A Review on Automated One-to-many Data Linkage

Shahid Anwar[1], Prof Rushi Longadge[2], Prof Deepak Kapgate[2]

[1]Department of Computer Science & Engineering
[1]G.H.Raisoni Academy College of Engineering, Nagpur, India.
[2] Department of Computer Science & Engineering
[2]G.H.Raisoni Academy College of Engineering, Nagpur, India.

Abstract- Data linkage identifies a different entity that belongs to different data sources .Data linkage is traditionally performed among the entities of same type. This technique is performed by using entities that may or may not share a common identifier. In this project we propose a new linkage method that performs linkage between matching entities of different data types. The proposed technique is based on one-class clustering tree that characterizes the entities which are to be linked. The tree is built in this way that it is easy to understand and can be transformed into association rules and the inner nodes of the tree consist of features of the first set of entities. The structure of the tree in such a way that leaves of the tree represent features of the second set that is matching. The data is split using splitting criteria. Also pruning methods are used for creating one-class clustering tree.

Keywords: Clustering, data linkage, data matching

*****

I. INTRODUCTION

The goal of the data linkage task is joining data sets that do not share a common identifier (i.e., a foreign key). Common data linkage scenarios include:linking data when combining two different databases [2].Data Linkage is divided into two types: one-to-oneand one to many [8]. In one toone data linkage, the goal is toassociate an entity from one data set with a single matching entity in another data set.One-to-many data linkage is an essential task in many domains. In one-to-many data linkage, thegoal is to associate an entity from the first data set with agroup of matching entities from the other data set. Most ofthe previous works focus on one-to-one data linkage.

Data linkage allows for the identification of distinct entities within datasets and betweendatasets.We propose a new data linkage method aimed at performing one-to-many linkage. In addition, while data linkage isusually performed among entities of the same type, this proposed data linkage technique can match entitiesof different types. For example, if there is a student database wemight want to link a student record with the coursesshe should take (according to different features thatdescribe 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 treein which each of the leaves contains a cluster instead of asingle classification and each cluster is generalized by a set ofrules that is stored inthe appropriate leaf [9], [10].

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 areused to Clustering tree.

In [1],TusharKhot, SriraamNatarajanand Jude Shavlik have usedrelational one-class classification approachbased on first-order trees.They defined a new distancemetric based of first order decision forest and density estimationmodel using the distance metric. We can efficientlyupdate the distance metric to improve the classifier's performance. Treebased distance is used tolearn 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 forgenealogical research. It is based on five attributes for data linkage name, gender, date of birth, location, and the relationships between theindividuals. It matches using specificattributes and, therefore, very hard to generalize. They are using data set which consists of names of people, relationships, and events. Event consists ofdate and a place.

In [4] Mohamed Yakout, Ahmed K. Elmagarmid, HazemElmeleegy, MouradOuzzani they present a new record linkage approach that uses an entity behaviorto 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 gainin recognizing behavior patterns as their matching score. An entity's behavior is extracted from a transaction log that records the actionsof 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 morecomplete after the merge.

In [5] Steven EuijongWhang, Hector Garcia &MolinaEntity resolution (ER) suggesteda 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 resolutionof other types of records. They proposed a flexible, modular resolution framework where existingER algorithms developed for a given record type canbe plugged in and used in concert with other ER algorithms. Their approach also makes it possible to runER on subsets of similar records at a time.

In[6] Karl Goiser and Peter Christen use record linkage technique whichconcerned with identifyingrecords from one or more datasets which referto the same underlying entities. Where the entity-unique identifiers arenot available and errors occur, the processis non-trivial. They used one supervised and two unsupervised classificationmethods were chosen. The supervised method requirestraining data, and, being partitioning clusteringtechniques, the unsupervised methods require thespecification of the number of clusters. As the aimis to have a cluster of matches and a cluster of nonmatches,this value is fixed at two. Being fixed, thevalue doesn't change meaning it is not supplied as aparameter.

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

III. EVALUATION AND DISCUSSION

The advantages and disadvantages of all the above discussed techniques are described in the Table-I.

Table 1: Comparison between various techniques of Data linkage and Clustering

Sr.no.	Paper Title	Authors	Approach	Advantage	Disadvantage
1	Relational One-Class Classification: A Non-Parametric Approach	TusharKhot, SriraamNatarajan and Jude Shavlik	relational one-class classification	Efficient way to calculate the relational distances	Relational data is required.
2	A Metric-Based Machine Learning Approach to Genealogical Record Linkage	S. Ivie, G. Henry, H. Gatrell and C. Giraud-Carrier	Genealogical Record Linkage (GRL)	Easy process to determine whether two pedigrees refer to the same base individual.	perform matches using specific attributes
3	Learning From positive and Unlabeled examples	Denis, Gilleron and Letouzey	POSC4.5	It learns from positive example.	Only Binary classification can consider.

4	Matching of catalogues by probabilistic pattern classification	D. J. Rohde, M. R. Gallagher, M. J. Drinkwater and K. A. Pimbblet	Maximum Likelihood Estimation (MLE)	Simple to implement	Require same entity
5	Towards Automated Record Linkage	Karl Goiser and Peter Christen	C4.5	providing high quality results.	predefined and only one or two attributes are usually used.
6	Joint Entity Resolution on Multiple Database	Steven EuijongWhang and Hector Garcia Molina	Entity Resolution	It matches the different entities	It is very complex process

IV.CONCLUSION

All the above techniques tried to cover different issues of data linkage and clustering treesuch as maintaining the cost of implementation and implementation complexities as low as possible. Data linkage is an essential task in many domains. One to many data linkagemethod which differs from clustering trees mainly inits ability to link two different types.

REFERENCES

[1] TusharKhot, SriraamNatarajan and Jude Shavlik, "Relational One-Class Classification: A Non-Parametric Approach" 28th AAAI Conference 2014.

[2] Ma'ayanDror, AsafShabtai, LiorRokach, and Yuval Elovici, "OCCT: A One-Class Clustering Tree forImplementing One-to-Many Data Linkage" IEEE transactions on knowledge and data engineering, vol. 26, no. 3, march 2014.

[3] S. Ivie, G. Henry, H. Gatrell, and C. Giraud-Carrier, "A Metric Based Machine Learning Approach to Genealogical RecordLinkage," Proc. Seventh Ann. Workshop Technology for FamilyHistory and Genealogical Research, 2007

[4] 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.

[5] S.E. Whang and H. Gercia-Molina,"Joint Entity Resolution,"technical report, Stanford Univ., 2009.

[6] P. Christen and K. Goiser, "Towards Automated Data Linkageand Deduplication," technical report, Australian Nat'l Univ., 2005.

[7] TusharKhot and SriraamNatarajan_ and Jude Shavlik"Relational One-Class Classification: A Non-Parametric Approach" 28th AAAI Conference 2014.

[8] J. Domingo-Ferrer and V. Torra, "Disclosure Risk Assessment in Statistical Microdata Protection via Advanced Record Linkage,"Statistics and Computing, vol. 13, no. 4, pp. 343-354, 2003.

[9] L. Gu and R. Baxter, "Decision Models for Record Linkage," 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] I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-TheoreticCo-Clustering," Proc. Ninth ACM SIGKDD Int'l Conf. KnowledgeDiscovery and Data Mining, pp. 89-98, 2003

[12] C. Li, Y. Zhang, and X. Li, "OcVFDT: One-Class Very FastDecision Tree for One-Class Classification of Data Streams," Proc.Third Int'l Workshop Knowledge Discovery from Sensor Data, pp. 79-86, 2009.

[13] J. Struyf and S. Dzeroski, "Clustering Trees with Instance LevelConstraints," Proc. 18th European Conf. Machine Learning, pp. 359-370, 2007.

[14] P. Christen, "A Survey of Indexing Techniques for Scalable RecordLinkage and Deduplication," IEEE Trans. Knowledge and Data Eng.,vol. 24, no. 9, pp. 1537-1555, Sept. 2012, doi:10.1109/TKDE. 2011.

[15] V. Torra and J. Domingo-Ferrer, "Record Linkage Methods forMultidatabase Data Mining," Studies in Fuzziness and SoftComputing, vol. 123, pp. 101-132, 2003.

ISSN: 2321-8169
054– 055

[16] S. Guha, R. Rastogi, and K. Shim, "Rock: A Robust ClusteringAlgorithm for Categorical Attributes," Information Systems, vol. 25,no. 5, pp. 345-366, July 2000.

[17] A. Gershman et al., "A Decision Tree Based RecommenderSystem," Proc. 10th Int'l Conf. Innovative Internet CommunityServices, pp. 170-179, 2010.

[18] F. Provost and P. Domingos, "Tree Induction for Probability Based Ranking," Machine Learning, vol. 52, no. 3, pp 199-215, 2003.

[19] [19] C. Ferri, P. Flach, and J. Herna´ndez-Orallo, "Learning DecisionTrees Using the Area under the ROC Curve," Proc. Ninth Int'lConf. Machine Learning, pp. 139-146, 2002.

[20] [20] C.A. Metz, "ROCKIT Software," http://metz-roc.uchicago.edu/MetzROC, 2003.