Automated One-to-Many Data Linkage
Shahid Anwar¹ Prof. Deepak Kagpate²
¹,²G.H.Raisoni Academy College of Engineering, Nagpur, India

Abstract—One-to-many information linkage is an important task in several domains nevertheless solely a couple of previous publications have addressed this issue. Moreover, whereas historically data linkage is performed among entities of constant sort, it's very necessary to develop linkage techniques that link between matching entities of various varieties further. We have a tendency to propose a replacement one-to-many data linkage technique that links between entities of various natures. The projected technique is predicated on a one-class clustering tree (OCCT) that characterizes the entities that ought to be joined along. The tree is made such it's simple to know and remodel into association rules, i.e., the inner nodes consist solely of options describing the primary set of entities, whereas the leaves of the tree represent features of their matching entities from the second data set. We have a tendency to propose four splitting criteria and two totally different pruning ways that can be used for causation the OCCT. The strategy was evaluated mistreatment data sets from three totally different domains. The results affirm the effectiveness of the projected technique and show that the OCCT yields higher performance.

Key words: Clustering, Data Linkage, Data Matching

I. INTRODUCTION

Common data linkage scenarios include: linking data when combining two different databases, data deduplication (a data compression technique for eliminating redundant data), which is commonly done as a preprocessing step for data mining tasks identifying individuals across different census data sets, linking similar DNA sequences and matching astronomical objects from different catalogues. It is common to divide data linkage into two types: one-to-one and one-to-many. In one-to-one data linkage, the goal is to associate an entity from one data set with a single matching entity in another data set. In one-to-many data linkage, the goal is to associate an entity from the first data set with a group of matching entities from the other data set. Most of the previous works focus on one-to-one data linkage. The proposed method is based on a one-class clustering tree (OCCT) that characterizes the entities that should be linked together. The tree is built such that it is easy to understand and transform into association rules, i.e., the inner nodes consist only of features describing the first set of entities, while the leaves of the tree represent features of their matching entities from the second data set. We propose four splitting criteria and two different pruning methods which can be used for inducing the OCCT. The method was evaluated using data sets from three different domains.

The OCCT was evaluated using data sets from three different domains: data leakage prevention, recommender systems, and fraud detection. In the data leakage prevention domain, the goal is to detect abnormal access to database records that might indicate a potential data leakage or data misuse. The goal is to match an action, performed by a user within a specific context, with records that can be legitimately retrieved within that context. In the recommender systems domain, the proposed method is used for matching new users of the system with the items that they are expected to like based on their demographic attributes. In the fraud detection domain, the goal is to identify online purchase transactions that are executed by a fraudulent user and not the legitimate user (i.e., identity theft). The results show that the OCCT performs well in different linkage scenarios.

II. RELATED WORK

The following sections explain the survey of various papers regarding this concern. Different methods are used for that have been proposed for having data linkage for different. Following section also explain different methods that are used to Clustering tree.

In [2], Tushar Khot, Sriraam Natarajan and Jude Shavlik have used relational one-class classification approach based on first-order trees. They defined a new distance metric based of first order decision forest and density estimation model using the distance metric. We can efficiently update the distance metric to improve the classifier’s performance. Tree based distance is used to learn a first-order tree for calculating relational distances with the help of lowest common ancestor (LCA). They are using density estimation model which combines the distances from multiple trees. Use the distance function to perform One Class Classification. Tree learning updates the distance measure & adding to the set of trees. Weight learning updates the weights.

In [3] S. Ivie, G. Henry, H. Gatrell and C. Giraud-Carrier suggested a genealogical record linkage (GRL) process which is used to check that two pedigrees refer to the same base individual. They use one-to-many data linkage for genealogical research. It is based on five attributes for data linkage name, gender, date of birth, location, and the relationships between the individuals. It matches using specific attributes and, therefore, very hard to generalize. They are using data set which consists of names of people, relationships, and events. Event consists of date and a place.

In [4] Mohamed Yukout, Ahmed K. Elmagarmid, Hazem Elmeleegy, Mourad Ouzzani they present a new record linkage approach that uses an entity behavior to decide if potentially different entities are in fact the same. The aim of this approach is a technique that merges the behavior of two possible matched entities and computes the gain in recognizing behavior patterns as their matching score. An entity’s behavior is extracted from a transaction log that records the actions of this entity with respect to a given data source. The idea is that if it obtains a well-recognized behavior after merge, then the original two behaviors belong to the same entity as the behavior becomes more complete after the merge.

In [5] Steven Euijong Whang, Hector Garcia & Molina Entity resolution (ER) suggested a technique to
identifying which records in a database represent the same entity. Sometimes records of different types are involved (e.g., institutions, venues, authors, publications), and resolving records of one type can impact the resolution of other types of records. They proposed a flexible, modular resolution framework where existing ER algorithms developed for a given record type can be plugged in and used in concert with other ER algorithms. Their approach also makes it possible to run ER on subsets of similar records at a time.

In [6] Karl Goiser and Peter Christen use record linkage technique which concerned with identifying records from one or more datasets which refer to the same underlying entities. Where the entity-unique identifiers are not available and errors occur, the process is non-trivial. They used one supervised and two unsupervised classification methods were chosen. The supervised method requires training data, and, being partitioning clustering techniques, the unsupervised methods require the specification of the number of clusters. As the aim is to have a cluster of matches and a cluster of non-matches, this value is fixed at two. Being fixed, the value doesn’t change meaning it is not supplied as a parameter.

In [7] Parag and Pedro Domingos proposed one technique that mainly focuses on the Multi-Relational Record Linkage, Record linkage or de-duplication, is identifying which records in a database refer to the same entities. This problem is traditionally solved separately for each candidate record pair. We propose to use instead a multi-relational approach, performing simultaneous inference for all candidate pairs, and allowing information to propagate from one candidate match to another via the attributes they have in common. Parameters are learned using a voted perceptron algorithm.

III. PROPOSED METHODOLOGY

A. Acquiring Data
This is the module which used to enter the data into our process. Here we had used sample dataset as product details and online Buyer details. The data can be Entered and view in a structured format in this module. The product dataset contain the detail of the product like Product Id, prize, Product Name and Product Description. In the same way the Buyer dataset contains the Buyer Id, Product Name, prize, and sight id. All the detail of the both dataset will be acquired and viewed in easy way.

B. Associate the Elements
A typical data linkage problem consists of two data tables that do not share a unique identifier. This module is used to solve the problem. The records of the both dataset will be compared one by one using the method one-to-many data link. The linkage model encapsulates the knowledge of which records are expected to match each other. The induction process includes deciding the structure of the tree. Building the tree requires deciding which attribute should be selected at each level of the tree. Once the construction of the tree is completed, each leaf contains a cluster (or set) of records. A set of probabilistic models is induced for each of the leaves. During the linkage phase, each pair of records in the testing set is cross validated against the linkage model.

The output is a score representing the probability of the record pair being a true match.

C. Excruciating Data
After the data linked the next step is to Split the True matches and false matches. Using the true matches the splitting will performed under four methods like Coarse Grained Jaccard Coefficient (CGJC), Fine Grained Jaccard Coefficient (FGJC), Least Probable Intersections (LPI), and Maximum-Likelihood Estimation (MLE) each method contains some calculation the output will be generated according to the data true match values. Pruning operation will be processed which means that the child and the branch of the tree will removed, the Root element and the leaf node alone will be clustered. This module is used to find that either the true match value has a better result. Using the methods the true match value will be divided into two the subset. According to the subset the process will be happened.

D. Analysis Result
The Output of the four methods will be measure and it values will be compared. One of the best and easy methods to view the compression is display the result in chart format. In that way the output of the four methods will be displayed in the chart format. Using this we can find the best method easily.

Fig. 1: Flow Chart

The overall methodology of proposed work is shown in figure 1.

In proposed work there are two data sets which are product details and online Buyer details. These two data sets
will linked with each other. After that checking will performed whether it is positive or negative. If it is negative than that record will not further process. If it is positive then it will split.

There are four types of Splitting namely CGJ, FGJ, LPI and MLE. Each splitting criterion has their own method to calculate the similarities between two subsets. Pruning is an important task in the proposed method. In pruning method the child and the branch of the tree will remove, the Root element and the leaf node alone will be clustered. Each leaf holds a data set containing the matching records from table TB.

IV. PROBLEM DEFINITION

A. Motivation

Use of data mining increases enormously in these recent years which focuses researcher’s attention to challenges faced in data mining. The key challenges in data linkage are:

- Calculating the probability of a given record pair being a match.
- To associate an entity from the first data set with a group of matching entities from the other data set.

B. Problem Statements

- This project mainly works one to many data linkage.
- The four splitting criteria algorithm is implemented for inducing the linkage model.

V. RESULT & DISCUSSION

In this section, we will discuss the result analysis of the proposed method. In chart we are comparing splitting criteria with pruning method. When we do not apply any pruning method, we found that the result is significantly worse than all other methods. However when we apply LPI pruning than the result is increased significantly. In addition when we apply MLE pruning the result is increased as compare to no pruning but MLE pruning does not produce better result as compare to LPI. While in most cases LPI pruning is better than the MLE pruning.

REFERENCE


