An Effective Method for Matching Patient Records from Multiple Databases using Neural Network

Subitha.S
Research Scholar,
Department of Computer Science,
P.S.G.R. Krishnammal College for Women,
Coimbatore, India

S.C.Punitha
Assistant Professor, HOD,
Department of Computer Science,
P.S.G.R. Krishnammal College for Women,
Coimbatore, India

ABSTRACT
Record matching is the method of identifying records that denote the similar real world entity or item .The record matching method is helpful for matching health care data . Many problems occur while linking medical records from various databases. Comparing these medical data to other data is challenging because even small mistakes, for example data entry errors and lacking data. The earlier research proposed that estimate field matching represent a technique to solve the issue by finding similar string values in several representations. In our proposed system, we are proposing the Neural network based matching patient records in multiple databases. We can enhance the performance of the record matching method by introducing the Neural network approach. This technique is can improve the overall performance of the system. Among many Neural network techniques, we are using the Elman Back propagation network technique.

Keywords
Medical records, Record matching, Neural Network, Elman Back propagation.

1. INTRODUCTION
Record matching is the problem for identifying tuples in one or more relations that refer to the same real-world entity. This problem is also known as record linkage, merge-purge, duplicate detection and object identification. The need for record matching is evident.

The health care system has multiple legacy and information systems that support its health care professionals. The complexity of health care systems has necessitated the development of effective methods to manage the ever increasing volume of clinical, financial, demographic, and socioeconomic data[1]. Patient care data – all of it useful – are typically scattered across multiple departmental databases regardless of their size. Such as: Ex Emergency department admissions system, hospital's Admissions/Discharge/Transfer database, Pharmacy, Laboratory (onsite and offsite – contract labs), Heart Station – Cardiology Department (electrocardiographic and catheterization images), Billing, and Quality Improvement Department [2]. To make things worse, global applications require information from several databases in order to run. If these databases are independently managed, the same data is likely to be represented differently in these databases. Not only the values, but the semantics, the underlying assumptions and the integrity rules may differ as well. For this reason, databases can accrue a wide range of inaccuracies and inconsistencies such as misspelled names, inverted text, missing fields and outdated area or ZIP codes. For example, Kukich [3,4] found that the average error rate is 1–3% in typed data, 1–6% in optical character recognition (OCR) processed data, and 5–6% in data obtained by voice communication. Moreover, Elmagarmid et al. [5] reported that up to 25% of customer records are erroneous in a typical billing system, which results in substantial lost revenue opportunities. As a result, such data quality problems are a focus of increased attention [6].One common problem due to inconsistency in database is that data objects can exist in multiple variations of patients contacts or inconsistent text formats across multiple sources [11]. For example, a patient record may be saved in databases as ''Kate Simpson, Louisville, KY 40217'' and ''Kate Simson, Louisville, KY 40217''. This may cause duplicates in database systems and significantly increase direct costs, such as those associated with mailing. In addition, such inconsistencies may cause incorrect linkage of patient records With regard to the above linkage problem in database systems, many researchers have used record linkage or matching to create a frame, remove duplicates from files, or combine files, so that the relationships on two or more data elements from separate files can be obtained. Record matching can be divided into two categories: exact matching and statistical matching. Exact matching proposes to use identifiers such as 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 proposes to match a linkage of data for the same unit from different files based on similar characteristics rather than unique identifying information[7].

2. RELATED WORKS
In 1950s, New combe et al., developed concepts of record matching which were formalized in the mathematical type of Fellegi and Sunter [8].Fellegi and Sunter (1969) presented a specific mathematical structure for record linkage. To begin, notation is required. Two files A and B are matched. The concept is to categorize sets in an item space A × B from two files A and B into M, the list of true matches, and U, the pair of true no matches [8].

The standard probabilistic record linkage approach , as formalised in the 1960s by[8],possesses now a days been improved by applying the expectation maximization (EM) technique for best parameter valuation in record pair classification [9], and also by using estimated string evaluations to evaluate limited agreement weights while record attribute (field) values contain typographical differences[10], [11].Since the middle of the 1990s, research workers have examined a number of techniques to record linkage, deriving from artificial intelligence, database technology, information retrieval , machine learning , and data mining[12],[13],with the intension of improving the

linkage standard as well as the scalability of the record matching .

Three techniques for record pair classification have been established in TAILOR [13]: the initial is based on supervised decision tree method, the next is utilizing unsupervised k- means clustering approach (with three clusters, one each for matches, feasible matches and non-matches), and the lastly is a hybrid method that combination of the first two to deal with the issue of lack of training data. Unsupervised clustering methods are actually examined both to develop blocking and also for automatic record pair classification. Currently, unsupervised methods depending on relational clustering [14] have been discovered in the field of entity resolution of relational data. Xiaoyi Wang and Jiying Ling introduced multiple valued logic approach for matching medical records.

Though, non-relational data is still available in many real world applications, for example in databases that contain patient or customer information. In [15], [16], the PEBL and TC-WON techniques are proposed, which are both based on iteratively training a SVM using the positive and a selected set of strong negative examples.

3. PROPOSED METHODOLOGY

In our proposed system used Artificial Nerual Netowrk for matching the medical records. The datas are collected manually from the hospitals. Among many Nerual network methods the Elman Backpropogation network is used in this research work.Fig1 shows the process of the proposed work.

In the proposed system the first step is standardization. In this lookup table was created for equivalent names. Next step is blocking ,it is used to block the maximum number of values are used in the matching process. The next step is string comparator, here Levenshtein Edit Distance(LED) is used to compare the strings. Then calculate the fuzzy membership values, these membership values are used in the neural network. Here the Elman back propagation is used. Then the status of record is viewed. The status are “Favorable”, “Un Favorable”, “Somewhat Favorable” and “Somewhat Un Favorable”.

3.1 Standardization

The common methods of standardization are to change the many spelling differences of frequently occurring words with standard spellings like a fixed list of abbreviations or spellings .For example , In standardizing names , words of small distinguishing power such as "Corporation" or "Limited" are replaced with consistent abbreviations such as "CORP" and "LTD ," respectively . First name spelling variations such as "Rob" and "Bobbie" might be replaced with a consistent assumed original spelling such as "Robert" or an identifying root word such as "Robt" because "Bobbie" might refer to a woman with "Roberta" as her legal first name.

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.

3.2 Blocking/Searching

To minimize the large amount of feasible record pair comparisons, blocking is used to bring only potentially linkable record pairs together. This is accomplished by using one or more record attributes to separate the data sets into blocks. Only records having the similar value in the blocking variable are compared.

3.3 String Comparator-LED(Levenshtein Edit Distance)

The Levenshtein edit distance is a string metric for measuring the difference between two sequences. Levenshtein distance may also be referred to as edit distance. 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) [19]. Mathematically, the Levenshtein distance between two strings a,b is given by $lev_{a,b}(|a|,|b|)$ where

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(ai \neq bj)} \end{cases} & otherwise. \end{cases}$$

where $1_{(ai \neq bj)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise.The Levenshtein edit distance of two strings (s1, s2) can be denoted as LED (s1, s2). A similarity metric between two strings is constructed, ranging from 0 to 1.0 using a normalized formula:

$$\mathrm{Sim}(S_{1,}\, S_{2,}) = 1 - (\mathrm{LED}(S_{1,}S_{2}))/(\mathrm{MAXLEN}(S_{1,}S_{2}))$$

where MAXLEN denotes maximum numbers of characters in those two strings of length s1 and s2 and where LED is the Levnshtein edit distance, which is minimum number of deletions, insertions, and substitutions required to convert the contested string to presented on.

The maximum difference in this comparison of the two strings is the length of the longest string, the similarity is in scale of [0, 1].

3.4 Matching Process using ElmanNN

Elman models are two layered back propagation networks, with the addition of a feedback connection from the output of the hidden layer to its input [17]. This feedback path allows Elman networks to learn to recognize and generate temporal patterns, as well as spatial patterns. The ENN usually uses the Back-Propagation (BP) based algorithms to deal with the various problems Elman network are also known as "simple recurrent networks" (SRN).

The advantage of this feedback path is that it allows the ENN to recognize and generate temporal patterns and spatial patterns. This means that after training, interrelations between the current input and internal states are processed to produce the output and to represent the relevant past information in the internal states. As a result, the ENN has been widely used in various fields which includes classification, prediction and dynamic system identification.

3.4.1Elman Back propagation Algorithm:

Suppose have a fixed training set $\{(x^{(1)}, y^{(1)}), \ldots\ldots, (x^{(m)}, y^{(m)})\}$ of m training examples. Train our neural network using batch gradient descent. In detail, for a single training example (x,y), define the cost function with respect to that single example to be:

$$J(W, b; x, y) = \frac{1}{2}||h_{w,b}(x) - y||^2$$

The following describes the Back propagation algorithm

Initialize all weights with small random numbers, typically between -1 and 1

repeat
for every pattern in the training set
Present the pattern to the network

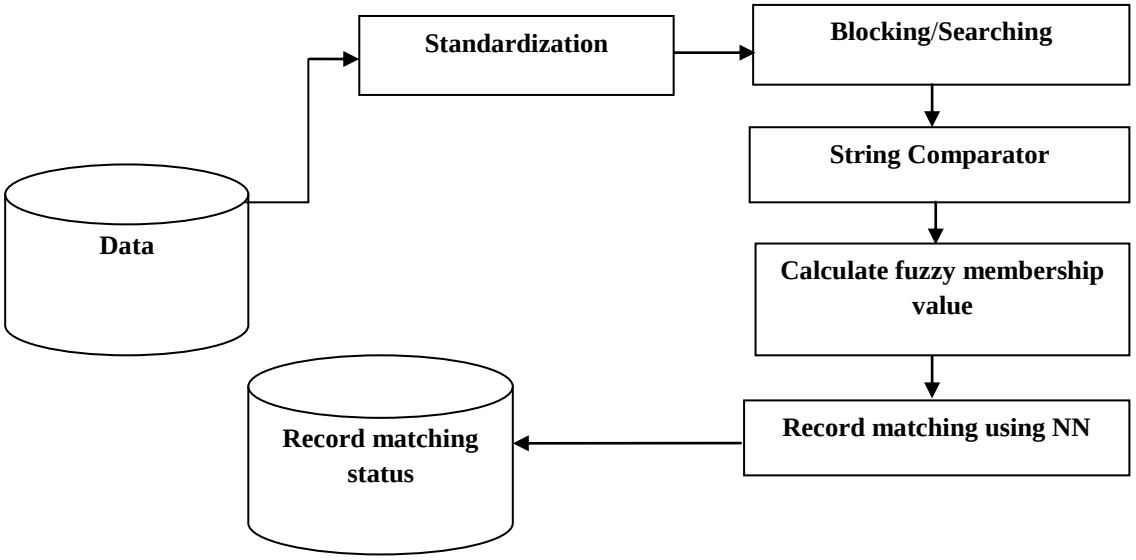

Fig 1: Neural network approach process diagram

// Propagated the input forward through the network:
for each layer in the network
for every node in the layer
1. Calculate the weight sum of the inputs to the node
2. Add the threshold to the sum
3. Calculate the activation for the node
end
end

// Propagate the errors backward through the network
for every node in the output layer
calculate the error signal
end

for all hidden layers
for every node in the layer
1. Calculate the node's signal error
2. Update each node's weight in the network
end
end

// Calculate Global Error
Calculate the Error Function

end

4. EXPREMENTAL RESULTS

Analyze and compare the performance offered by record matching using fuzzy logic and decision tree approach with Neural network method. 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.

4.1Accuray

Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

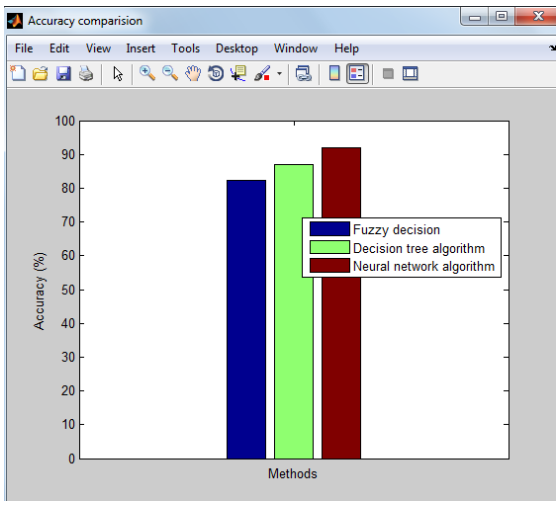

Fig2: Accuracy comparison chart

From the graph the accuracy of the system is reduced somewhat in existing system than the proposed system. From

this graph the accuracy of proposed system is increased which will be the best one.

4.2 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.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

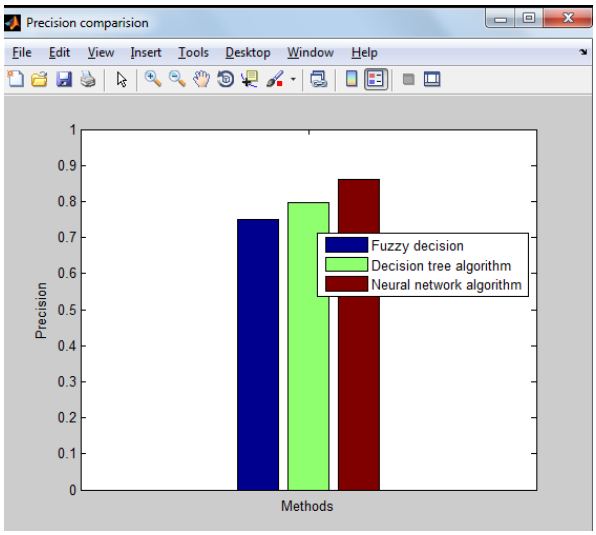

Fig3: Precision comparison chart

From the graph the precision of the system is reduced somewhat in existing system than the proposed system. From this graph the accuracy of proposed system is increased which will be the best one.

4.3 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,

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

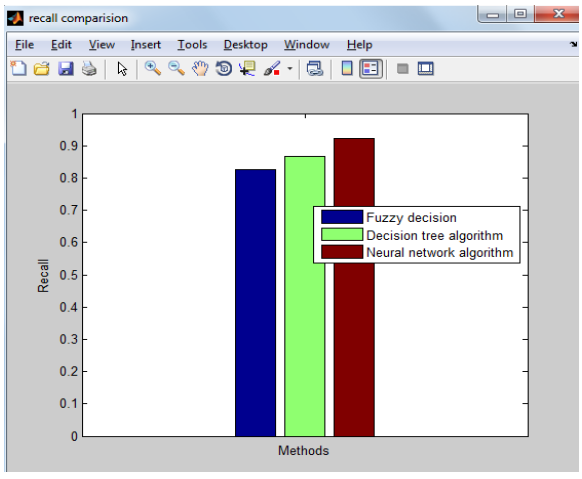

Fig4:Recall comparison chart

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 conclude as the proposed algorithm has more effective in recall performance compare to existing algorithms.

4.4 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:

$$\text{F} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

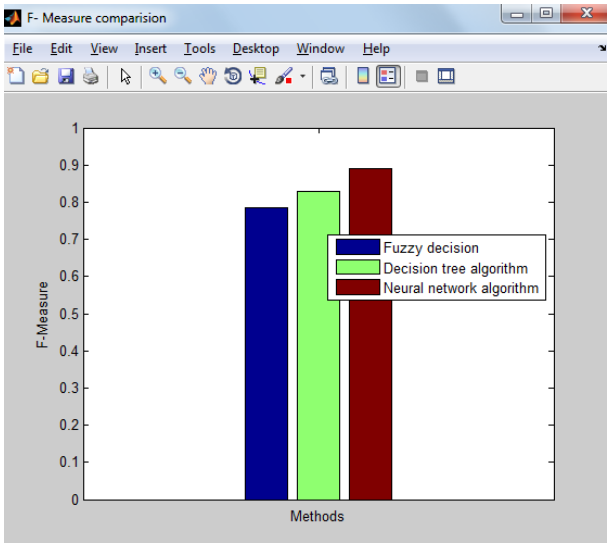

Fig5: F-Measure comparison chart

From view of this F-measure comparison graph conclude as the proposed algorithm has more effective in F-measure performance compare to existing system.

Table 1. Comparison table for the classifiers

Classifiers	Accuracy	Precision	Recall	F-measure
Fuzzy logic	83%	0.75	0.95	0.82
Decision tree	88%	0.79	0.96	0.88
Neural network	93%	0.85	0.98	0.92

5. 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, this work proposing the effective record matching system using decision tree approach. The performance of the resulting decision models were evaluated through extensive experiments and found to perform very well.

6. REFERENCES

[1] Xiaoyi Wang,Jiying Ling Multiple valued logic approach for matching patient records in multiple databases, Journal of Biomedical Informatics 45 (2012) 224–230.

[2] Verykios VS, Elmagarmid AK, Houstis EN. Automating the approximate record matching process. Inform Sci 2000;126:83–98.

[3] Kukich K. Techniques for automatically correcting words in text. ACM ComputSurv 1992;24(4):377–439

[4] Kukich K. Spelling correction for the tele-communications network for the deaf. Commun ACM 1992;35(5):80–90.

[5] Elmagarmid AK, Horowitz B, Karabatis G, Umar A. Issue in multisystem integration for achieving data reconciliation and aspects of solution. Technical report, Bellcore Research; 1996

[6] Trillium Software. How data profiling & analysis saves companies $millions.White paper in data integration and data quality management. <http://www.b-eye-network.com/view/4078>; 2003 [cited 11.08.11].

[7] Herzog TN, Scheuren FJ, Winkler WE, editors. Data quality and record linkage techniques. Washington, D.C.: Bureau of the Census; 2007.

[8] Fellegi and A. Sunter, A Theory for Record Linkage, Journal of the American Statistical Association 64 (1969), no. 328, 1183-1210.

[9] W. E. Winkler. Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. Technical Report RR2000/05, US Bureau of the Census, 2000.

[10] A. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering, 19(1):1–16, 2007

[11] W. E. Winkler. Overview of record linkage and current research directions. Technical Report RR2006/02, US Bureau of the Census, 2006.

[12] P. Christen and K. Goiser. Quality and complexity measures for data linkage and deduplication. In F. Guillet and H. Hamilton, editors, Quality Measures in Data Mining, volume 43 of Studies in Computational Intelligence. Springer, 2007

[13] M. Elfeky, V. Verykios, and A. Elmagarmid. TAILOR: A record linkage toolbox. In ICDE'02, pages 17–28, San Jose, 2002.

[14] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 2007.

[15] H. Yu, J. Han, and K. C.-C. Chang. PEBL: Positive example based learning for Web page classification using SVM. In ACM KDD'02, pages 239–248, Edmonton, 2002

[16] H. Yu, C. X. Zhai, and J. Han. Text classification from positive and unlabeled documents. In CIKM'03, pages 232–239, New Orleans, 2003

[17] Sotir Sotirov, Modelling the backpropagation algorithmof the Elman neural network by a generalized net, 13th Int.Workshop on Generalized Nets London, 29 October 2012, 49–55.