ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

MCCT: An Approach for Many-to-Many Data Linkage

Priyanka Hankare[1], Prof. A. B. Chougule[2]

Student ME-II, Dept. of CSE, D. Y. Patil College of Engineering and Technology, Shivaji University, Kolhapur, India[1]

Associate Professor, Dept. of CSE, Bharati Vidyapeeth's College of Engineering, Shivaji University, Kolhapur, India[2]

ABSTRACT: Many-to-many data linkage is a vital job in many areas, yet only a handful of former publications have addressed this subject. Besides, while customarily data linkage is performed among substances of the same sort, it is great degree important to create linkage mechanism between matching substances of different sorts as well. In this paper, we propose many-to-many data linkage technique that connects substances of various natures. The proposed strategy depends on a Multiple Clustering Tree (MCCT) that connects entities of various sorts to get a reasonable and justifiable relationship. We propose to utilize the Least Probable Intersection (LPI) technique to assemble the MCCT tree. This tree shall be a clustering tree whose inner nodes are set of records (i.e. cluster) from first substance and outer nodes are set of records (i.e. cluster) from second substance. The proposed technique gives better and preferable linkage results over past methodologies.

KEYWORDS: Clustering; data linkage; LPI; pruning

I. INTRODUCTION

In many applications for performing itemized investigation of information and information mining, there is a need to connect or join or incorporate the records from a few databases or to eliminate the duplicate records from a solitary database. The need of such linkages is to consolidate all records identifying with the same substance, such as, a patient, a client, or a business across different data sources. For example, a financial institution determines whether a person qualifies for a loan or mortgage by getting the information from several sources such as from person's credit card company, mortgage broker and bank, to get a better sense of a person's whole financial picture. The data linkage is progressively gaining prevalence in many applications. In the bank, for example, a bank account holder John may have accounts in various banks or credit cards of several companies. If he is seeking for a loan, then it is necessary that the financial institution that John has approached should be able to gain access to all his operational accounts and transactions. If this is not done in a systematic way, then there are chances that the user is doing a fraud transaction landing the bank in trouble.

The data linkage can be implemented for the various information sources like those of health systems, taxation offices, security organizations and crime detection frameworks, recommendation frameworks. Data linkage and de duplication can be used to enhance data quality, to allow re-utilization of existing information for new studies, and to diminish expenses in collecting the information for research work. Record linkage is done using either deterministic record linkage approach or probabilistic record linkage approach. Deterministic record linkage is a guideline based record linkage while probabilistic record linkage is a fluffy based record linkage. Probabilistic record linkage approach prompts preferable results over deterministic record linkage approach.

There are three sorts of data linkages: one-to-one, one-to-many and many to-many. In one-to-one data linkage, the record from information set A is connected with a single coordinating record in information set B. In one-to-many data linkage, the record from information set A is consolidated with a group of coordinating records from the information set B. One-to-many data linkage covers the one-to-one data linkage relationship. In many to-many data linkage relationship, any number of records from information set A are associated or connected with any number of coordinating records in information set B. Many-to-many data linkage is combination of one-to-one and one-to-many data linkage from information set A to B and information set B to A.

Customarily data linkage is performed among elements of same sort. A Multiple Class Clustering Tree (MCCT) will perform many-to-many data linkage among entities of different types. For example, in a patient database we might

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

want to link a patient record with the medicines he/she should take (according to different symptoms that describe the patient and features describing the medicines). MCCT will be a clustering tree with each node as well as each leaf is a cluster instead of a single classification. A cluster is a set of records satisfying the same condition. To perform many-to-many data linkage using MCCT model, firstly, a one-to-many data linkage using a One Class Clustering Tree (OCCT) model will be performed. OCCT is a clustering tree in which each of the leaves contains a cluster [1]. A one-to-one and one-to-many data linkage relationship is combined to provide better and strong linkage results.

II. RELATED WORK

Lots of research has been done in area of one-to-one and one-to-many linkage. Data linkage was implemented by utilizing techniques like SVM classifier, Maximum Likelihood Estimation (MLE) and by performing behavior investigation [2]. These strategies believe that same elements come into perspective in the two datasets to be connected and attempt to match records that refer to the same substance. Some noticeable work in area of data linkage is as follows:

Storkey et al. [3] did one-to-many data linkage using Expectation Maximization (EM) technique for the dataset comprising of number of items (stars or galaxies) which emit the radiation in far infra-red area. For connecting the matching records, the characteristics like infra-red area of the electromagnetic band, along with a few estimations from such objects, for example, position, flux etc. was utilized. No assessment was directed on this work. Ivie et al. [4] did one-to-many data linkage for genealogical research whose goal is to determine whether two records refer to the same base individual. Linking is done using the information of five features: an individual's name, gender, date of birth, location, and the relationship between the individuals. By using these attributes decision tree is built. They performed matches using specific attributes. V. Torra and J. Domingo-Ferrer [5] analyzed probabilistic and distance based record linkage methods for numerical and categorical information. They concluded that distance based record linkage is more appropriate for numerical information and probabilistic record linkage is more suitable for categorical information. Christen and Goiser [6] used a C4.5 decision tree to determine which records must be matched to each other. In their work, distinct string comparison techniques are used for matching the records. They performed linking using one or two attributes only. P. Christen [7] utilized indexing methods for making linkage process efficient and scalable. The goal of the indexing method is to remove the obvious non matching pairs and therefore, to minimize the quantity of record pairs to be compared in matching procedure. Six indexing procedures are used for data linkage. Dror et al. [1] used an OCCT approach for one-to-many data linkage to link records of different types of entities. Splitting and pruning techniques are used to build OCCT.

In all the previous classification techniques, the data linkage procedures are applied using one way traversal. The results obtained by applying data linkage in one way traversal may not be same with the results obtained by applying data linkage in reverse direction. So there is a need to develop a new technique performing deeper data linkage in two way traversal for obtaining more accurate linkage results for prediction. Also major weakness in decision trees as a prediction model is to provide a decision aka. 'match' class or a 'non-match' class for given input. The aim of MCCT model is to recommend the interested item-set for a given input instead of binary class output.

III. PROPOSED ALGORITHM

A. *Design Considerations:*

- Database is used for storing data of entities of different types.
- Weka tool is used for implementation of OCCT model.
- Categorical attributes are considered for implementation of OCCT model.

B. *Description of the Proposed Algorithm:*

The aim of proposed algorithm is to develop a MCCT model performing filtering to recommend accurate data linkage results. The proposed algorithm is consists of three main steps.

Step 1: Data Pre-processing:

Pre-processing is required to clean and standardize the data. If the class label for the records is missing, then such records are removed from the data set. In this step, firstly, the join of supervised data of entities of different types is taken. Secondly, the data is partitioned into match and non-match class by observing distribution of data and then

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

parsed to prune the non-matching and negative class instances from the dataset. The records with match or positive class label, called as set of matching instances (T_{ab}), are used for further processing.

Step 2: OCCT Formation:
In this step, the one-to-many data linkage is performed between T_a and T_b. For building the OCCT model, Least Probable Intersection (LPI) [8] splitting technique is used for selecting the internal attributes of OCCT model and the external nodes or leaf nodes of model are a cluster, formed by applying the Weka's J48 decision tree algorithm [9]. For example consider the OCCT formation from table T_a to table T_b:
In this process, for linking a record of table T_a with group of matching records of T_b, a tree is formed whose internal nodes are attributes from table T_a and leaf nodes are cluster i.e. set of records from table T_b. In order to select best attribute 'a' ϵ T_a, at each level of a tree, LPI splitting method is used. The LPI pre-pruning strategy is utilized to avoid overfitting of the tree. To form a matching set of records or cluster from table T_b at leaf node, J48 decision tree algorithm is applied on attributes of table T_b.
In the same manner OCCT model from table T_b to T_a is formed where internal nodes are attributes from table T_b and leaf nodes are cluster i.e. set of records from table T_a. Rest of the paper is explained on the basis of linkage from table T_a to T_b.

1. LPI Splitting Technique:

Consider 'q' numbers of records are present in set of matching instances (T_{ab}). For each attribute 'a' from table T_a, LPI divides T_{ab} into subsets according to values of chosen attribute. Suppose T_{ab} is divided into two subsets d1 and d2, consisting of 'k' and 'q-k' number of tuples respectively. The attribute that shares least amount of instances (from T_{ab}) between two subsets is chosen as a next splitting attribute. The probability of O_i appearances of record r_i belongs to subset d1 is calculated by using equation 1.

$$P(d1) = \left(\frac{k}{q}\right)^{oi} \quad \text{eq. (1)}$$

The probability of O_i appearances of record r_i belongs to subset d2 is calculated by using equation 2.

$$P(d2) = \left(\frac{q-k}{q}\right)^{oi} \quad \text{eq. (2)}$$

The probability P_i that record r_i in T_{ab} belongs to both subsets d1 as well as d2 is calculated by using the equation 3.

$$P_i = 1 - P(d1) - P(d2) \quad \text{eq. (3)}$$

The final score (Z) is calculated by using equation 4.

$$Z = \frac{j-\lambda}{\sqrt{\lambda}} \quad \text{eq. (4)}$$

Where λ = summation of P_i
j = | (records in subset d1) ∩ (records in subset d2) |
The attribute with the lowest LPI score (Z) is selected as the next splitting attribute.

2. Weka's J48 Algorithm

For forming a cluster of table T_b at leaf nodes of OCCT model, J48 decision tree algorithm is applied on each attribute of table T_b by setting it as a target attribute and all other attributes as an input attributes. Correct classes given by J48 decision tree algorithm are used as a cluster of OCCT model.

Step 3: MCCT Formation:
In this step, the OCCT models from table T_a to T_b and table T_b to T_a are combined. The records from the table T_a are linked with a group of matching records from the table T_b. The output of this module is a MCCT model whose internal nodes are cluster from table T_a and external nodes are a cluster from table T_b. Merging of trees is performed on the rule basis [10]. For performing many-to-many data linkage, each instance a_ib_j from T_{ab} is checked in both OCCT models. If an instance is present in both models, the respective cluster of the given instance is retrieved from them by traversing OCCT models.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

IV. PSEUDO CODE

A. *OCCT Formation from Table T_a to Table T_b:*

Step 1: Select the best attribute from set of attributes of table T_a (A) by LPI splitting technique.

best attribute ← selectAttributebyLPI (T_{ab}, A)

Step 2: Divide the T_{ab} according to each value of best attribute and apply the LPI technique on all other attributes except a best attribute until all the attributes of A gets selected.

for each value Vi of best attribute

child node of best attribute ← selectAttributebyLPI ($\sigma_{\text{best attribute} = Vi}$ (T_{ab}, A\{best attribute}))

end for

Step 3: Get the clusters of attributes of table T_b (B) by applying Weka's J48 algorithm.

clusters ← getClusters (T_{ab}, B)

Step 4: End.

B. *MCCT Formation:*

Step 1: Set MCCT model to NULL.

Step 2: For each record (a_ib_j) from $T_{ab,}$ check whether a_i part of record is present in OCCT model from table T_a to T_b and b_j part of record in OCCT model from table T_b to T_a.

Step 3: If present, then retrieve cluster of a_i part of record from OCCT model from table T_a to T_b and b_j part of record from OCCT model from table T_b to T_a.

Step 4: Get a MCCT model by merging clusters obtained in step 4.

MCCT model ← MCCT model ∪ cluster of a_i part of record ∪ cluster of a_i part of record

Step 5: End.

V. SIMULATION RESULTS

For experimental estimation, the vehicle insurance policy recommender domain is used. The aim of policy recommender system is to recommend the policy type and vehicle class of interest to the user by taking into consideration his features like gender, education, employment status. The features of entity policy are number of policies taken by user, policy type, sales channel, and vehicle class. The data set is partitioned randomly into training and testing sets; 80 percent of the data is used for training and the remaining 20 percent is used for testing. The same training set is given as input to both OCCT and MCCT systems. After implementation of both models, their correctness is analyzed on 300, 500, and 800 number of test records. Precision, recall and F-measure parameters are used for evaluating these linkage techniques.

Precision (or positive predictive value) is the number of correctly identified matching records divided by the number of pair of records that were identified as matching. Figure 1 demonstrates that precision metric remains the same irrespective of the number of records taken as input. Recall (or sensitivity) refers to the number of correctly identified matching records divided by the total number of matching records in the test set. For example, for 300 number of test records, 23 records are incorrectly identified by OCCT technique while MCCT technique identifies all the records correctly. Figure 2 depicts that MCCT technique yielded the best recall performance than OCCT technique. F-measure is a harmonic mean of precision and recall. F-measure graph is illustrated in figure 3. All the result figures shows that MCCT technique yielded in high accuracy results than OCCT technique.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

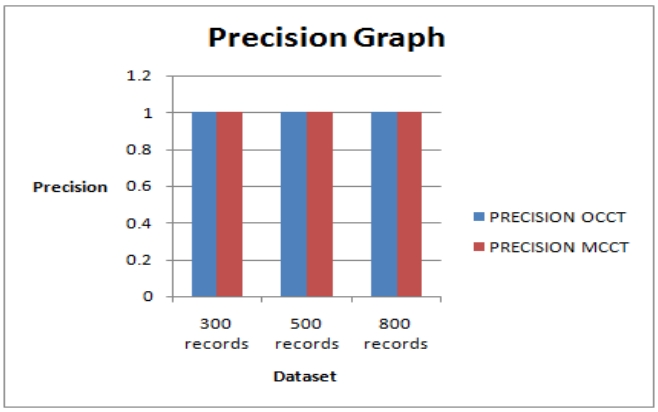

Fig.1. Precision Graph for OCCT and MCCT

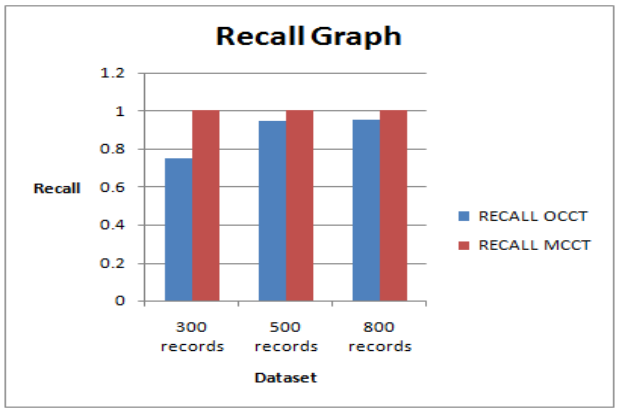

Fig.2. Recall Graph for OCCT and MCCT

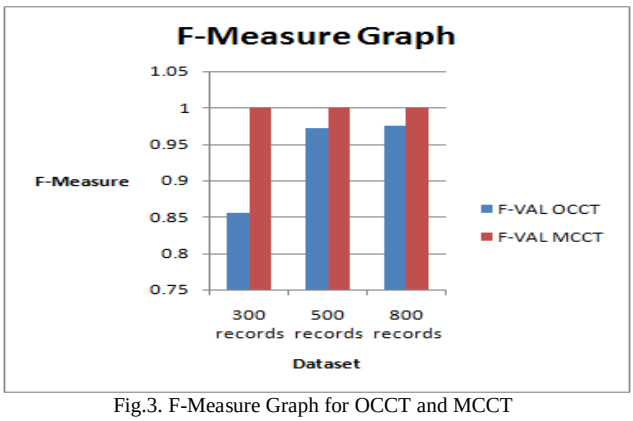

Fig.3. F-Measure Graph for OCCT and MCCT

VI. Conclusion and Future Work

The new approach of two way data linkage has been presented for producing the recommendation list. MCCT model is constructed by merging OCCT models to get advantages of one-to-one and one-to-many data linkage in a single representation which brings to many-to-many data linkage relationship. The results obtained for MCCT technique were compared against OCCT model and it was observed that new mechanism outperforms the high level of accurate linkage results than previous linkage approaches. In future work, a MCCT model is developed using one pass technique. Many-to-many data linkage can be implemented using other algorithms and compared with MCCT model.

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", IEEE Transactions from Knowledge and Data Engineering, vol. 26, no. 3, 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, nos. 1/2, pp. 439-448, 2010.
3. A.J.Storkey, C.K.I.Williams, E. Taylor, and R.G. Mann, "An Expectation Maximization Algorithm for One-to-Many Record Linkage", Univ. of Edinburgh Informatics Research Report, 2005.
4. S.Ivie, G.Henry, H. Gatrell, and C. Giraud-Carrier, "A Metric-Based Machine Learning Approach to Genealogical Record Linkage", Proc. Seventh Ann. Workshop Technology for Family History and Genealogical Research, 2007.
5. V. Torra and J. Domingo-Ferrer, "Record Linkage Methods for Multidatabase Data Mining", Studies in Fuzziness and Soft Computing, vol. 123, pp.101-132, 2003.
6. P.Christen and K. Goiser, "Towards Automated Data Linkage and Deduplication", technical report, Australian Nat'l Univ., 2005.
7. P. Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 9, pp. 1537-1555, 2012.
8. A.Gershman, Amnon Meisels, Karl-Heinz Luke, Lior Rokach, Alon Schclar, Arnon Sturm, "A Decision Tree Based Recommender System", in Proc. the 10th Int. Conf. on Innovative Internet Community Services, pp. 170-179, 2010.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

9. E. Frank, M.A. Hall, G. Holmes, R. Kirkby, and B. P fahringer, "WEKA - A Machine Learning Workbench for Data Mining", The Data Mining and Knowledge Discovery Handbook, pp. 1305-1314, Springer, 2005.
10. Hall, L. O., N. Chawla, and K. W. Bowyer, "Combining Decision trees learned in Parallel", Working Notes of the KDD-97 Workshop on Distributed Data Mining, pp. 10-15, 1998.

BIOGRAPHY

Priyanka Hankare is a ME-CSE student of D. Y. Patil College of Engineering and Technology, Shivaji University. She received the BE degree in Computer Science from Shivaji University in 2014. Her research interest is Data Mining.

Prof. Amit Chougule received the B.E. degree in Computer Science from Shivaji University in 2001 and the M.Tech degree in Computer Science and Technology from Shivaji University in 2008. He is currently working as Associate Professor in Shivaji University (Maharashtra, India). His research interests are Distributed System, Networking, and Data Mining.