Research Journal of Information Technology 4(2): 79-87, 2012
ISSN: 2041-3114

Submitted: April 17, 2012 Accepted: May 10, 2012 Published: June 30, 2012

Implementation and Comparison Clustering Algorithms with Duplicate Entities Detection Purpose in Data Bases

[1]Maryam Bakhshi, [2]Mohammad-Reza Feizi-Derakhshi and [3]Elnaz Zafarani
[1]Department of Computer, Zanjan Branch, Islamic Azad University, Zanjan, Iran
[2]Computer Department, University of Tabriz, Tabriz, Iran
[3]Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran

Abstract: The aim of study is finding appropriate clustering algorithms for iteration detection issues on existing data set. The issue of identifying iterative records issue is one of the challenging issues in the field of databases. As a result, finding appropriate algorithms in this field helps significantly to organize information and extract the correct answer from different queries of database. This study is a combination of the author's previous studies. In this study, 4 algorithms, K-Means, Single-Linkage, DBSCAN and Self-Organizing Maps have been implemented and compared. F1 measure was used in order to evaluate precision and quality of clustering that by evaluating the obtained results, the SOM algorithm obtained high accuracy. However, the base SOM algorithm due to using Euclidean distance has some defects in solving real problems. In order to solve these defects, Gaussian kernel has been used to measure Euclidean distance that by studying obtained results it was seen that KSOM algorithm has higher F1 measure than base SOM algorithm. Initializing weight vector in SOM algorithm is one of the main and effective problems in convergence of algorithm. In this research we presented a method that optimized initialing weight step. Presented method reduces the number of iteration in comparison than basic method as it increases the run rate.

Keywords: Clustering, DBSCAN, F1 measure, K-means, self-organizing maps, single-linkage

INTRODUCTION

Databases play an important role in the information age. Industries and systems in their operations depend on precision of databases. Therefore, the quality of data (or lack thereof) stored in the database have an important impact on the cost of information based systems. Often the data are not carefully controlled of quality and also they have not been defined compatible with various sources of data. Therefore, data quality is often compared with several factors such as input errors (e.g. Microsft instead of Microsoft), elimination of integrity constraints (e.g. permitting inputs such as age = 476) and various formats for data storage (e.g., St. A, Street A).

In worse conditions in the databases which are managed independently, rather than values, structure, concepts and assumptions about data may be different from each other. Consequently, in order to manage better and extract the correct answer from different queries of database, the problem of iteration detection is important. Iteration detection procedure includes three follow steps:

- Field matching
- Record matching
- Clustering

That this research is emphasized on the third stage, clustering and its ultimate goal is to find appropriate algorithms for iteration detection problem. After studying the existing algorithms in clustering field, Indicators for evaluating clustering algorithms in order to measure the accuracy of clustering results is given. Method is the following, after the first and two stages and get the degree of similarity between the records, clustering is done to similar records lie in a cluster. The ultimate goal of this research is to find the degree of the most suitable algorithms for existing data set. Used dataset include property information. There are different categories for clustering methods that the following is the most comprehensive (Ossama, 2008):

- Partitioning clustering
- Hierarchical clustering
- Density based clustering
- Grid based clustering
- Model based clustering
- Fuzzy clustering

In this study, four algorithms of Partitioning, Hierarchical, Density-based and Model-based categories were selected and compared.

Corresponding Author: Maryam Bakhshi, Computer Department, Zanjan Branch, Islamic Azad University, Zanjan-Iran

LITERATURE REVIEW

Some researchers have improved clustering algorithms. Some presented new algorithms. And some others studied and compared clustering algorithms. In this section, previous studies have been presented that studied the influence of different factors on efficiency of the number of clustering algorithms and compared the results.

Ling *et al*. (2007) provided a detailed survey of current clustering algorithms in data mining at first, then it makes a comparison among them, presented their scores (merits) and identified the problems to be solved and the new directions in the future according to the application requirements in multimedia domain. Rui and Wunsch (2005) presented the survey of clustering algorithms for data sets including in statistics, computer science and machine learning and explained their applications in some benchmark data sets, the traveling salesman problem and bioinformatics and also subjects like adjacent measures and evaluating clustering were discussed. Several tightly related topics, proximity measure and cluster validation, are also discussed. Treshansky *et al*. (2001) presented a survey of clustering algorithms and paid particular attention to those algorithms that require less amount of knowledge about the domain being clustered. Ossama (2008) studied and compared various clustering algorithms. Algorithms have been compared based on factors: Dataset size, number of clusters, type of dataset and used software. Dong-Jun (2006) and Ning and Hongyi (2009) first discussed about disadvantages of SOM in order to visualize prediction of financial time series and then he presented the Kernel SOM method in order to increase efficiency of visualization. Practical results show that in comparison with SOM, Kernel SOM is more appropriate in visualizing the prediction of financial time series. Maurizio (2009) and Dong-Jun *et al*. (2006) compared the efficiency of kernel clustering method on multiple data sets. These methods have been based on central clustering and also the results of authentication techniques were presented. Obtained results show that clustering on kernel space generally outperforms standard clustering. Although any of these methods never perform better than others. K-Means algorithm produces more compressed clusters compared with hierarchical method especially when clusters are in the shape of spherical.

IMPLEMENTED AND COMPARED ALGORITHMS

The main aim of this research is detection of suitable clustering algorithm for iteration detection procedure. For this purpose, from between existing algorithms in this context (clustering), of the four categories mentioned, an algorithm as the sample is selected and was compared.

Studied and compared algorithms include: K-means, Single-Linkage, self-organizing maps and DBSCAN.

K-means algorithm: Algorithm K-Means Glenn (2002) is one of the most popular iteration based clustering methods. This algorithm has application in some cases in which any data belongs only to one class. This algorithm is an unsupervised algorithm and has iteration in which the data set has been divided into k clusters and data points are randomly assigned to the clusters Rui and Wunsch (2005). Then for each point, the distance of point to the center of cluster has been calculated and target point is assigned to the closest cluster. These steps will be repeated until no point is shifted longer. Characteristics of this algorithm are as follows:

- Always there is k clusters.
- Always at least one point in each cluster is available.
- Clusters are not as hierarchical and do not overlap with each other.
- Each member of a cluster compared with other clusters has the lowest distance from cluster center.

Implementation steps of K-Means algorithm are expressed as follows:

- Establish primary centers of clusters with random selection of C point among all data points.
- Calculate the membership matrix U using the Eq. (1):

$$u_{ij} = \begin{cases} 1 & if \left\| x_j - c_i \right\|^2 \le \left\| x_j - c_k \right\|^2, \; for \; each \; k \ne i \\ 0 & otherwise \end{cases} \quad (1)$$

- Calculate membership function using the following equation:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, x_k \in G_i} \left\| x_k - c_i \right\|^2 \right) \quad (2)$$

- Calculate the new cluster centers using the following equation:

$$c_i = \frac{1}{|G_i|} \sum_{k, x_k \in G_i} x_k \quad (3)$$

And then return to Step 2.

It is noteworthy that the algorithm performance depends on the initial location of center of clusters. Therefore there is not any guarantee to reach the expected response by this algorithm.

Advantages of K-Means clustering algorithm:

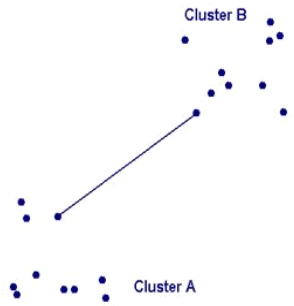

Fig. 1: Single-linkage algorithm

- In the case of large number of variables, this algorithm has higher computing rate than hierarchical approach (If k is small).
- K-Means algorithm produces more dense clusters than the hierarchical method especially when clusters are spherical.

Limitations of K-Means clustering algorithm:

- Difficulty of qualitative comparison of produced clusters
- Fixedness of the number of clusters that makes prediction of k value difficult
- It is not so suitable for non-spherical clusters
- It is sensitive to noisy data

Difference of the number of primary clusters leads to difference of final clusters. So it is better to run algorithm for different values and compare results with each other.

Single-linkage algorithm: This is one of the oldest and simplest methods of clustering methods and can be considered as a member of hierarchical and exclusive clustering methods. This clustering also called nearest neighbor technique (nearest neighbor). In this method for calculating the similarity between cluster A and B of following criteria are used (Fig. 1):

$$d_{AB} = \min_{i \in A, j \in B} d_{ij} \quad (4)$$

That i is a sample belonged to cluster A and j is a sample belong to a cluster B. In fact, the similarity between two clusters is the minimum distance between a member of one than from one another.

DBSCAN algorithm: Density based spatial clustering of applications with noise, DBSCAN; rely on a density-based notion of clusters, which is designed to discover clusters of arbitrary shape and also have ability to handle noise. The main task of this algorithm is class identification, i.e., the grouping of objects into meaningful subclasses Periklis (2002).
Two global parameters of DBSCAN algorithms are:

- **Eps**: Maximum radius of the neighbourhood.
- **MinPts:** Minimum number of points in an Eps-neighbourhood of that point.

Core Object: Object with at least MinPts objects within a radius 'Eps- neighbourhood.

Border Object: Object that on the border of a cluster
NEps (p): {q belongs to D | dist (p, q) <= Eps}

Directly Density-Reachable: A point p is directly density-reachable from a point q w.r.t Eps, MinPts
If p belongs to NEps (q)
|NEps (q)| > = MinPts

Density-Reachable: A point p is density-reachable from a point q w.r.t Eps, MinPts if there is a chain of points {p1... pn}, p1 = q, pn = p such that pi+1 is directly density-reachable from pi.

Density-Connected: A point p is density-connected to a point q w.r.t Eps, MinPts if there is a point 'o' such that both, p and q are density-reachable from 'o' w.r.t Eps and MinPts.

The algorithm of DBSCAN is as follows:

- Arbitrary selection of a point p.
- Retrieve all points density-reachable from p w.r.t Eps and MinPts.
- If p is a core point, then a cluster is formed.
- If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

Self-organizing algorithm: Teuvo Kohonen (1975) introduced new type of neural network that uses competitive, unsupervised learning Ling *et al*. (2007). This theory is based on WTA (Winner Takes All) and WTM (Winner Takes Most) algorithms. Therefore, these algorithms will be explained here briefly. The most basic competitive learning algorithm is WTA. When input vector (a pattern) is presented, a distance to each neuron's synaptic weights is calculated. The neuron whose weights are most correlated to current input vector is the winner. Correlation is equal to scalar product of input vector and considered synaptic weights. Only the winning neuron modifies it's synaptic weights to the point presented by input pattern. Synaptic weights of other neurons do not change. The learning process can be described by the following equation:

$$\| x - w_c \| = \min_j \{\| x - w_i \|\} \quad (5)$$

$$w_c (t + 1) = w_c(t) + a(t) [x(t) - w_c(t)] \quad (6)$$

where, $i \in [0..\text{numer of neurons}]$, W_i represents all synaptic weights of the winning neuron, $\alpha(t)$ is learning rate in the interval [0, 1] that linearly proportional with t inverse reduced and shows total weights attached to the winning cell and x stands for current input vector. In this section WTM strategy described that is an extension of WTA strategy. The difference between those two algorithms is that many neurons in WTM strategy adapt their synaptic weights in a learning iteration. In this case not only the winner, but also its neighbourhood adapts. The further the neighbouring neuron is from the winner, the smaller the modification which is applied to its weights. This adaptation process can be described as:

$$w_{i+1} = wi+\eta\, K(i, x)\times(x-w_i) \quad (7)$$

For all neurons (i) that belong to winner's neighbourhood. W_i stands for synaptic weights of neuron i and x is current input vector. η stands for learning rate and N(i, x) is a function that defines neighbourhood. where, w_i shows the weights attached to the cells and cells located in the neighborhood of winning. X vector is input pattern and w_i learning rate that have a positive value smaller than the unit. K (i, x) is neighborhood function that is Gaussian kernel that was a descent function and with a way of win cell and time decreases. And thus the cell in farthest neighborhood will have low change in weights. Neighborhood function can be described as:

$$N(i,x) = \begin{cases} \exp\left(-\left\|x - w_i\right\|^2 .t\right), & w_i \in \lambda(i,x) \\ 0 & , \; w_i \notin \lambda(i,x) \end{cases} \quad (8)$$

In order to train SOM network the Euclidean distance between input vector and weight vectors of all cells should be computed. The cell which has the lowest distance with input vector, in other words the cell which has the most similarity to input pattern is selected as a winner and its adjoined weights change in order to approach input pattern. In addition, adjacent cells are selected and according to their distance to winner cell their weights are modified in the same orientation. The movement of cells and the number of mobile cells is high in the beginning of algorithm and they reach their minimum value due to reducing the rate of learning and neighbor radius. This algorithm maps input vector on one line (in two-dimensional topological state). Figure 2 shows one two dimensional SOM neural network.

Input patterns that are similar to each other, have minimum Euclidean distance from each other, are also after mapped are placed together. In 1-D network each cell has 2 neighbors, a neighbor on the left and the other cell in the right placed. Two-dimensional network in each cell has four neighbors, which is on the left, right, top and bottom cell are placed.

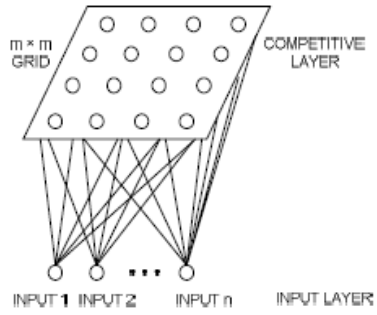

Fig. 2: SOM with the neighborhood of two-dimensional input vectors

SOM algorithm can be summarized as follows:

1 - Choose weight of all the cells randomly.
2 - Apply input pattern to network.
3 - Find win cells.
4 - Select the neighbor cells.
5 - Correct weights attached to the cells and the winner of the neighbor cells according to their Euclidean distance, learning rate and neighborhood radius.
6 - Repeat stages 2 to 5 for the number of distinct and pre-determined periods.

TESTS AND METHODOLOGY

Data set: Applied data set includes data relating to a property information system. This database has high volumes of data. Due to being in different parts, the system has high volumes of redundancy. System Database includes 162 main table and 33 base information table. So main tables include 2105 fields and base tables include 192 field. The effective fields in iteration detection includes: address, owner name, identification code, area, etc. Initially we have used owner name and address fields.

Evaluation measure: The goodness measure of one cluster is based on similarities of interior and exterior class. The similarity of within class should be high and this is when then similarity of out of class is so low. In other words, the data in one cluster should be identical and however they should be different from other cluster data. In order to evaluate the quality of clustering we need one qualitative measure of clustering. There are several measures for clustering and classifying that in this study we have used F-Measure Periklis (2002).

F1-measure: There are two states "belongs" and "not belongs" in the problems of binary classification in order

Table 1: Variables

	Positive predicated	Negative predicated
Real positive	a	b
Real negative	c	d

to assign data to one particular category. Data which belong to one category are called positive and data which do not belong to one category are called negative. Consider the variables of Table 1, if I was the dataset which is really positive and J is a dataset which was predicted as positive by clustering algorithm, according to Table 1, variables a, b, c and d are defined as follows:

$$a = I \cap J \quad b = I \cap \bar{J}$$
$$c = \bar{I} \cap J \quad d = \bar{I} \cap \bar{J} \tag{9}$$

In this section two variables are defined that are not effective separately. Variable P (Precision) that is defined as follows shows that how many data which have been predicted as positive are really positive:

$$a/(a+c) \tag{10}$$

and R (Recall) which is defined as follows shows that how many of the true positives have been predicted accurately:

$$a/(a+b) \tag{11}$$

According to the mentioned information, parameter P is obtained by dividing the number of common clusters into the number of computed clusters with clustering algorithm (system) and parameter R is obtained by dividing the number of common clusters into the number of accurate clusters. A good clustering algorithm should consider both parameters P and R. These two measures are combined and form F-Measure:

$$P = \frac{number\ of\ common\ clusters}{number\ of\ system\ calculated\ clusters} = \frac{\left|C_i \cap HC_j\right|}{\left|HC_j\right|} \tag{12}$$

$$R = \frac{number\ of\ common\ clusters}{number\ of\ correct\ tar\ get\ clusters} = \frac{\left|C_i \cap HC_j\right|}{\left|C_i\right|} \tag{13}$$

$$F_\beta = (1+\beta^2)*P*R/\beta^2*P+R \tag{14}$$

In Eq. (14), β is as weighting parameter. So that, with using of this parameter F1 measure is considered as Precision-oriented or Recall-oriented. Often, β replaced with 1 as follow:

$$\beta = 1 \rightarrow F_1 = 2*P*R/P+R \tag{15}$$

Therefore, for efficiency evaluation of implemented clustering algorithms Eq. (15) is used.

Tests: As was mentioned, duplicate detection includes three steps. The first step is field matching. At this stage to detect similarities between the fields, distance-based algorithms are used. Implemented algorithms in this phase are jaro and jaro winkler. That has achieved similar results almost. Basically jaro algorithm to compare names will do better. For this purpose, similarity between owner name and address fields achieved with using of jaro algorithm. And the degree of similarity obtained from the first stage, used as input of record matching. To perform record matching of the two follows methods used:

Mapping to two dimensional spaces: Owner name and address fields are considered as dimensions. Method is as follows that a record be considered as origin record and other records according to similarity degree in origin record lie in that space. In other words, each record according to the degree of similarity assigns a coordination of space.

Use statistical average: By accounting mean of field similarity degree of records, similarity degree of records was obtained.

After obtain the degree of similarity between the records, turn reach to the third stage. Third stage is clustering that in it four algorithms, K-Means, DBSCAN, Single-Linkage and SOM implemented and performed on existing data set. Method this was, 50 records of data set were selected and these records were manually clustered. The reason for doing this was to be using of the human accuracy factor. And in order to ease process, a few records (50 records) selected and manually clustered. This work Due to be using of accuracy human agent and selection of 50 records in order to ease manual clustering. Because with increase number of records, manual clustering time consumed and there will be Risk of confusion. The results of the algorithms used to calculate the quality of the clustering relationships were presented. Field values of Precision, Recall and F1 measure respectively were calculated using the Eq. (12)-(15). The resulting values are shown in Table 2. As is seen from Table Values of the parameters P and R are calculated according to the results of clustering algorithms. "Number of system cluster" shows produced clusters by algorithm (system). Mean of "number of correct cluster" is obtained cluster by human agent. "Number of common clusters" present produced common clusters by two methods. With the calculated values as is seen from Table SOM algorithm with a strategy to WMA has the highest F1 In other words, this algorithm has a high degree of accuracy in compared with other algorithms. In other words, The WMA is a better convergence than that of the WTA.

PARAMETER EVALUATION

Length parameter value estimation: SOM algorithm maps result over a length * length space. Since identifying the dimension length significantly impacts on the results of SOM algorithm. So choose a suitable value for this field has an important effect on results. As a result, different values of length parameter and its effect on the

Table 2: Evaluation results

Method	Used Alg.	System cluster no.	Correct clusters no.	Common clusters no.	Precision	Recall	F1 measure
Mapping to two dimensional space	K-means	38	38	22	0.66	0.66	0.58
	(DBSCAN) eps: 0.005, minpt = 1	23	38	10	0.43	0.26	0.33
	single-linkage	25	38	20	0.8	0.52	0.62
	SOM (WTA)	28	38	22	0.785	0.578	0.66
	SOM (WMA)	31	38	27	0.87	0.71	0.78
Use average	K-means	38	38	12	0.195	0.195	0.3
	(DBSCAN) eps: 0.005, minpt = 1	26	38	8	0.23	0.15	0.19
	single-linkage	18	38	10	0.55	0.26	0.35
	SOM (WTA)	22	38	12	0.54	0.31	0.39
	SOM (WMA)	25	38	17	0.68	0.447	0.53

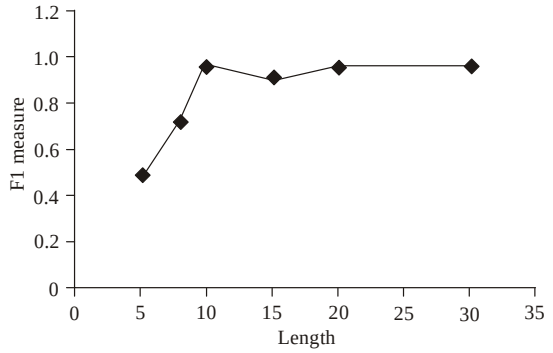

Fig. 3: Identify suitable values for length parameter

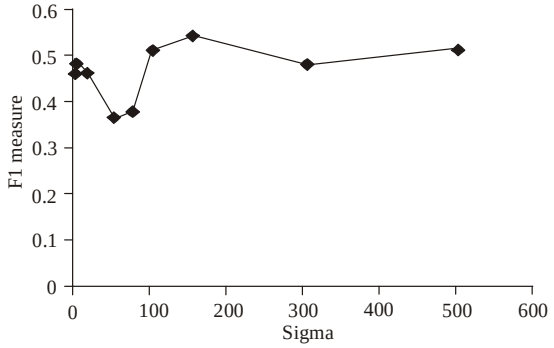

Fig. 4: Identify suitable values for sigma parameter

F1 field were evaluated that results are observable in Table 2. As Fig. 3 shows considering the number 10 for the length parameter leads to maximum value for F1 field.

Sigma parameter value estimation: The determination of value of parameter in the Gaussian kernel has Particular importance and has a significant impact on the results. In this study in the calculation of Euclidean distance values to different variables were Sigma and some in which the algorithm had the highest Sigma was selected as the appropriate variable. As can be seen from Fig. 4 Sigma is the best value for the variable number 150. Gaussian kernel function directly associated with the sigma value as low as in the interval [0, 10] will lead to Over fitting and very high values, for example, in the interval [300, 700] to be under fitting. Sigma 75-150 is the appropriate value for the variable Lipo wang (2005). According to studies conducted in this study the number 150 was chosen as the value for the parameter sigma.

PROPOSED ALGORITHM

Proposed algorithm:

- **Phase 1: Using kernel in euclidean distance calculation:** Using kernel functions in machine learning was introduced by Aizerman first in 1964 Zhexue (1998). The main idea was to create kernel functions, a relation to map data from a space with low dimensions to an inner product space with high dimensions using a non-linear conversion. In this space, the data which are linearly non separable can be separated linearly. Kernel function K: $X * X \rightarrow R$ is defined as follows:

$$K(x_i, x_j) = < \varphi(x_i), \varphi(x_j) > \quad (16)$$

To avoid complicated calculations, the function φ need not be known explicitly, only using a kernel function is sufficient. A Kernel used in this study is a RBF kernel which is defined as follows:

$$K(x, z) = e||x-z||^2/2\sigma^2 \quad (17)$$

In this study, basic SOM algorithm has been extended and since SOM algorithm has bugs in resolving real issues due to using Euclidean distance, in Euclidean distance calculation phase, Gaussian kernel has been used. One of the advantages of kernel is to calculate Euclidean distance without explicit knowledge of:

$$\begin{aligned} ||\Phi(x_i)-\Phi(x_j)||^2 &= ((\Phi(x_i)-\Phi(x_j)).(\Phi(x_i)-\Phi(x_j)) \\ &= \Phi(x_i).\Phi(x_i)+\Phi(x_j)(x_j)-2\Phi(x_i).\Phi(x_j) \\ &= K(x_i,x_i)+K(x_j,x_j)2K(x_i,x_j) \end{aligned} \quad (18)$$

On the other hand:

$$||\Phi(x_i)^2|| = \Phi(x_i).\Phi(x_i) = k_{ii} = 1 \quad (19)$$

With placing 1 Instead of kernel in Eq. (18) have:

$$= 1+1-2\exp[-||x_i - x^2_j|/\sigma^2]$$
$$= 2-2\exp[-||x_i - x^2_j|/\sigma^2] \quad (20)$$

So, Euclidean distance is calculated by using Eq. (20).

- **Phase 2: Changing the method of initializing weight vector with the aim of reducing the number iteration and increasing the rate in SOM algorithm:** In SOM algorithm, weight initializing is one of the main steps. Because, the proper initializing the weights has great influence on final convergence of network and it guides convergence toward local or global minimum. Generally weight initializing is selected traditionally and commonly randomly in the range of 0 and 1. The principle of SOM algorithm is to repeat steps until reaching to one convergent state and in these repetitions weights are being changed until they reach to a convergent and fixed state. In other words, the termination condition of algorithm is to reach to a convergent state. As a result in order to reduce the iteration number and increase the rate of algorithm, we have decided to improve the initializing step of weights. To do this, we selected one random sample from existing data set and the algorithm was run on this data set. In the last iteration of algorithm, the values of weight vector were extracted. In the next step, the weight vector which was obtained from first step has been allocated as primary values to weights and the algorithm was run over complete data set. The algorithm was run in two steps about 20 times that the results are presented in the table. Since the algorithm has reached one optimum state of weights values in the final state, extraction of these weights and applying them in the algorithm with complete data set will lead to reduction of iteration number and increment of rate.

RESULTS EVALUATION

Table 3 shows the results of the implementation of mentioned algorithms on existing data set. Using criteria P, R and F1 measure for each algorithm were computed and results were compared. As the results are observed, SOM algorithm includes high F1-measure.

As mentioned, SOM algorithm has a high F1 measure. Thus in this study attempt was taken place to improve the quality and efficiency of this algorithm. Since kernel has positive effect on the quality of algorithms, in weight correction phase of SOM algorithm Gaussian kernel was used. As it can be seen from Table 3, in Gaussian kernel mode, F1 measure was significantly increased (increasing F1 measure amount of 0.78 to 0.96).

As it can be seen of results in Table 4, the mean iteration in the presented method in this study (optimizing the initializing the weight vector) is less than the iteration number in the primary state of random initializing of weight vector (traditional method). In other words, the run rate of algorithm in the presented method is more than the random initializing method. As observed of Table 3 mean of iteration of algorithm in 'optimizing the initializing the weight vector' method rather to 'random initializing' method is decreased. By calculate mean of iteration no. for methods, observed that iteration no. for 'optimizing the initializing the weight vector' method is 210 times less than iteration no. for 'random initializing' method.

Comparison of record matching methods: A comparison performed between the results of two methods is mapped to two-dimensional space and the statistical average. As Fig. 5 shows mapping method has better results than the statistical average. Vector 'x' in the graph relates to precision field and vector 'y' relates to the F1 measure field of Table 2.

SUMMARY AND CONCLUDING REMARKS

The purpose of this study was to obtain the appropriate algorithms to detect iteration records in the available data set. For this purpose, from four clustering categories, that is described, an algorithm selected and was compared. The implemented algorithms are: K-Means, DBSCAN, Single-Linkage, Sel Organizing maps. For assessing the quality of clustering algorithms, F1 measure was used. Implemented algorithms on 50 selected records were performed. And results were compared based on F1 measure. According to the results, WMA SOM algorithm gained a higher F1 measure. The results shows, SOM algorithm was selected as the most appropriate clustering algorithm. Improve the efficiency of the algorithm was done in two phases. Since then, the base SOM algorithm has some defects in solving real problems due to using Euclidean distance. In order to

Table 3: Proposed algorithm result evaluation

Method	Algorithm	Num. of system clusters	Num. of correct clusters	Num. of common clusters	Precision	Recall	F1 measure
Mapping to two	SOM (WTA)	28	38	22	0.785	0.578	0.66
dimensional space	SOM (WMA)	31	38	27	0.87	0.71	0.78
Use average	SOM (WTA)	22	38	12	0.54	0.31	0.39
	SOM (WMA)	25	38	17	0.68	0.447	0.53
Kernel method	kernel SOM	37	36	38	0.947	0.972	0.958

Table 4: Weight optimization results

Iteration no.	Iteration no. of optimization method(phase1: random initialize with selected data set)	Iteration no. of optimization method (phase2:weight initializing with values of last iteration of phase1)	Iteration no. of traditional method (random initialize)
1	577	905	1533
2	665	829	1516
3	655	1046	1560
4	326	548	1422
5	523	972	1679
6	737	1078	1530
7	685	575	1525
8	665	611	1508
9	586	763	1491
10	441	917	1526
11	487	779	1569
12	600	893	1459
13	308	810	1556
14	882	589	1599
15	405	599	1526
16	337	699	1562
17	595	499	1472
18	595	499	1463
19	431	450	1475
20	596	753	1561
Mean	557.7	758.9	1526.6

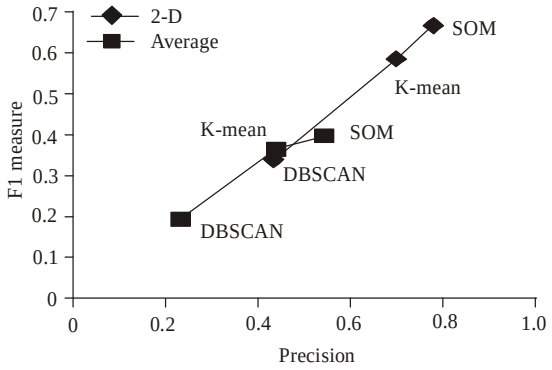

Fig. 5: Comparison of mapping and average methods

increase efficiency and quality of the algorithm in Euclidean distance calculation, kernel was used. And the results showed the efficiency of the algorithm significantly increased and F1 measure increased of 0.78 to 0.96. Too, the method to optimization weight vector initialization in SOM algorithm presented that results show number of iterations in presented method than to pervious state reduced (Table 4). According to Fig. 5 mapping method of record matching methods has better result of other.

REFERENCES

Dong-Jun, Y., Y. Qi, Y.H. Xu and J.Y. Ying, 2006. Kernel-SOM based visualization of financial time series forecasting. First International Conference on Innovative Computing, Information and Control, 2006. ICICIC '06. Aug. 30 2006-Sept. 1 2006, ch. of Comput. Sci. Technol., Nanjing Univ. of Sci. Technol., 2: 470-473.

Glenn, F., 2002. A comprehensive overview of basic clustering algorithms. The College of Information Sciences and Technology.

Ling, H.E., W.U. Ling-da and C. Yi-chao, 2007. Survey of clustering algorithms in data mining. Science China (Series E), 46(6): 627-638.

Maurizio, F., F. Masulli and S. Rovetta, 2009. An experimental comparison of kernel clustering methods. Proceedings of the 2009 conference on New Directions in Neural Networks: 18th Italian Workshop on Neural Networks: WIRN 2008, IOS Press, Amsterdam, pp: 118-126, ISBN: 978-1-58603-984-4.

Ning, C. and Z. Hongyi, 2009. Extended kernel self-organizing map clustering algorithm. Fifth International Conference on Natural Computation, 14-16 Aug. 2009, Mech. Eng. Coll., Jimei Univ., Xiamen, China, 3: 454-458, ISBN: 978-0-7695-3736-8.

Ossama, A.A., 2008. Comparisons between of data clustering algorithms. Int. Arab J. Inform. Techn., 5:(3): 320-325, Retrieved from: http: //www. ccis2k. org/iajit/PDF/vol.5,no.3/15-191.pdf.

Periklis, A., 2002. Data Clustering Techniques. University of Toronto.

Rui, X. and D. Wunsch, 2005. Survey of Clustering Algorithms. IEEE Transaction on Neural Networks, Dept. of Electr. and Comput. Eng., Univ. of Missouri-Rolla, Rolla, MO, USA, 16(3): 645-678.

Treshansky, A. and R.M. McGraw, 2001. Overview of clustering algorithms. Proceedings-Spie the International Society for Optical Engineering. International Society for Optical, pp: 41-51.

Zhexue, H., 1998. Extensions to the k-means algorithm for clustering large data set with categorical values. Data Min Knowledge Discovery, 2(3): 283-304, DOI: 10.1023/A:1009769707641.

Dong-Jun, Y., Y. Qi, Y.H. Xu and J.Y. Ying, 2006. Kernel-SOM based visualization of financial time series forecasting. First International Conference on Innovative Computing, Information and Control, 2006. ICICIC '06. Aug. 30 2006-Sept. 1 2006, ch. of Comput. Sci. Technol., Nanjing Univ. of Sci. Technol., 2: 470-473.

Glenn, F., 2002. A comprehensive overview of basic clustering algorithms. The College of Information Sciences and Technology.

Ling, H.E., W.U. Ling-da and C. Yi-chao, 2007. Survey of clustering algorithms in data mining. Science China (Series E), 46(6): 627-638.

Maurizio, F., F. Masulli and S. Rovetta, 2009. An experimental comparison of kernel clustering methods. Proceedings of the 2009 conference on New Directions in Neural Networks: 18th Italian Workshop on Neural Networks: WIRN 2008, IOS Press, Amsterdam, pp: 118-126, ISBN: 978-1-58603-984-4.

Ning, C. and Z. Hongyi, 2009. Extended kernel self-organizing map clustering algorithm. Fifth International Conference on Natural Computation, 14-16 Aug. 2009, Mech. Eng. Coll., Jimei Univ., Xiamen, China, 3: 454-458, ISBN: 978-0-7695-3736-8.

Ossama, A.A., 2008. Comparisons between of data clustering algorithms. Int. Arab J. Inform. Techn., 5:(3): 320-325, Retrieved from: http:// www. ccis2k. org/iajit/PDF/vol.5,no.3/15-191.pdf.

Periklis, A., 2002. Data Clustering Techniques. University of Toronto.

Rui, X. and D. Wunsch, 2005. Survey of Clustering Algorithms. IEEE Transaction on Neural Networks, Dept. of Electr. and Comput. Eng., Univ. of Missouri-Rolla, Rolla, MO, USA, 16(3): 645-678.

Treshansky, A. and R.M. McGraw, 2001. Overview of clustering algorithms. Proceedings-Spie the International Society for Optical Engineering. International Society for Optical, pp: 41-51.

Zhexue, H., 1998. Extensions to the k-means algorithm for clustering large data set with categorical values. Data Min Knowledge Discovery, 2(3): 283-304, DOI: 10.1023/A:1009769707641.