International Journal of Soft Computing 8 (3): 181-187, 2013
ISSN: 1816-9503


# Automatic Threshold Selection using PSO for GA based Duplicate Record Detection

[1]Deepa Karunakaran, [2]Rangarajan Rangaswamy and [1]Senthamil Selvi Marudavelu
[1]Department of Information Technology,
Sri Ramakrishna Engineering College, Coimbatore, India
[2]Indus College of Engineering, Coimbatore, India

**Abstract:** Normally setting the threshold is an important issue in applications where the similarity functions are used and it relies more on human intervention. The proposed research addressed two issues: first to find the optimal equation using Genetic Algorithm (GA) and next it adopts an intelligence algorithm, Particle Swarm Optimization (PSO) to get the optimal threshold to detect the duplicate records more accurately and also it reduces human intervention. Restaurant and Cora data repository are used to analyze the proposed algorithm and the performance of the proposed algorithm is compared against marlin method and the genetic programming with the help of evaluation metrics.

**Key words:** GA, PSO, similarity metrics, Threshold, Cora

## INTRODUCTION

Normally, organizations become conscious of practical precise disparities or inconsistencies while integrating data from diverse sources to implement a data warehouse. Such problems belong to the category called data heterogeneity (Elmagarmid *et al*., 2007). Erroneous duplication of data occurs when information from diverse data sources is integrated (Chaudhuri *et al*., 2007). But errors like spelling mistakes, conflicting customs across data sources, omitted fields, etc., normally exist in the data accepted at the data warehouse from external sources. Incoming data tuples from external sources need validation and refinement in order to provide high data quality (Chaudhuri *et al*., 2003).

Data cleaning, otherwise known as data cleansing or scrubbing is the process of identifying and eliminating errors and discrepancies from data so as to enrich the quality of data (Rahm and Do, 2000). Data cleaning can also be described as the process of recognizing and removing errors from a data warehouse. Data cleaning plays a significant role in the process of data mining. A number of organizations require quality data. It is necessary to improve the quality of data in a data warehouse prior to the data mining process (Chapman, 2005). Quality of data can be through data cleaning methods. Numerous data cleaning techniques are being employed for different purposes. Data cleaning methodologies in existence are employed to recognize record duplicates, missing values, record and field similarities and duplicate elimination (Sarawagi *et al*., 2002). Similarities among records and fields are identified using similarity functions (Monge and Elkan, 1996). Duplicate elimination functions are employed to identify if two or more records indicate the same real world object (Bitton and Dewitt, 1983).

The recent researches have given many methods for the deduplication purposes with many distinct features by their own. The proposed research addressed two issues: first to find the optimal equation using GA and next issue to get an optimal threshold using particle swarm optimization. Here, specific similarity metrics is used to calculate similarity values among the fields and each such value are combined to form a feature vector. This vector is able to identify whether two entries in a database are duplicates or not. Since, duplicate detection process is a time consuming process, the aim is to propose a method that finds a proper combination of the elements in the feature vector thus yielding a function that maximizes performance for training purposes. Then, this function can be used on the remaining testing data. Each expression requires the optimal threshold value in order to classify the duplicate and non duplicate entries. In this case, threshold definition is the main problem. Normally,

**Corresponding Author:** Deepa Karunakaran, Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, India

thresholds are set by the users based on the necessity of specific applications and the optimal thresholds are often found by either minimizing or maximizing an objective function regarding the values of the thresholds. In order to find the optimal threshold and to reduce the human intervention, the proposed research uses an intelligence algorithm, PSO. The performance is compared against a state of the Art Method Bilenko based on support vector machines and De Carvalho based on genetic programming.

**Literature review:** A structure to improve the duplicate record detection using trainable measures of textual similarity was proposed by Bilenko and Mooney (2003). They utilized a learnable text distance functions for each database field and illustrated that such measures are proficient enough to adapt to the precise notion of similarity that was suitable for the field's domain. An extended modification of learnable string edit distance and a vector-space based measure that utilizes a Support Vector Machine (SVM) for training were the two learnable text similarity measures applied in this approach. Their experimental results proved that the accuracy was improved over conventional methodologies.

De Carvalho *et al.* (2012) have proposed a genetic programming approach to record deduplication. This approach automatically proposes duplicate record detection function by combining several pieces of evidence taken from the data. This function is able to identify whether the two records in a repository are same or not.

Most of the deduplication process requires similarity function which address whether the two entries are duplicate or not by setting the threshold. Dos Santos *et al.* (2011) proposed a method to assess the quality of a similarity function at different threshold and choose an appropriate threshold for a specific application. The estimation process depends on on a clustering and silhouette coefficient is used to choose the similarity threshold. Results proved the effectiveness of the proposed approach.

Ye *et al.* (2008) employed a new method to select image threshold automatically based on PSO algorithm. The performance of this algorithm is compared with Otsu and results showed that PSO algorithm produce promising results in terms of the quality of solution found and the processing time required.

Isele and Bizer (2011) addressed an important problem in linked data which detect links between entities which identifies the same real world object. Normally, links are made based on manually by writing linkage rules. This approach automatically generates linkage rules from a set of reference links and uses genetic programming. De Freitas *et al.* (2010) presented the Active Learning GP (AGP), a semi-supervised GP for the data deduplication problem. AGP uses an active learning approach in which a committee of multi-attribute functions votes for categorizing record pairs as duplicates or not. When the voting is not enough to predict the class of the data pairs then a user is called in order to resolve the conflict. Results showed that AGP guarantees the quality of the deduplication.

Qingwei *et al.* (2010) have proposed a method to search the optimized partial contents which is the most similar in two documents using hybrid mutation PSO algorithm. A new related coefficient of strings is defined for strings similarity and the design of new evaluation function is based on the related coefficient function. Goncalves *et al.* (2010) proposed an approach based on a deterministic technique which automatically recommends training examples for a deduplication method based on genetic programming.

## MATERIALS AND METHODS

### Basic concepts of GA and PSO

**Genetic algorithm:** Genetic Algorithm (GA) is an adaptive heuristic search algorithm based on the evolutionary concepts of natural selection and genetics. GA exploit historical information to direct the search into the region of better performance within the search space. The basic techniques of the GAs are planned to put on the processes in natural systems required for evolution; especially those follow the principles first set by Charles Darwin of "survival of the fittest".

In this algorithm, a population of strings (called chromosomes or the genotype of the genome) which encode candidate solutions (called individuals, creatures or phenotypes) to an optimization problem, progresses toward better solutions. Traditionally, solutions are represented in binary as strings of 0's and 1's but other encodings are also possible. The evolution starts from a population of randomly generated individuals. In each generation, evaluate the fitness for each individual in the population. Based on the fitness value, multiple individuals are stochastically selected from the current population, crossover and randomly mutated to form a new population. The new population is then used in the next iteration of the algorithm. An algorithm terminates when either a maximum number of generations has been produced or a satisfactory fitness level has been reached.

**Simple genetic algorithm procedure:**

Choose the initial population of individuals
Evaluate the fitness of each individual in that population
Repeat until termination (time limit, sufficient fitness achieved, etc.)
Select the best-fit individuals for reproduction
Breed new individuals through crossover and mutation operation to give birth to offspring
Evaluate the individual fitness of new individuals
Replace least-fit population with new individuals

**Particle swarm optimization:** Particle Swarm Optimization (PSO) is an optimization technique which provides an evolutionary based search. This search algorithm was introduced by Dr. Russ Eberhart and Dr. James Kennedy in 1995. PSO is a computational method that optimizes a problem by iteratively trying to improve a candidate solution by a given measure of quality. PSO optimizes a problem by having a population of candidate solutions and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity. Each particle's movement is influenced by its local best known position and is also guided toward the best known positions in the search-space which are updated as better positions are found by other particles. This is expected to move the swarm toward the best solutions. The outline of PSO is stated as follows:

- Create a population of agents (called particles) uniformly distributed over X
- Evaluate each particle's position according to the objective function
- If a particle's current position is better than its previous best position, update it
- Determine the best particle according to the particle's previous best positions
- Update particles' velocities according to:

$$velocity_{ex} = velocity^{0} + \varphi.(pbest - pos^{0}) + \phi(gbest - pos^{0})$$

Where:

$velocity^{0}$ = Current velocity
pbest = Current best position
$pos^{0}$ = Current position
gbest = Global best position
$\varphi$, $\phi$ = Random values in range [0, 1]

- Move particles to their new positions according to:

$$pos = pos^{0} + velocity_{ex}$$

- Go to step 2 until stopping criteria are satisfied

**Proposed approach on duplicate record detection**

**Step 1 (Similarity computation for all pair of records):** Similarity functions compute the similarity of each field with other record field and assign a similarity value for each field. The similarity metrics used in the proposed research are Levenshtein distance and cosine similarity since researchers compared the results with the ones presented by De Carvalho *et al.* (2012) and Bilenko *et al.* (2003).

**Levenshtein distance:** The chosen name fields of the records are 1 and 2. The Levenshtein distance is computed by calculating the minimum number of operations that has to be made to transform one string to the other usually these operations are: replace, insert or deletion of a character. The Levenshtein distance between the records is finding out by considering the record as a whole.

**Cosine similarity:** The cosine similarity between the two records name field 1 and 2 are calculated as follows: first, the dimension of both strings are obtained by taking the union of two string elements in the record 1 and 2 as (word1, word2, ..., wordN) and then the frequency of occurrence vectors of the two elements are calculated, i.e., record 1 = (<vector value1>, <vector value2>, ..., <>) and record 2 = (<vector value1>, <vector value2>,..., <>). Finally, researchers obtain the dot product and magnitude of both strings.

**Step 2 (Generate feature vector):** In this approach, feature vectors represent the set of elements that is required for the detection of duplicate elements from the data repository. Each element represent the similarity function applied on the values of a specific attribute (e.g., <name, levenshtein>, <address, levenshtein> and <city, levenshtein>). Develop a set of expressions by using these vectors elements with the simple mathematical functions (+, ×, -, /).

**Step 3 (Optimized expression):** A set of such expression are supplied as input to GA to find best among the supplied inputs which is capable of providing better solution for the problem. The optimization algorithm PSO find the optimal threshold for each expression.

**Population:** Initialize the population with user provided individuals. Here, set of expressions is considered as an initial population which is (a + b) + (c + d), (a + b)×(c + d), (a-b) + (c + d), (a + b)×(c−d), (a + b)−(c−d). The a, b, c, d corresponds to similarity values defined on the attributes name, address, phone number and category, respectively.

**Fitness:** The fitness value is a value generated from the fitness function which is one of the most important components in this process. If the fitness function is badly chosen then it will surely fail to find the best expression. In this approach, researchers have used F1 metric as the fitness function and can be calculated as.

**Precision (P):** It is defined as the fraction of identified duplicate pairs that are correct:

$$\text{Precision (P)} = \frac{\text{Number of true duplicate records identified}}{\text{Total number of duplicate records identified}}$$

**Recall (R):** It is the fraction of actual duplicate pairs that are identified correctly in the input dataset:

$$\text{Recall(R)} = \frac{\text{Number of true duplicate records identified}}{\text{Total number of duplicate records present in the dataset}}$$

Recall can be seen as a measure of completeness whereas precision is a measure of exactness or fidelity.

**F1-measure (F):** It gives equal weight to both precision and recall and it is the harmonic mean of precision and recall. The traditional F-measure or balanced F-score is computed as:

$$F = \frac{(2\times R\times P)}{(R+P)}$$

Likewise find the fitness value for each expression in the population based on threshold value. Since, F1-value varies with different threshold, it is necessary to choose an optimized threshold to classify the dataset as duplicates and non-duplicates accurately. Hence, researchers applied one of the best intelligence swarm algorithm named Particle Swarm Optimization to find global optimal solution.

**Optimal threshold using PSO:** In PSO, a population starts with a random set of threshold (particle) on each expression. The position of the particle refers to the possible solutions to be optimized. Next, find the fitness value for each such particle using F1 metric and determine particle best (Pbest) and global best (Gbest). The particles move towards the optimal area by updating their position and the velocity as follows:

$$\text{velocity}_{ex} = \text{velocity}^0 + \varphi.(\text{pbest} - \text{pos}^0) + \phi(\text{gbest} - \text{pos}^0)$$

Where:

$\text{velocity}^0$ = Current velocity
pbest = Current best position
$\text{pos}^0$ = Current position
gbest = Global best position
$\varphi$, $\phi$ = Random values in range [0, 1]

$$\text{pos} = \text{pos}^0 + \text{velocity}_{ex}$$

Thus, PSO select the best threshold value on each such expression which classify the set of records as duplicate and non-duplicate.

**New population generation:** Select the best n expressions having high fitness value and apply crossover and mutation to generate new set of population. Repeat the process until termination criteria is reached. Once the optimal expression has obtained during the training phase, the duplicate detection of testing datasets is done with the help of the same expression.

## RESULTS AND DISCUSSION

**Dataset description:** The proposed approach used two real datasets namely restaurant and Cora which are commonly employed for evaluating duplicate record detection approaches.

**Restaurant:** This dataset contains 864 entries including 112 duplicates that were obtained by integrating records from Fodor and Zagat's guidebooks. Attributes used are names, address, city and speciality.

**Cora bibiliographic:** This dataset contains 864 entries including 112 duplicates that were taken from riddle repository. Attributes used are researchers names, year, title, venue and other information.

**Comparative analysis:** This study provides a comparative analysis of the proposed algorithm using Levenshstein distance and cosine similarity method with Marlin Method (Bilenko and Mooney, 2003) and the genetic programming method (De Carvalho *et al*., 2012). The analysis is based on F1-measure, training time and testing time of the algorithm.

**Experiment 1:** In experiment 1, the F1-measure of the proposed research using two similarity measures such as Levenshtein distance and cosine similarity is compared against Marlin, a state of the art of SVM Method (Bilenko and Mooney, 2003) and genetic programming method on restaurant dataset. From Fig. 1, it is clear that F1 value of the proposed method is 8% more than the Marlin and static tie on GP Method on both similarity measures.

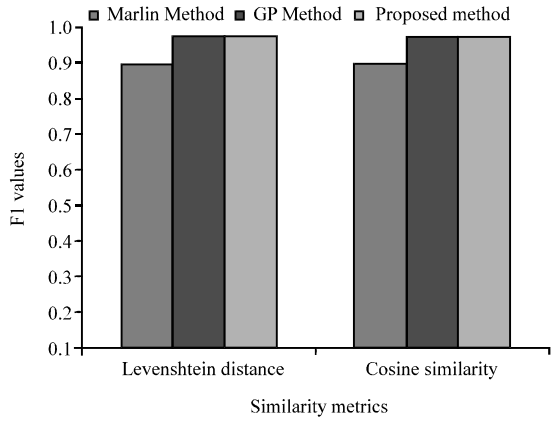


Fig. 1: F1 values vs. similarity metrics on restaurant

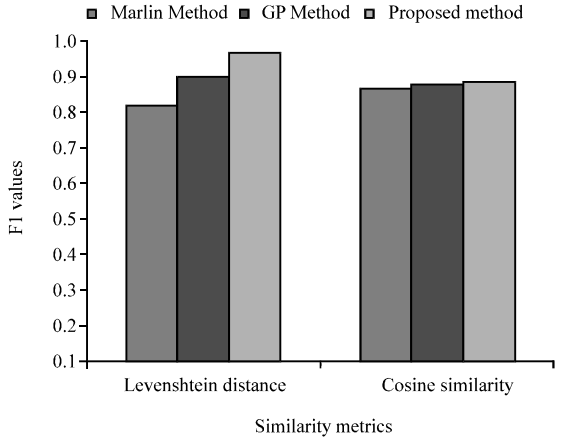


Fig. 2: F1 values vs. similarity metrics on Cora

**Experiment 2:** The F1 value of the proposed research using two similarity measures such as Levenshtein distance and cosine similarity is compared against Marlin, a state of the art of SVM Method and genetic programming method on Cora dataset. From Fig. 2, it is clear that F1 value of the proposed method out performed Marlin by 15% on Levenshtein and 2% on cosine metric. F1 value of proposed research is 7% more than GP on Levenshtein metric and tie on cosine.

**Experiment 3:** Figure 3 and 4 shows the precision and recall values for both measures on restaurant dataset and Cora dataset. It can be observed that the proposed duplicate detection approach is very efficient on both dataset by achieving high precision and recall.

**Experiment 4:** Time taken for duplicate records detection differ for different input records and also vary with the number of records. Hence, time incurred usually vary with the number of records in the input dataset. If more number

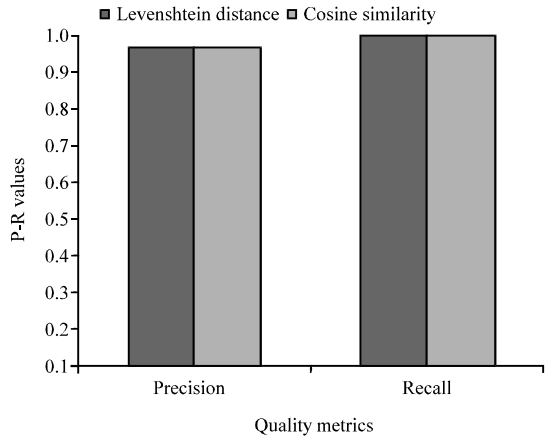


Fig. 3: Evaluation metric obtained for the proposed approach on restaurant

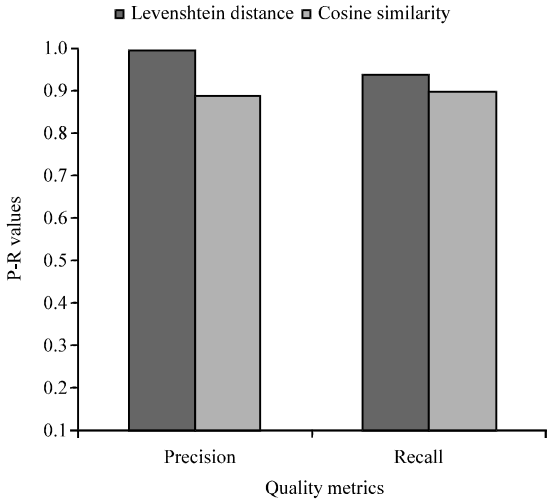


Fig. 4: Evaluation metric obtained for the proposed approach on Cora

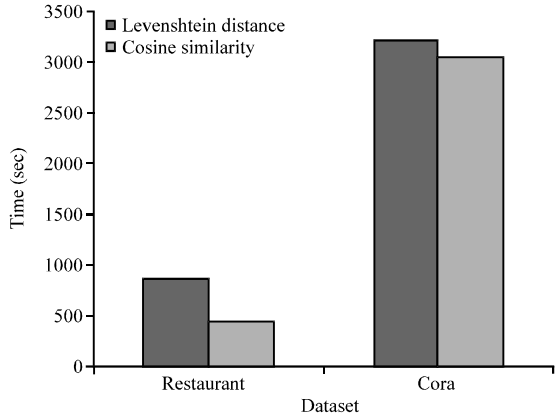


Fig. 5: Training time consumption of the proposed approach

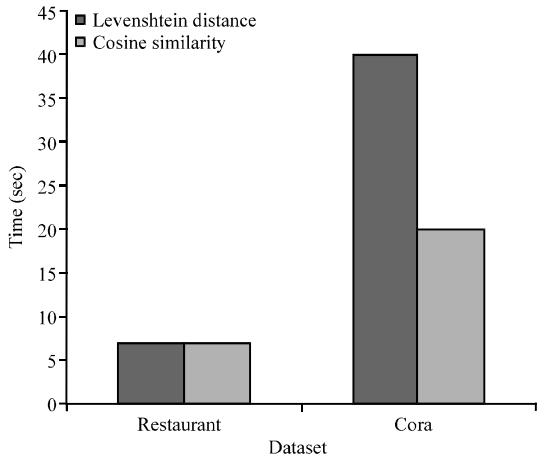


Fig. 6: Testing time consumption of the proposed approach

of records in the input dataset means it takes more time for comparison process. Figure 5 and 6 shows the proposed approach consumes minimum training and testing time on both dataset.

## CONCLUSION

The deduplication has been one of the most emerging techniques for data redundancy and duplication. The duplication creates lots of problems in the information retrieval system. This approach uses GA and PSO to find the optimal expression and optimal threshold, respectively in duplicate record detection problem. The experimentation of the proposed algorithms showed significant results. Both restaurant and Cora dataset have been used to evaluate the performance of the algorithm and the results showed that the proposed algorithm has better results than the Marlin Method and tie against genetic programming method.

## REFERENCES

Bilenko, M. and R.J. Mooney, 2003. Adaptive duplicate detection using learnable string similarity measures. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2003, Washington, DC., USA., pp: 39-48.

Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar and S. Fienberg, 2003. Adaptive name matching in information integration. IEEE Intell. Syst., 18: 16-23.

Bitton, D. and D.J. Dewitt, 1983. Duplicate record elimination in large data files. ACM Transac. Database Systems, 8: 255-265.

Chapman, A.D., 2005. Principles and methods of data cleaning-primary species and species-occurrence data. Version 1.0, Report for the Global Biodiversity Information Facility, Copenhagen.

Chaudhuri, S., A.D. Sarma, V. Ganti and R. Kaushik, 2007. Leveraging aggregate constraints for deduplication. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 12-14, 2007, New York, USA., pp: 437-448.

Chaudhuri, S., K. Ganjam, V. Ganti and R. Motwani, 2003. Robust and efficient fuzzy match for online data cleaning. Proceedings of the ACM SIGMOD International Conference on Management Data, June 9-12, 2003, San Diego, California, pp: 313-324.

De Carvalho, M.G., A.H.F. Laender, M.A. Goncalves and A.S. da Silva, 2012. A genetic programming approach to record deduplication. IEEE Trans. Knowledge Data Eng., 24: 399-412.

De Freitas, J., G.L. Pappa, A.S. da Silva, M.A. Goncalves and E. Moura *et al*., 2010. Active learning genetic programming for record deduplication. Proceedings of the IEEE Congress on Evolutionary Computation, July 18-23, 2010, Barcelona, Spain, pp: 1-8.

Dos Santos, J.B., C.A. Heuser, V.P. Moreira and L.K. Wives, 2011. Automatic threshold estimation for data matching applications. Inform. Sci., 181: 2685-2699.

Elmagarmid, A.K., P.G. Ipeirotis and V.S. Verykios, 2007. Duplicate record detection: A survey. IEEE Trans. Knowledge Data Eng., 19: 1-16.

Goncalves, G.S., M.G. de Carvalho, A.H.F. Laender and M.A. Goncalves, 2010. Automatic selection of training examples for a record deduplication method based on genetic programming. J. Inform. Data Manage., 1: 213-228.

Isele, R. and C. Bizer, 2011. Learning linkage rules using genetic programming. Proceedings of the 6th International Workshop on Ontology Matching, October 23-24, 2011, Bonn, Germany, pp: 13-24.

Monge, A.E. and C.P. Elkan, 1996. The field matching problem: Algorithms and applications. Proceedings of the 2nd SIGKDD, Portland, Oregon, USA.

Qingwei, Y., W. Dongxing, Z. Yu and W. Xiaodong, 2010. The duplicated of partial content detection based on PSO. Proceedings of the IEEE 5th International Conference on Bio-Inspired Computing: Theories and Applications, September 23-26, 2010, Changsha, China, pp: 350-353.

Rahm, E. and H.H. Do, 2000. Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23: 1-11.

Sarawagi, S., A. Bhamidipaty, A. Kirpal and C. Mouli, 2002. ALIAS: An Active learning led interactive deduplication system. Proceedings of the 28th International Conference on Very Large Databases, August 20-23, 2002, Hong Kong, China, pp: 1103-1106.

Ye, Z., H. Chen, W. Liu and J. Zhang, 2008. Automatic threshold selection based on particle swarm optimization algorithm. Proceedings of the International Conference on Intelligent Computation Technology and Automation, Volume 1, October 20-22, 2008, Hunan China, pp: 36-39.