Journal of Computer Sciences

Original Research Paper

Adaptive Classification Method Based on Data Decomposition

[1]Ayman E. Khedr, [2]Amira M. Idrees and [3]Ahmed I. El Seddawy

[1]Faculty of Computers and Information, Helwan University, Helwan, Cairo, Egypt
[2]Faculty of Computers and Information, Fayoum University, Fayoum, Egypt
[3]Arab Academy for Science Technology and Maritime Transport (AASTMT), Cairo, Egypt

Article history
Received: 30-11-2015
Revised: 31-12-2015
Accepted: 05-01-2016

Corresponding Author:
Ayman E. Khedr
Faculty of Computers and Information, Helwan University, Helwan, Cairo, Egypt
Postal Code: 11795
Tel: 002-01062638455
Email: ayman_khedr@helwan.edu.eg

Abstract: Knowledge discovery is one of the vital fields which strongly supports decision making by applying different techniques based on the targeted field and the required information. Focusing on clustering and classification techniques, this paper presents an approach for adapting one of the classification algorithms for supporting decision making procedure in radiology data analysis field. The proposed adaptation is based on dividing the analysis problem by data partitioning and individually examining against each cluster, with applying the classification algorithm in a parallel approach. The proposed approach has proved to produce higher results accuracy with minimization of time when compared with the traditional ID3.

Keywords: Data Mining, Supervised Learning, Classification, ID3

Introduction

Knowledge discovery aims to discover implicit information including the correlations among data items (Moore, 2009). While different tools are applied in business field (Domingos and Hulten, 2000) for extracting information, knowledge discovery is considered one of the powerful tools to be applied with the target of gaining the most related and useful information for decision making. Data classification is a task that can predict the class of the object in data which can further provide a solution for many diverse problems.

ID3 is an algorithm for classification that depends on building a decision tree (Ahire *et al*., 2015), its methodology is to calculate a weighting gain measure as a parameter for dividing information (Hulten *at al*., 2001). The main procedure of ID3 is to apply the required measures on the data in the learning phase. This methodology is very time consuming as ID3 provide sequence of iterations on all data which requires large volume of calculations. In this study, we focus on applying an adapted ID3 algorithm on a set of radiology data that is already clustered.

The remaining of the paper discuss the previous research in section 2, while the proposed methodology is discussed in section 3, then the experiment and results are presented in section 4 and finally the conclusion and the related work is presented in section 5.

Previous Research

ID3 has been presented in different research with modifications from different perspectives. The Very Fast Decision Trees (VFDT) algorithm presented in (Atallah, 2014) was based on ID3 algorithm with trying to reduce the learning phase time cost by applying a sampling methodology with a parallel approach which is one of our directions in this research. The algorithm could not provide accurate results with imbalanced clusters. Although an adaptation to the proposed algorithm is presented in (Hill, 2000), by introducing the Concept-adapting Very Fast Decision Trees learner (CVFDT) algorithm, the algorithm has not been evaluated as mentioned in (Quinlan, 1986) and moreover, the required parameters were user-dependent, in addition to the fixed values of these parameters in the clustering procedure.

Another perspective for enhancing ID3 is presented in (Patel, 2013), an integration of ID3 with other algorithms is proposed, they are Class-Attribute Interdependency Relation (CAIR) and Class-Attribute Interdependency Maximization (CAIM). Their research methodology was based on adapting the decision tree according to the selected criteria, the proposed approach provided a complexity in the whole task according to the need of applying a data pruning step.

Table 1. Comparison of previous research

Research	Pros	Cons
Atallah *et al*. (2014)	Reduce the learning phase time cost	Could not provide accurate results
Hill (2000)	Enhancing the work in Atallah *et al*. (2014)	Has not been evaluated
Patel (2013)	An integration of ID3 with other algorithms	Complexity in the whole task and the need of applying a data pruning
Khedr *et al*. (2014)	A weighting for the attributes to increase the overall output accuracy	Increase in the time computation
Lad *et al*. (2012)	Focused on the accuracy	Not considering the time cost

More research which considered ID3 by (Khedr *et al*., 2014) which provided a weighting for the attributes to increase the overall output accuracy, the proposed algorithm did not consider the complexity in the proposed methodology which provides an increase in the time computation. Although we consider the issue of the computational time is important to be considered, another adaptation is also discussed in (Lad *et al*., 2012) which focused on the accuracy with not considering the time cost.

Proposed Methodology

The proposed methodology is based on applying an adapted ID3 algorithm. Adapting ID3 algorithm is also based on two related approaches, they are data decomposition and parallel computation. In our methodology, data is decomposed into groups of data which is initially equal to the number of clusters. Applying the classification algorithm ID3 is then performed on all groups of data in a manner that each group is examined on a determined class. As a result of the previous step, some of the data objects are classified which are then annotated to be following the class and the remaining of the objects move to the next class and the cycle continue until there is no more objects to be classified.

To highlight the contribution of the proposed approach, we state the following points:

- The approach depends on the assumption that each object depends on only one class, therefore, when an object is already classified, then there is no need to re-examine the object against other classes
- The proposed approach also ensures less computation time due to the parallel computation which is performed
- It maintains the accuracy of the enhanced ID3 which produces higher accuracy results than the traditional ID3

A description of the main items in the proposed methodology is discussed in the next subsection and the main algorithm steps is illustrated in Fig. 1 and discussed in the following subsection.

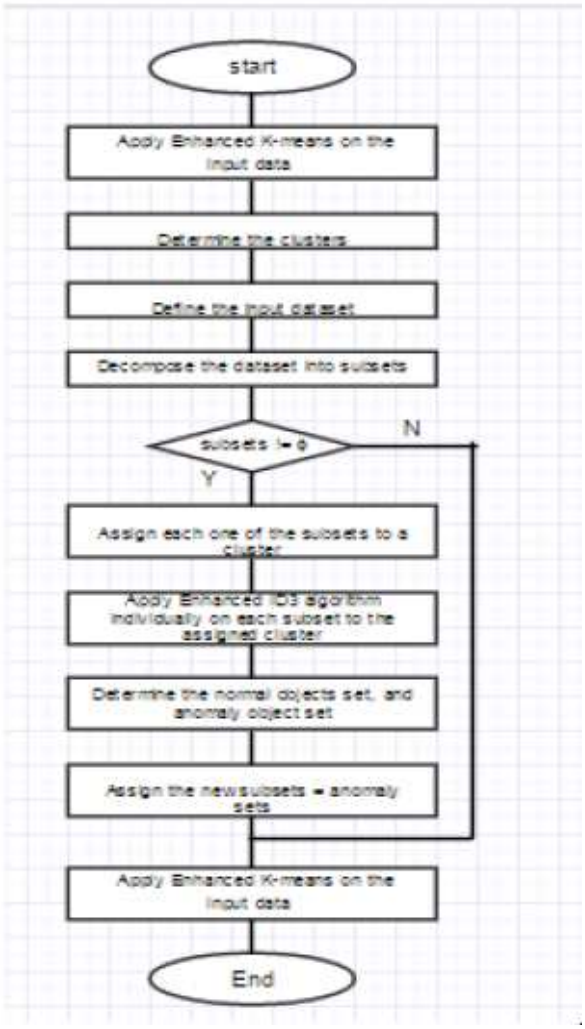

Fig. 1. Flow diagram for the proposed algorithm

Main Items Describing the Proposed Methodology

The input of the method is two inputs, they are:

- A set of the clusters “Clusters” that are produced by Adapted K-Means algorithm
- The data set is divided into N sub-datasets where N is the number of clusters
- Each group of the data set “Dset” includes a set of objects to be classified

After applying the proposed approach on the sub-datasets, two outputs are produced, they are:

- A set of objects that are already classified to follow the contributed class (Normal Objects) if its confidence is upper than the threshold which is defined in our experiment to be 50%
- A set of objects that do not belong to the examined class (Anomaly Objects)
- Each item in each objects is associated with its status, the normal objects are associated with the name of the class and the anomaly objects has a label "anomaly" to be further examined against other classes in the next rounds

The previous steps are continuously applied until one of the two situations reached, then the system stops:

- First, when all objects are examined against the defined classes
- Second, when all objects are already classified

Main Steps the Proposed Algorithm

To be further clarified, in this section we discuss the pseudo code of the proposed methodology as follows:

- Start Algorithm
- Apply Enhanced K-means on the input data
- Determine the clusters
- Define the input dataset
- Decompose the dataset into subsets
- Repeat
- Assign each one of the subsets to a cluster
- Apply Enhanced ID3 algorithm individually on each subset to the assigned cluster
- Determine the normal objects set and anomaly object set
- Assign the new subsets = anomaly sets
- Until subsets = ϕ or all subsets have been examined
- Integrate all remaining objects in the anomaly sets
- End Algorithm

Case Study Setup

The case study in this research have applied the proposed approach on radiology data, The aim of the experiment is to evaluate both algorithms and present different measures to prove he advance of the proposed algorithm over the ID3 algorithm. The dataset used is a radiology data which is collected from radiology center located in Egypt. The database stores values for four fields; patient name, the employed organization, scans date and scan type. The number of data records is 6700 transaction for about 487 patients from 40 different organizations and those patients requested 30 scan types. The received data is in the form of excel sheet. The following sections presents the results of the comparison between both researches.

Data Set Description

Data collected is from a radiology center that is located in Egypt and has several branches. The center serves more than 10000 customers per year, contracting with more than 500 organizations and provides more than 450 scan type. All patients' data is stored electronically using SQL database. The database stores values for four fields; patient name, the employed organization, scans date and scan. The number of data records is 6700 transaction for about 487 patients from 40 different organizations and those patients requested 30 scan types. The received data is in the form of excel sheet.

Three data sets A, B and C are constructed from the collected data. Data set A divides the original data into 90% for model calibration and 10% for model testing. Data set B divides the original data into 85% for model calibration and 15% for model testing. Data set C divides the original data into 80% for model calibration and 20% for model testing.

Applied Experiment

This section describe the steps for the applied experiment in the following subsections.

Preprocessing

The data collected undergoes four preprocessing steps and the data matrix is reduced from 600 rows and 4 columns, to 487 rows and 30 columns. It contains transactions for all patients in this period. The first step converts data from textual values to numeric ones in order to deal with identification numbers.

In the second step, the interesting attributes are selected which are Patient ID and Scan Type ID. The third step converts data from numeric matrix to binary matrix. In Table 2, the rows of the matrix represent patients ID while columns represent the scan type represented. Elements with value 1 indicate that the patient id did the scan type id at least once.

The Algorithm for Converting Data is as Follows

- Initialize Data matrix(number of patients, number of scan type ID) to zero
- Read Patient ID
- Read scan type ID done
- Data matrix (patient ID, Scan ID)=1

Table 2. Sample of Interesting attributes in numeric values

Patient ID	Scan ID
82803	12
82803	101
81205	190
81206	35
81207	12
81208	112

Table 3. Scaling of result and the evaluation of data

Serial number	Scaling result	Grade level
1	0	Not done
2	0.001-0.02	Very low
3	0.021-0.04	Low
4	0.041-0.06	Moderate
5	0.061-0.08	High-moderate
6	0.081-0.09	Very high
7	1	Done

The fourth step is a data filtering. It is needed as the data is a snapshot for a short period of time. Therefore, not all Patient IDs are expected to exist neither all Scan Type IDs. This is indicated in the binary matrix with either all zero row(s) or all zeros column(s), respectively. The algorithm for row elimination and is as follows while the algorithm for column elimination has similar steps:

1. % remove Patient Id who didn't request any service (scan)
2. If there is a row whose elements = 0
   Remove row
3. Otherwise
   Keep it

Classification

Training sets are used to calibrate the models using WEKA software and each classification model is then tested by the corresponding testing data. For each data set, three ID3 models are created with 10, 15 and 17 classification; this gives a total of 9 experiments. The experiments has been applied for the ID3 before and after adaptation and the results have been compared as will be shown in section 6.

Post Processing

In this study, seven quantification levels are used to quantify the classification centers shown in Table 3. The dimension of each classification is described by one of the seven quantification level. For example, Table 4 shows the transformation of one classification into the quantified level.

Table 4. Show the transformation of classification into quantified

Dimensions	value	Quantification
1	0.000	Not done
2	0.000	Not done
3	0.000	Not done
4	0.000	Not done
5	0.021	Low
6	0.000	Not done
7	0.000	Not done
8	0.000	NOT done
9	0.000	Not done
10	0.000	Not done
11	0.000	Not done
12	0.066	High moderate
13	1.000	Done
14	0.000	Not done
15	0.000	Not done
16	1.000	Done
17	0.000	Not done
18	0.000	Not done
19	0.000	Not done
20	0.081	Very high
21	0.000	Not done
22	0.000	Not done
23	0.000	Not done
24	0.000	Not done
25	0.000	Not done
26	0.000	Not done
27	0.044	Moderate
28	0.000	Not done
29	0.000	Not done
30	1.000	Done

Results

The results of running the adapted ID3 Algorithm system are presented as follows:

The result is represented by the patient's distribution percent in each model for each data set. Figure 2 shows the distribution percent for each data set in case of 10 classification model. It shows that the model for data set B is the most appropriate one as it describes an average between the other results.

Figure 3 describes the results of testing data. The results show that 28% of the patients exist in classification 0; nearly 16% of the patients exist in classification 9. It also shows that the percentage starts to decrease in the other classification which implies that there are a lot of patients having most of their scans in classification s 0 and 9.

For 15 classification model, the result is represented by the patient's distribution percent for each data set. Figure 4 shows the distribution percent for each data set in case of 15 classification s model. It shows that the model for data set B is the most appropriate one as it describes an average between the other results.

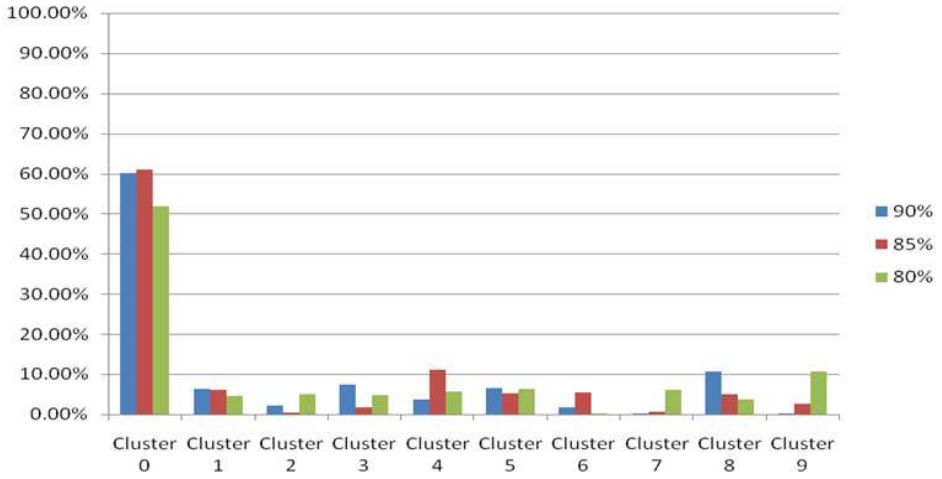

Fig. 2. Distribution percentage of patients in training set for experiments A, B and C for 10 classification model

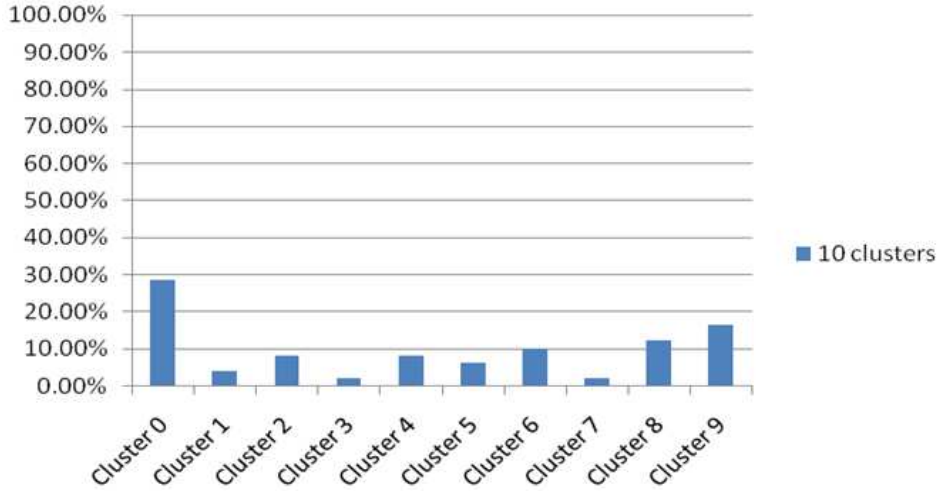

Fig. 3. Distribution percentage of patients in testing set for data set B for 10 classification model

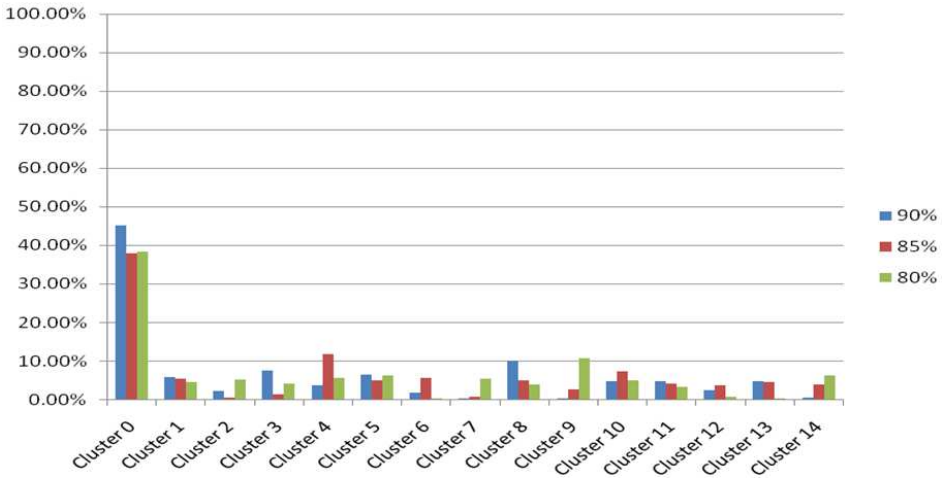

Fig. 4. Distribution percentage of patients in training set for experiments A, B and C for 15 classification model

Figure 5 describes the results of testing data. The results show that 22% of the patients exist in classification 0; nearly 16% of the patients exist in classification 9. It also shows that the percentage starts to decrease in the other classification which implies that there are a lot of patients having most of their scans in classification s 0 and 9.

For 18 classification model, the result is represented by the patient's distribution percent for each data set. Figure 6 shows the distribution percent for each data set in case of 18 classification s model. It shows that the model for data set B is the most appropriate one as it describes an average between the other results.

Figure 7 describes the results of testing data. The results show that 21% of the patients exist in classification 0; nearly 16% of the patients exist in classification 9 almost 1% in classification s 16, 10,8,5,4 and classification 2. It also shows that the percentage starts to decrease in the other classification which implies that there are a lot of patients having most of their scans in classification s 0 and 9.

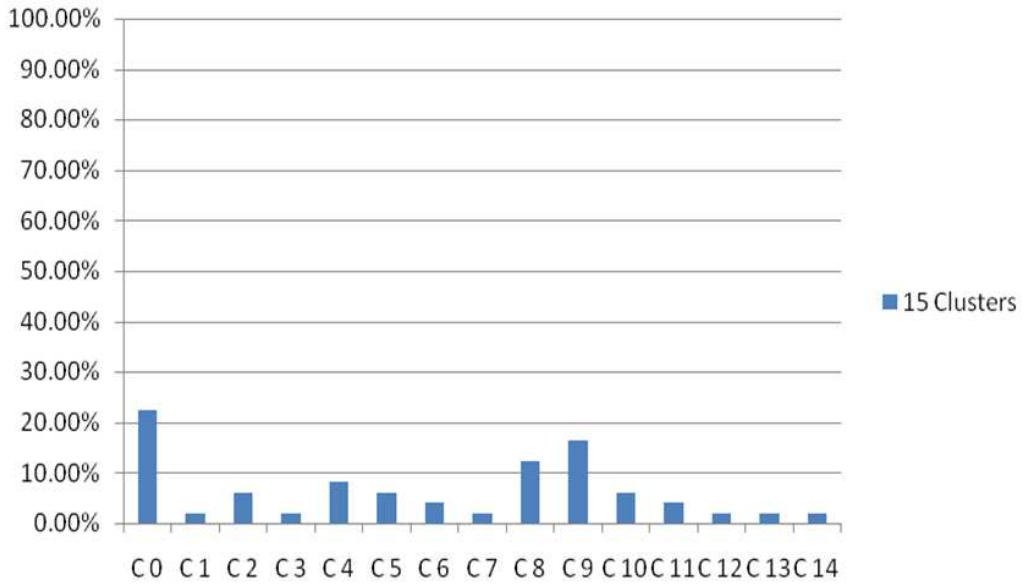

Fig. 5. Distribution percentage of patients in testing set for data set B for 15 classification model

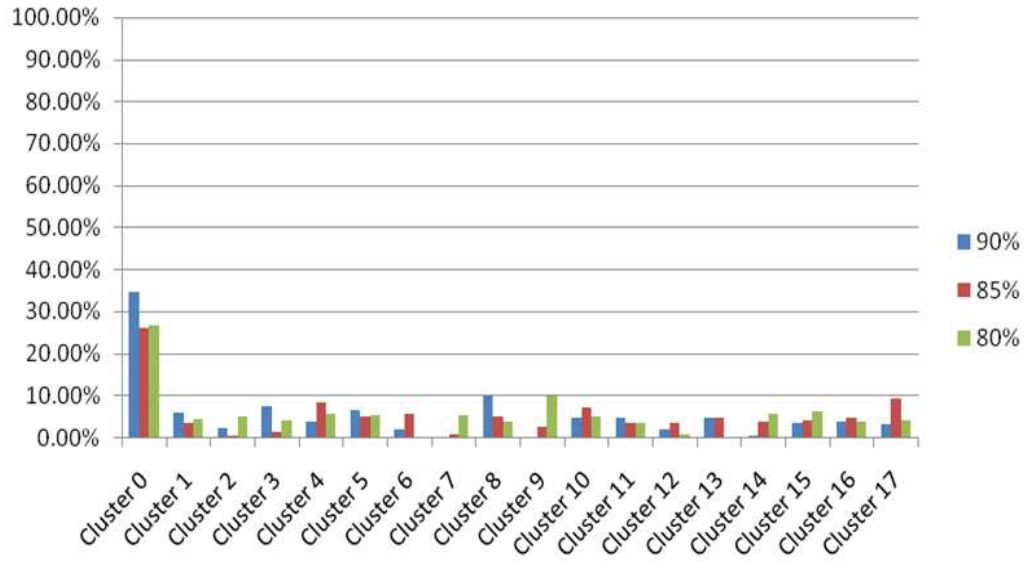

Fig. 6. Distribution percentage of patients in training set for experiments A, B and C for 18 classifications model

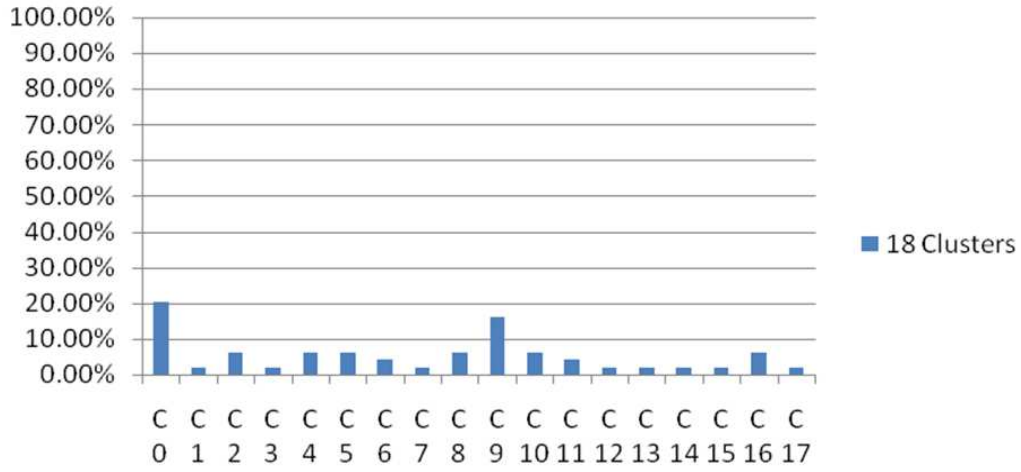

Fig. 7. Distribution percentage of patients in testing set for data set B for 18 classification model

Table 5. Evaluation of Experiments for 10. 15, 18 classifications

	1	2	3	4	5	6	7	8
A	79.5% (8)	10.3% (2)	3.2	0.4	0.1	0.3	20.3	65
B	19.9% (3)	21.9% (7)	4.8	0.9	0.2	0.6	21.3	65

Experimental Evaluation

As previously described, this research have applied the proposed adaptation of ID3 algorithm and ID3 over the radiology data and different measures have been calculated. We have applied two experiments, the first experiment is applied over 10 clusters, while the second used 15 clusters and the third used 18 clusters. The result of comparison for the three experiments are shown in Table 5 respectively. After applying the adapted framework, a comparison between the classifications of the ID3 before and after adaptation has been performed.

To clarify the contents of the three tables, the following abbreviations have been used:

A: ID3 Algorithm
B: Adapted ID3 Algorithm
1: Instances that are classified in the correct class %
2: Instances that are classified in the incorrect class %
3: The duration of time in seconds
4: Statistics measure (Kappa)
5: Mean Absolute Error
6: Root Mean Squared Error
7: Relative Absolute Error (%)
8: Root Relative Squared Error (%)

Conclusion and Future Research

This research presented an adaptation to the ID3 algorithm. The adaptation is based on two related approached, data decomposition and parallelism. The proposed methodology focused on raising the output accuracy of ID3 with ensuring the reduction of computation time which is one of the main focused topics currently. To prove the applicability of the proposed methodology, we have applied the method on the radiology dataset on three experiments with a variety of clusters' number and the results have been proposed with providing a comparison between the proposed approach and ID3.

However, further research can be performed by considering the multiple inheritance of the objects. Verifying the proposed methodology on different classification algorithms can be further examined and finally, applying the proposed method on other datasets from different domains can further improve the generality of the research.

Author's Contributions

Ayman E. Khedr: Contributed in all parts of the paper, the idea, the review part, the experiment, writing the paper and revising the paper.

Amira M. Idrees: Contributed in all parts of the paper, the idea, the review part, the experiment, writing the paper and revising the paper.

Ahmed I. El Seddawy: Contributed in all parts of the paper, the idea, the review part, the experiment, writing the paper and revising the paper.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Ahire, P.G., S. Kolhe, K. Kirange, H. Karale and A. Bhole, 2015. Implementation of improved ID3 algorithm to obtain more optimal decision tree. Int. J. Eng. Res. Develop., 11: 44-47.

Atallah, D., A. Eldesoky, H. Amira and M. Ghoneim, 2014. One-year renal graft survival prediction using a weighted decision tree classifier. Int. J. Eng. Technol. DOI: 10.14419/ijet.v3i3.2334

Domingos, P. and G. Hulten, 2000. Mining high-speed data streams. Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 20-23, Boston, MA, USA, pp: 71-80. DOI: 10.1145/347090.347107

Hill, L., 2000. Intelligent enterprise.

Hulten, G., L. Spencer and P. Domingos, 2001. Mining time-changing data streams. Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 26-29, San Francisco, CA, USA, pp: 97-106. DOI: 10.1145/502512.502529

Khedr, A.E., A.I. El Seddawy and A.M. Idrees, 2014. Performance tuning of k-mean clustering algorithm a step towards efficient DSS. Int. J. Innovative Res. Comput. Sci. Technol., 2: 111-118.

Lad, M.R., R.G. Mehta and D.P. Rana, 2012. A novel tree based classification. Int. J. Eng. Sci. Adv. Technol., 2: 581-586.

Moore, A., 2009. K-means and hierarchical clustering-tutorial slides.

Patel, K., 2013. Review on data stream classification algorithm (Hoeffding Tree, VFDT, CVFDT). In. J. Concept. Electrical Electron. Eng., 1: 30-35.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learn., 1: 81-106. DOI: 10.1023/A:1022643204877