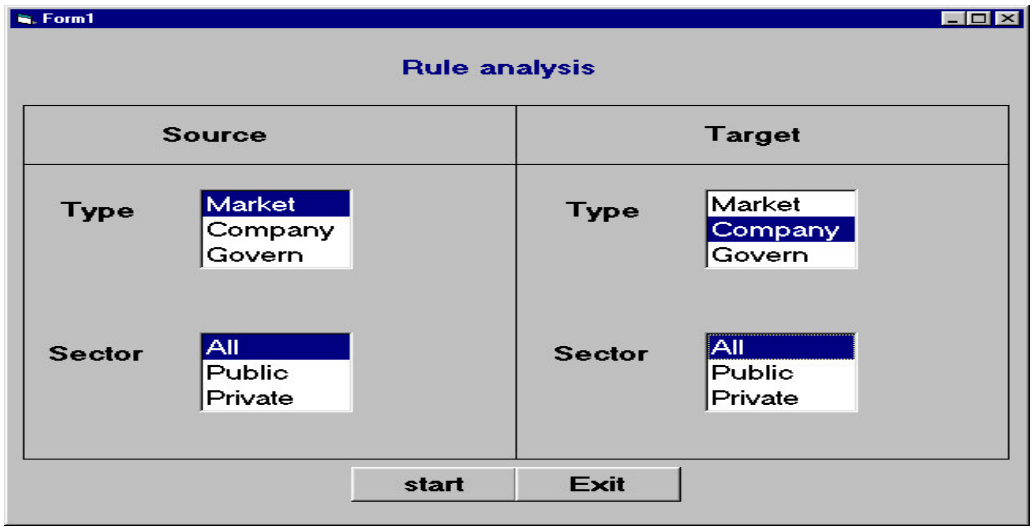


Figure 14 the screen layout of selection of rule scope analysis

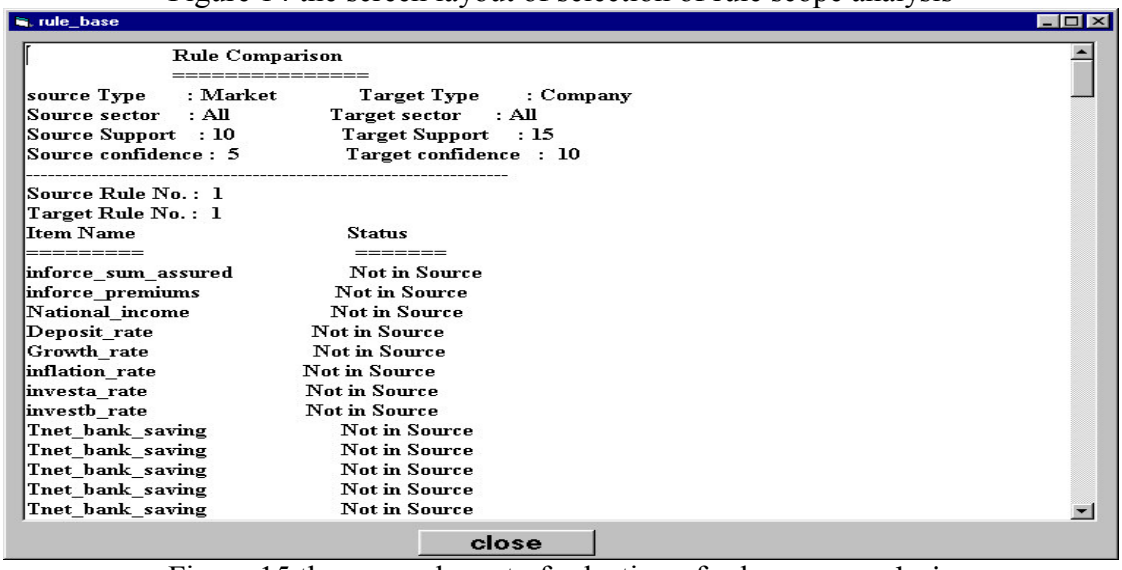


Figure 15 the screen layout of selection of rule scope analysis

Figure 16 depicts the screen layout of output post processing

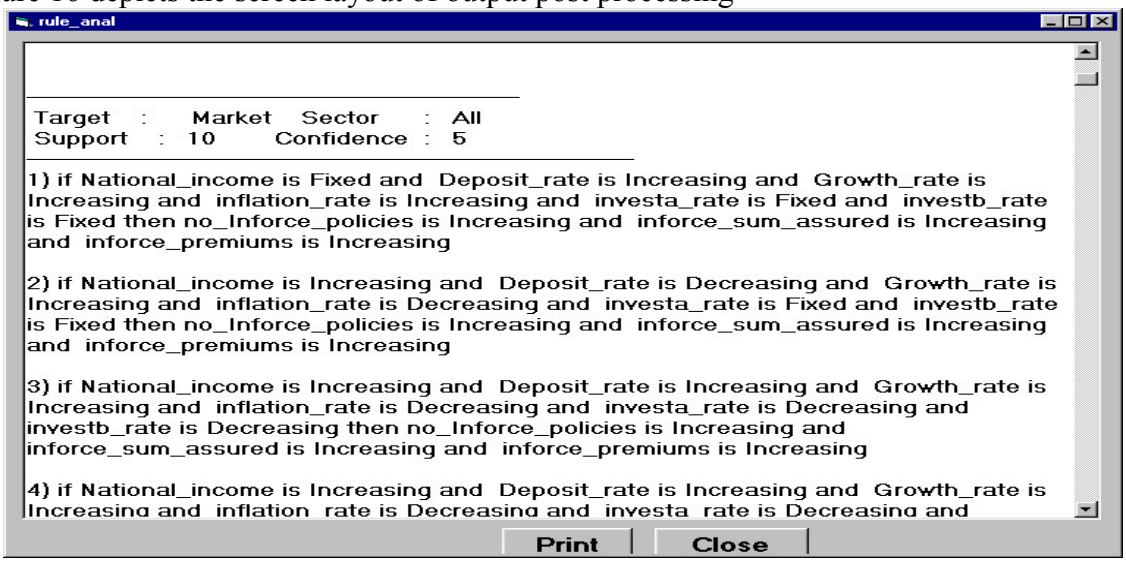


Figure 16 the screen layout of output post processing

6-2 Management of tool repository

In order to enable consistency, and interoperability between different information systems within the organization, the implemented tool libraries are stored at a repository as a meta data**.** The generation and management of Metadata serve two purposes: Minimize the efforts of the administration of a data warehouse and improve the extraction of information from it.

Levels of meta-model are depicted in figure 17. Level 0 contains data. Level 1 contains metadata, it represents database schema of stored data. Level 2 contains meta-model, it specifies the schema used to store metadata at the repository.

We note that an instantiation relationship exists between levels; each level contains instances from the schema of the level above.

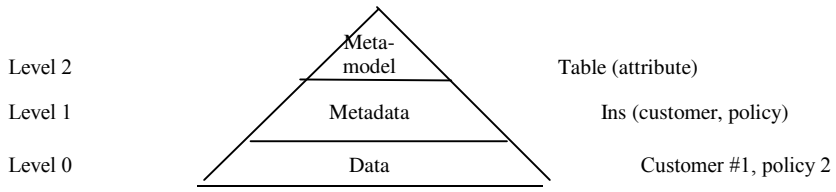
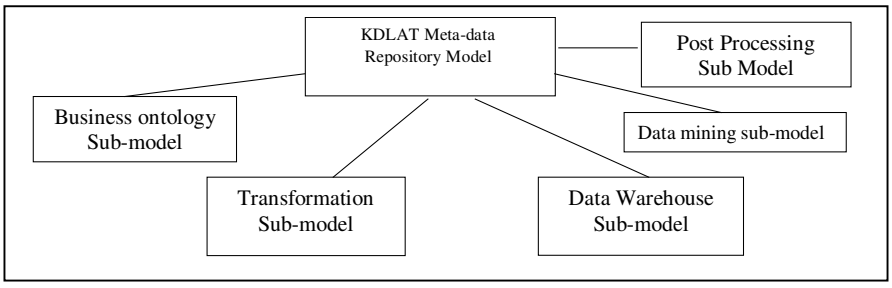


Figure 17 Meta-model levels

The architecture of meta-model as depicted in figure 18 is based on a proposed hybrid conceptual representation based on the framework of known standards: open information model (OIM), common warehouse meta-model (CWM) and Ontology model. Figure 18 depicts the architecture of proposed meta model. Meta-model consists of many sub-models as business ontology sub-model, preprocessing and transformation sub-model, data warehouse sub-model, data mining sub-model and post processing sub-model.

Figure 18 Architecture of meta-model

Repository sub-model is stored in a database, which is maintained by a dedicated system using the facilities of the underlying relational database management system. The function of the repository management system is to allow the expert user the capabilities of addition, updating, deletion and retrieval of the components of the sub-model.



7-Conclusion and Future Work

In this paper we propose techniques to build a domain specific life assurance knowledge discovery tool, dedicated to support both strategic and tactical decision making. In order to build such a tool the following methodology is applied: Devising ontology analysis that aims to identify domain conceptual elements, relations, and specify the requirements of insurance decision-makers. Insurance data as defined in ontology analysis are continuous, large, dynamic, multi dimensional database tables representing a sequence of events, where each event has an associated time of occurrence. By matching the goal of knowledge discovery tool in life assurance, we select a hybrid technique consisting of clustering partitioning and rule induction techniques based on Apriori algorithm and its extension. The domain expert can impose his interesting measures, by selecting appropriate attributes for each mining task to guide it.

This hybrid technique is proposed in order to solve rule induction problems. These problems are speed and clarity. Rule induction systems are often used in high dimensional spaces with high cardinality, they construct rules independently by extracting all possible patterns from the database. More dimensions slow down the process, the increase in time is usually linear with the number of predictors. The user is overwhelmed with obscure rules, which affects the clarity of the system, he is often required to review a large number of rules that are proposed as interesting in order to actually determine whether it is important. Speed and clarity problems are solved by data clustering partitioning and dimensions reduction, before rule induction process to eliminate redundant rules. We focus on mining rules for only one target item at a time, such rules could be mined more efficiently than rules with arbitrary heads [31]. We introduce algorithms for discovering knowledge in the form of association rules, suitable for data characteristics. These algorithms are classified according to its use into preprocessing to reconcile different types of data with mining goal, building a data warehouse, data mining and post processing. Association rules algorithms are classified according to level of aggregation of the used data to micro rule induction algorithm working in detailed database tables, and macro rule induction algorithm working in different level of aggregated database tables. A comparison of the rules antecedent and consequent generated by selecting different scope of time interval, aggregation levels support and confidence values are displayed. The proposed tool is managed by a repository meta-model instantiated by meta-data libraries specific to insurance domain.

The proposed tool is implemented on a PC running on Ms Windows 2000. Samples of life assurance data are extracted from different geographical locations of an insurance company covering ten years, and some published national economic and demographic statistics.. By using the induced rules, the decision-maker can analyze the impact of economic and demographic changes on the development of insurance business. In this paper, we focus on macro level rule induction to support decision making in strategic planning.

Future work has two perspectives, business perspective and technical perspective. For business perspective, expand the scope of the tool to add new entities to the ontology holding general insurance emission and claim data. Integrate the tool to some available decision support systems, and to be integrated with the organizational information systems.

Technical perspective issues are the expansion of incremental updating to study the impact of actions taken by the decision makers as a result of generated rules in previous session, to evaluate the action-ability of generated rules. Expand the functionality of induced rules stored into the rule base table to hold an engine executable file to be used as a knowledge base of insurance expert system, to overcome the problem of sparse domain expertise. Improve the time complexity of the data mining algorithms by experimenting other techniques such as genetic algorithm, neural networks, decision trees, to be used as a hybrid of techniques suitable for financial data characteristics, and compare the induced rules and performance measures of the new technique with the used technique. Build an interface with e-commerce applications to provide insurance services tailored to the discovered patterns for different customers and market channels.

REFERENCES

[1] Alex Berson and Stephen J. Smith, “Data Warehousing, Data Mining, And OLAP”, MC Graow–Hill, 1997.

[2] Bigus and Joseph P, “Data Mining With Neural Networks”, MC Graw–Hill, New York 1996.

[3] Christopher J. Matheus, Gregory Piatetshy–Shapiro and Dwight Mcneill”, Selecting and Reporting what is Interesting The Kefir Application to Health Care Data”, Advances in Knowledge Discovery and Data Mining, AAA1 Press/The MIT Press, 1996.

[4] Dasrathy B. V., Ed, “Nearest Neighbor Norms: NN Pattern Classification Techniques”, IEEE, Computer Society Press, Calif. 1990.

[5] David Cheung, Vincent T., Ada W. Fu and Yongjian Fv, “Efficient Mining of Association Rules in Distributed Databases”, IEEE, 1996.

[6] Graig Silverstein, Sergey Brin and Rajeev Montwani, “Beyond Market Baskets: Generalizing Association Rules to Dependence Rules”, Data Mining and Knowledge Discovery, Vol. 2, No. 1, Jan 1998, Kluwer Academic Publishers.

[7] Hongjun LU, Ling Feng and Jiawei Han, “Beyond Intratransaction Association Analysis: Mining Multidimensional Intertransaction Association Rules”, ACM Transactions on Information Systems, Vol. 18, October 2000.

[8] Huan Liu, Farhad Hussain, Chew Lim Tan and Manoranjan Dash, “Discretization: An Enabling Technique", Data Mining and Knowledge Discovery”, vol. 6 No. 4, October 2002.

[9] J. Date, "An Introduction to Database Systems", Addition Wesley Longman, Seven Edition, 2000.

[10] Jiawei Han, Laks V. S. Lakshmanan and Raymond T.NG, “Constraint-Based Multidimensional Data Mining”, IEEE, August 1999.

[11] Jorg-Uwe Kietz, Regina Zucker and Anca Vaduva, “Mining Mart: Combining Case-Based-Reasoning and multi-Strategy Learning Into a Frame For Reusing KDD-Applications”, Proc 5th Workshop on Multi-Strategy Learning (MSL 2000) Portugal, June 2000, Kluwer Academic Publishers.

[12] Ken Orr, “Data Warehouse Technology”, Copyright. The Ken Or Institute, 1997.

[13] Krzysztof J. Cios, Witold Pedryez and Roman W. Surniarski, “ Data Mining Methods for Knowledge Discovery“, Kluwer Academic Publishers 1998 Second Printing 2000.

[14] Mariano Fernendez Lopez, Asuncion Gomez-Perez, Juan Pazos Sierra, Polytechnic and Alejandro Pazos Sierra, “Building a Chemical Ontology Using Methontology and the Ontology Design Environment”, IEEE Intelligent System. Jan / Feb 1999.

[15] Martin Staudt, Anca Vaduva and Thomas c, “Metadata Management and Data Warehouse", Technical Report, Information System Research, Swiss Life, University of Zurich, Department of Computer Science, July 1999. vaduva@ifi.unizh.ch

[16] Ming-Syan chen, Jiawei Han and Philip S. Yu, “Data Mining: An Overview From a Database Perspective”, IEEE Transactions on Knowledge and Data Engineering Vol. 8, No. 6, Dec. 1996.

[17] Natalya Friedman Noy and Carole D. Hafner, “The State of The Art in Ontology Design”, AI Magazine Vol. 18, No. 3, Fall 1997.

[18] Rakesh A. grawal, “Parallel Mining of Associations Rule”, IEEE, Dec 1996.

[19] Ramakrishnan Srikant and Rakesh A. Grawal, “Mining Quantitative Association Rules in Large Relational Tables”, Proc Sigmod ‘96, 6/96 Montreal Canada, 1996 ACM.

[20] Ramakrishnan Srikant and Rakesh A. Grawal, “Mining Generalized Association Rules", Proceedings of The ‘21st VLDB Conference”, Zurich, Switzerland, 1995.

[21] Raymond T. Ng, Laks V. S. Lakshmanan, Jiawei Hon and Alex Pany, “Exploratory Mining and Pruning Optimizations of Constrained Associations Rules”, ACM 1998 page 13.

[22] Ronald J. Brachman, “The Process of Knowledge Discovery in Database”, Advances in Knowledge Discovery and Data Mining, AAA1 Press / The MIT Press, 1996.
[23] Ronald J. Brachman, Tom Khabaza, Willie Kloegsgen, Gregoy Piatetsky – Shadiro and Evangelos Simoudis, “Mining Business Databases” Communications of the ACH, November 1996, Vol. 39, No. 11.
[24] Thomas Vertterly, Anca Vaduva and Marten Staudt, “Metadata Standards for Data Warehouse: Open Information Model as Common Warehouse Metadata” vaduva@ifi.unizh.ch, 2000.
[25] Tomaz imielinski, Leonid Khachiy and Amin Abdulghani, “Cubegrades: Generalizing Association Rules", Data Mining and Knowledge Discovery”, Vol. 6, No. 3, Kluwer Academic Publisher, July 2002.
[26] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, “The KDD Process For Extracting Useful Knowledge From Volume F Data”, Communication of ACM, Nov 1996 / Vol. 39, No. 11.
[27] Ussama M. Fayyad, Gregory Piatetsky – Shapiro and Padhraic Smylh, “From Data Mining to Knowledge Discovery, an Overview”, Advanced in Knowledge Discovery & Data Mining, AAA1 Press / The MIT Press, Massachusetts Inst. of Tech, 1996.
[28] Valery Guralnik, Dumindo Wijesekera and Jaideep Srivastava, “Pattern Directed Mining of Sequence Data”, 4th International Conference on Knowledge Discovery & Data Mining, August 1998, New York.
[29] Vasant Dhar, Dashin Chou and Foster Provost, “Discovering Interesting Patterns For Investment Decision Making With Glower–Agenetic Learner Overlaid With Entropy Reduction”, Data Mining and Knowledge Discovery Vol. 4, No. 4, Kluwer Academic Publishers, October 2000.
[30] Wei-Min Shen and Bing Leng, “A Meta-pattern Based Automated Discovery Loop For Integrated Data Mining–Unsupervised Learning of Relational Patterns”, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, Dec. 1996.
[31] Weiyang Lin, Sergio Alvarez and Carolina Ruiz, “Efficient Adaptive Support User Association Rule Mining for Recommender System", Data Mining and Knowledge Discovery, Jan. 2002, Vol. 6, No.1.
[32] Widrow and Rumelhart, “Neural Networks: Application in Industry”, Business and Science Communication of the ACM 37(3) 1994.
[33] Y. Balaji Padmanabham and Alexander Tuzhili, "A Belief Driven Method for Discovering Unexpected Patterns", the 4th International Conference Knowledge Discovery and Data Mining, August 1998.
[34] Y. Gauten Das, King-Ip Lin, Heikki Mannila, Gopal Renganathen and Padhrik Smyth, "Rule Discovery From Time Series", Proceedings of The 4th International Conference on Knowledge Discovery and Data Mining, November 1998.
[35] Z. S. Leigh, "Underwriting - A Dying Art", Journal of The Institute of Actuaries, Vol. 117, part III, The Alden Press Oxford, 1990.