"DATA ANALYSIS OF COVID19 PANDEMICARTIFICIAL INTELLIGENCE BASED SOLUTIONS FOR INOVATIVE TREATMENT AND DISEASE DETECTION SOLUTIONS"

A Dissertation

Submitted

In partial fulfillment of the requirements for the award of the degree of

Master of Technology

in

Computer Science Engineering

By

RAMCHANDRA SHARMA
Enrollment Number – 0179CS10MT15
Under the guidance of

RAJEEV RAGHUWANSHI

Asst. Prof. & HOD

Department of Computer Science Engineering

MILLENNIUM INSTITUTE OF TECHNOLOGY & SCIENCE, BHOPAL

Affiliated to

RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

(M.P)

SESSION 2010-12

TABLE OF CONTENTS

TITLE OF DISSERTATION

"A rough set approach for association rule mining"

Abstract

Here we propose a new approach for finding association rule mining .In our proposed methodology we apply Rough set theory for finding the association rule. In previous paper a soft set approach is used to find the association rule. In soft set approach first the transactional database is transferred into the Boolean valued information system because soft set is apply only on the Boolean valued information system and then uses the concept of parameter co-occurrence in a transaction we define the association rule. In this method the input is transactional database that contains the redundant and missing information therefore the final result generated is not accurate and contains some redundant information. To deal with such condition and to produce result that is free from such redundant result we propose a new approach which is based on Rough set theory that is used to deal with redundant and uncertain data that is present in transaction database.

INTRODUCTION

Association rule mining

Association rule is one of the most popular data mining techniques and has received considerable attention, particularly since the publication of the Apriori algorithms. They are particularly useful for discovering relationships among data in huge databases and applicable to many different domains including market basket and risk analysis in commercial environments, epidemiology, clinical medicine, fluid dynamics, Astrophysics , and crime prevention. The association rules are considered interesting if it satisfies certain constraints, i.e. predefined minimum support (minsupp) and minimum confidence (minconf) thresholds. Many algorithms of association rules mining have been proposed. The association rules method was developed particularly for the analysis of transactional databases.

Let $i = \{i_1, i_2, \ldots\ldots i_{|A|}\}$ for $|A|>0$ refers to the set of literals called set of items and the set $D= \{t_1, t_2, \ldots\ldots t_{|U|}\}$, for $|U|>0$ refers to the transaction dataset, where each transaction $t \in D$ is a list of distinct items $t = \{ i_1, i_2, \ldots\ldots i_{|M|}\}$, $1<=|M|<=|A|$ and each transaction can be identified by a distinct identifier TID. Let a set $X \subseteq t \subseteq I$ called an itemset. An itemset with k-items is called k-itemset. The support of an itemset X is denoted as sup(X) is defined as number of transaction contains X. An association rule between X and Y is implication of the form $X \rightarrow Y$ where $X \cap Y = \emptyset$. The itemset X and Y called antecedent and consequent, respectively. The support of an association rule

X ⟶Y, denoted of the form sup (X⟶Y) is defined as the number of transaction in D contains X Ụ Y. The confidence of an association rule X⟶Y, denoted as cfi (X⟶Y) is defined as a ratio that number of transaction in D that contains X Ụ Y to the number of transaction that in D that contain X. Thus

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

A huge number of association rules can be found from a transactional dataset. To find the interesting association rules in a transactional dataset, we must define a specified minimum support (called minsup) and specified minimum confidence (called minconf).

The itemset $Y \subseteq I$ is called frequent itemset if sup (X) => minsup. It is known that a subset of frequent itemset is a frequent itemset any superset of infrequent itemset is not a frequent itemset. Finally, the association rule X→Y holds if conf (X→Y)>= minconf.

The association rules are said to be strong if it meets the minimum confidence threshold. However, while association rules provide means to discover many interesting associations

MINING ASSOCIATION RULES—AN EXAMPLE

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Min. support 50%
Min. confidence 50%

Frequent Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

For rule $A \Rightarrow C$:

support =2/4=.5

support($\{A \cup C\}$) = 50%

Confidence=2/3=.66

confidence = support($\{A \cup C\}$)/support($\{A\}$) = 66.6

LITERATURE SURVEY

In literature survey study the two basic algorithm Apriory algorithm and FP growth algorithm which is used to find the frequent pattern in the transaction database.Then study the paper in which new approach for association rule mining is proposed .In this approach a soft set approach is applied to find the association rule.

1. Apriori algorithm

The Apriori principle:

Any subset of a frequent itemset must be frequent.

Find the *frequent itemsets*: the sets of items that have minimum support

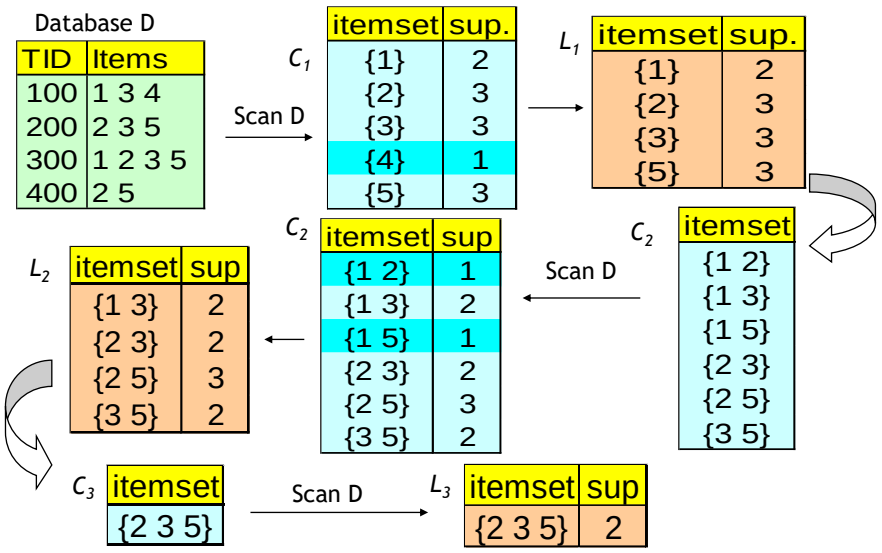
A subset of a frequent itemset must also be a frequent itemset

i.e., if {*AB*} is a frequent itemset, both {*A*} and {*B*} should be a frequent itemset

Iteratively find frequent itemsets with cardinality from 1 to *k (k*-itemset)

Join Step: C_k is generated by joining L_{k-1} with itself

Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset.



2. FP growth algorithm

Compress a large database into a compact, Frequent-Pattern tree (FP-tree) structure

highly condensed, but complete for frequent pattern mining

Avoid costly database scans

Develop an efficient, FP-tree-based frequent pattern mining method

A divide-and-conquer methodology: decompose mining tasks into smaller ones

Avoid candidate generation: sub-database test only

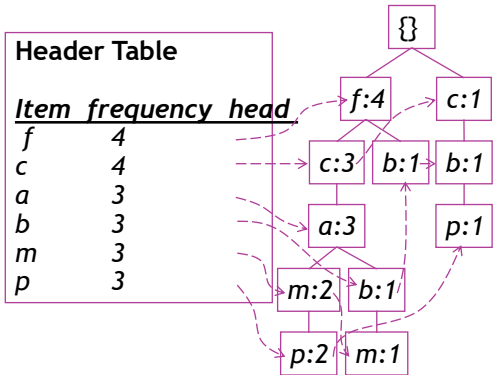
CONSTRUCT FP-TREE FROM A TRANSACTION DB

TID	Items bought	frequent item
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

min_support=0.5

Steps:

1. Scan DB once, find frequent 1-itemset (single item pattern)
2. Order frequent items in frequency descending order
3. Scan DB again, construct FP-tree



3. A soft set approach for association rules mining

It is a new approach for finding the association rule from transactional database using soft set theory. This approach is started by a transformation of a transaction database into Boolean valued information system. Since the ‘‘standard” soft set deals with such information system, thus a transactional dataset can be represented as a soft set. Using the concept of parameters co-occurrence in a transaction

Boolean valued information system

The syntax of an information system is very similar to relations in relational database. Entities in relational databases are also represented by tupples of attribute values. An information system as is a 4-tuple (quadruple) S = (U, A ,V, f).

In an information system S = (U, A, V, f)., if V={0,1} then S is called Boolean valued information system.

Soft set theory

A pair (F,E) is called a soft set over U, where F is a mapping given by

$$F: E \rightarrow P(U)$$

In other words, a soft set over U is a parameterized family of subsets of the universe U. For e 2 E, F(e) may be considered as the set of e-elements of the soft set (F,E) or as the set of e-approximate elements of the soft set. Clearly, a soft set is not a (crisp) set.

Problem Definition

In the paper “A soft set approach for association rule mining “is a new approach for finding frequent pattern and then association rule mining is proposed. In this approach first the transaction dataset is transferred into Boolean valued information system and then soft set is apply to find the association rule. This approach is produce association rule similar to previous approach but with faster in execution.

This approach is good but cause some problem .The transformation process of transaction dataset into Boolean valued information system is complex and there is possibilities of producing result is not accurate.

In this transaction database is used as input to our process therefore there are some type of redundant information is present in the transaction database for ex the two transaction has the same value but the soft set can’t distinguish these type of transaction and consider these transaction two distinct transaction and produce result for both transaction which is actually similar. Therefore these types of redundant information are not handling by soft set due to which the result produce by this approach is not accurate.

It suffer from a problem that the process to translate a transaction database into Boolean valued information system and then apply soft set is complex.

Proposed methodology

A soft set approach is facing a problem to deal with the redundancy pr esent in the transaction database such as duplicate value, missing information and does not produce accurate result .Therefore to deal with such type situation we proposed new method to deal with such type of redundancy.In our method we apply a rough set theory on the transaction database to deal with such type of redundancy. Rough set theory can be regarded as a new mathematical tool for imperfect data analysis. The theory has found applications in many domains, such as decision support, engineering, environment, banking, medicine and others.

Rough set philosophy is founded on the assumption that with every object of the universe of discourse some information (data, knowledge) is associated. Objects characterized by the same information are *indiscernible (similar)* in view of the available information about them. The *indiscernibility relation* generated in this way is the mathematical

basis of rough set theory. Any set of all indiscernible (similar) objects is called an *elementary set*, and forms a basic *granule (atom)* of knowledge about the universe. Any union of some elementary sets is referred to as a *crisp (precise)* set – otherwise the set is *rough (imprecise, vague)*. Each rough set has boundary-line cases, i.e., objects which cannot be with certainty classified, by employing the available knowledge, as members of the set or its complement. Obviously rough sets, in contrast to precise sets, cannot be characterized in terms of information about their elements. With any rough set a pair of precise sets, called the *lower* and the *upper approximation* of the rough set, is associated. The lower approximation consists of all objects which *surely* belong to the set and the upper approximation contains all objects which *possibly* belong to the set. The difference between the upper and the lower approximation constitutes the *boundary region* of the rough set. Approximations are fundamental concepts of rough set theory.

REFRENCES:

1. A soft set approach for association rules mining by Tutut Herawan , Mustafa Mat Deris in Elsevier in 2010.

2. A collaborative educational association rule mining tool by

Enrique García, Cristóbal Romero ∗, Sebastián Ventura, Carlos de Castro. in Elsevier in 2010.

3. Apriori algorithm for frequent pattern mining by R. Agrawal and Srikant.

4. FP growth algorithm for mining frequent pattern without candidate set.