Fuzzy motivated data linkage for Self-Adaptive Systems

G. Roja Ramani, M. SrujanKumar Reddy

[1]PG Student, Department of Computer Science and Engineering, Madanapalle Institute of Technology & Science, India.

[2]Assistant Professor, Department of Computer Science and Engineering, Madanapalle Institute of Technology & Science, India.

Abstract: *Data-linkage is an extremely important task in several domains. Data-linkage means that, identifying the different data entries that are of same entity from different data sources. Joins (i.e., matches) related data entries or data sets which do not share a common identifier as it's goal. Data linkage commonly divided into two types. Namely One-to-One and One-to-many. In first, we relate one row of tableA with a single matching row of tableB. In case of second, we relate one row of A with many rows of B. In these scenarios data linkage is done across the same type of entities. It is necessary and important to match entities of different types too. For that, we are proposing a new Many-to-Many data-linkage scheme for entities of different types in self-adaptive systems using Fuzzy motivated data linkage. Each and every entity of A is weighted with each and every entity of B. New method uses, Clustering Tree technique and the tree characterizes that entities should be linked together. The tree can be easily transformed into association rules and understandable.*

Keywords*:* *Data-linkage, Many-to-Many linkage, fuzzy motivation, self adaptive systems*

I. INTRODUCTION

Data linkage is the process of identifying dissimilar data entries, which refers to the same entity among different data sources. The aim of this task is, joining the data sets which do not share a common identifier (i.e., a foreign key). Common data linkage scenarios include: linking data when combining the two different databases; data de-duplication, linking related DNA sequences and matching astronomical objects from different catalogues. Data linkage classified into two types: one-to-one and one-to-many. In an earlier, the goal is to link an entity from one data set with a single matching entity in another data set. In the later scenario, the goal is to link an entity from the first data set with a group of matching entities from the other data set.

In the proposed Fuzzy motivated linkage, the linkage takes place between the entities of different types(heterogeneous data sets) of entities. The relation between each and every entity of table A is weighted with each and every entity of table B using one-class clustering tree method developed by fuzzy logic. This method is different from the traditional decision tree method because, in clustering tree every leaf contains a cluster rather than a single classification and the inner nodes represents the features of the entities of first dataset. Every leaf of the tree represents features of the second set that are matching. Whereas decision trees used for regression tasks and classification. Training set which we used for inducing the tree must be labeled. But acquiring a labeled data set is an expensive task. Tree can be easily transformed into association rules and understandable. For inducing tree, we are using Prepruning method.

By an appropriate reorganizing themselves internally, the fuzzy control systems are able to function within a changing outside world. Fuzzy logic used by designers for developing sophisticated control systems are discovering support form associated technologies. By including adaptive control in their planning, fuzzy control systems are proficient to design the systems which can adjust to environmental changes - an important factor in wide array of applications.

Adaptive systems have capability to explain their reasoning and to learn, in additional they get capacity to be modified and extended. Due to the balance among explicit human knowledge and learned responsiveness capabilities, these make the systems extensible, very robust and suitable for solving a variety of troubles.

II. EXISTING SYSTEM

One-to-Many data linkage, an important task in several domains. Here linkage is performed between entities of same or different types using one-class clustering tree method, where in the linkage problem only two tables are involved. We relate one row of table A with many rows of table B. Here OCCT(one class clustering tree) is used for the induction of tree. But it is also very important to develop linkage techniques that link among matching entities of different types as well.

Drawbacks:

- ✓ Data linkage is performed between the entities of the identical or different types.
- ✓ Performs one-to-many data linkage only.

III. PROPOSED SYSTEM

We are propose a new Many-to-Many data linkage scheme using Fuzzy motivated data linkage, which links among the entities of different types. The proposed scheme is based on a clustering tree that characterizes the entities that should be linked together. Tree is constructed to understand and transform entities into association rules easily, i.e., the interior nodes consists of features describing the first dataset, and the tree leaves describes features of their related entities from the second data set. For inducing tree we are using four splitting methods and one pruning criteria i.e., Prepruning.

Fuzziness is an alternative to randomness. Randomness describes the uncertainty of event occurrences. It describes the event ambiguity and measures the degree to which an event can occur, but not about the event i.e., event occurs or not. If an event occurs, that is "random" and to which degree it occurs is “fuzzy”.

For this purpose we are using the real Weather Information data received from the revenue department as the data for the tree induction as training data and the corresponding disease cases(i.e., Malaria occurrence for corresponding weather conditions). As the weather conditions changes the occurrence of disease cases also changes i.e., adaptive nature. For the given data weights of each data item will be calculated so that the items which are having less threshold than expected threshold are pruned using Miner Algorithms and Fuzzy Membership Function given in modules.

Advantages:

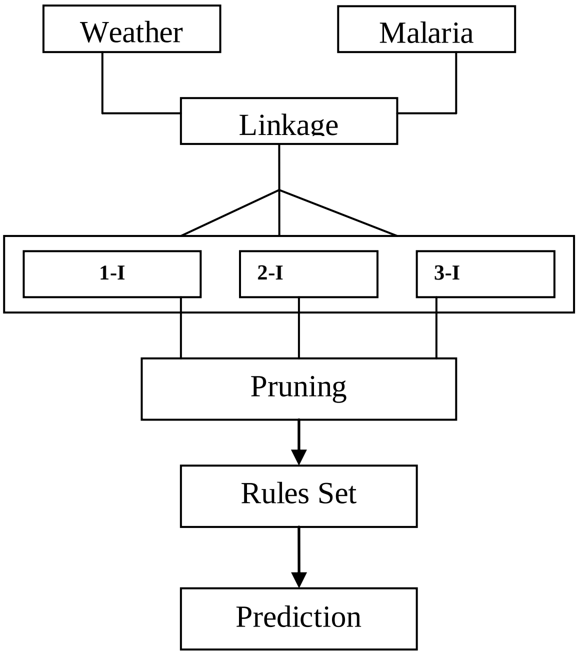
- Links between entities of different types
- More efficient and Scalable
- Convenience and Portability
- Flexibility
- Global Opportunities

Algorithm:

```
BEGIN
   1.Initialiaze Ftree
   2.Read Itemset
   3.Calculate WI-Support
   4.Insert all items into Ftree
   5.Join prefix items
   6.Identify prunable items
       7. If    FIWI-Support(i)>MST  then
          8.insert I to Ftree
             9.Else
                 10.prunable items

       11.Return tree
   12.End if
END
```

IV. ARCHITECTURE DIAGRAM



V. IMPLEMENTATION

Modules

1. Item Set Mining
2. Weighted Transaction Equivalence
3. The Infrequent Weighted Item set Miner Algorithm
4. The Minimal Infrequent Weighted Item set Miner Algorithm

Modules Description

1. Item Set Mining:

Item set mining is an exploratory data mining technique widely used for discovering valuable correlations among data. The first attempt to perform item set mining was focused on discovering frequent item sets, i.e., patterns whose observed frequency of occurrence in the source data (the support) is above a given threshold. Frequent Item sets find application in a number of real-life contexts (e.g., market basket analysis , medical image processing , biological data analysis). However, many traditional approaches ignore the influence/interest of each item/transaction within the analyzed data. To allow treating items/transactions differently based on their relevance in the frequent item set mining process, the notion of weighted item set has also been introduced . A weight is associated with each data item and characterizes its local significance within each transaction.

2. Weighted Transaction Equivalence

The weighted transaction equivalence establishes an association between a weighted transaction data set T, composed of transactions with arbitrarily weighted items within each transaction, and an equivalent data set TE in which each transaction is exclusively composed of equally weighted items. To this aim, each weighted transaction tq corresponds to an equivalent weighted transaction set, which is a subset of TE's transactions. Item weights in tq are spread, based on the irrelative significance, among their equivalent transactions in TEq. The proposed transformation is particularly suitable for compactly representing the original data.

3. The Infrequent Weighted Item set Miner Algorithm

A weighted transactional data set and a maximum IWI-support (IWI-support-min or IWI-support-max)threshold, the Infrequent Weighted Item set Miner algorithm extracts all IWIs whose IWI-support satisfies. Since the IWI Miner mining steps are the same by enforcing either IWI-support-min or IWI-support-max thresholds, we will not distinguish between the two IWI support measure types in the rest of this section. IWI Miner is a FP-growth-like mining algorithm that performs projection-based item set mining. Hence, it performs the main FP-growth mining steps:

(a) FP-tree creation and

(b) Recursive item set mining from the FP tree index. Unlike FP-Growth, IWI Miner discovers infrequent weighted item sets instead of frequent (unweighted) ones. To accomplish this task, the following main modifications with respect to FP-growth have been introduced:

(i) A novel pruning strategy for pruning part of the search space early and

(ii) A slightly modified FP tree structure, which allows storing the IWI-support value associated with each node.

4. *The Minimal Infrequent Weighted Item set Miner Algorithm*

A weighted transactional data set and a maximum IWI-support (IWI-support-min or IWI-support-max) threshold , the Minimal Infrequent Weighted Item set Miner algorithm extracts all the MIWIs that satisfy . The pseudo code of the MIWI Miner algorithm is similar to the one of IWI Miner. The MIWI Mining procedure is invoked instead of IWI Mining. The MIWI Mining procedure is similar to IWI Mining. However, since MIWI miner focuses on generating only minimal infrequent patterns, the recursive extraction in the MIWI Mining procedure is stopped as soon as an infrequent item set occurs. In fact, whenever an infrequent item set I is discovered, all its extensions are not minimal.

VI. RESULTS

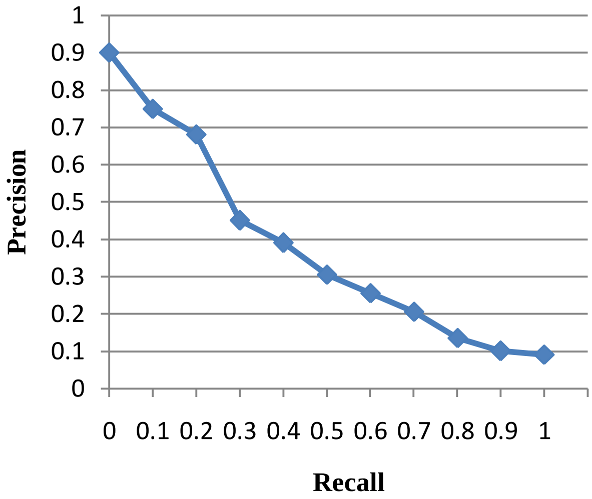
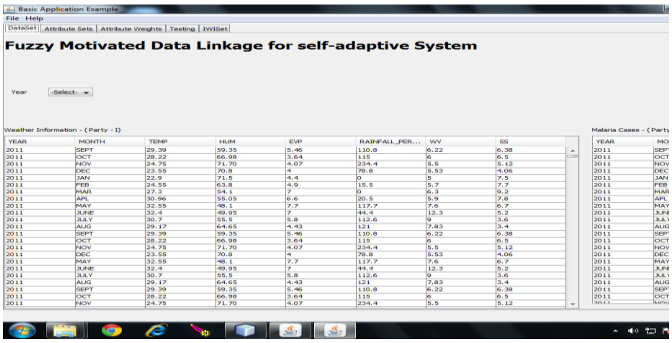


Fig 1 Precision and Recall graph for Fuzzy linkage in self-adaptive systems

Result analysis: As we know that precision and recall shows the best performance of a system. Hence, the above graph shows the best performance for self-adaptive systems by using fuzzy linkage when compared to traditional OCCT: One-to-Many data linkage method.


Basic Application Example
File Help
DataSet | Attribute Sets | Attribute Weights | Testing | IWISet

Fuzzy Motivated Data Linkage for self-adaptive System

Year -Select-

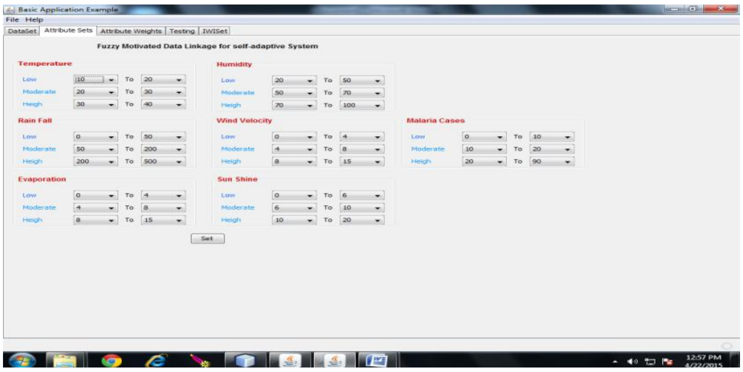
Weather Information - (Party - I)

YEAR	MONTH	TEMP	HUM	EVP	RAINFALL_PER...	WV	SS
2011	SEPT	29.39	59.35	5.46	110.8	6.22	6.38
2011	OCT	28.22	66.98	3.64	115	6	6.5
2011	NOV	24.75	71.70	4.07	234.4	5.5	5.12
2011	DEC	23.55	70.8	4	78.8	5.53	4.06
2011	JAN	22.9	71.5	4.4	0	5	7.5
2011	FEB	24.55	63.8	4.9	15.5	5.7	7.7
2011	MAR	27.3	54.1	7	0	6.3	9.2
2011	APL	30.96	55.05	6.6	20.5	5.9	7.8
2011	MAY	32.55	48.1	7.7	117.7	7.6	6.7
2011	JUNE	32.4	49.95	7	44.4	12.3	5.2
2011	JULY	30.7	55.5	5.8	112.6	9	3.6
2011	AUG	29.17	64.65	4.43	121	7.83	3.4
2011	SEPT	29.39	59.35	5.46	110.8	6.22	6.38
2011	OCT	28.22	66.98	3.64	115	6	6.5
2011	NOV	24.75	71.70	4.07	234.4	5.5	5.12
2011	DEC	23.55	70.8	4	78.8	5.53	4.06
2011	MAY	32.55	48.1	7.7	117.7	7.6	6.7
2011	JUNE	32.4	49.95	7	44.4	12.3	5.2
2011	JULY	30.7	55.5	5.8	112.6	9	3.6
2011	AUG	29.17	64.65	4.43	121	7.83	3.4
2011	SEPT	29.39	59.35	5.46	110.8	6.22	6.38
2011	OCT	28.22	66.98	3.64	115	6	6.5
2011	NOV	24.75	71.70	4.07	234.4	5.5	5.12

Malaria Cases - (Party

YEAR	MO
2011	SEP
2011	OCT
2011	NOV
2011	DEC
2011	JAN
2011	FEB
2011	MAR
2011	APL
2011	MAY
2011	JUN
2011	JUL
2011	AUG
2011	SEP
2011	OCT
2011	NOV
2011	DEC
2011	MAY
2011	JUN
2011	JUL
2011	AUG
2011	SEP
2011	OCT

Fig2. Reading data items


Basic Application Example
File Help
DataSet | Attribute Sets | Attribute Weights | Testing | IWISet
Set
12:57 PM
4/22/2015

Fuzzy Motivated Data Linkage for self-adaptive System

Temperature: Low 10 To 20; Moderate 20 To 30; Heigh 30 To 40

Humidity: Low 20 To 50; Moderate 50 To 70; Heigh 70 To 100

Rain Fall: Low 0 To 50; Moderate 50 To 200; Heigh 200 To 500

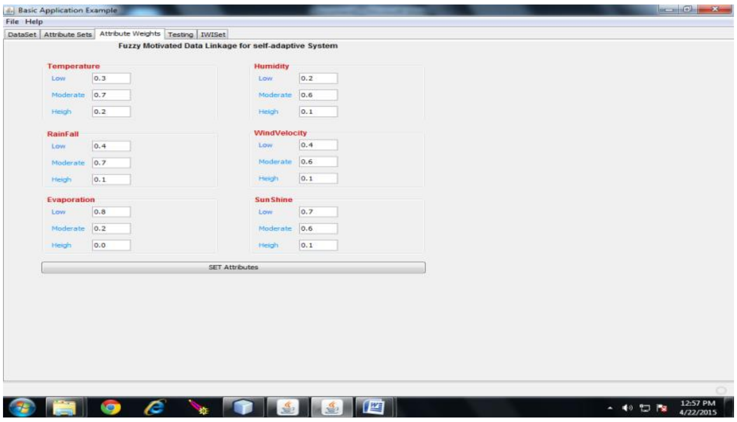
Wind Velocity: Low 0 To 4; Moderate 4 To 8; Heigh 8 To 15

Malaria Cases: Low 0 To 10; Moderate 10 To 20; Heigh 20 To 90

Evaporation: Low 0 To 4; Moderate 4 To 8; Heigh 8 To 15

Sun Shine: Low 0 To 6; Moderate 6 To 10; Heigh 10 To 20

Fig 3. Setting attributes for data items


Basic Application Example
File Help
DataSet | Attribute Sets | Attribute Weights | Testing | IWISet
SET Attributes
12:57 PM
4/22/2015

Fuzzy Motivated Data Linkage for self-adaptive System

Temperature: Low 0.3; Moderate 0.7; Heigh 0.2

Humidity: Low 0.2; Moderate 0.6; Heigh 0.1

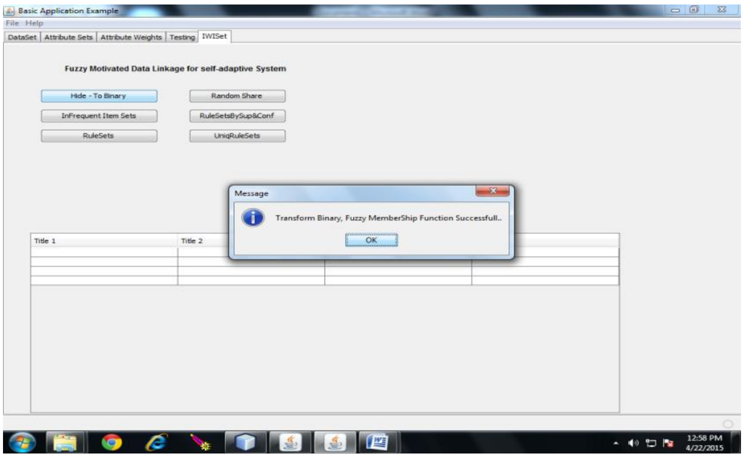
RainFall: Low 0.4; Moderate 0.7; Heigh 0.1

WindVelocity: Low 0.4; Moderate 0.6; Heigh 0.1

Evaporation: Low 0.8; Moderate 0.2; Heigh 0.0

SunShine: Low 0.7; Moderate 0.6; Heigh 0.1

Fig 4. After setting attributes weighted will be allocated to them

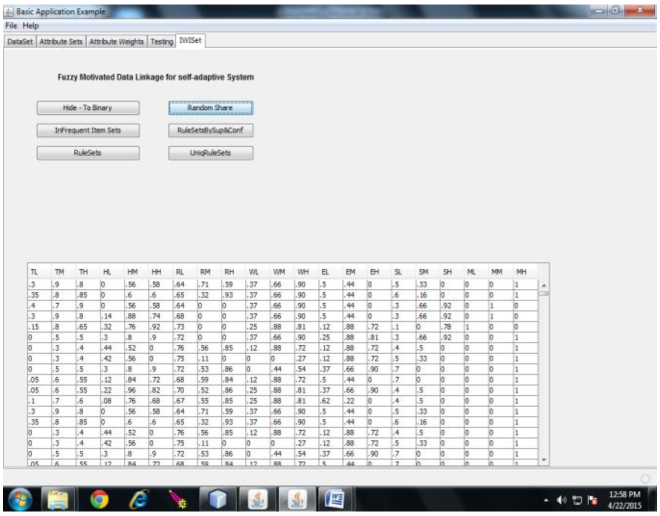


Fuzzy Motivated Data Linkage for self-adaptive System

Transform Binary, Fuzzy MemberShip Function Successfull..

Fig 5. Fuzzy function

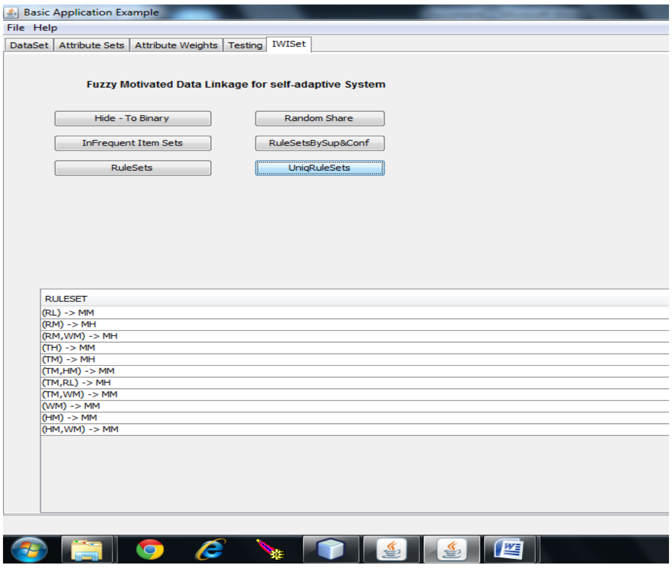
When the weights were set, the actual Fuzzy membership function works means that the weights randomly shared for many-to-many linkage among data items.



Fuzzy Motivated Data Linkage for self-adaptive System

TL	TM	TH	HL	HM	HH	RL	RM	RH	WL	WM	WH	EL	EM	EH	SL	SM	SH	ML	MM	MH
.3	.9	.8	0	.56	.58	.64	.71	.59	.37	.66	.90	.5	.44	0	.5	.33	0	0	0	1
.35	.8	.85	0	.6	.6	.65	.32	.93	.37	.66	.90	.5	.44	0	.6	.16	0	0	0	1
.4	.7	.9	0	.56	.58	.64	0	0	.37	.66	.90	.5	.44	0	.3	.66	.92	0	1	0
.3	.9	.8	.14	.88	.74	.68	0	0	.37	.66	.90	.5	.44	0	.3	.66	.92	0	1	0
.15	.8	.65	.32	.76	.92	.73	0	0	.25	.88	.81	.12	.88	.72	.1	0	.78	1	0	0
0	.5	.5	.3	.8	.9	.72	0	0	.37	.66	.90	.25	.88	.81	.3	.66	.92	0	0	1
0	.3	.4	.44	.52	0	.76	.56	.85	.12	.88	.72	.12	.88	.72	.4	.5	0	0	0	1
0	.3	.4	.42	.56	0	.75	.11	0	0	0	.27	.12	.88	.72	.5	.33	0	0	0	1
0	.5	.5	.3	.8	.9	.72	.53	.86	0	.44	.54	.37	.66	.90	.7	0	0	0	0	1
.05	.6	.55	.12	.84	.72	.68	.59	.84	.12	.88	.72	.5	.44	0	.7	0	0	0	0	1
.05	.6	.55	.22	.96	.82	.70	.52	.86	.25	.88	.81	.37	.66	.90	.4	.5	0	0	0	1
.1	.7	.6	.08	.76	.68	.67	.55	.85	.25	.88	.81	.62	.22	0	.4	.5	0	0	0	1
.3	.9	.8	0	.56	.58	.64	.71	.59	.37	.66	.90	.5	.44	0	.5	.33	0	0	0	1
.35	.8	.85	0	.6	.6	.65	.32	.93	.37	.66	.90	.5	.44	0	.6	.16	0	0	0	1
0	.3	.4	.44	.52	0	.76	.56	.85	.12	.88	.72	.12	.88	.72	.4	.5	0	0	0	1
0	.3	.4	.42	.56	0	.75	.11	0	0	0	.27	.12	.88	.72	.5	.33	0	0	0	1
0	.5	.5	.3	.8	.9	.72	.53	.86	0	.44	.54	.37	.66	.90	.7	0	0	0	0	1
.05	.6	.55	.12	.84	.72	.68	.59	.84	.12	.88	.72	.5	.44	0	.7	0	0	0	0	1

Fig 6. Random sharing of weights

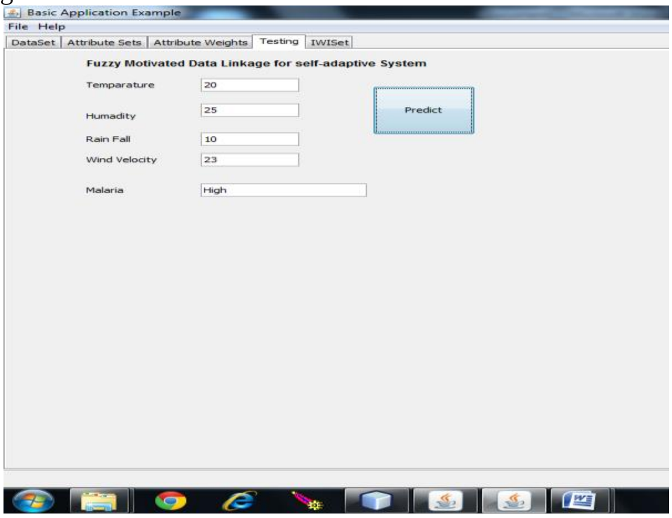


Fuzzy Motivated Data Linkage for self-adaptive System

RULESET
(RL) -> MM
(RM) -> MH
(RM,WM) -> MH
(TH) -> MM
(TM) -> MH
(TM,HM) -> MM
(TM,RL) -> MH
(TM,WM) -> MM
(WM) -> MM
(HM) -> MM
(HM,WM) -> MM

Fig 7. Unique rule sets

When the random share is done, then rules will be generated as per support & confidence. There after unique rules will be generated.



Fuzzy Motivated Data Linkage for self-adaptive System

Temparature 20
Humadity 25
Rain Fall 10
Wind Velocity 23
Malaria High

Fig 8. Figure showing final result

After generating unique rules, the testing takes place i.e., gives prediction results for disease occurrence may be like high/medium/low for the given weather information.

VII. CONCLUSION AND FUTURE WORK

We are providing Fuzzy mechanism for developing many to many data linkage in heterogeneous datasets using clustering tree for the induction of tree. The linkage takes place between the entities of different types(heterogeneous data sets) of datasets. The relation between each and every entity of A is weighted with each and every entity of B and linkage is done among entities by fuzzy logic. This method is different from the traditional decision tree method. On including adaptive control systems to this, we design the system which can adjust to environmental changes - an important factor in wide array of applications. And it improves efficiency of linkage capacity over different data sets.

In future, we are also trying to expand the scope of this system involving different environments.

VIII. REFERENCES

[1]. Ma'ayan Dror, Asaf Shabtai, Lior Rokach, and Yuval Elovici, "OCCT: A One-Class Clustering Tree for Implementing One-to-Many Data Linkage" IEEE tractions on knowledge and Data Engineering, vol. 26, no. 3, March 2014.

[2]. Hendrik Blockeel, Luc De Raedt**,** Jan Ramon, ""Top-Down Induction of Clustering Trees," ArXiv Computer Science e-prints, pp. 55-63, 1998.

[3]. M. Yakout, A.K. Elmagarmid, H. Elmeleegy, M. Quzzani, A. Qi, "Behavior Based Record Linkage," Proc. VLDB Endowment, vol. 3, nos. 1/2, pp. 439-448, 2010.

[4]. I.P. Fellegi and A.B. Sunter, "A Theory for Record Linkage," J. Am.Statistical Soc., vol. 64, no. 328, pp. 1183-1210, Dec. 1969.

[5]. A.J. Storkey, C.K.I. Williams, E. Taylor, and R.G. Mann, "An Expectation Maximisation Algorithm for One-to-Many Record Linkage," Univ. of Edinburgh Informatics Research Report, 2005.

[6]. Dick, S.; Dept. of Electr. & Comput. Eng., Univ. of Alberta, Edmonton, Canada "Towards complex fuzzy logic" IEEE tractions on Fuzzy Systems,Volume:13, Issue:3, June 2005.

[7]. Giannini, J.A. ; Appl. Res. Lab., Pennsylvania State Univ., University Park, PA, USA ; Kilgus, C. "A fuzzy logic technique for correcting climatological ionospheric models" IEEE transactions on Fuzzy Systems, (Volume:35, Issue:2), March 1997.