

Efficient Data Linkage Technique Using One Class Clustering Tree For Database Misuse Domain

Mandilkar Kalyani¹, Prof.Panhalkar A.R²

¹Student MEIT, Dept. of Information Technology, Amrutvahini college of engineering, Sangamner, Maharashtra.

²Professor, MEIT, Dept. of Information Technology, Amrutvahini college of engineering, Sangamner, Maharashtra.

Abstract - The task of data linkage is to join datasets that do not share a common identifier. The process of Data Linkage is implemented with entities which are of same and different types. One-to-many data Linkage is an necessary Data Model. Two types of data linkages techniques are one-to-one and one-to-many. In this paper, we propose a one-to-many data linkage. it is used to accomplish link between different types in the matching sets. This present method is depends on One-Class Clustering Tree (OCCT). The tree is constructed in such a way that it is easily understandable and can be transformed into association rules. The inner nodes of the tree include features of the first set of entities. The leaves of the tree appear as features of the second set that are matching. The data is split using four splitting criteria, and two pruning methods are used. The result shows that OCCT gives better performances in terms of precision and recall. A threshold is calculated which plays a main role in the categorization of the users. A pruning technique is used to reduce the tree by avoiding the unnecessary branches. Thus, the proposed system is enhanced in terms of time complexity and accuracy.

Key Words: Pruning, Splitting criteria, Record Linkage

1. INTRODUCTION

There is a need to found the techniques to link the datasets that does not share the common identity; it comes under the data linkage process. Data linkage is used when information from two or more records of independent sources are brought together, when they are perceived to belong to the same individual, family, event or place. It is the technique that is used to connect the information across several disparate data sources. The major task of data linkage is to identify the different objects which were used to refer the same entity across the different source of data.

The data linkage can be divided into two types: one to one and one to many. The goal is to associate an entity from one data set with a single matching entity in another data set is in one to one linkage. Most of the previous works focus on one-to-one data linkage. we propose a new data linkage method aimed at performing one to many linkages. The proposed data linkage technique can match entities of

different types, while data linkage is usually performed among entities of the same type.[1]

In this paper, we propose a new data linkage method one-to-many linkage that can match entities of different types. For example, Let TA and TB are the two tables of different types. The inner nodes of the tree consist of attributes referring to both of the tables being matched (TA and TB). The leaves of the tree will determine whether a pair of records described by the path in the tree ending with the current leaf is a match or a nonmatch. We propose a new data linkage method aimed at performing one-to-many linkage. In addition, while data linkage is usually performed among entities of the same type, this proposed data linkage technique can match entities of different types. For example, if there is a student database we might want to link a student record with the courses she should take (according to different features that describe the student and features describing the courses). The proposed method links between the entities using a One-Class Clustering Tree (OCCT). A clustering tree is a tree in which each of the leaves contains a cluster instead of a single classification and each cluster is generalized by a set of rules that is stored in the appropriate leaf [9], [10].

Detecting data misuse poses a great challenge for organizations. Whether caused by malicious intent or an inadvertent mistake, data leakage/misuse can diminish a company's brand, reduce shareholder value, and damage the company's goodwill and reputation. This challenge is intensified when trying to detect and/or prevent data leakage/misuse performed by an insider with legitimate permissions to access the organization's systems and its critical data.

The goal of using the OCCT in this domain is to link a set of records, representing the context of the request (i.e., the actual access to certain data), with a set of records representing the data that can be legitimately retrieved within the specific context. Thus, the inner nodes of the OCCT represent contextual attributes in which the request occurs (table TA), and the probabilistic models in the leaves represent the data that can be legitimately retrieved in the specific context (table TB). Pairs which are detected as nonmatching are assumed to be illegitimate (i.e., malicious), and will trigger an alert to the organization's security officer. By analyzing both the context of the request as well as the

data that the user is exposed to, the OCCT method can improve the detection accuracy and better distinguish between a normal and abnormal request

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 employed 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.

The proposed method is implemented using the database misuse domain used to identify the common and malicious users. In this domain, the goal is to identify anomalous access to database records that may indicate a possible data leakage, loss of data integrity or data mishandling.

2. RELATED WORK

Record linkage is the process of matching entities from two different data sources that may or may not contain a common element. Shabtai et al. [1] used one-to-many linkage in different domains like fraud detection, recommender systems and data leakage prevention by developing a One-Class Clustering Tree. Four splitting criteria and pruning methods are implemented to construct the tree and to avoid the needless branches of the tree correspondingly. The shortcoming of this approach is that it is difficult to reduce the linkage computation time and it is a one-class approach.

Christen and Goiser [2] used a C4.5 decision tree to determine which records must be matched to one another. In their work, various string comparison methods are used and compared by constructing different decision trees. However, their method performs the matching of attributes that are only predefined. Moreover only one or two attributes are usually used.

Henry et al. [3] used one-to-many linkage for genealogical research. Record linkage was performed using five attributes: name of the person, birth date, place, gender and the relationships between the persons. Using these five attributes a decision tree was induced. The drawback of this approach is that it performs matching using the specific attributes and therefore it is very hard to generalize.

F.De Comite, F.Denis, R. Gilleron and F.Letouzey introduced POSC4.5 algorithm for record linkage of positive and unlabeled examples. They considered binary classification and hence this method is not generalized. They require

not only the data set but also the information of positive examples out of whole data set. The attraction of their work is that they presented modified entropy formula which that considers weight of positive examples in a given data set.

They assumed that negative examples are in unlabeled data set as per given distribution [4].

[3] PROPOSED SYSTEM

A. Problem Statement

The aim of this paper is to link a record from a table TA with records from another table TB. The generated model is in the form of a tree in which the inner nodes represent attributes from TA and the leafs hold a compact representation of a subset of records from TB.

B. Feature

The OCCT yields better performance in terms of precision and recall. It will improve the efficiency of classifiers by improvement in the pre and post processing techniques.

The simplicity of the model, which can easily be transformed to rules of the type A \rightarrow B.

The prepruning approach which is used to reduce the time complexity of the algorithm.

C. Scope

The OCCT can used in different domains like:

FraudDetection- To find the fraudulent users. In Recommender systems the proposed system can be used for matching new users with their product expectations. In Data leakage prevention.- to detect the abnormal access to the database records that indicates data leakage or data misuse.

The process flow of the proposed system is as follows: First consider the customer dataset and the context dataset for linkage process. Then preprocess both the data by removing the null values and error data. After preprocessing, create a training table based on both the context and customer dataset. For training table creation, consider the attributes such as customer id, requesting location, requesting day, requesting time from the first database and in the other table consider the attributes such as user location and the business type.

Then measure the similarity score based on the attributes of requesting location, requesting day, requesting time with user location and the business type. Based on that, construct the tree by applying the splitting criteria. Then apply the post pruning method to remove the unnecessary branches that are not used in the classification process. Then apply the SEMANC method which is used to analyze the normal user or the malicious user from the linked data. At last, the

performance graph is drawn based on both the existing and the proposed system.

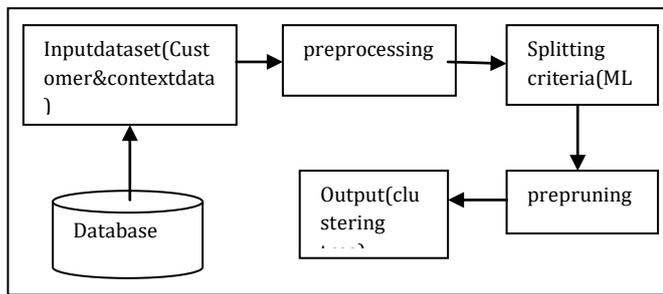


Figure 1:- Architecture diagram

4. SPLITTING CRITERIA

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 value are having more than the threshold value that record is a exact match. Those value are having less than the threshold value that records are false 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.

The domain that is going to be implemented is database misuse domain. Then the attributes are split up by the similarity values between the clusters. Finally tree will be constructed. The internal nodes of the tree consist of attributes referring to both of the tables being matched (TA and TB). The leaves of the tree will determine whether a pair of records described by the path in the tree ending with the current leaf is a match or a non-match.

5. PRUNING

Pruning is considered to be an important activity during the tree construction process. The necessity of pruning is to build a tree with accuracy and also to avoid over fitting. Pruning can be classified into two types: pre-pruning and post-pruning. In pre-pruning, the branches are pruned during the induction process if there are no possible splits found. In post-pruning, the tree is built completely followed by a bottom-up approach to determine which branches are not beneficial. Pruning is the process to trim unnecessary branches to improve accuracy of model. Thus tree is induced using matching examples only. The pruning process is use to give compact representation of tree i.e. it contains small

number of attributes. It also avoids over fitting and improve the time complexity.

In our proposed system we are using pre-pruning process to reduce the time complexity. The decision whether to prune the branch taken once best splitting attribute is chosen. We propose MLE for our system. In Maximum Likelihood Estimation (MLE), a MLE score is calculated for each of splitting attribute. If none of the candidate attribute achieve MLE score greater than current node then branch is pruned and current node becomes leaf node.

6. IMPLEMENTATION DETAILS

6.1 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 Table1 and Table2 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 T1 and T2. The goal is to match the records of the table T1 and T2. 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 $|TA| \times |TB|$. The advance indexing technique can used for an efficient linkage of records. TAB is denoted for matching records and TAB 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 T1 is match with the matching records of T2. 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.

For implementation purpose we create customer database of an organization are shared with a business partner. Therefore, the simulated data include requests for customer records of an organization ,submitted by a business partner of the organization.

6.2 THE PROCESS FLOW OF THE PROPOSED SYSTEM:-

- First the user dataset and the server log dataset for linkage process is taken into account.
- Then preprocess both the data by removing the null values or missing values and those attributes that have error data.
- After preprocessing, create a training table based on both the server log and user dataset. For training

table creation, consider the elements such as requesting location, requesting day, requesting time from the first database and in the other table consider the elements such as user location and the business type.

- Then measure the similarity score based on the elements of requesting location, requesting day, requesting time with user location and the business type by using the Maximum likelihood estimation technique.
- Then apply the pre pruning method to remove the unnecessary branches that are not used
- A total probability value (sum all possible values of an attribute) is calculated which is compared with the probability calculated for each value of an attribute and that is used in the classification of the users as the common user and the malicious users.
- At last, the performance of the existing and the proposed system is evaluated in terms of execution time and the accuracy of the result.

6.3 PREPARING THE INPUT

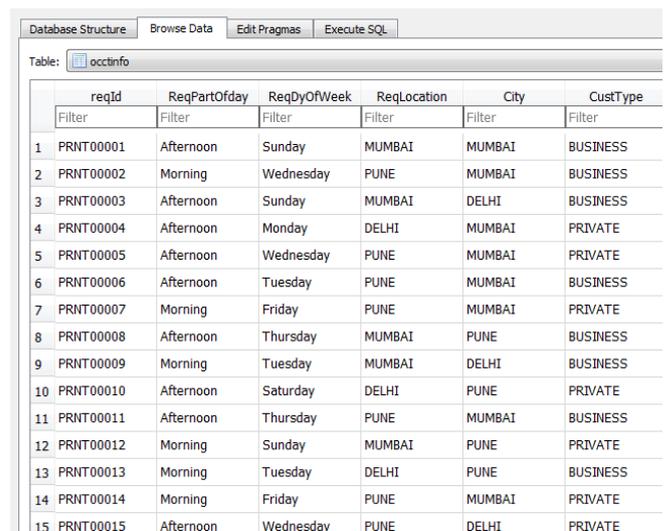
The training and also the testing input data for the OCCT algorithm should be a single dataset TAB which contains only matching instances from $TA \subseteq TB$. Each record, r , of the dataset is actually a pair of records $(r(a), r(b))$, one from the table TA and one from the table TB, so that: $r = (r(a); r(b)) \in TAB \subseteq TA \times TB$. In the current version of the implementation, due to circumstances, we require all the attributes of TA to appear in the input dataset, followed by all the attributes of TB. By this way, the attributes of TB (City) can be pointed by a single index which refers to the first attribute of TB in the input dataset.

Given a record $a \rightarrow TA$, a leaf of the tree is extracted by following the values of the record a to the correct path of the tree. Each leaf of the tree represents a set of records from the second table (TB) which are the most likely to be linked with record a . In order to represent this set in a compact way and avoid overfitting, a set of probabilistic models is induced for each leaf. Each model is used for deriving the probability of a value of some attribute from table TB, given all other attributes from that table.

In our implementation there is no need to save models for all possible attributes of TB. Thus, a feature selection process is executed on the leaf dataset in order to choose the attributes that will be represented by the leaves. In our implementation it needs only examples of matching pairs in order to learn and build the model. So this leads to much more accuracy as only matching pairs are train so it accurately choose matching pairs. This feature is important since in many domains it is difficult to obtain nonmatching examples.

6.4 RESULT AND DISSCUSSION

- First consider the customer dataset and the context dataset for linkage process. Then preprocess both the data by removing the null values and error data. After preprocessing, create a training table based on both the context and customer dataset. For training table creation, consider the attributes such as customer id, requesting location, requesting day, requesting time from the first database and in the other table consider the attributes such as user location and the business type.



	reqId	ReqPartOfDay	ReqDyOfWeek	ReqLocation	City	CustType
1	PRNT00001	Afternoon	Sunday	MUMBAI	MUMBAI	BUSINESS
2	PRNT00002	Morning	Wednesday	PUNE	MUMBAI	BUSINESS
3	PRNT00003	Afternoon	Sunday	MUMBAI	DELHI	BUSINESS
4	PRNT00004	Afternoon	Monday	DELHI	MUMBAI	PRIVATE
5	PRNT00005	Afternoon	Wednesday	PUNE	MUMBAI	PRIVATE
6	PRNT00006	Afternoon	Tuesday	PUNE	MUMBAI	BUSINESS
7	PRNT00007	Morning	Friday	PUNE	MUMBAI	PRIVATE
8	PRNT00008	Afternoon	Thursday	MUMBAI	PUNE	BUSINESS
9	PRNT00009	Morning	Tuesday	MUMBAI	DELHI	BUSINESS
10	PRNT00010	Afternoon	Saturday	DELHI	PUNE	PRIVATE
11	PRNT00011	Afternoon	Thursday	PUNE	MUMBAI	BUSINESS
12	PRNT00012	Morning	Sunday	MUMBAI	PUNE	PRIVATE
13	PRNT00013	Morning	Tuesday	DELHI	PUNE	BUSINESS
14	PRNT00014	Morning	Friday	PUNE	MUMBAI	PRIVATE
15	PRNT00015	Afternoon	Wednesday	PUNE	DELHI	PRIVATE

Fig 2:- Customer dataset for database misuse domain

- After applying splitting criteria, it gives result of four splitting methods. For implementation we consider MLE as a splitting method.

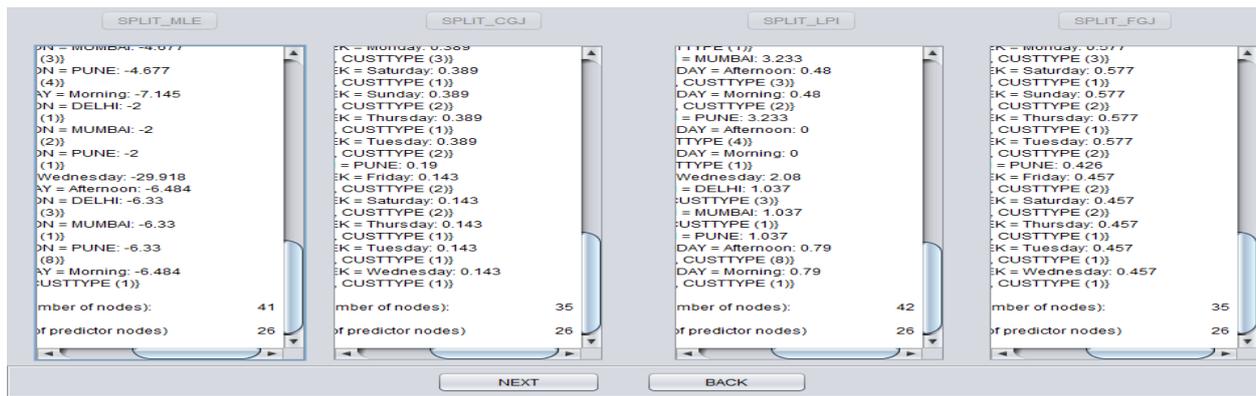


Fig 3:-Using different splitting criteria result shows the tree size and leaves for dataset database misuse

III. Now result of pruning method.we use prepruning approach.

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", iccc transactions on knowledge and data engineering, vol. 26, no. 3, march 2014.
- [2] 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.
- [3] m. yakout, a.k. elmagarmid, h. elmeleegy, m. quzzani, and a. qi, "behavior based record linkage", proc. vldb endowment, vol. 3, nos. 1/2, pp. 439-448, 2010.
- [4] f. de comite', f. denis, r. gilleron, and f. letouzey, "positive and unlabeled examples help learning", proc. 10th int'l conf. algorithmic learning theory, pp. 219-230, 1999.
- [5] a.j.storkey, c.k.i.williams, e.taylorandr.g.mann"an expectation maximisation algorithm for one-to-many record linkage," university of edinburgh informatics research report, 2005

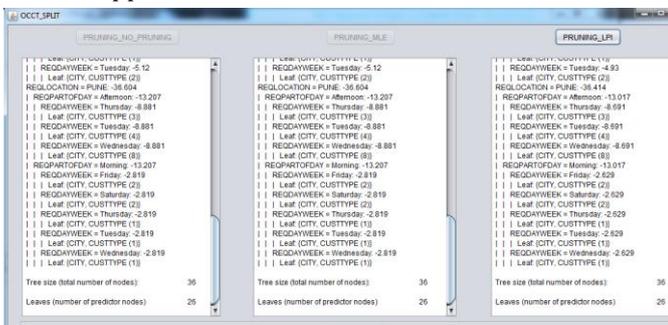


Fig 4:- After applying pruning method it shows tree size and leaves

IV. Precision and Recall value comparison for different methods

Result Evaluation						
	PRECISION_NO_PRUN.	PRECISION_LPI	PRECISION_MLE	RECALL_LPI	RECALL_MLE	RECALL_NO_PRUN.
FGJ	0.87234	0.90791	0.87126	0.89042	0.89046	0.8906
CGJ	0.87146	0.90814	0.87126	0.8895	0.89977	0.89015
LPI	0.87059	0.90909	0.87234	0.87529	0.89448	0.89041
MLE	0.87077	0.90785	0.87116	0.88943	0.88992	0.89046

Fig 5: Precision and Recall for Database Misuse domain

- [6] S.Ivie, G.Henry, H.Gatrell and C.Giraud-Carrier, "A Metric Based Machine Learning Approach to Genealogical Record Linkage," in Proc. of the 7th Annual Workshop on Technology for Family History and Genealogical Research, 2007.
- [7] P.Christen and K.Goiser, "Towards Automated Data Linkage and Deduplication," Australian National University, Technical Report, 2005.
- [8] A. Gershman et al., "A Decision Tree Based Recommender System," Proc. 10th Int'l Conf. Innovative Internet Community Services, pp. 170-179, 2010.

- [9] L. Gu and R. Baxter, "Decision Models for RecordLinkage," *DataMining*, vol. 3755, pp. 146-160, 2006.
- [10] O. Benjelloun, H. Garcia, D. Menestrina, Q. Su, S. Whang, and J. Widom, "Swoosh: A Generic Approach to Entity Resolution," *TheVLDB J.*, vol. 18, no. 1, pp. 255-276, 2009.
- [11] G. Henry, S. Ivie, H. Gatrell and C. Giraud-Carrier, "A Metric Based Machine Learning Approach to Genealogical Record Linkage," in *Proc. of the 7th Annual Workshop on Technology for Family History and Genealogical Research*, 2007.