IJCSN - International Journal of Computer Science and Network, Volume 3, Issue 6, December 2014
ISSN (Online): 2277-5420 **www.IJCSN.org**
Impact Factor: 0.274

An Improved Data Linkage Technique Based on Clustering Tree for Decision Making

[1] Sathya. T, [2] Nithya. K

[1] ME Student, Department of Computer Science and Enginerring
KSR College of Engineering, Anna University,
Namakkal, Tamilnadu 637215, India

[2] Assistant Professor, Department of Computer Science and Enginerring
KSR College of Engineering, Anna University,
Namakkal, Tamilnadu 637215, India

Abstract - The present state-run of the fine art in the data linkage is to match the entities from the different data sources which do not contain the common identifier. In that here one-to-many data linkage is considered to obtain the decision process based on the clustering tree. In prior work, there is no one-to-many data linkage tasks instead the issue addressed are to link among the same type of the entities. In this paper, two new splitting criterion are introduced to enhance the performance of the linkage process for the best split at each node during the decision tree construction process and securing the linked data from the unauthorized usage. Pruning techniques are implemented to remove the anomalies of the clustering tree.

Keywords - *Data linkage, clustering tree, splitting criteria, pruning.*

1. Introduction

Nowadays, 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. The data linkage's major task is to identify the different objects which were used to refer the same entity across the different source of data. It is much need in the combining different databases or like the preprocessing step process in the dataset oriented process. Data linkage is split up into two types they are one-to-one and one-to-many linkage. One-to-one type is used to combine the single object with the one data set with the single matching. Other type is one-to-many type which is used to combine one data set with the group of objects to the other data set. Data linkage is simply defined as the combining together of information from two records that is supposed to relate to the same entity—for example, the same individual, family or the business. This might involve the linkage of records within a single database to identify duplicate case records. And also, record linkage might also does the process of linking records across two or more databases. Such work might be undertaken to combine these databases into a single database with an improved coverage or scope. The record linkage work is easiest when unique identification numbers such as Social Security Numbers are readily available. The work is more challenging when only quasi-identifiers such as given name, surname, date of birth, and address are available. In combination, quasi-identifiers may uniquely identify an individual.

Linkage allows the combination of different databases into one extensive data set for analysis. The insertion of address and telephone changes into a mailing or telephone list and the removal of duplicate entries from a mailing list are basic examples of record linkage. Historically record linkage was assigned to clerks who would search and review lists to bring together the appropriate pairs of records for comparison, seek additional information when there were questionable matches, and finally make decisions regarding the linkages based on established rules. Two frequently applied strategies in record linkage are deterministic (DRL) and probabilistic (PRL) record linkage that differ fundamentally in their approach. In deterministic record linkage all or a predefined subset of linking variables have to agree (corresponding values on a linking variable are the same within a pair) to consider a pair as a link. In probabilistic linkage, weights for agreement (reward) or disagreement (penalty) are estimated for each variable based on the difference in probability that a variable agrees among matches and non-matches. The term match refers to the situation that two records in reality belong to the same person, while a link refers to the outcome of the record linkage procedure. The first probability reflects the reliability of the variable (error rate) and the second probability the discriminating power of the variable (chance agreement). If the total sum of

weights is above a certain threshold value, the pair is considered a link.

Data linkage is a challenging problem because of errors, variations and missing data on the information used to link records, differences in data captured and maintained by different databases, e.g. age versus Date Of Birth, regularly and routinely change over time (for example, name changes due to marriage), often no unique entity identifiers are available and no training data in many linkage applications (no record pairs with known true match status). A clustering tree is a decision tree where the leaves do not contain the labels. Each internal nodes as well as each leaf corresponds to a cluster or a concept and thus it is termed as a clustering tree. That is, each of the leaves contains a cluster instead of a single classification. The tree as a whole describes a hierarchy (e.g., a taxonomy).Each leaf of the tree is characterized by a logical expression (e.g., conjunction of literals) representing the instances belonging to it. According to the main advantage of using clustering trees is that they provide a description for each of the clusters using a logical expression.

In this paper. we propose a new data linkage method aimed at performing one-to-many linkage that can match entities of different types. For example, Let T_A and T_B are the two tables of different types. The inner nodes of the tree consist of attributes referring to both of the tables being matched (T_A and T_B). 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. The proposed method was evaluated using the data leakage prevention domain. 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.

2. Related Work

Record linkage is a process of matching entities from two different data sources that may or may not share a common identifier (i.e., foreign key). One-to-one record linkage was implemented using algorithms like SVM classifier, Maximum Likelihood Expectation and performing behavior analysis [3]. These methods assume that entities in the datasets are linked and try to match records that refer to the same entity. Only a few previous works have dealt about one-to- many record linkage.

Dror et al. [1] used one-to-many linkage in different domains like fraud detection, recommender systems and data leakage prevention. Four splitting criteria is used and pruning methods are implemented to develop One-Class Clustering Tree (OCCT). The drawback of this approach is that it is difficult to reduce the linkage computation time and it is a one-class approach.

Ivie et al. [4] used one-to-many linkage for genealogical research. In that work, data linkage was performed using five attributes: a person's name, gender, date of birth, location 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 specific attributes and therefore it is very hard to generalize. Christen and Goiser [5] used a C4.5 decision tree to determine which records must be matched to one another. In their work, different string comparisons methods are built and compared using different decision trees. However, their method performs the matching of attributes that are only predefined. Moreover only one or two attributes are usually used.

3. Improved Data Linkage Model

In the proposed method, linkage model induction is the first step. The linkage model gets the knowledge about records that are expected to match each other. The process includes deriving the structure of the tree. The tree building requires the decision of which attributes must be selected at each level of the tree. The inner nodes of the tree consist of attributes from table T_A. The leaf contains the cluster that is matching attributes from table T_B to T_A. The selection of attributes is actually done by using any one of the splitting criteria. The splitting criteria ranks the attributes based on their clustering of matching examples.
A pre-pruning approach is implemented in this proposed method. When using this approach, the algorithm stops expanding a branch whenever the sub-branch does not improve the accuracy of the given model. The inducer is actually trained with matching examples only.

The can be derived using any one of the splitting criteria. The splitting criterion is used to determine which attribute must be used in each step of constructing the tree. Our main goal is to achieve a tree that contains less number of nodes, as smaller trees easily generalize the data by avoiding over fitting. The two types of splitting criteria used in this system are hybrid jaccard coefficient and jaro-winkler technique. The process flow of the proposed system is as follows:

Step1. First have to consider the customer dataset and the context dataset.

Step2. Then have to preprocess both the data.

Step3. After preprocessing, create the training table based on both the table.

Step4. For training table creation, consider three attributes as customer id, requesting location, requesting day, requesting time, in the other table have to consider two attributes as user location and the business type.

Step5. Then measure the similarity score based on the attributes of requesting location, requesting day, requesting time with user location and the business type.

Step6. Based on that score make the splitting criterion applied and constructs the tree. Here for existing we apply MLE method (Maximum Likelihood estimation) and for proposed system consider the Jaccard – Jaro coefficient method.

Step7. Then have to apply the pruning method as MLE pruning and the similarity pruning method.

Step8. Then apply the SEMANC method which is used to analyze the normal user or the malicious user from the linked data.

Step9. Then consider the performance graph based on both existing and proposed system.

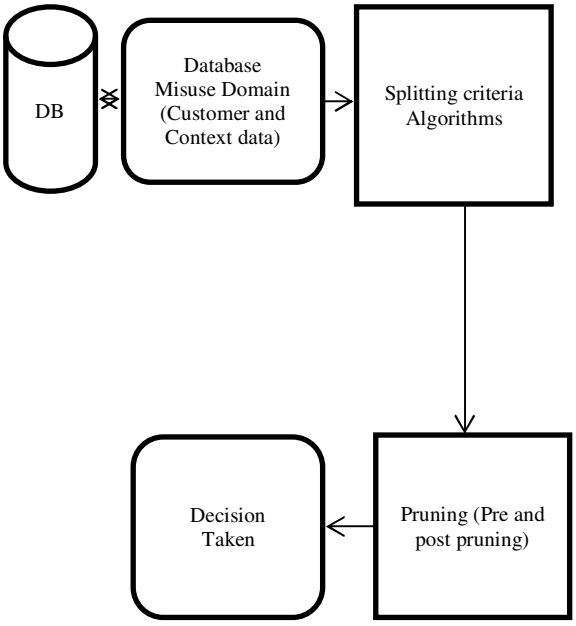

Fig.1 Architecture diagram

4. Splitting Criteria

The goal is to achieve a tree which contains a minimum number of nodes. Smaller trees generalize the data in a better way, avoid over fitting, and will be simpler for the human eye to understand. Therefore, it is crucial to use an effective splitting criterion in order to build the tree. The user will select the splitting criteria method which should be applied in the decision making process by linking the data.

4.1 Hybrid Jaccard Coefficient

The Hybrid Jaccard Coefficient Technique is used to attain the combination of the fine grained and coarse grained Jaccard coefficient technique for measuring the similarity between the clusters.

The Jaccard similarity coefficient of the Coarse-grained Jaccard (CGJ) coefficient algorithm, a measure that is commonly used in clustering, measures the similarity between the clusters .The goal is to choose the splitting attribute which leads to the smallest possible similarity between the subsets (i.e., an attribute that generates subsets that are different from each other as much as possible).

The fine-grained Jaccard coefficient is capable of identifying partial record matches, as opposed to the coarse-grained method, which identifies exact matches only. It not only considers records which are exactly identical, but also checks to what extent each possible pair of records is similar.

4.2 Jaro-Winkler Technique

It is used to measure the similarity between two strings. The Jaro measure is the weighted sum of percentage of matched characters from each file and transposed characters. Winkler increased this measure for matching initial characters, then rescaled it by a piecewise function, whose intervals and weights depend on the type of string (first name, last name, street, etc.).

The domain we are going to implement is database misuse domain they are called as context dataset and the customer dataset. The context dataset contains the following attributes they are Time of execution, Day of execution, geographical location of the request, user's role, and type of request. The customer dataset contains the following attributes they are Customer id, Customers first name, Customers last name, address, zip code, place of work, customer type. Then the attributes are split up by the similarity values between the clusters. Finally tree will be constructed.

The inner nodes of the tree consist of attributes referring to both of the tables being matched (T_A and T_B). 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

In a tree induction process, pruning is considered to be an important activity. The necessity of using pruning is to build a tree with accuracy and also to avoid over fitting. Pruning can be done in two ways: 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.

In our system we have followed a pre-pruning approach. It was chosen for the reason that it reduces the time complexity of the algorithm. The decision made to prune the branch or not is taken once the next attribute for split is chosen. In this proposed system, two pre-pruning methods are used.

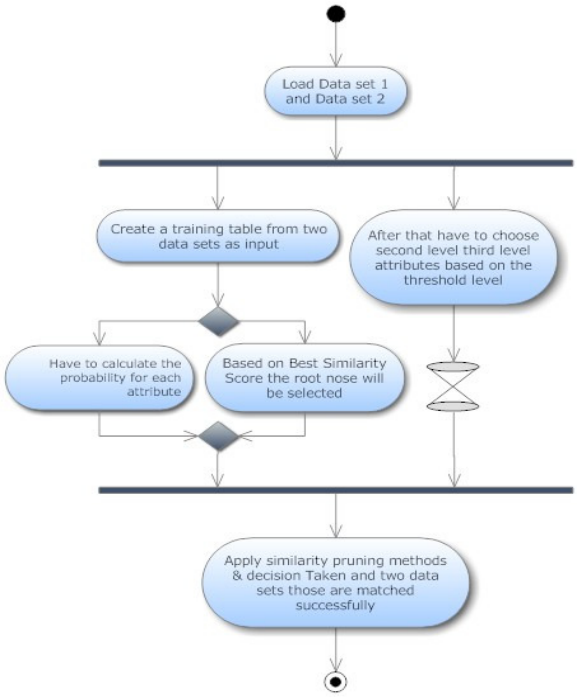

Fig.2 Activity diagram

6. Decision Taken

In this model we are going to analyze the decision using the tree constructed by the above process and then the decision is extracted from the constructed tree. The SEMANC method is going too be applied here for the decision result checking process by the real criteria and after that we can match the performance of the each process. Then the rules will be extracted from the constructed pruned tree. The proposed method is evaluated using the ROC graph. The graph plots the True positive Rate versus the False Positive Rate. Recall and Precision measurements are calculated for evaluating the data linkage methods.

7. Conclusion

In this paper, we are proposing the novel method to link the data which does not have the common entity and also based on the clustering tree. Based on this we can match the two different cluster of data from the different dataset. That was the main challenge in the data linkage here we are applying the one to many data linkage technique with the one class clustering tree in particularly database misuse domain. That was takes place by the decision tree technique. Here each node will be considered as the cluster of nodes and the whole data on the different dataset will be matched as the result. Here we attain the improved efficiency of the data linkage process.

References

[1] M.Dror, A.Shabtai, L.Rokach, Y. Elovici, "OCCT: A One-Class Clustering Tree for Implementing One-to-Many Data Linkage," IEEE Trans. on Knowledge and Data Engineering, TKDE-2011-09-0577, 2014.

[2] M.Yakout, A.K.Elmagarmid, H.Elmeleegy, M.Quzzani and A.Qi, "Behavior Based Record Linkage," in Proc. of the VLDB Endowment, vol. 3, no 1-2, pp. 439-448, 2010.

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

[4] S.Ivie, G.Henry, H.Gatrell and C.Giraud-Carrier, "A Metric Based Machine Learning Approach to Genea-Logical Record Linkage," in Proc. of the 7th Annual Workshop on Technology for Family History and Genealogical Research, 2007.

[5] P.Christen and K.Goiser, "Towards Automated Data Linkage and Deduplication," Australian National University, Technical Report, 2005.

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

[7] S.Guha, R.Rastogi and K.Shim, "Rock: A Robust Clustering Algorithm for Categorical Attributes," Informat- ion Systems, vol. 25, no. 5, pp. 345-366, July 2000.

[8] D.D.Dorfmann and E.Alf, "Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals-RatingMethod Data," Journal of Math Psychology, vol. 6, no. 3, pp. 487-496, 1969.

[9] A.Gershman et al., "A Decision Tree Based Recommender System," in Proc. the 10th Int. Conf. on Innovative Internet Community Services, pp. 170-179, 2010.

[10] J.R.Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81-106, March 1986.

[11] C. Li, Y. Zhang, and X. Li, "OcVFDT: One-Class Very Fast Decision Tree for One-Class Classification of Data Streams," in Proc. the 3rd Int. Workshop on Knowledge Discovery from Sensor Data, pp. 79-86, Paris, France, 2009.

[12] P.Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication," IEEE trans. on knowledge and data engineering, DOI:10.1109, TKDE.2011.127, 2011.

[13] N. Golbandi, Y. Koren, and R. Lempel, "Adaptive Boot-strapping of Recommender Systems Using Decision Trees," in Proc. the 4th ACM Int. Conf. on Web search and data mining, pp.595-604, Honk Kong, 2011.

[14] M. Gafny, A. Shabtai, L. Rokach, and Y. Elovici, "Detecting Data Misuse By Applying Context-Based Data Linkage," in Proc. ACM CCS Workshop on Insider Threats, Chicago, USA, 2010.

[15] S. Mathew, M. Petropoulos, H. Ngo, S. and Upadhyaya, "A Data-Centric Approach to Insider Attack Detection in Data-base Systems," Recent Advances in Intrusion Detection, Spring-er, vol. 6307, pp. 382-401, 2009.