International Journal For Technological Research In Engineering
Volume 4, Issue 10, June-2017 *ISSN (Online): 2347 - 4718*

RECORD LINKAGE USING NON-STOCHASTIC METHOD

Ranjana G[1], Dr. K. Thippeswamy[2]
[1]PG Student, [2]Professor & Head
Department of computer Science and Engineering, VTU PG Center, Mysore, India

Abstract: Record Linkage has wide variety of application in various research fields like knowledge discovery in databases, data warehousing, system integration and e-services. The process of identifying the record pairs that represent the same entity (duplicate records), is known as record linkage. It is one of the essential elements of data cleaning. In this paper, we address the record linkage problem by adopting a Non-stochastic methods from Machine Learning approach.
Keywords: Non-stochastic, supervised, clustering,

I. INTRODUCTION

Record linkage is the process of comparing the records from two or more data sources in an effort to determine which pairs of records represent the same real-world entity. Record linkage may also be defined as the process of discovering the duplicate records in one file. What makes record linkage a problem in its own right, (i.e., different from the duplicate elimination problem [1]), is the fact that real-world data is "dirty". In other words, if data were accurate, record linkage would be similar to duplicate elimination, since the duplicate records would have the same values in all fields. Yet, in real-world data, duplicate re cords may have different values in one or more fields. For example, more than one record may correspond to the same person in a customer database because of a misspelled character in the name field. Record linkage is related to the similarity search problem, which is concerned with the retrieval of those objects that are similar to a query object. In particular, record linkage may use similarly search techniques in order to search for candidate similar records. From these candidate similar records, record linkage should determine only those that are actually duplicates Record linkage can be considered as part of the data cleansing process, which is a crucial first step in The knowledge discovery process [2]. Data cleansing, also called data cleaning, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. In 1969,Fellegi and Sunter [3] were the first to introduce the formal mathematical foundations for record linkage, following a number of experimental papers that were published since 1959 [4]. The model proposed by Fellegi and Sunter, is characterized as a probabilistic model since it is entirely based on probability theory. Winkler [5] surveys the research that extends and enhances the model proposed by Fellegi and Sunter. The record linkage problem can be viewed as a pattern classification problem. The goal is to correctly assign patterns to one of a finite number of classes. By the same token, the goal of the record linkage problem is to determine the matching status of a pair of records brought together for comparison. Machine learning methods, such as decision tree induction, neural networks, instance-based learning, clustering, etc., are widely used for pattern classification. Specifically, given a set of patterns, a machine learning algorithm builds a model that can be used to predict the class of each unclassified pattern. Machine learning methods are categorized into two main groups: supervised learning and unsupervised learning. A method is supervised if a training set is available; otherwise the method is unsupervised [6]. Cochinwala et al. [7], and Verykios et al. [5] were the first to exploit the use of decision tree induction for the solution of the record linkage problem. A typical and emerging area that involves access to both databases and applications is Digital Government. The aim of digital government is to provide computer-based systems that allow dynamic management and access of a large number of governmental databases and services. The government data is so critical that it should be designed, analyzed and managed with data quality as a guiding principle and not as an afterthought.

1.1 Paper Organization

The rest of this paper is organized as follows. In Section 2, the record linkage problem is introduced along with the notation that is used throughout the paper. In section 3 various machine learning approaches have been presented. Section 4 summarizes the various methods.

II. RECORD LINKAGE PROBLEM

2.1 Definition and Notation

For two data sources A and B, the set of ordered record pairs A X B :::{(a, b):a€ A, b E B} is the union of two disjoint sets, M where a::: band U where a'* b . We call the former set matched and the latter set unmatched. The problem, then, is to determine in which set each record pair belongs to. Having in mind that it is always better to classify a record pair as a possible match than to falsely decide on its matching status with insufficient information, a third set P, called possible mate/led, is introduced. In the case that a record pair is assigned to P, a domain expert should manually examine this pair. We assume that a domain expert can always identify the correct matching status (M or U) of a record pair.

III. MACHINE LEARNING APPROACH

One of the disadvantages of the probabilistic record linkage model is its ability to handle only binary or categorical comparison vector attributes. Our goal is to overcome this disadvantage using new machine learning approach. The proposed machine learning record linkage models can handle all comparisons types, including the continuous ones. Another disadvantage of the probabilistic record linkage model is that it relies on the existence of a training set.

Although the proposed induction record linkage model has the same disadvantage, both the clustering and the hybrid record linkage models do not.

3.1 Induction Record Linkage Model

In supervised machine learning, a training set of patterns in which the exact class of each pattern is known a priori, is used in order to build a classification model that can be used afterwards to predict the class of each unclassified pattern. A training instance has the form < x, f{x)> where x is a pattern, and J{x} is a discrete-valued function that represents the class of the pattern x, i.e., f(X) € $\{L_1L_2,...,L_m..\}$,where m is the number of the possible classes. The classification model can be defined as an approximation to *f* that is to be estimated using the training instances. A supervised learning technique can be called a classifier, as its goal is to build a classification model. Induction of decision trees [8] and instance based learning [10], which are called inductive learning techniques, are two examples of classifiers. These techniques share the same approach to learning. This approach is based on exploiting the regularities among observations, so that predictions are made on the basis of similar, previously encountered situations. The techniques differ, however, in the way of how similarity is expressed: decision trees make important shared properties explicit, whereas instance-based techniques equate (dis)similarity with some measure of distance. By itself, the induction of decision trees technique does feature selection that decreases the cost of prediction.

The proposed induction record linkage model is illustrated in Figure 1. The training set consists of instances of the form < c, J{c) > where c is a comparison vector and J(c) is its corresponding matching status, i.e., j{C)€ {M,U} where M denotes a matched record pair and U denotes an unmatched one. A classifier is employed to build a classification model that estimates the function J and is able to predict the matching status of each comparison vector of the whole set of record pairs.

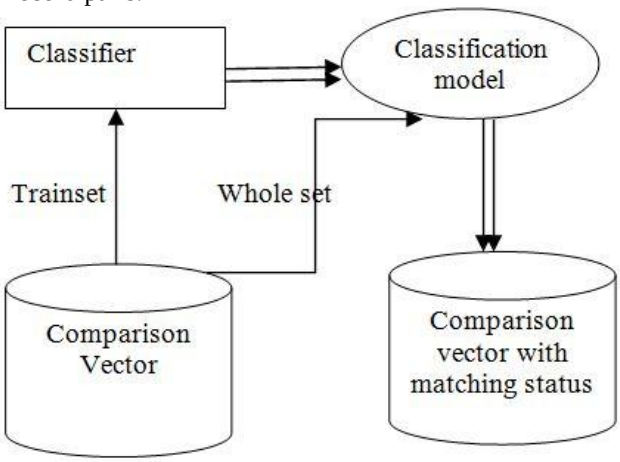

Figure 1. Induction Record Linkage Model

3.2 Clustering Record Linkage Model

The disadvantage of the previous model, as well as of the probabilistic record linkage model, is that it relies on the existence of a training set. Such a training set is not readily available for most real-world applications. In unsupervised learning methods, the notion of a training set does not exist. The whole set of patterns is given as input to the unsupervised learning algorithm to predict the class of each unclassified pattern, or in the record linkage case, the matching status of each record pair. Following the same notation used in the previous section, unsupervised learning tries to approximate the function without having any training instances. Clustering is the only known way for unsupervised learning, and so the model proposed can be called clustering record linkage model. The fundamental clustering problem involves grouping together those patterns that are similar to each other [9]. In other words, if each pattern is represented as a point in the space, clustering algorithms try to cluster these points into separate groups in the space. A specific technique, called k-means clustering, tries to cluster the points into k clusters. This is used specifically when the number of classes of the data items is known.

The clustering record linkage model considers each comparison vector as a point in n-dimensional space, where n is the number of components in each record. A clustering algorithm, such as k-means clustering, is used to cluster those points into three clusters, one for each possible matching status, matched, unmatched, and possibly matched. After applying the clustering algorithm to the set of comparison vectors. The issue is to determine which cluster represents which matching status.

Let c_{ij}= [c^{ij}_1, c^{ij}_2 …$c^{ij}_{n]}$ be the comparison vector resulting from component-wise comparison of the two records r_i and r_j. Assuming that all the comparison functions are defined in such a way that the value 0 means a perfect agreement between the two compared values, then $c_k^{i,j}$ =0 means that the two compared values $r_i.f_k$ and $r_j.f_k$ agree perfectly. Therefore, a perfectly matched record pair that agrees in all fields results in a comparison vector that has zeros in all of its components, i.e., its location coincides with the origin in n-dimensional space. Similarly, a completely unmatched record pair results in a comparison vector that has I's in all its components. Hence, in order to determine which cluster represents which matching status, the central point of each cluster in the space is determined. The nearest cluster to the origin is considered to be the cluster that represents the matched record pairs, where as the farthest cluster from the origin is considered to be the one that represents the unmatched record pairs. The remaining cluster is considered the one that represents the possibly matched record pairs.

3.3 Hybrid Record Linkage Model

The third model proposed in this paper is the hybrid record linkage model. Such a model combines the advantages of both the induction and the clustering record linkage models. Supervised learning gives more accurate results for pattern classification than unsupervised learning. However, supervised learning relies on the presence of a training set, which is not available in practice for many applications. Unsupervised learning can be used to overcome this limitation by applying the unsupervised learning on a small set of patterns in order to predict the class of each unclassified pattern. i.e., a training set is generated.

The proposed hybrid record linkage model proceeds in two steps. In the first step, clustering is applied to predict the matching status of a small set of record pairs. A training set is formed as {<c,J(c)>} where C is a comparison vector and f(c) is the predicted matching status of its corresponding record pair, i.e., f(C) €{M,U,P} where P denotes a possible matched record pair and M and U are as before. In the second step, a classifier is employed to build a classification model just like the induction record linkage model.

IV. CONCLUSION

From the paper it can be concluded non stochastic based machine learning approaches can be used to apply record linkage of large dataset. As supervised learning expects labels, it is little bit confusing to assign label. Some techniques like crowdsourcing can be used to label the data and results of already matching records can be considered as trainset. Unsupervised methods will be suitable for simple attributes in data. Combination of both supervised and unsupervised can be give best results.

ACKNOWLEDGEMENTS

I am using this opportunity to express gratitude to my guide and all the members for their help and support. I render my sincere gratitude to my family for their support.

REFERENCES

[1] D. Bitton and D. DeWitt. Duplicale Record Elimination in Large Data Files. ACM Transactions 011 Database Systems. 8(2), pages 255-265, June 1983.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery in Databases (a survey). AT Magazille, 17(3), pages 37·54,1996.

[3] I. Fellegi and A. Sunter. A Theory for Record Linkage. JOIITllal a/tile American Statistical Association. 64,pages 1183-1210, 1969.

[4] H. Newcombe, J. Kennedy, S. Axford, and A. James. Automatic Linkage ofVital Records. Science, 130, pages 954-959,1959. W. Winkler. The Stale of Record Linkage and Current Research Problems. U.S. Bureau of the Census, Research Report, 1999.

[5] V. Verykios, A. Elmagarmid, and E. Houslis. Automating the Approximate Record Matching Process. Journal of Information Sciences, 126(1-4), pages 83·98, July 2000

[6] T. Mitchell. Machine Learning. McGraw Hill, 1997.

[7] M. Cochinwala, V. Kurien, G. Lalk, and D. Shasha. Efficient Dala Reconciliation. Technical Report, Telcordia Technologies, February 1998.

[8] J. Quinlan. Induction of Decision Trees Machine Learning, 1, pages 81-106, 1986.

[9] P. Bradley and U. Fayyad. Refining Initial Points for K-Means Clustering. In Proc. Of the 15th Int. Conf Machine Learning, pages 91-99, San Francisco, CA, 1998

[10] D. Aha. Tolerating Noisy, Irrelevant and Novel Attributes in Instance Based Learning Algorithms. International Journal of man-machine studies, 36(1), pages 267-287,1992.

