Journal of Computer Science 7 (3): 341-347, 2011
ISSN 1549-3636

A Comparative Study in Classification Techniques for Unsupervised Record Linkage Model

Mohammadreza Ektefa, Fatimah Sidi, Hamidah Ibrahim,
Marzanah A. Jabar and Sara Memar
Department of Computer Science,
University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Abstract: Problem statement: Record linkage is a technique which is used to detect and match duplicate records which are generated in data integration process. A variety of record linkage algorithms with different steps have been developed in order to detect such duplicate records. To find out whether two records are duplicate or not, supervised and unsupervised classification techniques are utilized in different studies. In order to utilize the supervised classification algorithms without consuming a lot of time for labeling data manually, a two step method which selects the training data automatically has been proposed in previous studies. However, the effectiveness of different classification techniques is the issue which should be taken into accounts in record linkage systems in order to classify records more accurately. **Approach:** To determine and compare the effectiveness of different supervised classification techniques in an unsupervised manner, some of the prominent classification methods are applied in duplicate records detection. Duplicate detection and classification of records in two real world datasets, namely Cora and Restaurant is experimented by Support Vector Machines, Naïve Bayes, Decision Tree and Bayesian Networks which are regarded as some prominent classification techniques. **Results:** As experimental results show, while Support Vector Machines outperforms with F-measure of 96.27% in Restaurant dataset, for Cora dataset, the effectiveness of Naïve Bayes is the best and it leads to an improvement with F-measure of 89.7%. **Conclusion/Recommendation:** The result of detecting duplicate records with different classification techniques tends to fluctuate depending on the dataset which is used. Moreover, Support Vector Machines and Naïve Bayes outperform other methods in our experiments.

Key words: Record linkage, duplicate detection, classification techniques, Optical Character Recognition (OCR), Longest Common Subsequence (LCS), data integration, support vector machines, heterogeneous data, ID3 algorithm, Bayesian network

INTROCUDTION

Data integration is defined as the process of merging data from various databases and sources such as flat files, data cube and databases into a coherent source like data warehouse. Data integration is using vastly in current information systems. Since heterogeneous data sources have different formats and standards, a real world entity may be presented with different styles in each of these sources. Moreover, data entry mistakes such as typing errors or utilizing Optical Character Recognition (OCR) can also cause different presentations of the same object. So, these matters lead to duplication which is considered as one of the major problems of data quality. Hence, finding such duplicates in order to make a proper decision to handle them is an essential requirement of information systems. This task is also known as record linkage.

Record linkage also known as citation matching (McCallum *et al*., 2000), authority control (Warnner and Brown, 2001), object matching (Surajit *et al*., 2003) and entity resolution (Sarawagi and Bhamidipaty, 2002), is a difficult and heavy step of data integration. The goal of record linkage is to find, match and aggregate duplicate tuples in an integrated database. Record linkage is vastly used in different contexts including Digital Libraries, bioinformatics and business customer information. Moreover, it is also a common pre-processing step of mining projects.

Web datasets are often lack proper quality; so, finding duplicate records in such databases is a challenging task. The bibliographic entities in online digital libraries can be mentioned as an example.

Corresponding author: Mohammadreza Ektefa, Department of Computer Science, University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Table 1: A sample of duplicate data

Journal name	Automating the Approximate record matching process
Authors	
1.	Vassilios S.Verykios, Ahmed K.Elmagarmid, Elias N.Houstis
2.	Verykios V.S., Elmagarmid A.K., Houstis E.N.
3.	VS Verykios, AK Elmagarmid
4.	VS Verykios, AK Elmagarmid, EN Houstis
5.	Vassilios S. Verykios , Ahmed K. Elmagarmid
Journal name	
1.	INF. SCI.
2.	Information Sciences-Informatics and Computer Science: an International Journal
3.	Information Sciences

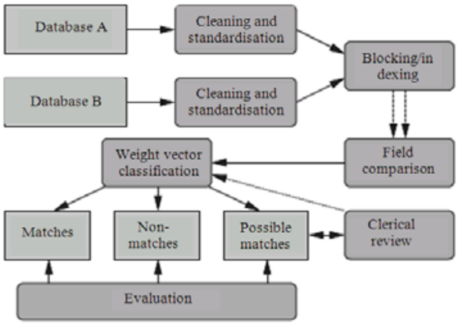

Fig. 1: Record Linkage Steps (Christen, 2008)

Table 1, which is the bibliography result of same entity collected from different web sources, illustrates this problem. As can be seen from Table 1, there are different values pointing to the same entity.

Several record linkage models have been proposed in different studies. Figure 1 shows some common steps of record linkage process. This model consists of several steps such as data cleaning, blocking, field comparison and classification. Some usual problems of real-world databases are noisy, incomplete and incorrect data (Churches *et al*., 2002). Data cleaning is an initial pre-processing step to modify such mentioned problems.

Comparison is an important step which has been applied in all record linkage frameworks. In order to detect duplicate records in two input dataset, all records from the first dataset should be compared with the second one. This task can involve a bulk amount of comparisons which makes the task inefficient. To solve such a mentioned problem, blocking techniques are used in record linkage frameworks to decrease the number of comparisons by putting more similarly duplicate records in the same block. Only the content of each block will be compared together in the next step.

Determining a similarity function and matching records are two important steps applied in most record linkage frameworks. Records which are in the same block are compared together by similarity functions. Comparison similarity functions can be a simple string function or a complicated combination of several functions. The results of applying comparison functions are scores (also named weight vectors) which show the degree of similarity between record pairs.

Classification techniques are applied on the results of the previous step in order to classify the records in three classes: match, non-match and possibly match. To classify the records, supervised and unsupervised classification techniques have been utilized in various studies. One of the technique for classifying the record pairs is to separate the training data by selecting the record pairs with the highest and the lowest similarity scores as match and non-match classes, respectively. Other pairs are considered as possibly match records and are classified by data mining techniques based on known match and non-match samples.

Initially, the concept of record linkage was presented by Newcombe *et al*. (1959) in the context of medical records. Fellegi and Sunter (1969) proposed an EM-Based method to determine error rates and set matching parameters. Their theory was followed by (Winkler, 1999) in which EM-based methods were utilized for setting optimal matching rules.

One of the significant aspects of record linkage is related to blocking. In order to decrease the number of comparisons between record pairs and come up with faster execution time, a variety of blocking strategies have been proposed. To mention a few, standard blocking (Jaro, 1989), sorted neighborhood method (Hernández and Stolfo, 1998) and canopy clustering algorithm (McCallum *et al*., 2000) are considered as some popular ones.

Since 1990s, the usage of techniques which were related to such areas as machine learning, artificial intelligence, data mining and information retrieval have been explored in record linkage and duplicate detection. Most of these strategies are supervised. It means that classifying is done based on available training samples which are labeled manually. Two prominent machine learning techniques which have been applied for classifying record pairs in the area of record linkage are decision tree (Elfeky *et al*., 2002) and Support Vector Machines (Nahm *et al*., 2002). However, labeling data manually can be a costly task. Furthermore, unsupervised techniques have also been employed providing that training samples are not available or sufficient enough. In (Gu and Baxter. 2006), one of the clustering techniques, namely k-means was utilized for classifying record pairs into match and non-match classes. Elfeky *et al*. (2002) proposed a hybrid approach in which supervised and unsupervised techniques were combined. This approach performed well in facing lack of training samples.

Christen in (2007) proposed a two step classification approach for classifying record pairs. In this approach, after computing similarity between record pairs, training examples are selected based on their similarity scores. Then, other instances are classified based on training samples by Support Vector Machines.

In this study, we follow the approach of (Christen, 2007) with different classifiers in order to determine their effectiveness in detecting duplicate records. Then, applied classification techniques are compared together.

MATERIALS AND METHODS

We follow a two step classification method presented in (Christen, 2007). The data used in our paper is Restaurant and Cora datasets. The details of these datasets will be described later. Sorted neighborhood (Hernández and Stolfo, 1995) is used as a blocking algorithm and Longest Common Subsequence (LCS) (Allison and Dix, 1986) is utilized as a similarity function. In the next step, SVM, C4.5, Naïve bayes and Bayesian network classifiers are applied on selected training records in order to train and build the models. Finally, testing set is classified based on training results to see the effectiveness of each classifier in this dataset.

Blocking technique: Nearest neighborhood (Hernández and Stolfo, 1995) blocking algorithm is used in this study. In Restaurant dataset, firstly, a blocking key is produced for sorting by combination of first three characters of Name, Address and City attributes of this dataset. The blocking key of Cora dataset is composed of first three characters of Title, Author and Venue fields of this dataset. Then, all records of each dataset are sorted based on blocking key by considering the window size as three. Each three records in a same window are compared to each other with comparison function.

Similarity function: Next step is computing the similarity between records within the same block. Longest Common Subsequence (LCS) is an algorithm proposed in (Allison and Dix, 1986) and is used to find the longest subsequences which are common in two strings. It has been successfully experimented in several contexts such as record linkage. A normalized version of LCS in which the result is normalized by considering the length of both input strings is proposed in (Islam and Inkpen, 2008) as follow:

$$\mathrm{NLCS}(s1,s2)=\frac{\mathrm{lenght}(\mathrm{LCS}(s1,s2))^{2}}{\mathrm{lenght}(s1)\times\mathrm{lenght}(s2)} \tag{1}$$

where, S1 and S2 are two input strings. In this study, the normalize version of LCS is applied in order to calculate the similarity of fields of two records. The output of this task is the similarity scores of compared records for each field, also known as weight vectors.

Classification: Each weight vector consists of several values. These values are the result of comparing two fields of each record and are considered to be in the range of 0 and 1. In the classification step, firstly, the distances of all weight vectors are computed from two vectors with the values of 1 and 0, respectively by Euclidean distance measure. Afterwards, some of the weight vectors which have the nearest values to 1 and 0, are selected as match and non-match classes, respectively. These weight vectors are considered as training set for a classifier. The remaining weigh vectors are regarded as test set and will be classified by different supervised classification techniques based on known training samples.

Support Vector Machines (SVM) are a set of techniques which investigate data to recognize patterns. SVM is used as a classification tool to build hyperplane or some hyperplanes to separate instances into two classes: -1 and +1. The more distance of hyperplane to the nearest training data-points, the less classification errors for unseen data instances. A separating hyperplane can be written as:

$$W. X + b = 0 \tag{2}$$

where, $W = \{w_1, w_2,\ldots, w_n\}$are weight vectors for n attributes $A = \{A_1, A_2,\ldots, A_n\}$; b is a scalar and $X = \{x_1, x_2,\ldots, x_n\}$ are values of attributes (Han and Kamber, 2006). There are more details on SVM in (Han and Kamber, 2006; Pugazhenthi and Rajagopalan, 2009; Lee *et al.*, 2010; 2011).

Decision Tree is one of the significant data mining techniques for classification. This technique facilitate the decision making process by dividing it to several steps. It uses labeled training instances to classify unseen data. The most common algorithm for building decision trees is the C4.5 algorithm (Quinlan, 1992; Kusrini *et al.*, 2010) which is an extension of ID3 algorithm (Quinlan, 1979).

A Bayes classifier is a simple probabilistic and statistical classifier which can predict class membership probabilities and is based on applying Bayes's rule of conditional probability. Naïve Bayesian classifiers assume that all predictor variables are independent (Han and Kamber, 2006). Naïve Bayes is utilized in several studies (Al-Salemi and Aziz, 2011; Wagner, 2010).

A Bayesian Network or Bayesian belief network or directed acyclic graphical model is represented as a directed acyclic graph in which each node holds a random variable and each variable corresponds to a particular attribute in the data. These variables may have continuous or discrete values. Mainly, a Bayesian network is based on this assumption that each variable as a parent node is conditionally independent of its non-decedents in the graph (Han and Kamber, 2006). Bayesian Network is used in a variety of domains (Ting and Phon-Amnuaisuk, 2009; Mustapha *et al.*, 2011; Mehdi *et al.*, 2007).

RESULTS

The experiments are done on two real world datasets, namely Cora and Restaurant. In the classification step, the nearest 1, 5 and 10 percent weight vectors to one and zero are selected as the training set in different experiments. The rest of weight vectors in each step are considered as test set. Finally, Weka classifier package has been used as a tool in order to classify test set instances with different classification techniques.

Restaurant dataset: Restaurant is a standard dataset which is used in several record linkage studies (Christen, 2008; Kopcke and Rahm, 2010; Stoermer *et al.*, 2010). It was created by merging the information of some restaurants from two websites: Zagat (331 non-duplicate restaurants) and Fooders (533 non-duplicate restaurants). There are 864 records in this dataset and 112 of them are duplicates. Name, Address, City, Phone and Type of restaurants are attributes of this dataset.

Cora dataset: The second applied dataset is Cora. Cora is a real world dataset which contains 1295 citations of 112 computer science papers which were gathered from the Cora Computer Science Research Paper Engine. The attributes of the citation are as follow: Author, Volume, Title, Institution, Venue, Address, Publisher, Year, Pages, Editor, Note and Month. Moreover the attribute of Class in this dataset is also used for determining whether two records are duplicates or not. It also used in several record linkage studies (Kopcke and Rahm, 2010; Ojokoh *et al.*, 2011; Christen, 2008; Hassanzadeh and Miller, 2009).

Evaluation metrics: The effectiveness of each classifier can be measured by precision, recall and F-score metrics. The following measures are required in order to calculate evaluation metrics:

- True Positive (TP): Corresponds to the number of matched detected when it is really match
- True Negative (TN): Corresponds to the number of non-matches detected when it is really non-match
- False Positive (FP): Corresponds to the number of matches detected when it is really non-match
- False Negative (FN): Corresponds to the numbers of non-matches detected when it is really match

Table 2: Effectiveness of Different Classifiers on Restaurant Dataset

Evaluation Metric	Training Size (%)	SVM (%)	Decision Tree (%)	Naïve Bayes (%)	Bayesian Network (%)
Accuracy	1	95.79	93.86	92.16	93.85
	5	96.04	74.76	97.19	93.05
	10	95.88	73.49	83.98	92.79
	Average	95.90	80.70	91.11	93.23
Precision	1	97.40	88.10	91.80	88.10
	5	96.80	94.50	97.60	96.70
	10	97.00	94.50	95.60	96.70
	Average	97.07	92.37	95.00	93.83
Recall	1	95.80	93.90	92.20	93.90
	5	96.00	74.80	97.20	93.00
	10	95.90	73.50	84.00	92.80
	Average	95.90	80.73	91.13	93.23
F-Measure	1	96.30	90.90	92.00	90.90
	5	96.30	81.30	97.30	94.20
	10	96.20	80.40	87.80	94.00
	Average	96.27	84.20	92.37	93.03

The definitions of effectiveness measures are as follow:

Precision: Precision is the fraction of true matches over the all number of candidate pairs which are classified as matches by the classifier and the formula is defined as:

$$\text{Precision} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|} \quad (3)$$

Recall: Recall is the fraction of matches correctly classified over the all number of matches and is defined as below:

$$\text{Recall} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|} \quad (4)$$

F-measure: F-measure is regarded as the mean of precision and recall values and it is defined as below:

$$\text{F} - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

Results and analysis: The results of Precision, Recall and F-measure for different classifiers are shown in Table 2 and 3 for Restaurant and Cora datasets, respectively.

In Restaurant dataset, as statistical results indicate, the F-measure values for SVM, Decision tree, Naïve bayes and Bayesian network are 96.27, 84.20, 92.37 and 93.03%, respectively. Furthermore, SVM outperforms

Table 3: Effectiveness of different classifiers on Cora dataset

Evaluation Metric	Training Size (%)	SVM (%)	Decision Tree (%)	Naïve Bayes (%)	Bayesian Network (%)
Accuracy	1	83.22	67.62	90.95	89.07
	5	85.44	69.50	89.75	88.00
	10	86.57	70.99	89.70	89.70
	Average	85.08	69.37	90.13	88.92
Precision	1	88.50	64.70	91.30	89.30
	5	89.40	67.60	89.60	90.50
	10	88.80	70.70	90.40	89.60
	Average	88.90	67.67	90.43	89.80
Recall	1	83.20	67.60	90.90	89.10
	5	85.40	69.50	89.80	88.00
	10	86.60	71.00	89.70	89.70
	Average	85.07	69.37	90.13	88.93
F-Measure	1	84.00	65.70	90.60	89.20
	5	86.10	68.30	89.60	88.50
	10	87.10	70.90	88.90	89.70
	Average	85.73	68.30	89.70	89.13

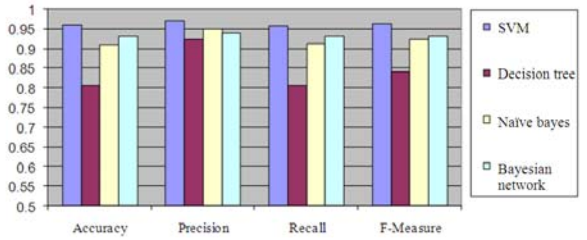

Fig. 2: Effectiveness of different classifiers on restaurant dataset

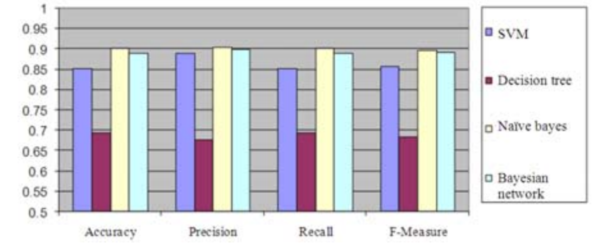

Fig. 3: Effectiveness of different classifiers on cora dataset

other algorithms in all terms of precision, recall and f-measure in this dataset. After SVM, the effectiveness of Bayesian network classifier is better than two others.

Figure 2 shows a comparison of the effectiveness of different classifiers for Restaurant dataset. As can be seen from Fig. 2, the SVM outperforms other techniques in all of the evaluation metrics.

For Cora dataset, unlike Restaurant, the Naïve bayes method outperforms others in all evaluation metrics. As statistical results show, the F-measure values for SVM, Decision tree, Naïve bayes and Bayesian network are 85.73, 68.30, 89.70 and 89.13, respectively. The results of this classifier is slightly better than Bayesian network.

Figure 3 Compares the effectiveness of different classifiers in Cora dataset.

DISCUSSION

Finding and matching duplicate records is an essential task to improve data quality. In this study, some prominent classification techniques were utilized in order to detect duplicate records in two integrated real world datasets. As The experimental results show, the effectiveness of classifiers in detecting duplicate records is different based on the input dataset. While SVM outperforms other methods in detecting duplicate objects in Restaurant dataset, Naïve bayes comes up with the best results in Cora dataset. However, the Precision of SVM is still noticeable in the Cora dataset.

CONCLUSION

Considering the results, there is no best classification technique for all datasets. Users could try different classification techniques on a new dataset in order to detect the best classification technique for it. However, SVM which is known as a robust and prominent classification technique is a good option for the classification task.

Applying record linkage task in data integrating improves the quality of data significantly. This matter leads to more accurate decisions in information systems. Finding other methods to enhance the effectiveness of detecting duplicate records, such as combining similarity measures in classification or finding more proper similarity measures will be examined in future study.

REFERENCES

Allison, L. and T.I. Dix, 1986. A bit-string longest-common-subsequence algorithm. Inform. Process. Lett., 23: 305-310. DOI: 10.1016/0020-0190(86)90091-8

Al-Salemi, B. and M.J.A. Aziz, 2011. Statistical bayesian learning for automatic arabic text categorization. J. Comput. Sci., 7: 39-45. DOI: 10.3844/jcssp.2011.39.45

Christen, P., 2007. A two-step classification approach to unsupervised record linkage. Proceedings of the 6th Australasian Conference on Data Mining and Analytics (AusDM'07), Australian Computer Society, Inc. Darlinghurst, Australia, pp: 111-119.

Christen, P., 2008. Automatic record linkage using seeded nearest neighbour and support vector machine classification. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD'08), ACM New York, NY, USA., pp: 151-159. DOI: 10.1145/1401890.1401913

Churches, T., P. Christen, K. Lim and J.X. Zhu, 2002. Preparation of name and address data for record linkage using hidden Markov models. BMC Med. Inform. Decision Mak., 2: 9-9. DOI: 10.1186/1472-6947-2-9

Elfeky, M.G., A.K. Elmagarmid and V.S. Verykios, 2002. Tailor: A record linkage toolbox. Proceedings of the 18th International Conference on Data Engineering, Feb. 26-1 Mar., San Jose, CA, USA., pp: 17-28. DOI: 10.1109/ICDE.2002.994694

Fellegi, I.P. and A.B. Sunter, 1969. A theory for record linkage. J. Am. Stat. Assoc., 64: 1183-1210. DOI: 10.2307/2286061

Gu, L. and R. Baxter, 2006. Decision Models for Record Linkage. Data Min., 3755: 146-160. DOI: 10.1007/11677437_12

Han, J. and M. Kamber, 2006. Data mining: concepts and techniques. 2nd Edn., Morgan Kaufmann, USA., ISBN-10: 1558609016, pp: 800.

Hassanzadeh, O. and R.J. Miller, 2009. Creating probabilistic databases from duplicated data. VLDB J., 18: 1141-1166. DOI: 10.1007/s00778-009-0161-2

Hernández, M.A. and S. Stolfo, 1995. The merge/purge problem for large databases. Proceedings of the International Conference on Management of Data, (ICMD'95), ACM New York, NY, USA., pp: 127-138. DOI: 10.1145/223784.223807

Hernández, M.A. and S.J. Stolfo, 1998. Real-world data is dirty: Data cleansing and the merge/purge problem. Data Min. Know. Discovery, 2: 9-37. DOI: 10.1023/A:1009761603038

Islam, A. and D. Inkpen, 2008. Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans. Knowl. Discov. Data, 2: 1-25. DOI: 10.1145/1376815.1376819

Jaro, M.A., 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa. Florida. J. Am. Stat. Assoc., 84: 414-420. DOI: 10.2307/2289924

Kopcke, H. and E. Rahm, 2010. Frameworks for entity matching: A comparison. Data Knowl. Eng., 69: 197-210. DOI: 10.1016/j.datak.2009.10.003

Kusrini, S. Hartati, R. Wardoyo and A. Harjoko, 2010. Differential diagnosis knowledge building by using CUC-C4.5 framework. J. Comput. Sci., 6: 180-185. DOI: 10.3844/jcssp.2010.180.185

Lee, L.H., C.H. Wan, T.F. Yong and H.M. Kok, 2011. A review of nearest neighbor-support vector machines hybrid classification models. J. Applied Sci., 10: 1841-1858. DOI: 10.3923/jas.2010.1841.1858

Lee, L.H., D. Isa, W.O. Choo and W.Y. Chue, 2010. Tournament structure ranking techniques for Bayesian text classification with highly similar categories. J. Applied Sci., 10: 1243-1254

McCallum, A., K. Nigam and L.H. Ungar, 2000. Efficient clustering of high-dimensional data sets with application to reference matching. Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD'00), ACM New York, NY, USA., pp: 169-178. DOI: 10.1145/347090.347123

Mehdi, M., S. Zair, A. Anou and M. Bensebti, 2007. A bayesian networks in intrusion detection systems. J. Comput. Sci., 3: 259-265. DOI: 10.3844/jcssp.2007.259.265

Mustapha, A., M.N. Sulaiman, R. Mahmod and M.H. Selamat, 2011. Dynamic bayesian networks in classification-and-ranking architecture of response generation. J. Comput. Sci., 7: 59-64. DOI: 10.3844/jcssp.2011.59.64

Nahm, U.Y., M. Bilenko and R.J. Mooney, 2002. Two approaches to handling noisy variation in text mining. Proceedings of the Papers from the 19th International Conference on Machine Learning Workshop on Text Learning, (ICML-2002), Sydney, Australia, pp: 18-27.

Newcombe, H.B., J.M. Kennedy, S.J. Axford and A.P. James, 1959. Automatic linkage of vital records. Science, 130: 954-959. DOI: 10.1126/science.130.3381.954

Ojokoh, B., M. Zhang, J. Tang, 2011. A trigram hidden Markov model for metadata extraction from heterogeneous references. Inform. Sci., 181: 1538-1551. DOI: 10.1016/j.ins.2011.01.014

Pugazhenthi, D. and S.P. Rajagopalan, 2009. Unbalance quantitative structure activity relationship problem reduction in drug design. J. Comput. Sci., 5: 764-772. DOI: 10.3844/jcssp.2009.764.772

Quinlan, J.R., 1979. Discovering Rules by Induction from Large Collections of Examples. In: Expert Systems in the Micro-Electronic Age, D. Michie, (Ed.). Edinburgh: Edinburgh University Press, pp: 168-201.

Quinlan, J.R., 1992. C4.5: programs for machine learning. 1st Edn., Morgan Kaufmann, USA., ISBN-10: 1558602380, pp: 302.

Sarawagi, S. and A. Bhamidipaty, 2002. Interactive deduplication using active learning. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD'02), ACM New York, NY, USA., pp: 269-278. DOI: 10.1145/775047.775087

Stoermer, H., N. Rassadko and N. Vaidya, 2010. Feature-based entity matching: The FBEM model, implementation, evaluation. Advanced Inform. Syst. Eng., 6051: 180-193. DOI: 10.1007/978-3-642-13094-6_15

Surajit, C., K. Ganjam, V. Ganti and R. Motwani, 2003. Robust and efficient fuzzy match for online data cleaning. Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, (SIGMOD'03 ACM New York, NY, USA., pp: 313-324. DOI: 10.1145/872757.872796

Ting, C.Y. and S. Phon-Amnuaisuk, 2009. Log data approach to acquisition of optimal Bayesian learner model. Am. J. Applied Sci., 6: 913-921. DOI: 10.3844/ajassp.2009.913.921

Wagner, M., 2010. Forecasting daily demand in cash supply chains. Am. J. Econ. Bus. Admin., 2: 377-383. DOI: 10.3844/ajebasp.2010.377.383

Warnner, J.W. and E.W. Brown, 2001. Automated name authority control. Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, (JCDL'01), ACM New York, NY, USA., pp: 21-22. DOI: 10.1145/379437.379441

Winkler, W.E., 1999. The State of Record Linkage and Current Research Problems. Statistical Research Division, U.S. Census Bureau. http://www.census.gov/srd/papers/pdf/rr99-04.pdf