ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Efficient Data Linkage Using One Class Decision Tree.

Santha Sheela AC*.

Faculty of Computing, Sathyabama University, Chennai, Tamil Nadu, India.

ABSTRACT

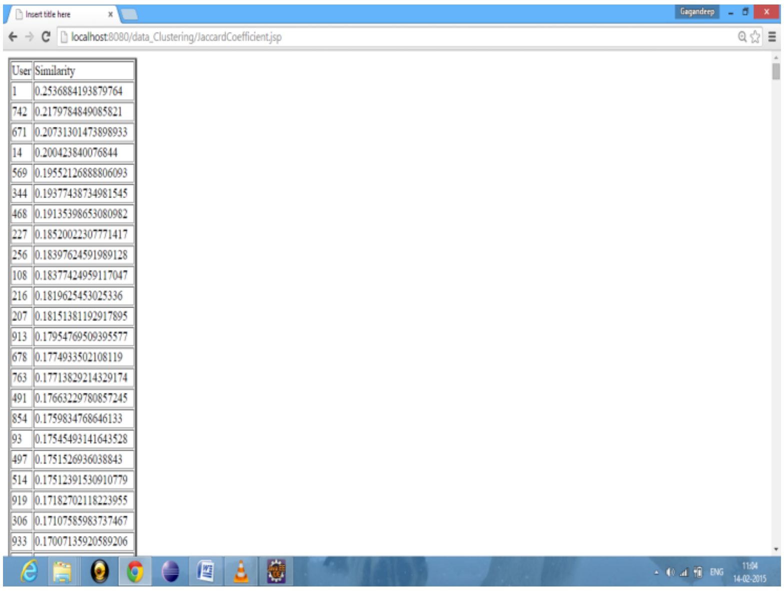
Record linkage is done between the entity of related category. It is done on the same or different entity which can share or not share the same entity. In this paper an proposed of new linkage method is been performs to make an linkage between matching entities of dissimilar data set. In this proposed system we use a one class decision tree that describes the entities which linked to each other. Decision trees are in tree shaped structure represents sets of dataset. The internal nodes which having the entities of the first tables and the terminal node having the entity of second table those are identical. The pruning technique is utilizes to devlop the decision tree.

Keywords: Record linkage, decision tree, Classification, pruning.

****Corresponding author***

INTRODUCTION

Record linkage is process of identifying different data sets that fit ins to the similar entity surrounded by altered data resource. Record linkage usually performed to shrink the large data into smaller data. . It helps to removing duplicate of records in the datasets. This technique is known as data de-duplication. Record linkage can be isolated into: one-to-one and one-to-many record linkage. In one-to-one record linkage, a solitary matching substance of an attribute from one set of record to the alternate dataset. In one to many records linkage, a record of one dataset has a set of matching record from an alternate dataset, illustrated in Fig1.



User	Similarity
1	0.2536884193879764
742	0.2179784849085821
671	0.20731301473898933
14	0.200423840076844
569	0.19552126888806093
344	0.19377438734981545
468	0.19135398653080982
227	0.18520022307771417
256	0.18397624591989128
108	0.18377424959117047
216	0.1819625453025336
207	0.18151381192917895
913	0.17954769509395577
678	0.1774933502108119
763	0.17713829214329174
491	0.17663229780857245
854	0.1759834768646133
93	0.17545493141643528
497	0.1751526936038843
514	0.17512391530910779
919	0.17182702118223955
306	0.17107585983737467
933	0.17007135920589206

Figure 1: summarize of record linkage process

In this paper, a new record linkage method is introduced which performs many-to-many record linkage. In many-to-many record linkage the dataset from different tables match with the entity of different tables. Decision tree employd for decide which records are similar to each-other.Decision trees are usually regarded as representing for classification. The leaves of the tree contain the classes and the branches from the root to a leaf contain sufficient conditions for classification.

RELATED WORK

The objective of methodology is a procedure that unions the conduct of two conceivable matched substances and registers to recognize conduct designs according to their matching score. The thought is that if a well matching conduct then undoubtedly element according to the score.

Mohamed Yak out, Ahmed K. [1] Here they make use of the support vector machine algorithm in two ways, first they calculating a probability value for the first records and then they try to matches with the other records.

Storkey [2] a decision tree is employd for checking wether the records are should match each other or un-match. Here they purposed different string evaluations methods are used and evaluated using decision trees.

Christen and Goosier [3]A way to grouping is exhibited that adjusts the fundamental top-down incitement of choice trees strategy towards bunching. To this point, its taking into account the standards of case based learning.

Hendricks Blockier, Luc Deraedt [4] This technique is actualized inthe TIC (Top down Induction of Clustering trees) framework for grouping. The TIC framework objective is the first request legitimate choice tree representation of the inductive rationale programming system.

Record linkage is the knowledge about matches or duplicates within or crosswise files. Records, identifiers might not match up closely. In the computer science literature, data cleaning regularly refers to methods of judgment duplicates. William E.Winkler. [5].

PROPOSED WORK

A decision tree is a flowchart shaped formation, the inside node consist of a set of attributes. The results of the test are symbolizes by the branch of the tree. Every terminal node haves a class label. The classification and generalize of data are very simple to understand. Generally Pruning is an essential task. The pruning method is divided in two types: pre-pruning and post-pruning. Here we majorly consider with the pre-pruning method that stop the growth of a tree if there are no appropriate split. The bottom-up methods used to conclude the branches are valuable or not in the post-pruning.

The major objective of the pre-pruning method is to minimize the time complexity. In this project we mainly used the pre-pruning approach.

ONE-TO-MANY DATA LINKAGE PROCESS

A characteristic record linkage trouble consists of two data tables that do not split an ordinary identifier. Consider two tables Table_1 and Table_2 as an example. The first table is viewer table and second one as movie table. Here A and B are taken as attribute for table T_1 and T_2. The goal is to match the records of the table T_1 and T_2. Generally here we define the same entity for the both table. Here each dataset of the both table match each other. here we proposed an many-to-many linkage model also. In many-to-many record linkage the dataset from different tables match with the entity of different tables.

Here the problem is define as $|T_A|$ x $|T_B|$. The advance indexing technique can used for an efficient linkage of records.T_{AB} is denoted for matching records and T_{AB} is denoted for non-matching records. The objective of the algorithm is to accurately recognize the true identical pairs (true positive) and identify non-identical pairs (False Positives).

Each possible records pair of T_1 is match with the matching records of T_2. Every probable records pair are allot a value that describe the probability of two identical tables. For Each class one probability value should be provides. A threshold value has to be declare from the starting. If the value surpasses the define threshold, the records measure as true match or link otherwise it declare as non-match.

APPLYING DECISION TREE FOR ONE-TO-MANY DATA LINKAGE

For identify the proper matched records we tested the each records in opposition to representation. We check whether it's a match pair or un-match pair. This method generates a probability values that demonstrating the matched pair. An early value is calculated using the MLE algorithm.

A one class decision tree used to match the data set of the two tables T_1 and T_2. The entity of the table T_1 and table T_2 may be varies from each-other. The internal nodes of the tree denote the attribute of table T_1. The leaves of the tree have a demonstration of the matched records from the both tables.

CLEANING AND STANDARDIZATION WITH DATA SETS

A datasets means that which want to merge as a combination and form a single table with efficient data sets for the effective mining. If consider in to two major parts means one is useful data extraction from the data warehouse or database. The another one is data arrangement in this concept we are majorly concentrate and achieve the data clustering of different datasets for the better and esfficient retrieval of data mining.

INDEXING OF RECORD DATA ADDING

This module of our application, in this module what to achieve means, record wise data adding of each and every table to mention table lists But before this module the pre requisite thing is need to give very clean and standardized data. Why to perform this module means   need to form a add table for efficient mining. In this module compare the one table to another table.

SIMILARITY VECTOR CLASSIFICATION

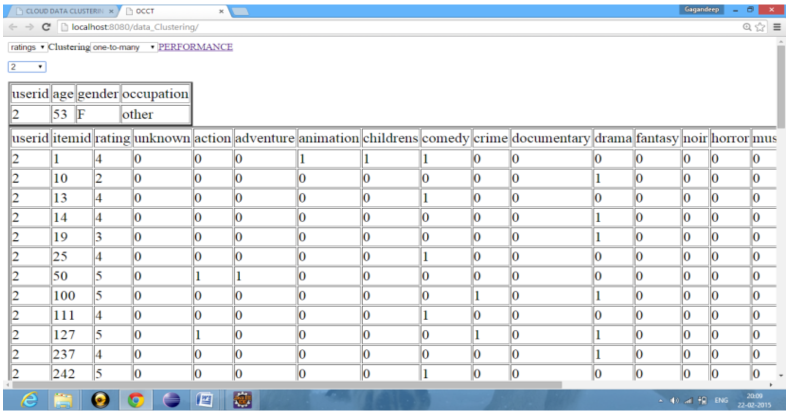
In this module similar data of the tables can consider for the indexing of their keywords and separate table which will yield the better results compared to the existing mechanisms. Because of this number of user queries which they are requesting data will reduce the time complexity when compare to existing work .The user given query should be verify with those tables and evolve results but now there is only one efficient table which can reduce the query response time.

MAXIMUM LIKELIHOOD ESTIMATION

This algorithm is used for choosing the suitable splitting methods. This splitting methods help to find out the suitable attributes that are not split until now. We choose those attributes those are having the uppermost value as compare to our threshold value. Those vlaue are having more than the thresold value that record is a exact match. Those value are having less than the threshold value that records are flase match. Its depends upon the time complexity for calculating the all the values for their respective attribute. . Its help to calculate the time complexity..Its help to calculate the early threshold value and the choose the next which feature going to be split.

JACCARD COEFFICIENT

Generally, the jaccard coefficient is utilized for grouping the similar type of data. It examines the similarity among the clusters. This coefficient is mainly applyed for select the splitting attribute. The main purpose of this method is finds out the likely similarity between the subsets. So we have to scan all the attributes to evaluate the similarity. The attribute those are created are not similar to each other. It is very expensive to examine the value of splitting attributes. It try to match the record of leaf node to the records of another leaf node. It examine as a standard of sequence binary splits. The objective of jaccard coefficient in Fig2, that generates subsets from the attribute, that are not similar to each other.



ratings Clustering one-to-many PERFORMANCE

2

userid	age	gender	occupation
2	53	F	other

userid	itemid	rating	unknown	action	adventure	animation	childrens	comedy	crime	documentary	drama	fantasy	noir	horror	mus
2	1	4	0	0	0	1	1	1	0	0	0	0	0	0	0
2	10	2	0	0	0	0	0	0	0	0	1	0	0	0	0
2	13	4	0	0	0	0	0	1	0	0	0	0	0	0	0
2	14	4	0	0	0	0	0	0	0	0	1	0	0	0	0
2	19	3	0	0	0	0	0	0	0	0	1	0	0	0	0
2	25	4	0	0	0	0	0	1	0	0	0	0	0	0	0
2	50	5	0	1	1	0	0	0	0	0	0	0	0	0	0
2	100	5	0	0	0	0	0	0	1	0	1	0	0	0	0
2	111	4	0	0	0	0	0	1	0	0	0	0	0	0	0
2	127	5	0	1	0	0	0	0	1	0	1	0	0	0	0
2	237	4	0	0	0	0	0	0	0	0	1	0	0	0	0
2	242	5	0	0	0	0	0	1	0	0	0	0	0	0	0

Figure 2: Jaccard coefficient similarity.

IMPLEMENTION

Consider the two table, Here Table1 consider as viewer table and Table2 consider as movie table. With the help of decision tree and MLE linkage will happen. Consider one class decision tree because smaller tree better generalizes the data. For identify the proper matched records test the each records in opposition to linkage representation. This method generates an early probability value for the true match pairs. The above algorithm is used for calculating the probabiltics value. For the movie table generate a probability score and set the threshold value. In Fig 3, As per the threshold value we compare which table is match. if it exceed from the threshold value then it’s a true match or if it not exceed then it’s a false match.

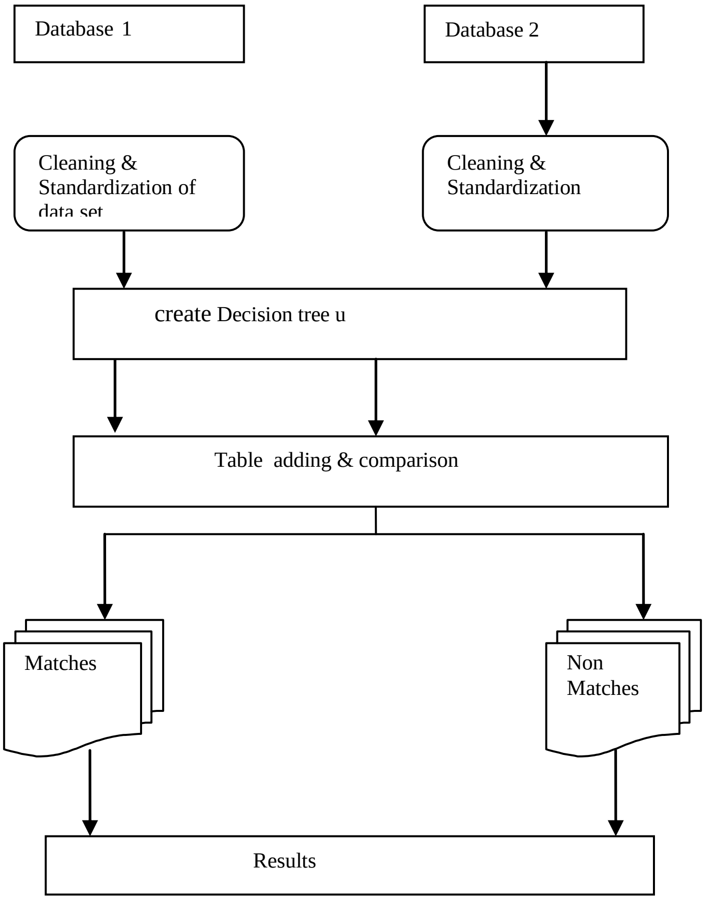

Database 1
Database 2
Cleaning & Standardization of data set
Cleaning & Standardization
create Decision tree u
Table adding & comparison
Matches
Non Matches
Results

Table 1: Viewer table

VIewer ID	Age	Gender	Occupation
1	23	Male	Technician
2	45	Female	Critics
3	35	Male	Actor

Table 2: Movie table

Movie ID	Action	thriller	Comedy	Romance
1	1	0	1	0
2	0	1	1	1
3	0	1	1	0
4	1	1	0	1
5	1	0	0	1

RESULTS

The first step it recognize the appropriate surrounding where we examines the records. The objective of our methodology is a procedure that unions the conduct of two conceivable matched substances and registers to recognize conduct designs according to their matching score.Then further for records linkage process we test and excutes in different area for getting the better resulst set.

The user given query should be verify with those tables and evolve results but now there is only one efficient table which can reduce the query response time. On the basis of the threshold values varys we observe the identical pairs and non-identical pairs.

CONCLUSION AND FUTURE WORK

In this project represented a one class decision tree which did the job for record linkage. Here we used one class decision tree because smaller tree better generalize the data and its gives idea about the records which are connected each other.

This method allows performing one-to-many linkage while the established methods follow one-to-one linkage. Then, we have used a one-class approach which results are better generalizes matching pairs and as more number of pairs those are un-matched will confound the model and it will not be the exact model. . The thought is that if a well matching conduct then undoubtedly element according to the score.

REFERENCES

[1] Dror M, Shabtai A, Rokach L, Elovici Y. "OCCT: A One- Class Clustering Tree for Implementing One-to- Many Data Linkage," IEEE Trans. on Knowledge and Data Engineering, 2013.

[2] Yakout M, Elmagarmid AK, Elmeleegy H, Quzzani M, and Qi A. "Behavior Based Record Linkage," Proceedings of the VLDB Endowment, 2010;3 (1).

[3] Storkey AJ, Williams CKI, Taylorand E Mann RG. "An Expectation Maximisation Algorithm for One-to-Many Record Linkage, University of Edinburgh Informatics Research Report ,2005.

[4] Ivie S, Henry G, Gatrell H and Giraud-Carrier C. "A Metric Based Machine Learning Approach to Genea- Logical Record Linkage," Proc. Seventh Ann. Workshop Technology for Family History and Genealogical Research, 2007

[5] Christen P and Goiser K. "Towards Automated Data Linkage and Deduplication," technical report, Australian Nat'l University, 2005.

[6] Langley P. Elements of Machine Learning, San Franc-Isco, Morgan Kaufmann, 1996.

[7] Guha S, Rastogi R and Shim K. "Rock: A Robust Clustering Algorithm for Categorical Attributes," Information Systems, 2000; 25(5):345-366.