Chapter

Data Mining and Fuzzy Data Mining Using MapReduce Algorithms

Poli Venkata Subba Reddy

Abstract

Data mining is knowledge discovery process. It has to deal with exact information and inexact information. Statistical methods deal with inexact information but it is based on likelihood. Zadeh fuzzy logic deals with inexact information but it is based on belief and it is simple to use. Fuzzy logic is used to deal with inexact information. Data mining consist methods and classifications. These methods and classifications are discussed for both exact and inexact information. Retrieval of information is important in data mining. The time and space complexity is high in big data. These are to be reduced. The time complexity is reduced through the consecutive retrieval (C-R) property and space complexity is reduced with blackboard systems. Data mining for web data based is discussed. In web data mining, the original data have to be disclosed. Fuzzy web data mining is discussed for security of data. Fuzzy web programming is discussed. Data mining, fuzzy data mining, and web data mining are discussed through MapReduce algorithms.

Keywords: data mining, fuzzy logic, fuzzy data mining, web data mining, fuzzy MapReduce algorithms

1. Introduction

Data mining is an emerging area for knowledge discovery to extract hidden and useful information from large amounts of data. Data mining methods like association rules, clustering, and classification use advanced algorithms such as decision tree and k-means for different purposes and goals. The research fields of data mining include machine learning, deep learning, and sentiment analysis. Information has to be retrieved within a reasonable time period for big data analysis. This may be achieved through the consecutively retrieval (C-R) of datasets for queries. The C-R property was first introduced by Ghosh [1]. After that, the C-R property was extended to statistical databases. The C-R cluster property is a presorting to store the datasets for clusters. In this chapter, C-R property is extended to cluster analysis. MapReduce algorithms are studied for cluster analysis. The time and space complexity shall be reduced through the consecutive retrieval (C-R) cluster property. Security of the data is one of the major issues for data analytics and data science when the original data is not to be disclosed.

The web programming has to handle incomplete information. Web intelligence is an emerging area and performs data mining to handle incomplete information. The incomplete information is fuzzy rather than probability. In this chapter, fuzzy web programming is discussed to deal with data mining using fuzzy logic. The fuzzy algorithmic language, called FUZZYALGOL, is discussed to design queries in data mining. Some examples are discussed for web programming with fuzzy data mining.

2. Data mining

Data mining [2–5] is basically performed for knowledge discovery process. Some of the well-known data mining methods are frequent itemset mining, association rule mining, and clustering. Data warehousing is the representation of a relational dataset in two or more dimensions. It is possible to reduce the space complexity of data mining with consecutive storage of data warehouses.

The relational dataset is a representation of data with attributes and tuples.

Definition: A relational dataset R or cluster dataset is defined as a collection of attributes $A_1, A_2, \ldots, A_m$ and tuples $t_1, t_2, \ldots, t_n$ and is represented as

$R = A_1 \text{ x } A_2 \text{ x } \ldots \text{ x } A_m$

$t_i = a_{i1} \text{ x } a_{i2} \text{ x } \ldots \text{ x } a_{im}$ are tuples, where i =1,2,.., n

or

$R(A_1. A_2. \ldots A_m)$. R is a relation.

$R(t_i) = (a_{i1}. a_{i2} \ldots . a_{im})$ are tuples, where i =1,2,.., n

or instance, two sample datasets “price” and “sales” are given in **Tables 1** and **2**, respectively.

INo	IName	Price
I005	Shirt	100
I007	Dress	50
I004	Pants	80
I008	Jacket	60
I009	Skirt	100

Table 1.
Sample dataset “price.”

INo	IName	Sales
I005	Shirt	80
I007	Dress	60
I004	Pants	100
I008	Jacket	50
I009	Skirt	80

Table 2.
Sample dataset “sales.”

The lossless join of the datasets "price" and "sales" is given in **Table 3**.

INo	IName	Sales	Price
I005	Shirt	80	100
I007	Dress	60	50
I004	Pants	100	80
I008	Jacket	50	60
I009	Skirt	80	100

Table 3.
Lossless join of the price and sales datasets.

In the following, some of the methods (frequency, association rule, and clustering) are discussed.

Consider the "purchase" relational dataset given in **Table 4**.

CNo	INo	IName	Price
C001	I005	shirt	100
C001	I007	Dress	50
C003	I004	pants	80
C002	I007	dress	80
C001	I008	Jacket	60
C002	I005	shirt	100

Table 4.
Sample dataset "purchase."

2.1 Frequency

Frequency is the repeatedly accrued data.

Consider the following query:

Find the frequently customers purchase more than one item.

SELECT P.CNo, P.INo, IName, COUNT(*)
FROM purchase P
WHERE COUNT(*)>1.

The output of this query is given in **Table 5**.

CNo	INo	COUNT
C001	I005	2
C002	I005	2

Table 5.
Frequency.

2.2 Association rule

Association rule is the relationship among the data.

Consider the following query:

Find the customers who purchase shirt and dress.

<shirt⇔ dress>

SELECT P.CNo, P.INo

FROM purchase P
WHERE IName=”shirt” and IName=”dress”.
The output of this query is given in **Table 6**.

CNo	INo
C001	I005
C002	I005

Table 6.
Association.

2.3 Clustering

Clustering is grouping the particular data.
Consider the following query:
Group the customers who purchase dress and shirt.
The output of this query is given in **Table 7**.

CNo	INo	IName	Price
C001	I007	Dress	50
	I005	shirt	100
C002	I007	dress	80
	I005	shirt	100

Table 7.
Clustering.

3. Data mining using C-R cluster property

The C-R (consecutive retrieval) property [1, 3] is the retrieval of records of database consecutively. Suppose $R = \{r_1, r_2, \ldots, r_n\}$ is the dataset of records and $C = \{C_1, C_2, \ldots, C_m\}$ is the set of clusters.

The best type of file organization on a linear storage is one in which records pertaining to clusters are stored in consecutive locations without redundancy storing any data of R.

If there exists on such organization of R for C said to have the Consecutive Retrieval Property or C-R cluster property with respect to dataset R. Then C-R cluster property is applicable to linear storage.

The C-R cluster property is a binary relation between a cluster set and dataset.

R	C_1	C_2	….	C_m
r_1	1	0	…	1
r_2	0	1	;;;	0
-	-	-	…	-
-	-	-	…	-
=	-	-	…	-
r_n	1	1	…	1

Table 8.
Incidence matrix.

Suppose if a cluster in a cluster set *C* is relevant to the data in a dataset *R*, then the relevancy is denoted by 1 and the irrelevancy is denoted by 0. Thus, the relevancy between cluster set *C* and dataset *R* can be represented as (*n* x *m*) matrix, as shown in **Table 8**. The matrix is called dataset-cluster incidence matrix (CIM).

Consider the dataset for customer account given in **Table 9**.

R	CNo	IName	Sales
r_1	70001	Shirt	150
r_2	70002	Dress	30
r_3	70003	Pants	100
r_4	60001	Dress	50
r_5	60002	Jacket	75
r_6	60003	Shirt	120
r_7	60004	Dress	40

Table 9.
Storage of sales.

The dataset given in **Table 9** is reorganized in ascending order based on sorting, as shown in **Table 10**.

R	CNo	IName	Sales
r_1	70001	Shirt	150
r_6	60003	Dress	120
r_3	70003	Pants	100
r_5	60002	Dress	75
r_4	60001	Jacket	50
r_7	60004	Shirt	40
r_2	70002	Dress	30

Table 10.
Reorganizing for C-R cluster.

Consider the following clusters of queries:

C1 = Find the customers whose sales is greater than or equal to 100.

C2 = Find the customers whose sales is less than 100.

C3 = Find the customers whose sales is greater than or equal average sales.

C4 = Find the customers whose sales is less than average sales.

The CIM is given in **Table 11**.

The dataset given in **Table 11** is reorganized with sort on C_1 in descending order, as shown in **Table 12**. Thus, C_1 has C-R cluster property.

The dataset given in **Table 11** is reorganized with sort on C_2 in descending order, as shown in **Table 13**. Thus, C_2 has C-R cluster property.

The dataset given in **Table 11** is reorganized with sort on C_3 in descending order, as shown in **Table 14**. Thus, C_3 has C-R cluster property.

The dataset given in **Table 11** is reorganized with sort on C_4 in descending order, as shown in **Table 15**. Thus, C_4 has a C-R cluster property.

R	C_1	C_2	C_3	C_4
r_1	1	0	1	0
r_2	0	1	0	1
r_3	1	0	1	0
r_4	0	1	0	1
r_5	0	1	1	0
r_6	1	0	1	0
r_7	0	1	0	1

Table 11.
Cluster incidence matrix.

R	C_1
r_1	1
r_3	1
r_6	1
r_2	0
r_4	0
r_5	0
R_7	0

Table 12.
Sorting on C_1.

R	C_2
r_1	0
r_3	0
r_6	0
r_2	1
r_4	1
r_5	1
r_7	1

Table 13.
Sorting on C_2.

R	C_3
r_1	1
r_3	1
r_5	1
r_6	1
r_2	0
r_4	0
r_7	0

Table 14.
Sorting on C_3

R	C_4
r_1	0
r_3	0
r_5	0
r_6	0
r_2	1
r_4	1
r_7	1

Table 15.
Sorting on C_4.

The dataset is given for $C_1 \bowtie C_2$ has C-R cluster property (**Table 16**).

R	$C_1 \bowtie C_2$
r_1	1
r_3	1
r_6	1
r_2	1
r_4	1
r_5	1
r_7	1

Table 16.
$C_1 \bowtie C_2$.

The dataset is given for $C_3 \bowtie C_4$ has C-R cluster property (**Table 17**).

R	$C_3 \bowtie C_4$
r_1	1
r_3	1
r_5	1
r_6	1
r_2	1
r_4	1
r_7	1

Table 17.
$C_3 \bowtie C_4$.

The dataset is given for $C_1 \bowtie C_3$ has C-R cluster property (**Table 18**).
The dataset is given for $C_2 \bowtie C_4$ has C-R cluster property (**Table 19**).
The dataset is given for $C_2 \bowtie C_3$ has C-R cluster property (**Table 20**).
The cluster sets $\{C_1 \bowtie C_2, C_3 \bowtie C_4, C_1 \bowtie C_3, C_2 \cup\bowtie C_4, C_2 \cup\bowtie C_3\}$ has C-R cluster property. Thus, the cluster sets have C-R cluster properties with respect to dataset *R*.

3.1 Design of parallel C-R cluster property

The design of parallel cluster shall be studied through the C-R cluster property. It can be studied in two ways: the parallel cluster design through graph

R	$C_1 \bowtie C_3$
r_1	1
r_3	1
r_6	1
r_2	1
r_4	0
r_5	0
r_7	0

Table 18.
$C_1 \bowtie C_3$.

R	$C_2 \bowtie C_4$
r_1	0
r_3	0
r_6	0
r_2	1
r_4	1
r_5	1
r_7	1

Table 19.
$C_2 \bowtie C_4$.

R	C_2 U C_3
r_1	1
r_3	1
r_6	1
r_2	1
r_4	1
r_5	1
r_7	1

Table 20.
$C_2 \bowtie C_3$.

theoretical approach and the parallel cluster design through response vector approach.

The C-R cluster property between cluster set C and dataset R can be stated in terms of the properties of vectors. The data cluster incidences of cluster set C with C-R cluster property may be represented as response vector set V. For instance the cluster set $\{C_1, C_2, C_3, C_4\}$ has response vector set $\{V_1=(1,1,1,0,0,0,0), V_2=(0,0,0,1,1,1,1), V_3=(1,1,1,0,0,0), \text{and } V_4=(0,0,0,0,1,1,1)\}$ (**Tables 21–23**).

For instance, the response vector of the cluster $C1$ is given by column vector (1,1,1,0,0,0,0).

Suppose C_i and C_j are two clusters. If the two vectors V_i and V_j of C_i and C_j and the intersection $V_i \cap V_j = \Phi$, then the cluster set $\{C_i, C_j\}$ has a parallel cluster

R	C_1	C_2
r_1	1	0
r_3	1	0
r_6	1	0
r_2	0	1
r_4	0	1
r_5	0	1
r_7	0	1

Table 21.
{C_1, C_2}.

R	C3	C_4
r_1	1	0
r_3	1	0
r_6	1	0
r_2	1	0
r_4	0	1
r_5	0	1
r_7	0	1

Table 22.
{C_3, C_4}.

R	C_2	C_3
r_1	0	1
r_3	0	1
r_6	0	1
r_2	1	1
r_4	1	0
r_5	1	0
r_7	1	0

Table 23.
{C_2, C_3}.

property. Consider the vectors V_1 and V_2 of C_1 and C_2. The intersection of $V_1 \cap V_2 = \Phi$, so that the cluster set $\{C_1, C_2\}$ has parallel cluster property. Similarly the cluster set $\{C_3, C_4\}$ has parallel cluster property. The cluster set $\{C_2, C_3\}$ does not have parallel cluster property because $V_1 \cap V_2 \# \Phi$ and r_2 depending on C_1 and C_2.

3.2 Visual design for parallel cluster

The C-R cluster property is studied with graphical approach. This graphical approach can be studied for designing parallel cluster processing (PCP).

Suppose V_i is the vertex of RICM of C. The G(C) is defined by vertices V_i, i=1,2, … , and n, and two vertices have an edge E_{ij} associated with interval I_i={V_i, V_{i+1}} i=1, … ,n-1.

If G(C) has C-R cluster property, the vertices of G(C) have consecutive 1's or 0's.

Consider the cluster set {C_1, C_2}. The G(C1) has the vertices (1,1,1,0,0,0,0), and the G(C_2) has the vertices (0,0,0,1,1,1,1), G(C_3) has the vertices (1,1,1,1, 0,0,0), and G(C_4) has vertices (0,0,0,0,1,1,1).

The parallel cluster property exists if G(C_i) ∩G(C_j)=Φ.

For instance, consider the G(C_1) and G(C_2). G(C_1) ∩G(C_2)=Φ, so that the cluster set {C_1, C_2} has parallel cluster property. The graphical representation is shown in **Figure 1**.

Similarly the cluster set {C_3, C_4} has the parallel cluster property (PCP). The cluster set {C_3, C_4} has no PCP because it is G(C_2) ∩ G(C_3) # Φ

The graph G(C_1) ∩ G(C_2) = Φ have consecutive cluster property.

The graph G(C_3) ∩ G(C_4) = Φ have consecutive cluster property. The graphical representation is shown in **Figure 2**.

The graph G(C_2) ∩ G(C_3) # Φ does not have consecutive cluster property. The graphical representation is shown in **Figure 3**.

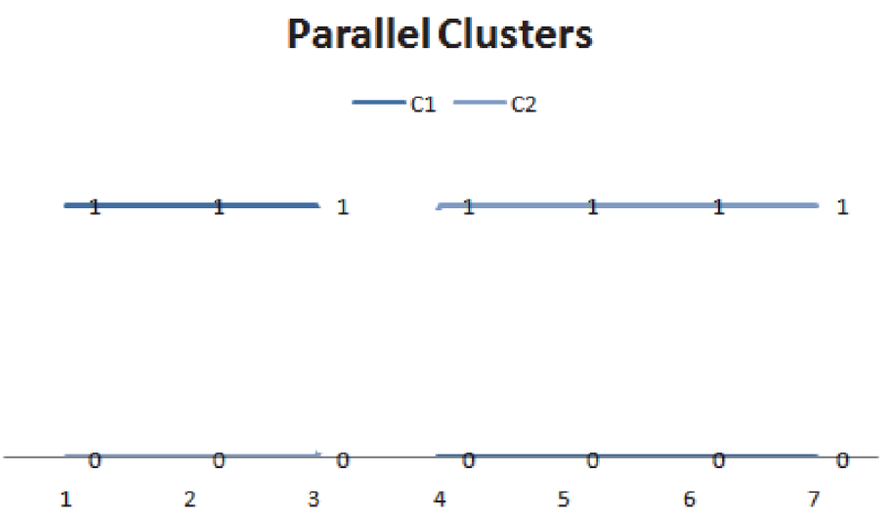

Figure 1.
{C_1, C_2}.

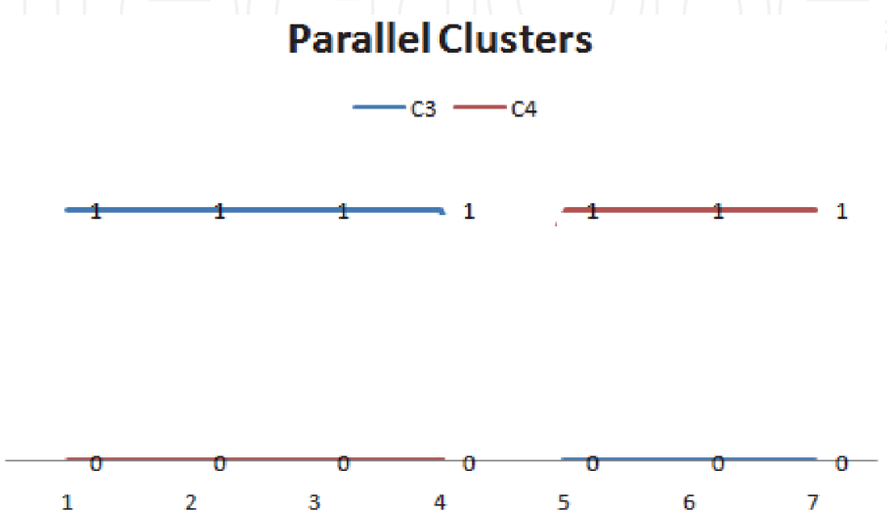

Figure 2.
{C_3, C_4}.

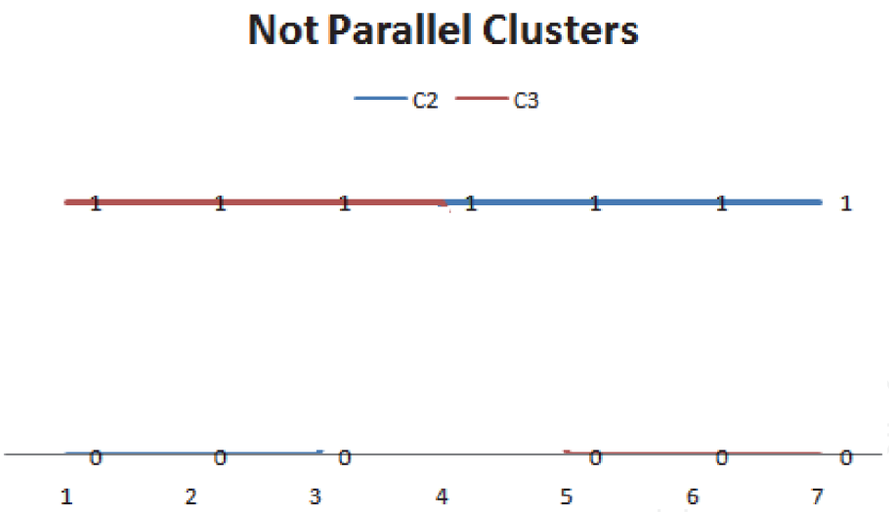

Figure 3.
{C_2, C_3}.

3.3 Parallel cluster design through genetic approach

Genetic algorithms (GAs) were introduced by Darwin [6]. GAs are used to learn and optimize the problem [7]. There are four evaluation processes:

- Selection
- Reproduction
- Mutation
- Competition

Consider the following crossover with two cuts:
Parent #1 00001111
Parent #2 11110000
The parent #1 and #2 match with crossover.
The C-R cluster property is studied through genetical study. This study will help for designing parallel cluster processing (PCP).
Definition: The gene G of cluster G(C) is defined as incidence sequence.
Suppose G(C_1) is parent and G(C_2) child genome of cluster incidence for C_1 and C_2.
Suppose the G(C_1) has (1,1,1,0,0,0,0) and the G(C_2) has the v(0,0,0,1,1,1,1).
The parallel cluster property may be designed using genetic approach with the C-R cluster property.
Suppose C is cluster set, R is dataset and G(C) is genetic set.
The parallel cluster property exists if G(C_i) and G(C_j) matches with crossover.
For instance,
G(C_1) = 11110000
G(C_2) = 00001111
G(C_1) and G(C_2)matches with the crossover.
The cluster set {C_1, C_2} has parallel cluster property.
Similarly the cluster set {C_3, C_4} has the parallel cluster property. The cluster set {C_3, C_4} has no PCP because G(C_2) and G(C_3) are not matched with crossover.

3.4 Parallel cluster design cluster analysis

Clustering is grouping the particular data according to their properties, and sample clusters C_1 and C_2 are given in **Tables 24** and **25**, respectively.

R	C_1
r_1	1
r_3	1
r_6	1

Table 24.
Cluster C_1.

R	C_2
r_2	1
r_4	1
r_5	1
r_7	1

Table 25.
Cluster C_2.

Thus, the C_1 and C_2 have consecutive parallel cluster property (**Tables 26** and **27**).

R	C_3
r_1	1
r_3	1
r_5	1
r_6	1

Table 26.
Cluster C_3.

R	C_4
r_2	1
r_4	1
r_7	1

Table 27.
Cluster C_4.

Thus, the C_3 and C_4 have consecutive parallel properly. C_2 and C_3 do not have consecutive parallel cluster property because r_2 is common.

4. Design of retrieval of cluster using blackboard system

Retrieval of clusters from blackboard system [8] is the direct retrieval of data sources. When the query is being processed, the entire database has to bring to main

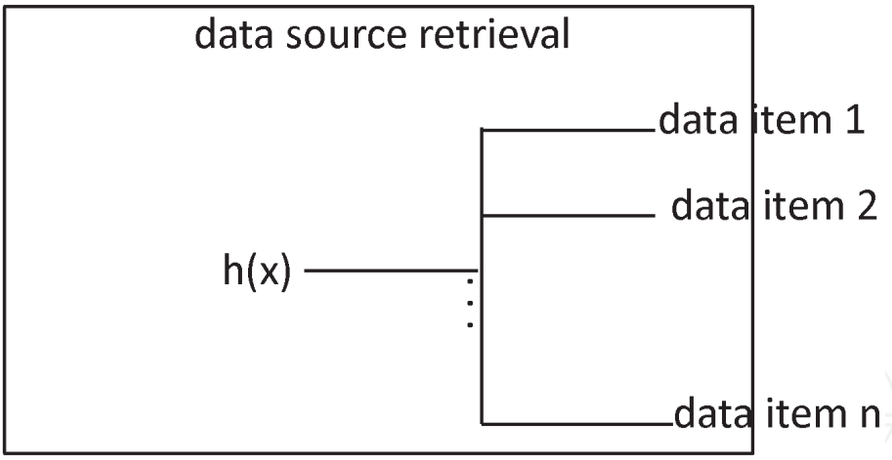

Figure 4.
Blackboard system.

memory but in blackboard architecture, the data item source is direct from the blackboard structure. For the retrieval of information for a query, data item is directly retrieved from the blackboard which contains data item sources. Hash function may be used to store the data item set in the blackboard.

The blackboard systems may be constructed with data structure for data item sources.

Consider the account (AC-No, AC-Name, AC-Balance)

Here AC-No is key of datasets.

Each data item is data sourced which is mapped by h(x).

These data items are stored in blackboard structure.

When the transaction is being processed, there is no need to take the entire database into the main memory. It is sufficient to retrieval of particular data item of particular transaction from the blackboard system (**Figure 4**).

The advantage of blackboard architecture is highly secured for blockchain transaction. The blockchain technology has no third-party interference.

5. Fuzzy data mining

Sometimes, data mining is unable to deal with incomplete database and unable to combine the data and reasoning. Fuzzy data mining [6, 7, 9–18] will combine the data and reasoning by defining with fuzziness. The fuzzy MapReducing algorithms have two functions: *mapping* reads fuzzy datasets and *reducing* writes the after operations.

Definition: Given some universe of discourse X, a fuzzy set is defined as a pair $\{t, \mu_d(t)\}$, where t is tuples and d is domains and membership function $\mu_d(x)$ is taking values on the unit interval [0,1], i.e., $\mu_d(t) \rightarrow [0,1]$, where $t_i \in X$ is tuples (**Table 28**).

R1	$\mathbf{d_1}$	$\mathbf{d_2}$	.	$\mathbf{d_m}$	**μ**
t_1	a_{11}	a_{12}	.	a_{1m}	$\mu_d(t_1)$
t_2	a_{21}	a_{22}		A_{2m}	$\mu_d(t_2)$
.	.	.	.	.	.
t_n	a_{1n}	a_{1n}	.	A_{nm}	$\mu_d(t_n)$

Table 28.
Fuzzy dataset.

The sale is defined intermittently with fuzziness (**Tables 29–32**).

CNo	INo	IName	Demand
C001	I005	shirt	0.9
C001	I007	Dress	0.65
C003	I004	pants	0.85
C002	I007	dress	0.6
C001	I008	Jacket	0.65
C002	I005	shirt	0.9

Table 29.
Fuzzy demand.

$\mu_{\text{Demand}}(x)$=0.9/90+0.85/80+0.8/75+0.65/70
or
Fuzziness may be defined with function
$\mu_{\text{Demand}}(x)$= (1+(Demand-100)/100) $^{-1}$ Demand <=100
=1 Demand>100

A. *Negation*

CNo	INo	IName	Negation of price
C001	I005	shirt	0.3
C001	I007	Dress	0.5
C003	I004	pants	0.4
C002	I007	dress	0.5
C001	I008	Jacket	0.4
C002	I005	shirt	0.3

Table 30.
Negation of price.

A. *Union*

CNo	INo	IName	Sales U price
C001	I005	Shirt	0.8
C001	I007	Dress	0.5
C003	I004	Pants	0.6
C002	I007	Dress	0.5
C001	I008	Jacket	0.6
C002	I005	Shirt	0.7

Table 31.
Sales U price.

Union of 1105 = max{0.8,0.7}=0,8
Fuzzy semijoin is given by sales ⋈ items-sale as shown in **Table 33**.

INo	IName	Sales
I005	Shirt	0.8
I007	Dress	0.5
I004	Pants	0.6
I007	Dress	0.5
I008	Jacket	0.6

Table 32.
Items-sales.

CNo	INo	IName	Sales
C001	I005	shirt	0.8
C001	I007	Dress	0.5
C003	I004	pants	0.6
C002	I007	dress	0.5
C001	I008	Jacket	0.7
C002	I005	shirt	0.7

Table 33.
Fuzzy semijoin.

The fuzzy k-means clustering algorithm (FKCA) is optimization algorithm for fuzzy datasets (**Table 34**).

CNo	INo	IName	Sales
C001	I005⇔I007	Shirt⇔Dress	0.4
C003	I004	pants	0.6
C002	I007⇔I005	Dress⇔shirt	0.5

Table 34.
Association.

Fuzzy k-means cluster algorithm (FKAC) is given by, using FAD
best=R
K=means=best
for i range(1,n)
for j range(1,n)
t_i=fuzzy union(r_i.RU r_i.R_j), if r_i.R=r_j.R
C reduce best
k-means < best
return

The fuzzy multivalued association property of data mining may be defined with multivalued fuzzy functional dependency.

The fuzzy multivalued association (FMVD) is the multivalve dependency (MVD). The association multivalve dependency (FAMVD) may be defined by using Mamdani fuzzy conditional inference [3].

If EQ($t_1(X)$,$t_2(X)$,$t_3(X)$) then EQ(t_1(Y) ,$t_2(Y)$) or EQ($t_2(Y)$,$t_3(Y)$) or EQ($t_1(Y)$, $t_3(Y)$)

= min{EQ($t_1(Y)$,$t_2(Y)$) EQ(t_2(Y) ,t_3(Y)) EQ($t_1(Y)$,$t_3(Y)$)}

= min{min($t_1(Y)$,t_2(Y)) , min(t_2(Y) ,t_3(Y)) , min($t_1(Y)$,$t_3(Y)$)}

$= \min(t_1(Y), t_2(Y). t_3(Y))$

The fuzzy k-means clustering algorithm (FKCA) is the optimization algorithm for fuzzy datasets (**Table 35**).

CNo	INo	IName	Sales
C001	I005⇔I007 ⇔I008	Shirt⇔Dress ⇔Jacket	0.8 0.4 0.5
C003	I004	Pants	0.6
C002	I007⇔I005	Dress⇔shirt	0.5 0.7

Table 35.
Association using AFMVD.

Fuzzy k-means cluster algorithm (FKAC) is given by, using FAMVD

best=R

K=means=best

for i range(1,n)

for j range(1,n)

for k range(1,n)

t_i=fuzzy union($r_i.R$ U $r_j.R$ U $r_k.R$), if $r_i.R$=$r_j.R$=$r_k.R$

C reduce best

k-means<best

return

The fuzzy k-means clustering algorithm (FKCA) is the optimization algorithm for fuzzy datasets.

K=means=n

for i range(1,n)

for j range(1,n)

t_i=fuzzy union(r_i.R U $s_i.S_j$), if $r_i.R$=$s_j.S$

C =best

k-means < best

return

For example, consider the sorted fuzzy sets of **Table 5** is given in **Table 36**.

CNo	INo	IName	Sales ⋈ Price⋈ Demand
C001	I005	Shirt	0.8
C001	I007	Dress	0.5
C003	I004	Pants	0.6
C002	I007	Dress	0.5
C001	I008	Jacket	0.6
C002	I005	Shirt	0.7

Table 36.
Fuzzy join.

6. Fuzzy security for data mining

Security methods like encryption and decryption are used cryptographically. These security methods are not secured. Fuzzy security method is based on the mind and others do not descript. Zadeh [16] discussed about web intelligence, world knowledge, and fuzzy logic. The current programming is unable to deal question answering containing approximate information. For instance "which is the best car?" The fuzzy data mining with security is knowledge discovery process with data associated.

The fuzzy relational databases may be with fuzzy set theory. Fuzzy set theory is another approach to approximate information. The security may be provided by approximate information.

Definition: Given some universe of discourse X, a relational database $R1$ is defined as pair $\{t, d\}$, where t is tuple and d is domain (**Table 37**).

R1	$\mathbf{d_1}$	$\mathbf{d_2}$	.	$\mathbf{d_m}$
t_1	a_{11}	a_{12}	.	a_{1m}
t_2	a_{21}	a_{22}		A_{2m}
.	.	.	.	.
t_n	a_{1n}	a_{1n}	.	A_{nm}

Table 37.
Relational database.

Price = 0.4/50+0.5/60+07/80+0.8/100

The fuzzy security database of price is given in **Table 38**.

INo	**IName**	**Price**
I005	Benz	0.8
I007	Suzuki	0.4
I004	Toyota	0.7
I008	Skoda	0.5
I009	Benz	0.8

Table 38.
Price fuzzy set.

Demand = 0.4/50+0.5/60+0.7/80+0.8/100

The fuzzy security database of demand is given in **Table 39**.

INo	**IName**	**Demand**	**μ**
I005	Benz	80	0.7
I007	Suzuki	60	0.5
I004	Toyota	100	0.8
I008	Skoda	50	0.4
I009	Benz	80	0.7

Table 39.
Demand fuzzy set.

The lossless natural join of demand and price is union and is given in **Table 40**.

ino	Iname	Demand	price	μ
I005	Benz	80	100	0.8
I007	Suzuki	60	50	0.5
I004	Toyota	100	80	0.8
I008	Skoda	50	60	0.5
I009	Benz	80	100	0.8

Table 40.
Lossless join.

The actual data has to be disclosed for analysis on the web. There is no need to disclose the data if the data is inherently define with fuzziness.

"car with fuzziness >07" may defined as follows:

For instance,
XML data may be defined as
<CAR>
<COMPANY>
<NAME> Benz <NAME>
<FUZZ> 0.8 <FUZZ>
</COMPANY>
<COMPANY>
<NAME> Suzuki <NAME>
<FUZZ> 0.9<FUZZ>
</COMPANY>
<COMPANY>
<NAME> Toyoto<NAME>
<FUZZ> 0.6<FUZZ>
</COMPANY>
<COMPANY>
I<NAME> Skoda<NAME>
<FUZZ> 0.7<FUZZ>
</COMPANY>

Xquery may define using projection operator for demand car is given as
Name space default = http:\www.automoble.com/company
Validate <CAR> {
For $name in COMPANY/CAR
where $company/ Max($demand>0.7)}
return <COMPANY> {$company/name, $company/fuzzy}</COMPANY>
</CAR>
The fuzzy reasoning may be applied for fuzzy data mining.
Consider the more demand fuzzy database by decomposition (**Tables 41** and **42**).

The fuzzy reasoning [14] may be performed using Zadeh fuzzy conditional inference

The Zadeh [14] fuzzy conditional inference is given by
if x is P_1 and x is P_2 …. x is P_n then x is Q =
min 1, {1-min($\mu_{P1}(x)$, $\mu_{P2}(x)$, … , $\mu_{Pn}(x)$) +$\mu_Q(x)$}

INo	IName	Demand
I005	Benz	0.8
I007	Suzuki	0.9
I004	Toyota	0.6
I008	Skoda	0.7
I009	Benz	0.9

Table 41.
Demand.

INo	IName	Price
I005	Benz	0.7
I007	Suzuki	0.4
I004	Toyota	0.6
I008	Skoda	0.5
I009	Benz	0.7

Table 42.
Price.

The Mamdani [7] fuzzy conditional inference s given by
if x is P_1 and x is P_2 …. x is P_n then x is Q =
min $\{\mu_{P1}(x), \mu_{P2}(x), \ldots, \mu_{Pn}(x), \mu_Q(x)\}$
The Reddy [12] fuzzy conditional inference s given by
= $\min(\mu_{P1}(x), \mu_{P2}(x), \ldots, \mu_{Pn}(x))$
If x is Demand then x is price
x is more demand

---

x is more Demand o (Demand ➔ Price)
x is more Demand o min{1, 1-Demand+Price}Zadeh
x is more Demand o min{Demand, Price} Mamdani
x is more Demand o {Demand} Reddy
"If x is more demand, then x is more prices" is given in **Tables 43** and **44**.
The inference for price is given in **Table 45**.
So the business administrator (DA) can take decision to increase the price or not.

INo	IName	More demand
I005	Benz	0.89
I007	Suzuki	0.95
I004	Toyota	0.77
I008	Skoda	0.84
I009	Benz	0.95

Table 43.
More demand.

INo	IName	Zadeh	Mamdani	Reddy
I005	Benz	0.9	0.7	0.7
I007	Suzuki	0.5	0.4	0.4
I004	Toyota	1,0	0.6	0.6
I008	Skoda	0.8	0.5	0.5
I009	Benz	0.8	0.7	0.7

Table 44.
Demand ➔ Price.

INo	IName	Zadeh	Mamdani	Reddy
I005	Benz	0.89	0.7	0.7
I007	Suzuki	0.5	0.4	0.4
I004	Toyota	0.77	0.6	0.6
I008	Skoda	0.8	0.5	0.5
I009	Benz	0.8	0.7	0.7

Table 45.
Inference price.

7. Web intelligence and fuzzy data mining

Let C and D be the fuzzy rough sets (**Tables 46–51**).

	d_1	2_2	.	d_m	μ
t_1	a_{11}	a_{12}	.	a_{1m}	$\mu_d(t_1)$
t_2	a_{21}	a_{22}		A_{2m}	$\mu_d(t_2)$
.	.	.	.	.	.
t_n	a_{1n}	a_{1n}	.	A_{nm}	$\mu_d(t_n)$

Table 46.
Fuzzy database.

INo	IName	Price	μ
I005	Shirt	100	0.8
I007	Dress	50	0.4
I004	Pants	80	0.7
I008	Jacket	60	0.5
I009	Skirt	100	0.8

Table 47.
Price database.

The operations on fuzzy rough set type 2 are given as
1-C= 1- $\mu_C(x)$ Negation
CVD=max{$\mu_C(x)$, $\mu_D(x)$} Union
CΛD=min{$\mu_C(x)$, $\mu_D(x)$} Intersection

INo	IName	Demand	Price	μ
I005	Shirt	80	100	0.7
I007	Dress	60	50	0.4
I004	Pants	100	80	0.7
I008	Jacket	50	60	0.4
I009	Skirt	80	100	0.7

Table 48.
Intersect of demand and price.

INo	IName	Demand	μ
I005	Shirt	80	0.8
I007	Dress	60	0.5
I004	Pants	100	0.8
I008	Jacket	50	0.5
I009	Skirt	80	0.8

Table 49.
Lossless decomposition of demand.

INo	IName	Price	μ
I005	Shirt	100	0.8
I007	Dress	50	0.5
I004	Pants	80	0.8
I108	Jacket	60	0.5
I009	Skirt	100	0.8

Table 50.
Lossless decomposition of price.

Company	μ
IBM	0.8
Microsoft	0.9
Google	0.75

Table 51.
Best software company.

XML data may be defined as
<SOFTWARE>
<COMPANY>
<NAME> IBM <NAME>
<FUZZ> 0.8 <FUZZ>
</COMPANY>

<COMPANY>
<NAME> Microsoft <NAME>
<FUZZ> 0.9<FUZZ>
</COMPANY>
<COMPANY>
<NAME> Google<NAME>
<FUZZ> 0.75<FUZZ>
</COMPANY>

Xquery may define using projection operator for best software company is given as

Name space default = http:\www.software.cm/company

Validate <SOFTWARE> {For $name in COMPANY/SOFTWARE where $company/ Max($fuzz)}

return <COMPANY> {$company/name, $company/fuzzy} </COMPANY>

</SOFTWARE>

Similarly, the following problem may be considered for web programming.

Let P is the fuzzy proposition in question-answering system.

P=Which is tallest buildings City?

The answer is "x is the tallest buildings city."

For instance, the fuzzy set "most tallest buildings city" may defined as

most tallest buildings city = 0.6/Hoang-Kang + 0.6/Dubai + 0.7/New York +0.8/ Taipei+ 0.5/Tokyo

For the above question, output is "tallest buildings city"= 0.8/Taipei by using projection.

The fuzzy algorithm using FUZZYALGOL is given as follows:

BEGIN

Variable most tallest buildings City = 0.6 / Hoang-Kang + 0.6 / Dubai + 0.7 / New York + 0.8 / Taipei + 0.5 / Tokyo

most tallest buildings City =0.8 / Taipei

Return URL, fuzziness=Taipei, 0.8

END

The problem is to find "most pdf of type-2 in fuzzy sets"

The Fuzzy algorithm is

Go to most visited fuzzy set cites

Go to most visited fuzzy sets type-2

Go to most visited fuzzy sets type -2 pdf

The web programming gets "the most visited fuzzy sets" and put in order

The web programming than gets "the most visited type-2 in fuzzy sets"

The web programming gets "the most visited pdf in type-2"

8. Conclusion

Data mining may deal with incomplete information. Bayesian theory needs exponential complexity to combine data. Defining datasets with fuzziness inherently reduce complexity. In this chapter, fuzzy MapReduce algorithms are studied based on functional dependencies. The fuzzy k-means MapReduce algorithm is studied using fuzzy functional dependencies. Data mining and fuzzy data mining are discussed. A brief overview on the work on business intelligence is given as an example.

Most of the current web programming studies are unable to deal with incomplete information. In this chapter, the web intelligence system is discussed for fuzzy data mining. In addition, the fuzzy algorithmic language is discussed for design

fuzzy algorithms for data mining. Web intelligence system for data mining is discussed. Some examples are given for web intelligence and fuzzy data mining.

Acknowledgements

The author thanks the reviewer and editor for revision and review suggestions made in this work.

Author details

Poli Venkata Subba Reddy
Department of Computer Science and Engineering, Sri Venkateswara University, Tirupati, India

*Address all correspondence to: pvsreddy@hotmail.co.in

References

[1] Ghosh SP. File organization: The consecutive retrieval property. Communications of the ACM. 1972; **15**(9):802-808

[2] Chin FY. Effective Inference Control for Range SUM Queries, Theoretical Computer Science, 32,77-86. North-Holland; 1974

[3] Kamber M, Pei J. Data Mining: Concepts and Techniques. New Delhi: Morgan Kaufmann; 2006

[4] Ramakrishnan R, Gehrike J. Data Sets Management Systems. New Delhi: McGraw-Hill; 2003

[5] Tan PN, Steinbach V, Kumar V. Introduction to Data Mining. New Delhi: Addison-Wesle; 2006

[6] Zadeh LA. Fuzzy logic. In: IEEE Computer. 1988. pp. 83-92

[7] Tanaka K, Mizumoto M. Fuzzy programs and their executions. In: Zadeh LA, King-Sun FU, Tanaka K, Shimura M, editors. Fuzzy Sets and Their Applications to Cognitive and Decision Processes. New York: Academic Press; 1975. pp. 47-76

[8] Englemore R, Morgan T. Blackboard Systems. New Delhi: Addison-Wesley; 1988

[9] Poli VSR. On existence of C-R property. Proceedings of the Mathematical Society. 1989;**5**:167-171

[10] Venkta Subba Reddy P. Fuzzy MapReduce Data Mining Algorithms, 2018 International Conference on Fuzzy Theory and Its Applications (iFUZZY2018), November 14-17; 2108

[11] Reddy PVS, Babu MS. Some methods of reasoning for conditional propositions. Fuzzy Sets and Systems. 1992;**52**(3):229-250

[12] Venkata Subba Reddy P. Fuzzy data mining and web intelligence. In: International Conference on Fuzzy Theory and Its Applications (iFUZZY); 2015. pp. 74-79

[13] Reddy PVS. Fuzzy logic based on belief and disbelief membership functions. Fuzzy Information and Engineering. 2017;**9**(9):405-422

[14] Zadeh LA. A note on web intelligence, world knowledge and fuzzy logic. Data and Knowledge Engineering. 2004;**50**:91-304

[15] Zadeh LA. A note on web intelligence, world knowledge and fuzzy logic. Data and Knowledge Engineering. 2004;**50**:291-304

[16] Zadeh LA. Calculus of fuzzy restrictions. In: Zadeh LA, King-Sun FU, Tanaka K, Shimura M, editors. Fuzzy Sets and Their Applications to Cognitive and Decision Processes. New York: Academic Press; 1975. pp. 1-40

[17] Zadeh LA. Fuzzy algorithms. Information and Control. 1968;**12**: 94-104

[18] Zadeh LA. Precipitated Natural Language (PNL). AI Magazine. 2004; **25**(3):74-91