iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 Issue: III Month of publication: March 2018

DOI: http://doi.org/10.22214/ijraset.2018.3566

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Enhanced the Performance of Medical Data Based on Data Farming Techniques

Abhilash Raghuwanshi[1], Gagan Sharma[2], Rajesh Sharma[3]
[1, 2, 3] Sri Satya SAI College of Engineering, Bhopal M.P., India

Abstract:* We can see that data farming is an emerging field of research because decision making is an important issue in the competitive business environment. The appropriateness of data and the data collection cost become the goals of data farming. Data farming is emerging fields of research in the current scenario, where data collection cost and time consumed in data collection is significant to reduce. We can farm the data where we have narrow data set and then apply the data mining algorithm to extract the useful knowledge. We proposed an algorithm for data farming steps data plantation & harvesting. We farm sufficient data from the available little seed data by applying the proposed algorithm of data farming. Classification results of J48 classification, for farmed data is achieved better than classification results for the seed data, which proves that the proposed data farming algorithm has produced effective data. In this paper, we present an algorithm for data farming which farms the data with the help of the seed data on a predefined error threshold rate. Proposed algorithm has implemented on farmed datasets are verified for the classification accuracy on the weka open source data mining tool.

Keywords: Keywords: -J48, decision tree, WEKA, data farming.

I. INTRODUCTION

Data farming is a process of growing sufficient data with the help of various statistical and heuristic techniques. As data collection cost is high, so many times data mining projects uses existing data collected for various other purposes, such as daily collected data to process and data required for monitoring & control. Sometimes, the dataset available might be large or wide dataset and sufficient for extraction of knowledge but sometimes the dataset might be narrow and insufficient to extract meaningful knowledge or the data may not even exist [6]. Mining from wide datasets has received wide attention in the available literature. Many models and algorithms for feature selection have been developed for wide datasets. Determining or extracting knowledge from a narrow dataset or partial availability of data has not been sufficiently addressed in the literature. The data farming methodology provides a more comprehensive understanding of all possible outcomes of the mining results, and offers the opportunity to discover outliers, surprises in the narrow dataset. In this chapter, we cover an introduction to data farming and also we survey various data farming methodologies & approaches.

II. DATA PRE-PROCESSING VS DATA FARMING

Data pre-processing and data farming (feature definition) represent the opposite ends of the data spectrum. Data pre-processing deal with a redundant number of features and the data farming begins with a potentially empty set of features that is gradually transformed into a set of features satisfying the selected performance criteria. Figure 2.3 depicts the difference between data farming and data preprocessing. Preprocessing is a ‘Push’ approach to data mining as selected feature determines the quality of knowledge on the other hand data farming ‘Pulls’ the necessary data for knowledge extraction [6].

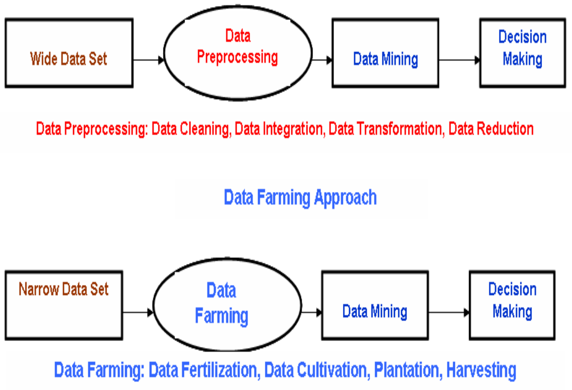

Wide Data Set
Data Preprocessing
Data Mining
Decision Making
Data Preprocessing: Data Cleaning, Data Integration, Data Transformation, Data Reduction
Data Farming Approach
Narrow Data Set
Data Farming
Data Mining
Decision Making
Data Farming: Data Fertilization, Data Cultivation, Plantation, Harvesting

Figure 1. Data Farming vs Data Preprocessing

III. PROBLEM FORMULATION

Objective of data farming is to improve the mining accuracy as well as reduce the data collection cost. Classification accuracy, cluster density and rule support or confidence is a measure of the data mining results. Data farming is used to improve the results in terms of these performance measures. The goal of data farming is given below.

A. Maximize performance measure (e.g., classification accuracy, cluster density, rule support and confidence)
B. Minimize or reduce the data collection cost

These criteria directly affect accuracy and cost savings. High accuracy on low price increased competitiveness. Various other criteria for data farming may be evaluated in real life.

IV. PROPOSED ALGORITHM

In the step of data fertilization, we load input seed data to the model; if input seed data have some missing values, filling of these missing values is done by applying appropriate missing data estimation methods. After that, we predict some attributes to refine the quality of the seed data i.e., reduce the error between actual and predicted values of some attributes by applying regression . In cultivation we get min-max (lower bound and upper bound) of each attribute. Now in the step of data plantation, we use this fully completed (no missing) & updated version of the seed data to farm more data. In this chapter, we assume that data fertilization is already done & input seed is complete and satisfactory to perform cultivation. Cultivation initiates with getting the lower bound & upper bound of each attribute. In cultivation step, we apply the error threshold on the lower bound & upper bound i.e., Min-Max range of each attributes. Hence, cultivation step is completed. Now, plantation is to be done with this cultivated Min-Max Range (*lb*, *ub*). For plantation, we generate values between this cultivated Min-Max ranges for each attribute in the seed dataset. Finally, collection of these generated values is done in harvesting step.

A. Algorithm: Data_farming (seed_dataset, k, error_thresold)

```
// seed_dataset, it contains seed data in n attribute (a, $a_{2,\ a3\ ...}$ $a_n$) & m
tuples.
// k, Number of the tuples to be generated.
// error_thresold, permissible error in the actual seed data range & farmed
data set values of attributes.
// farmed_data, it contains the farmed data set of each iteration
{
   Fill missing Values & prediction by regression (if any) // Fertilization
          Farmed_data[k][n];

for i = 1 to n
      {
         $L_i$= Minimum of column i in seed_data;      // Cultivation
             $M_i$= Maximum of column i in seed_data;
             $diff_i$= $L_i$ - $M_i$ ;
             $lb_i$ = $L_i$ – ($diff_i$* error_thresold/100);
             $ub_i$ = $L_i$ + ($diff_i$* error_thresold/100);
         }

      for i=1 to k        // Plantation
          {
        for j=1 to n
                 {
                  farm_data (i,j) = randomly generate the data item with bounded range
                 [$lb_i$ , $ub_i$] for column j;
```

```
                }
            }
        return farmed_data;          // Harvesting
}
```

V. EXPERIMENTAL RESULT ANALYSIS

we enumerate the various experiments of farming data on different combination of threshold values (2, 5 and 10), number of seed instances (50 and100) & number of farmed data instances (500, 1k, 2k, 5k and 10k). Seed data used in this paper, is related to the cardiac patent. This seed data has 20 attributes as given above. We have performed total 30 numbers of experiments to analyze the proposed algorithm. In this table, we gave the time required in each experiment & save the farmed data with .csv file name as naming convention described earlier.

Table 1; contains the results of the 10 experiments with error threshold value 2 and two samples of the seed data of size 50 & 100. Column 4 of table 3.2 shows the time taken by the experiment i.e., to farm the dataset. We can see that, time is directly affected by the number of farmed tuple. In table 3.2 we can see 1.094 seconds, minimum time is required to farm 500 tuple from the seed data size 50 and 31.922 seconds, maximum time is required to farm 10000 tuple from the seed data size 100.

Table 1: Data Farming Result with time for error threshold 2

S. No.	Error Threshold	No. of Seed Tuple	No. of Farmed Tuple	Time Taken	Farmed Dataset
1	2	50	500	1.094	farmed_2_50_500
2	2	50	1000	2.109	farmed_2_50_1K
3	2	50	2000	4.266	farmed_2_50_2K
4	2	50	5000	12.172	farmed_2_50_5K
5	2	50	10000	31.328	farmed_2_50_10K
6	2	100	500	1.11	farmed_2_100_500
7	2	100	1000	2.172	farmed_2_100_1K
8	2	100	2000	4.36	farmed_2_100_2K
9	2	100	10000	31.922	farmed_2_100_10K

Figure 2 shows the graph of the time required to farm the data by the proposed algorithm for different value of threshold & seed data size.

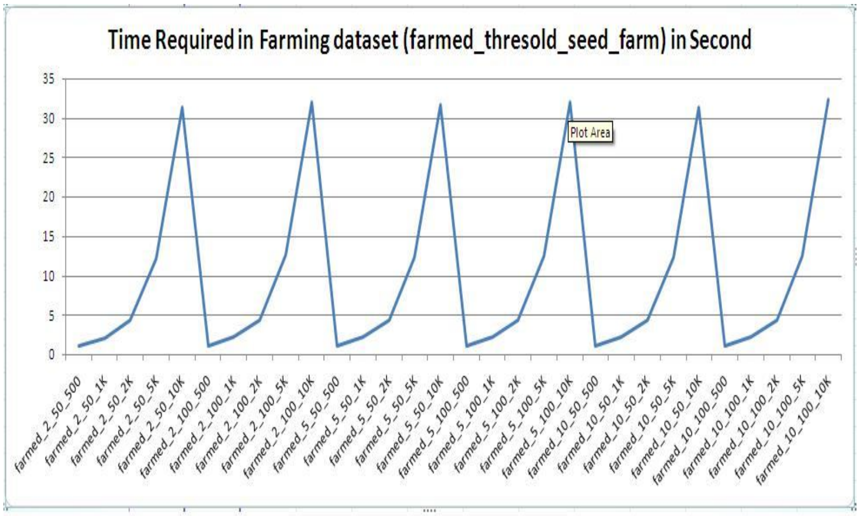


Figure 2. Plot of time required by the proposed algorithm

Figure 3. The time required to farm the dataset with the number of farmed tuple in thousands. We can see in this graph, the time required to farm the data set by the proposed algorithm, is directly affected by the number of farmed tuple.

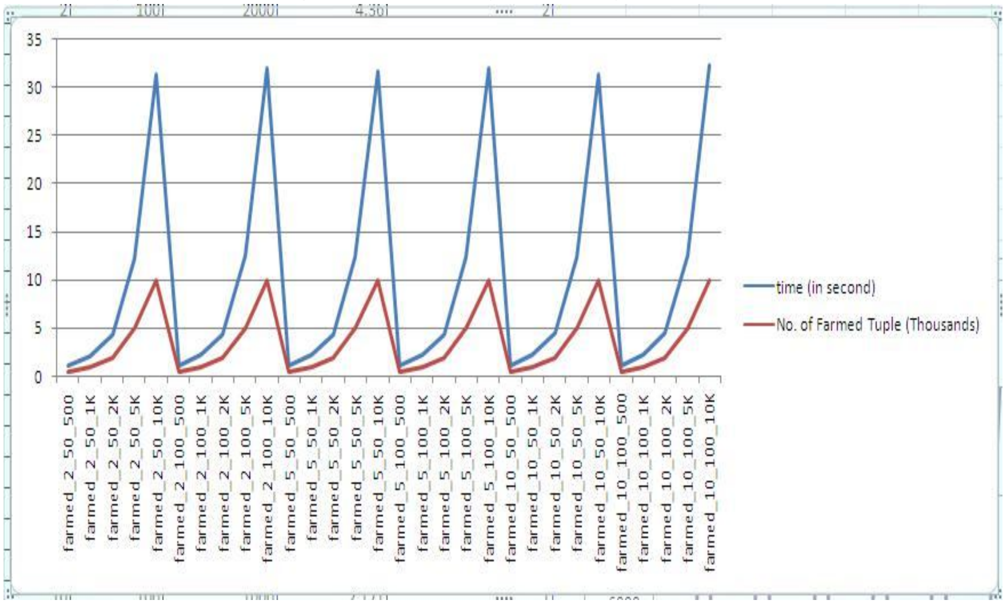


Figure 3. Plot of time required with number on instances farmed by algorithm

Analysis of the proposed algorithm and factor affecting the performance of the proposed algorithm described here. We can conclude from the results achieved from the figure & graphs that -

A. We can observe from the table 1 and figure 2 & 3 that time required to farm a dataset is highly dependent on the factor that how much instances are to be farmed (number of farmed instances). As more instances to be farmed, as much time is required.

B. Time required to farm a dataset is lightly dependent on the factor that how much seed data instances are used in farming. As the number of seed data instances increases, the time required to farm the data also increases.

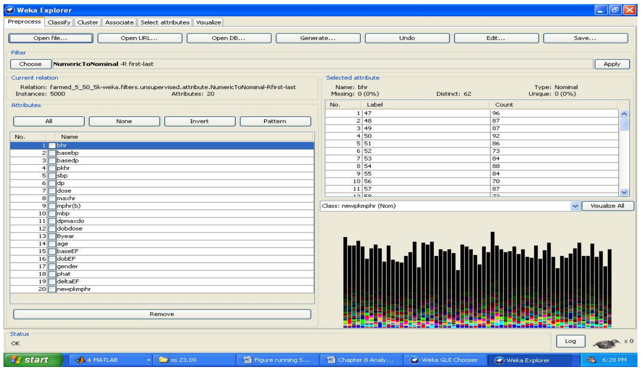
C. Time required to farm a dataset is lightly dependent on, error threshold permissible in farming. As the error threshold increases, the time required to farm the data also increases.

To check the quality of the farmed datasets, we performed the classification & compare the classification accuracy among the original dataset, sample datasets and farmed datasets.

Table 2. J48 Classification Results on Original Dataset & sample data of size 50 & 100.

Name	Factor	Original Data	Samdata 50	Samdata 100
CCI	Correctly Classified Instances	68.10%	82%	79%
ICI	Incorrectly Classified Instances	31.90%	18%	21%
KS	Kappa statistic	0.5715	0.7106	0.71
MAE	Mean absolute error	0.1128	0.079	0.0947
RMSE	Root mean squared error	0.2375	0.1987	0.2176
RAE	Relative absolute error	59.07%	37.24%	41.01%
RRSE	Root relative squared error	76.99%	62.06%	64.41%
INSTANCE	Total Number of Instances	558	50	100

Incorrectly classified instances (ICI) for the original dataset, sample data_50 & sample data_100 are 31.90%, 18% & 21% respectively, which are increased for the farmed datasets. It indicates the farmed data is more appropriate compared to the sample datasets for mining purposes.



NumericToNominal -R first-last

Relation: farmed_5_50_5k-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances: 5000 Attributes: 20

No.	Name
1	bhr
2	basebp
3	basedp
4	pkhr
5	sbp
6	dp
7	dose
8	maxhr
9	mphr(b)
10	mbp
11	dpmaxdo
12	dobdose
13	Byear
14	age
15	baseEF
16	dobEF
17	gender
18	phat
19	deltaEF
20	newpkmphr

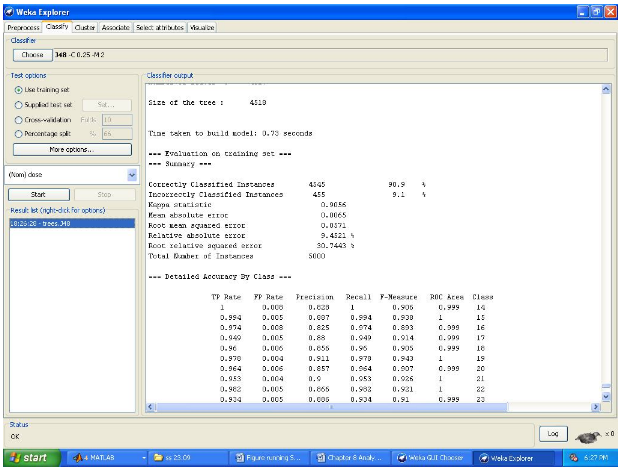
Name: bhr Type: Nominal
Missing: 0 (0%) Distinct: 62 Unique: 0 (0%)

No.	Label	Count
1	47	96
2	48	87
3	49	87
4	50	92
5	51	86
6	52	73
7	53	84
8	54	88
9	55	84
10	56	70
11	57	87

Class: newpkmphr (Nom)

OK

Figure 4. Numeric to nominal conversion by weka of farmed_5_50_5k dataset



J48 -C 0.25 -M 2

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) dose

18:26:28 - trees.J48

```
Size of the tree :      4518


Time taken to build model: 0.73 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        4545               90.9    %
Incorrectly Classified Instances       455                9.1    %
Kappa statistic                          0.9056
Mean absolute error                      0.0065
Root mean squared error                  0.0571
Relative absolute error                  9.4521 %
Root relative squared error             30.7443 %
Total Number of Instances             5000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.008      0.828     1         0.906      0.999    14
                 0.994     0.005      0.887     0.994     0.938      1        15
                 0.974     0.008      0.825     0.974     0.893      0.999    16
                 0.949     0.005      0.88      0.949     0.914      0.999    17
                 0.96      0.006      0.856     0.96      0.905      0.999    18
                 0.978     0.004      0.911     0.978     0.943      1        19
                 0.964     0.006      0.857     0.964     0.907      0.999    20
                 0.953     0.004      0.9       0.953     0.926      1        21
                 0.982     0.005      0.866     0.982     0.921      1        22
                 0.934     0.005      0.886     0.934     0.91       0.999    23
```

OK

Figure 5. Weka J48 Classification screen shot on farmed_5_50_5k dataset.

Figure 5 depicts the numeric to nominal conversion of the farmed_5_50_5k dataset and figure 3.9 shows running snapshot of the J48 classification on this farmed dataset, while Figure 7.10 depicts the numeric to nominal conversion of the farmed_5_50_5k dataset and figure 3.11 shows running snapshot of the J48 classification on this farmed dataset.

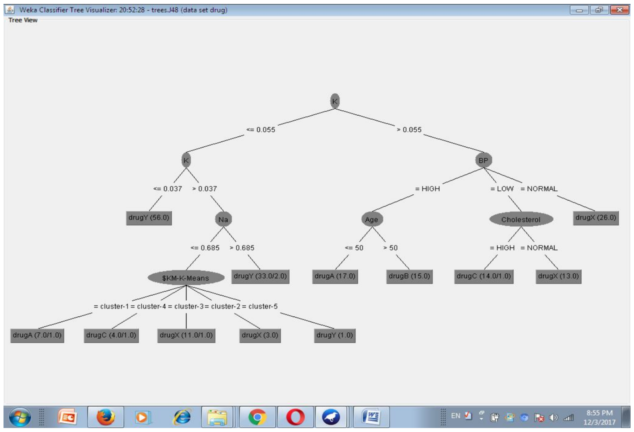


Figure 6. Decision tree on the basis of dataset chosen

The figure shown above is the proposed horizontal partition based J48 decision tree algorithm in which the gain of each attribute is calculated and the attribute having highest gain is the root node of the tree and then second time the gain of each attribute is calculated and the full decision tree is computed

Name	farmed_2_50_500	farmed_2_50_1k	farmed_2_50_2k	farmed_2_50_5k	farmed_2_50_10k
CCI	99.20%	98.30%	95.70%	89.64%	81.55%
ICI	0.80%	1.70%	4.30%	10.36%	18.45%
KS	0.9917	0.9823	0.9553	0.8924	0.8084
MAE	0.0006	0.0013	0.0032	0.0077	0.0137
RMSE	0.0172	0.0251	0.04	0.0621	0.0828
RAE	0.83%	1.77%	4.48%	10.80%	19.24%
RRSE	9.12%	13.30%	21.18%	32.87%	43.86%
INSTANCE	500	1000	2000	5000	10000

Table 3. J48 Classification Result on farmed data on error threshold 2 & seed tuple 50

Table 3 contains the results obtained from the weka software by applying J48 classification on permissible threshold value 2, seed data size 50 and farmed tuple 500, 1 k, 2 k, 5 k and 10 k. Incorrectly classified instances are 0.80, 1.70, 4.30, 10.36 and 18.45% respectively while kappa statistics are 0.9917, 0.9823,0.9553, 0.8924 & 0.8084 respectively.

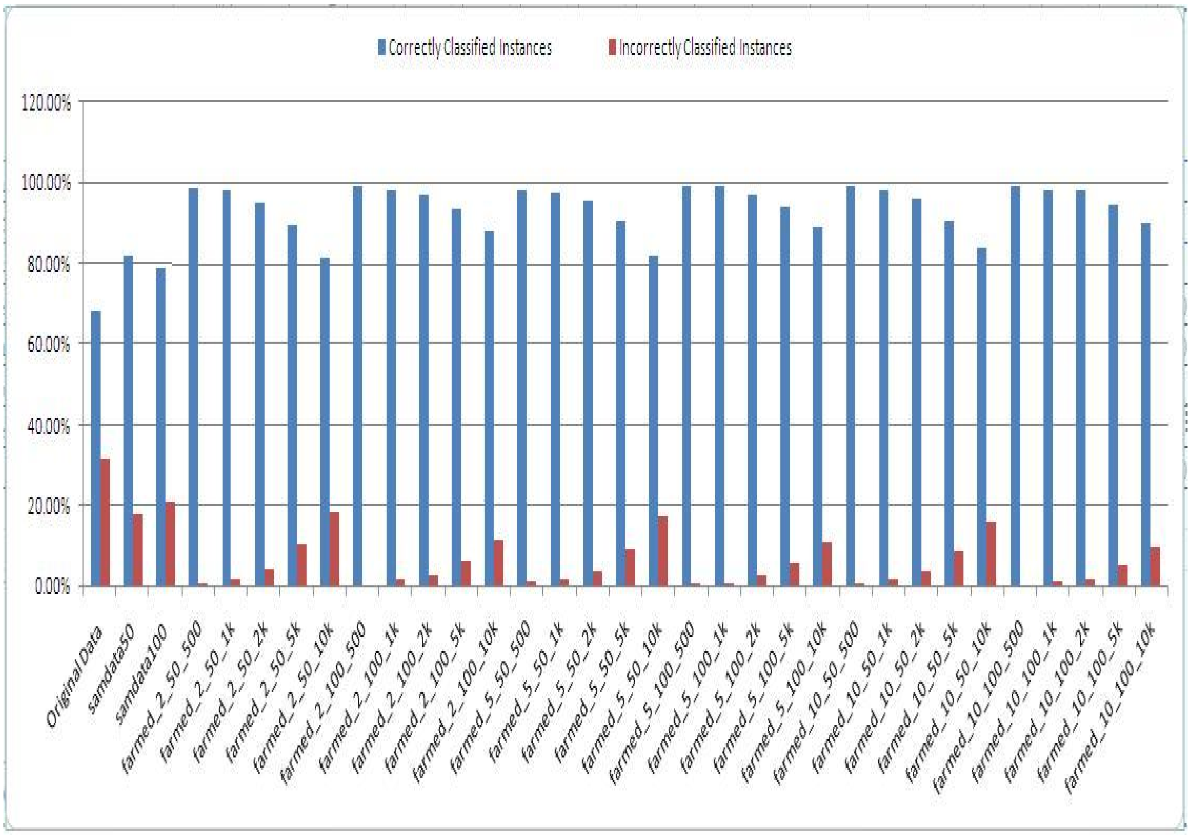


Figure 7. Plot of correctly & incorrectly classified instances by J48 Classification on original, sample & farmed Data

Figure 7. shows the percentage of correctly & incorrectly classified instances for the original, sample and farmed datasets. It can be seen that percentage of correctly classified instances is increased & percentage of incorrectly classified instances is decreased.

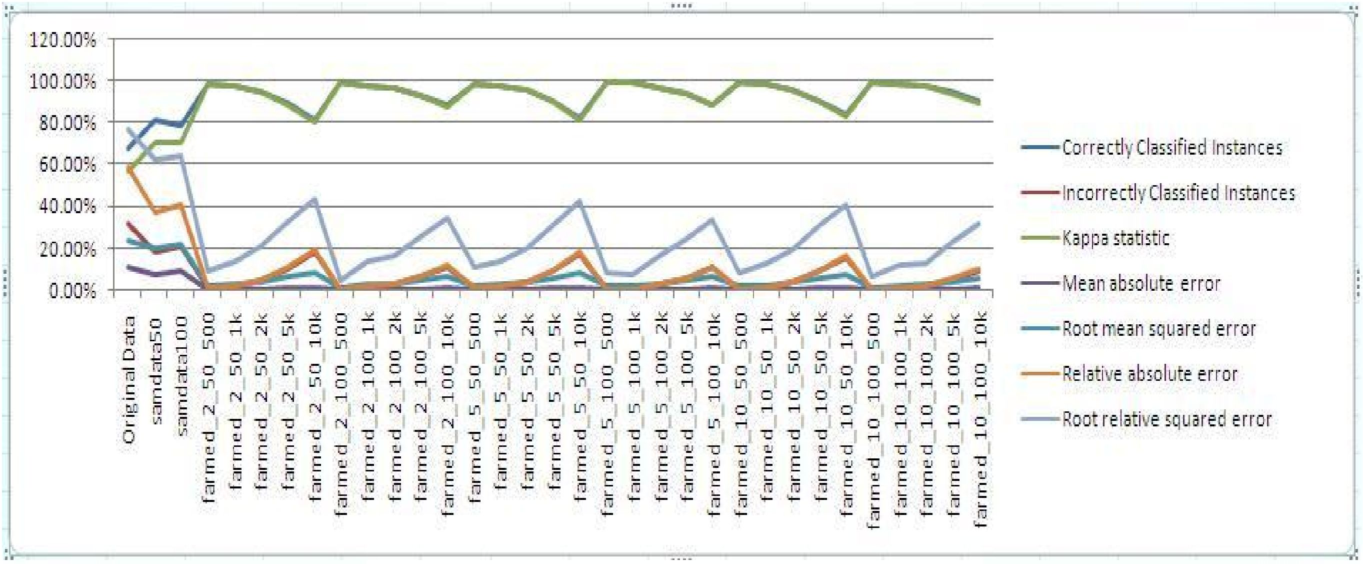


Figure 8. Plot of classification result on original, sample & farmed Data

Figure 8 shows that correctly classified instances (CCI) & Kappa statistics (KS) are increased & incorrectly classified instances (ICI), Mean absolute error (MAE), Root mean squared error (RMSE), Relative absolute error (RAE), Root relative squared error (RRSE) are decreased for the farmed data compared to the original dataset and sample datasets. The time complexity of the proposed algorithm is O (mn), where m is the number of data to be farmed and n is the number of attributes in the seed dataset. It is a quadratic time complexity algorithm.

VI. CONCLUSION AND FUTURE WORK

Data farming is an emerging field of research in the current scenario, where data collection cost and time consumed in data collection are significant to reduce. we proposed an algorithm for data farming steps data plantation & harvesting. We farm sufficient data from the available little seed data by applying the proposed algorithm of data farming.

Proposed algorithm farmed sufficient data with improved adequateness of the available seed dataset for mining. By filling up of missing data & updating predicted values of few attributes, we get fertile seed dataset & by cultivation we prepare the environment for plantation. Proposed algorithm plants these fertile seed in a cultivated environment & harvests the crops in the form of farmed data. We can see that the farmed data is sufficient to perform various mining techniques and find out the hidden knowledge while seed data is not sufficient. Classification accuracy of the farmed data proved, that it is better compared to the sample datasets. Farming time required is highly dependent on the instances to be farmed and lightly on the number of seed data & error threshold. Correctly classified instances (CCI) & kappa statistics (KS) are increased & incorrectly classified instances (ICI), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative absolute error (RAE), Root relative squared error (RRSE) are decreased for the farmed data when compared to the original dataset and sample datasets. This variation shows that, the farmed data is more effective compared to the sample datasets.

This thesis provides the overall conclusion of the research work done in this thesis as well as limitations and future scope of the work. This will be helpful in the further research in this field. The current work can be enhanced in the future with the concept of cloud computing environment.

REFERENCES

[1] Gary E. Horne, Ted E. Meyer, Data Farming: Discovering Surprise, Proceedings of the 2004 Winter Simulation Conference, R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

[2] C.S. Choo, E.C. Ng, Dave Ang, C.L. Chua, Data Farming In Singapore: A Brief History , Proceedings of the 2008 Winter Simulation Conference S. J. Mason, R. R. Hill, L. Monch, O. Rose, T. Jefferson, J. W. Fowler eds. http://www.researchgate.net/publication/221525990_Data_Farming_in_Singapore_A_brief_history

[3] Philip Barry, Mathew Koehler, Simulation in Context: Using Data Farming for Decision Support, Proceedings of the 2004 Winter Simulation Conference, R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

[4] Gary E. Horne, Klaus Peter, Schwierz, Data Farming Around the World Overview, Proceedings of the 2008 Winter Simulation Conference, S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J.W.Fowleds. http://www.researchgate.net/publication/221525532_Data_Farming_around_the_world_overview

[5] Adam J. Forsyth, Gary E. Horne, Stephen C. Upton, Marine Corps Applications of Data Farming, Proceedings of the 2005 Winter Simulation Conference, M. E. Kuhl, N. M. Steiger, F. B. Armstrong, And J. A. Joines, ed

[6] Andrew Kusiak, Data Farming Methods for Temporal Data Mining, Intelligent Systems Laboratory, 2139 Seamans Center, The University of Iowa, Iowa City, Iowa 52242 http://www.sigkdd.org/kdd2001/Workshops/kus.pd

[7] D. Burnell, A.Al-Zobaidie, G.Windall, A.Butler. Self-Optimising Data Farming for Web Applications. Proceedings of the 15th International Workshop on Database and Expert Systems Applications (Dexa'04) 1529-4188/04 IEEE

[8] Gary E. Horne, Ted E. Meyer, Data Farming: Discovering Surprise, Proceedings of the 2005 Winter Simulation Conference, R. G. Ingalls,M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

[9] Jian Lin and Minjing Peng 2007, SVR-Based Data Farming Technique for Web Application. In Ifip International Fedration for Information Processing, Volume 254, Research and Practical Issues of Enterprises Information Systems II Volume I, eds. L.Xu, Tjoa A.,Chaudhry S. (Boston: Springer),pp 433-441.

[10] M.Fleury, A.C.Downton and A.F.Clark, Scheduling Schemes for Data Farming, IEEE Proc. Computer & Digital Tech., Vol. 146, No. 5, September 1999.

[11] http://www.cs.uiuc.edu/homes/hanj/bk2/toc.pdf

[12] Dariusz Krola, Bartosz Kryzaa, Michal Wrzeszcza, Lukasz Dutka, Jacek Kitowski, Elastic Infrastructure for Interactive Data Farming Experiments, International Conference on Computational Science, ICCS 2012

[13] Henrik Friman, Gary E.Horne, Using Agent Models and Data Farming to Explore Network Centric Operations. Proceedings of the 2005 Winter Simulation Conference.

[14] C.L. Chua, W.C. Sim, Automated Red Teaming: An Objective-Based Data Farming Approach for Red Teaming. Proceedings of the 2008 Winter Simulation Conference.

[15] Dr. Alfred G. Brandstein, Dr. Gary E. Horne, Data Farming: A Meta-Technique for Research in the 21st Century, Maneuver Warfare Science 1998.

[16] Dr. Gary E. Horne, Beyond Point Estimates: Operational Synthesis and Data Farming, Maneuver Warfare Science 2001.

[17] Gary E.Horne, Henrik Friman. "Analysis of the Military Effectiveness of Future C2 Concepts and Systems", Held at NC3A, The Hague, the Netherlands, 23-25 April 2002, in RTO-MP-117.

[18] Andrew Kusiak, Member, IEEE, "Feature Transformation Methods in Data Mining", IEEE Transactions on Electronics Packaging Manufacturing, Vol. 24, No. 3, July 2001.

[19] Jun Zheng, Ming-Zeng Hu, Hong-Li Zhang, A New Method of Data Preprocessing and Anomaly Detection, Proceedings of the third international Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004.

[20] Fang Yuan, Li-Juan Wang, Ge Yu, Study on Data Pre-processing Algorithm in Web Log Mining, Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003.

[21] Srivatsan Laxman And P.S. Sastry, A Survey of Temporal Data Mining, Sadhana Vol. 31, Part 2, April 2006, pp. 173–198.

[22] Andrew Kusiak, Data Farming: A Primer, International Journal of Operations Research Vol. 2, No. 2, 48−57 (2005) 1527 http://www.orstw.org.tw/ijor/vol2no2/Paper-6-IJOR-Vol2_2_-Kusiak.pdf

[23] Brian F. Tivnan, Data Farming Co evolutionary Dynamics in Repast, Proceedings of the 2004 Winter Simulation Conference R. G. Ingalls,M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.