AN OPTIMIZED TECHNIQUE FOR MATCHING PATIENT RECORDS IN MULTIPLE DATABASES USING DECISION TREE

Ms.Subitha.S 1, Mrs.S.C.Punitha 2

1Research Scholar, Department of Computer Science, P.S.G.R.Krishnammaal College for Women, Coimbatore, India
2HOD, Department of Computer Science, P.S.G.R.Krishnammaal College for Women, Coimbatore, India

ABSTRACT:

Record matching is the process of identifying records that refer to the same real world entity or object. The record matching process can be used for matching medical records. Multiple ration cards and multiple credit application fraud is a specific case of identity crime. In literature non-data mining detection system of business rules and core cards, and known fraud matching has limitations. To tackle these issues and combat identity crime in real time and also many problems arise when linking patient records from multiple databases. In the existing work, the fuzzy logic is used for construct the membership values of string similarity and based on this rules are build to find the record matching. However in this system, many drawbacks are still there to address such as uniqueness weight of an attribute are not analyzed in this system. Due to this, the time complexity of the system is increased as well as accuracy of record matching system is reduced. Undoubtedly, this is degrades the efficiency of the system. In order to overcome these issues, we are proposing the effective record matching system using decision tree approach. The main aim of this proposed system is to discover the effectiveness of a medical record matching system using a decision tree framework. This system considers quantitative measures of the elements in medical records, and decision tree rule is applied to link to the linguistic concepts. Experimental results are shown that our proposed system called record matching using decision tree approach is well effective than the existing system. Our proposed system is offers an effective solution for dealing with linkage problems.

Keywords: Health care system, Patient records, Record matching, Decision tree approach, Levenshtein edit distance

[1] INTRODUCTION

The health care system has multiple legacy and information systems that support its health care experts. These effective systems enable the integration and streamlining of healthcare delivery process with the intension of enhances the quality of US health care [1]. Though, the implementation of comprehensive information systems in health care system has proven to be a path ridden with risks and dangers owing to insufficient design of the database management systems and poor performance [2]. Furthermore, the complexity of health care systems has required the development of effective methods to manage the ever increasing volume of clinical, financial, demographic, and socio-economic data. Patient care data – all
of it useful – are typically scattered across multiple departmental databases regardless of their size.

In addition, the lack of a common data model, errors in data flows, errors during data entry, or situations where updates are not reflected into the database can cause inconsistencies to arise. Today, these inconsistencies are common in information systems and are the cause of considerable revenue loss, as it has been expected that information errors in the US medical care system contribute up to 98,000 deaths in hospitals and costs approximately $38 billion per year [3]. As a result, such data quality problems are a focus of increased attention [4]. All these data quality problems can prevent the accurate recording of patients’ information and even can obstruct efficient data access when occurring in identifying attribute domains [5]. Consequently, obtaining the metadata (data descriptions, business rules) represent a important challenge by reason of the lack of consistencies in databases and the fact that databases are regularly too large and difficult for manual inspection [6]. One common issue because of irregularity in database is that data objects can exist in multiple variations of patients’ contacts or incompatible text formats across multiple sources [7]. As well, such irregularities may cause incorrect linkage of patient records.

Locating matches across a pair of list not having unique identifiers such as social security number is often difficult. In general available identifiers such as first name, last name, date of birth, gender and address components may not uniquely identify matches because of legitimate variations [8]. A typical problem of medical records linkage is, for example, matching service recipients to Medicare eligibility. Given a service recipient’s record from a hospital database (the “presented” record) which of several, possible numerous, records (“contested” records) in the Medicare Common Working File (MCWF) mostly closely match the recipient if based on the matching element such as the HIC (Health Insurance Code) number, last name, first name, date of birth and gender.

In connection with the above linkage difficulty in database systems, many authors have used record linkage or matching to create a frame, eliminate duplicates from files, or combine files, so that the relationships on two or more data elements from separate files can be attained. Record matching can be classified into two categories: exact matching and statistical matching. Exact matching is used the identifiers like name, address, social security number or tax unit number to match a linkage of data for the same unit from different files, while statistical matching is used to match a linkage of data for the same unit from different files according to similar characteristics rather than unique identifying information [9].

[2] RELATED WORKS

The traditional probabilistic record linkage technique, as formalised in the 1960s by [10], has in recent years been enhanced by using the expectation-maximization (EM) method for good parameter estimation in record pair classification [11], and by applying approximate string comparisons to evaluate partial agreement weights when record attribute (field) values have typographical variations [12], [13]. Because the mid 1990s, researchers have analyzed a several methods to record linkage, deriving from artificial intelligence, database technology, information retrieval, machine learning, and data mining [12], [13], [14], with the intention of enhancing the linkage quality and the scalability of the record matching. Most of these
techniques are according to supervised learning approaches and necessitate training data (record pairs with known true match or true non-match status). Such training examples, on the other hand, are regularly not available in real world applications, or they have to be built manually. This is costly and high time complexity procedure that can be quite inaccuracy prone, as even humans sometimes are not able to clearly identify whether two records are a true match or not without having access to further information.

Three approaches for record pair classification have been developed in TAILOR [15]: the first is according to supervised decision tree method, the second is using unsupervised k-means clustering approach (with three clusters, one each for matches, possible matches and non-matches), and the third is a hybrid technique that combination of the first two to address the issue of lack of training data. Active learning is a technique which has the goal to address the difficulty of lack of training data. A system that presents a complicated to classify record pair to a user for manual classification is discussed in [16].

Unsupervised clustering approaches have been investigated both to improve blocking and for automatic record pair classification. In recent years, unsupervised techniques based on relational clustering [17] have been discovered in the field of entity resolution of relational data. The result of this work showed that relational entity resolution does better than non-relational entity resolution based only on record attribute similarities. Though, non-relational data is still available in many real world applications, for example in databases that contain patient or customer information. In [18], [19], the PEBL and TC-WON techniques are proposed, which are both based on iteratively training a SVM (Support Vector Machine) using the positive and a selected set of strong negative examples.

[3] EXISTING WORK

In record linkage problem, there are two approaches for record matching. The first approach is called exact or deterministic and it is mainly used when there are unique identifiers for each record. Deterministic algorithms use a set of rules founded on exact agreement or disagreement results between corresponding fields in record pairs. The second technique to record matching is classified as approximate or probabilistic. Approximate methods generally utilize likelihood scores which are evaluated from rates of identifier agreement and disagreement in the middle of fields from potentially matched and not-matched records. The methods evaluated in this work, fall under the second category. The two principle steps in the record matching development are the searching of potential pairs of records, the searching step, and the decision whether a given pair is correctly matched, the matching. The aim of searching step is to decrease the number of failures to bring linkable records together for comparison. For the matching stage, the main aim is to let the computer score the closeness of a match when some attributes of records match exactly and others do not.

This system effort recommends a solution using a multiple valued logic technique integrating fuzzy matching approaches to overcome issues related with matching patient records. The technique presenting in this work can successfully be applied to match patient record in multiple resources or eliminate duplicates in a single database. In a manual record matching process, knowledgeable users build up a “rule of thumb” to judge how well a
contested record similar to the presented record. These rules may be conceptualized as
collection weightings of several individual or attribute comparisons between the contested
and presented records. Fuzzy logic enables the mapping of similarity values of two
 corresponding attributes in a contested-presented pair of records to linguistic concepts, such
as “matched”, “possible matched”, and “not-matched”. The selected attributes or features
in the records may contain data elements such as last and first name, and date of birth. The
fuzzy logic method process is used as the framework in this system.

[4] PROPOSED METHODOLOGY

In our proposed system, we are proposing the decision tree based matching patient
records in multiple databases. We can enhance the efficiency of the record matching process
by introducing the decision tree approach. Undoubtedly, this system is can improve the
performance of the system. Among many decision tree methods, we are using the effective
decision tree approach called as ID3 tree technique.

[4.1] Standardization

Standardization included in the framework is a set of general domain-independent
transformation functions to resolve the different text formats of attributes or fields in the
records. For instance, abbreviation transformation replaces token with corresponding
abbreviation (e.g., Blvd, Boulevard), and Soundex transformation converts a token into a
Soundex code. Tokens that sound similar have the same code, etc. The transformation functions
are applied between sets of attribute values individually, i.e., first name with first name, HIC
(Health Insurance Code) number with HIC number. A lookup table for equivalent names can be
applied to help avoid not matching records when an equivalent name is used. The first name
can be looked up in the table to determine the comparable name.

[4.2] Blocking/Searching

If two data sets A and B are to be linked, the number of possible comparison equals the
product of the number of records in the two data sets. For example, if data set A contains
1,000,000 and data set B has 50,000,000 records. The total number of possible comparisons
would be 50,000,000,000,000. Assuming each comparison takes 0.01 seconds, it would take
500,000,000,000 seconds for all possible comparisons. The example demonstrates that it is
computationally intractable to consider all pairs when the data sets are large. To reduce the
large amount of possible record pair comparisons, blocking is used to bring only potentially
linkable record pairs together. This is obtained by using one or more record attributes to split
the data sets into blocks. Only records having the same value in the blocking variable are
compared. For text attributes, various phonic codes have been derived to avoid effects of
spelling and aural errors in recording names. This technique, however, becomes problematic
if a value in the blocking variable is recorded erroneously, and the corresponding record is
inserted into an incorrect block. To address this trouble, a number of iterations with different
blocking variables are normally performed.

[4.3] Levenshteine distance (LED) String Comparator

This method uses edit distance to compare the similarity of two strings. Edit distance, a
common measure of textural similarity, determines the minimum number of insertions,
deletions, and substitutions of single character required to change one string into another (i.e., make two strings equal). The edit distance is symmetric, it holds \(0 \leq d(x,y) \leq \max(|x|, |y|)\), where \(x, y\) represents the number of characters in the two strings & \(d(x,y)\) is the distance measure.

For instance: quickly
qucehkly

A simple character-wise comparison suggests that all letters after the ‘‘u’’ are incorrect. However, the final three (‘‘kly’’) appear correct, despite misalignment. The minimum string distance is 3. In fact, there are often multiple answers, because more than one minimum set of transformation may exist for the computed LED. Each transformation is called an ‘‘alignment’’, and represents a possible explanation of the error made. A dynamic programming algorithm is used to find the optimal edit distance. The time complexity of this algorithm can be an issue for large databases. The Levenshtein edit distance of two strings \((s_1, s_2)\) can be denoted as LED \((s_1, s_2)\). A similarity metric between two strings is constructed, ranging from 0 to 1.0 using a normalized formula:

\[
Sim(s_1, s_2) = 1 - \frac{\text{LED}(s_1, s_2)}{\text{MAXLEN}(s_1, s_2)}
\]

where MAXLEN denotes maximum numbers of characters in those two strings of length \(s_1\) and \(s_2\) and where LED is the Levenshtein edit distance, which is minimum number of deletions, insertions, and substitutions required to convert the contested string to presented on. From this formula, the similarity value of 1 represents the two compared strings are exactly the same, a perfect match, while value of zero indicates little similarity.

4.2 Record matching using Decision Tree

The basic idea of ID3 algorithm is to create a decision tree of given set, by using top-down greedy search to check each attribute at every tree node. To select the most useful attribute using classification technique, we present a metric—information gain and to catch an optimal way to classify an erudite set, we need to minimize the depth of the tree. Thus, we need some function which should be able to measure the most balanced splitting. The information gain metric is such a function that we can use for efficient balanced splitting. We have to use this concept of gain to rank attributes to build decision trees where at each node is located the attribute with utmost gain among the attributes that not yet considered in the path from the root. The purpose of this ordering is to create small decision trees so that records can be identified after only a few decision tree splitting and match a hoped for plainness of the process of decision making.

Each decision tree learner creates its own set of mapping rules from training examples labeled by the user. The mapping rules classify an example as mapped or not mapped. These classifications are used by the transformation weight learner for increasing the accuracy of the transformation weights, and are also needed for deciding which training examples should be labeled. Mapping rules contain information about which combination of attributes are important for determining the mapping between two objects, in addition to, the thresholds on
the similarity scores for each attribute. A number of mapping rules may be necessary to properly classify the objects for a specific domain application.

Algorithm 1: Record Matching using ID3 tree algorithm

The ID3 algorithm is used to build a decision tree, given a set non-categorical attributes C1, C2, .., Cn, the categorical attribute C, and a training set T of records.

Function ID3
(R: a set of non-categorical attributes,  
C: the categorical attribute,  
S: a training set)

Returns a decision tree;

begin

If S is empty, return a single node with value Failure;

If S consists of records all with the same value for the categorical attribute, return a single node with that value;

If R is empty, then return a single node with as value the most frequent of the values of the categorical attribute that are found in records of S; [note that then there will be errors, that is, records that will be improperly classified];

Let D be the attribute with largest Gain(D,S) among attributes in R;

Let {dj| j=1,2, .., m} be the values of attribute D;

Let {Sj| j=1,2, .., m} be the subsets of S consisting respectively of records with value dj for attribute D;

Return a tree with root labeled D and arcs labeled d1, d2, .., dm going respectively to the trees ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);

End ID3.

[5] EXPERIMENTAL RESULTS

We analyze and compare the performance offered by record matching using fuzzy decision and record matching using decision tree approach. The performance is evaluated by the parameters such as accuracy, precision, recall and f-measure. Based on the comparison and the results from the experiment show the proposed approach works better than the existing system.
[5.1] **Accuracy**

Accuracy can be calculated from formula given as follows

\[
\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{false positive} + \text{false negative} + \text{true negative}}
\]

From the graph we can see that, accuracy of the system is reduced somewhat in existing system than the proposed system. From this graph we can say that the accuracy of proposed system is increased which will be the best one.

[5.2] **Precision**

Precision value is calculated is based on the retrieval of information at true positive prediction, false positive. In healthcare data precision is calculated the percentage of positive results returned that are relevant.

\[
\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{false positive}}
\]
From the graph we can see that, precision of the system is reduced somewhat in existing system than the proposed system. From this graph we can say that the accuracy of proposed system is increased which will be the best one.

[5.3] Recall

Recall value is calculated is based on the retrieval of information at true positive prediction, false negative. In healthcare data precision is calculated the percentage of positive results returned that are Recall in this context is also referred to as the True Positive Rate. Recall is the fraction of relevant instances that are retrieved.

\[
Recall = \frac{\text{True positive}}{\text{True positive} + \text{false negative}}
\]

Recall means information retrieval. It is mathematically calculated by using formula. As usual in the graph X-axis will be methods such as existing and proposed system and Y-axis will be recall rate. From view of this recall comparison graph we obtain conclude as the proposed algorithm has more effective in recall performance compare to existing algorithms.

[5.4] F-measure

F-measure distinguishes the correct classification of document labels within different classes. In essence, it assesses the effectiveness of the algorithm on a single class, and the higher it is, the better is the clustering. It is defined as follows:

\[
F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]
From view of this F-measure comparison graph we obtain conclude as the proposed algorithm has more effective in F-measure performance compare to existing system.

[6] CONCLUSION

A major application focus was the patient record matching in third party payer databases. The literature provides numerous solution methods for quantifying the differences between two strings. On the other hand, selecting the best for patient record matching problem in the context of an integrated multiple valued logics have not been done. In our proposed system, with the intension of overcome the drawbacks of existing system such as time complexity of the system as well as lower accuracy of record matching system, we are proposing the effective record matching system using decision tree approach. This decision tree approach called ID3 method is used for our medical record matching application. The performance of the resulting decision models were evaluated through extensive experiments and found to perform very well.

REFERENCES


